CAD TOOLS AND METHODOLOGIES FOR RELIABLE 3D IC DESIGN, ANALYSIS, AND OPTIMIZATION

A Dissertation
Presented to
The Academic Faculty

by

Yarui Peng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2016

CAD TOOLS AND METHODOLOGIES FOR RELIABLE 3D IC DESIGN, ANALYSIS, AND OPTIMIZATION

Approved by:

Dr. Sung Kyu Lim, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Madhavan Swaminathan
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Hyesoon Kim
College of Computing
Georgia Institute of Technology

Date Approved: November 4, 2016

ACKNOWLEDGEMENTS

Finishing Ph.D. has been an amazing journey in my life, and it wouldn't have been possible without the support and guidance I received. I wish to express my sincere appreciation to ones who contributed to this thesis and supported me during this journey.

Firstly, I would like to thank my advisor, Dr. Sung Kyu Lim, for guiding me throughout my Ph.D research. I am very grateful for his scientific advice and knowledge and I hope that I could be as wise, supportive, and passionate as Dr. Lim someday. I wish him to continue his success with great people.

I would like to thank Dr. Saibal Mukhopadhyay and Dr. Arijit Raychowdhury for suggestions and guidance on my research. In addition, I thank Dr. Madhavan Swaminathan and Dr. Hyesoon Kim for serving on my dissertation defence committee.

As a major part of my research is performed with the collaboration of Mentor Graphics, I would like to thank Dr. Dusan Petranovic, for his insightful guidance on my research and kind help during my summer internship. I would like to thank Mr. Ray Juan and Mr. Michael Buehler for providing a great opportunity to perform research at Mentor Graphics. I would like to thank Kendall Hiles, Christopher Clee, Sandeep Koranne, John Ferguson, Chase Yun, Lei Ling, Lychung Ma, Valeriy Sukharev and Xin Huang for their help and many brilliant comments.

I am also grateful for having the opportunity to collaborate with Intel, Samsung, and Qualcomm in research projects. I would like to thank Dr. Paul Fischer, Younsik Park, Dr. Kambiz Samadi, Pratyush Kamal, Dr. Yang Du for providing valuable feedbacks.

I would like to thank the past and current members of the GTCAD lab: Dr. Dae Hyun Kim, Dr. Daniel Limbrick, Dr. Xin Zhao, Dr. Krit Athikulwongse, Dr. Young-Joon Lee, Dr. Moongon Jung, Dr. Taigon Song, Dr. Shreepad Panth for all the guidance and advice

they provided, and Sandeep Samal, Neela Lohith, Kyung Wook Chang, Bon Woong Ku, Kartik Acharya, and Kwang Min Kim for many insightful discussions on research ideas and sharing their knowledge with me.

I would also like to thank Dr. Seung-Ho Ok, Dr. Woongrae Kim, Yang Zhang, Yang Wan, Yiyu Zhang, Zhenxuan Zhang, Yang Li, Hourieh Atarzadeh, Tianchen Wang and Can Rao for their helpful comments and collaborations. I also thank David Webb, Keith May, Peter Huynh, and Pamela Halverson for their help to resolve numerous requests from me.

Lastly, I would like to thank my family. Words cannot express how grateful I am to my parents, and grandparents for their support throughout my life. I wish them long and healthy lives. I would also like to thank all of my friends who supported me and incented me to strive towards my goal.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **iii**

LIST OF TABLES . . . . . **xi**

LIST OF FIGURES . . . . . **xiv**

SUMMARY . . . . . **xix**

I INTRODUCTION **1**

1.1 Introduction to 3D ICs 1

1.2 Existing Work 2

1.2.1 3D Power Integrity Analysis and Optimization 2

1.2.2 TSV-to-TSV Coupling Extraction 3

1.2.3 TSV-to-wire Coupling Extraction 3

1.2.4 Face-to-face Inter-die Coupling Extraction 5

1.2.5 Signal Integrity Analysis and Optimization 7

1.3 Contributions of This Work 8

1.3.1 Power Integrity Analysis and Optimization for 3D DRAM 8

1.3.2 TSV-to-TSV Coupling Extraction and Optimization 8

1.3.3 TSV-to-Wire Coupling Extraction and Optimization 9

1.3.4 Inter-die Coupling Extraction in Face-to-Face 3D ICs 9

1.3.5 Impact of Physical Design and Technology Scaling on F2F Inter-die Coupling 10

II POWER INTEGRITY ANALYSIS AND OPTIMIZATION FOR 3D DRAM 11

2.1 DRAM Benchmarks 11

2.1.1 Stacked DDR3 12

2.1.2 Wide I/O 13

2.1.3 HMC 13

2.2 CAD Platform for 3D DRAM Power Integrity 13

2.2.1 R-Mesh Model and Validation 14

2.2.2 Inter-die Impact 16
2.2.3 Memory Controller Simulator 16
2.3 Design Solutions 18
2.3.1 Stand-alone vs. Mounted on a Logic Die 18
2.3.2 Impact of TSV Count and Alignment 19
2.3.3 Impact of TSV Location and RDL 20
2.4 Packaging Solutions 20
2.4.1 Impact of Dedicated TSVs and Wire Bond 20
2.4.2 Impact of PDN Sharing with F2F Bonding 22
2.4.3 Impact of Inter-Die Spatial Locality 23
2.5 Architectural Solutions 24
2.5.1 Impact of Memory State and I/O Activity 24
2.6 Impact of the Read Scheduling Policy 25
2.6.1 Impact of IR drop on DRAM Performance 26
2.7 Cross-Domain Co-optimization 27
2.7.1 Cost and IR drop Model 27
2.7.2 Putting it Altogether: Best Solutions 29
III TSV-TO-TSV COUPLING AND OPTIMIZATION METHODOLOGIES . 31
3.1 Models for TSV-to-TSV Coupling 31
3.1.1 Two-TSV Model 31
3.1.2 Multi-TSV Model 33
3.2 Silicon and E-Field Distribution Impacts 35
3.2.1 Impact of Silicon Depletion Region 35
3.2.2 Impact of Substrate Resistance 38
3.2.3 Impact of Electrical Field Distribution 42
3.3 Full-chip Analysis 44
3.3.1 Models Used for Full-chip Analysis 44
3.3.2 Full-chip Analysis Strategies and Flow 47
3.3.3 Designs Specification 50

3.3.4 Worst case Analysis v.s. Average Case Analysis 51
3.3.5 Full-chip Substrate and Field Impact 54
3.4 TSV-to-TSV Coupling Noise Reduction Using Guard Ring 54
3.4.1 Guard Ring Model 54
3.4.2 Optimization Flow and Results 56
3.5 TSV-to-TSV Coupling Noise Reduction Using Differential TSV pair . . . 57
3.5.1 Differential TSV Impact on Modeling 57
3.5.2 Full-chip Optimization Flow and Analysis With Differential TSVs 59
3.6 TSV Noise Optimization Method Comparison 62

IV TSV-TO-WIRE COUPLING EXTRACTION AND OPTIMIZATION METHODOLOGIES 65

4.1 E-field Sharing Impact 65
4.1.1 TSV Influence Region 65
4.1.2 Multi-Wire Impact 66
4.1.3 Wire Coverage Impact 69
4.2 TSV-to-Wire Extraction Technique 69
4.2.1 Pattern Matching Technique 69
4.2.2 Line Library 72
4.2.3 Ring Library 73
4.2.4 Corner Library 74
4.3 Pattern Matching Algorithm 75
4.3.1 Single-TSV Validation 76
4.4 Extraction With Multi-TSV 79
4.4.1 E-Field Sharing With Multi-TSV 79
4.4.2 Multi-TSV Libraries 81
4.4.3 Multi-TSV Validation 82
4.5 Full-Chip TSV-to-Wire Impact 84
4.5.1 Design Specification and Analysis Flow 84
4.5.2 Full-Chip TSV-to-Wire Impact 85

4.6 Coupling Minimization 87
4.6.1 Increasing Keep-Out-Zone 88
4.6.2 Guard Ring Protection 89

V INTER-DIE COUPLING EXTRACTION METHODOLOGIES IN FACE-TO-FACE 3D ICS 98

5.1 F2F Extraction Methodologies 98
5.1.1 Die-by-die Extraction 98
5.1.2 Holistic Extraction 99
5.1.3 In-context Extraction 101
5.2 Field Sharing Analysis 102
5.2.1 Die-to-die Distance Impact 103
5.2.2 Wire Spacing Impact 104
5.3 Full-Chip Extraction Flows 105
5.3.1 Die-by-die Extraction 105
5.3.2 Holistic Extraction 106
5.4 In-Context Extraction 109
5.4.1 Technology and Design Generation 109
5.4.2 Double Counting Correction 112
5.4.3 Surface Layer Correction 113
5.5 Full-chip Extraction Results 115
5.5.1 Inter-die vs. Intra-die Breakdown 115
5.5.2 Die-by-die vs. Holistic Extraction 116
5.5.3 In-Context vs. Holistic Extraction 118
5.5.4 Impact of Interface Layer Handling 120
5.6 Full-chip Power, Performance, and Noise Analysis 122
5.6.1 Impact of Inter-die Coupling on 3D Nets 122
5.6.2 Full-Chip Power, Performance, and Noise 122
5.6.3 Summary of Various Methodologies 124

VI TOWARDS FUTURE TECHNOLOGY 128
6.1 Extraction for Heterogeneous 3D ICs 128
6.1.1 Methodology 128
6.1.2 Routing Direction Impact 131
6.1.3 Full-chip Extraction of Heterogeneous Technologies 133
6.2 Physical Design Impact 135
6.2.1 F2F Bonding Technology Settings 135
6.2.2 Design Hierarchy Choice 136
6.2.3 Routing Blockages by F2F Vias 138
6.2.4 Coupling Impact on Power Net 140
6.2.5 Coupling Impact on Clock Net 142
6.3 Logic-Memory Extraction 143
6.3.1 Context Creation Methodology 143
6.3.2 Extraction of Logic-Memory Design 145
6.4 Technology Scaling Impact 146
6.4.1 Logic-Logic Design 146
6.4.2 Logic-Memory Design 149
VII SUMMARY AND FUTURE WORK 152
7.1 Summary and Conclusions 152
7.1.1 Power Integrity Analysis and Optimization for 3D DRAM 152
7.1.2 TSV-to-TSV Coupling Extraction and Optimization 152
7.1.3 TSV-to-Wire Coupling Extraction and Optimization 153
7.1.4 Inter-die Coupling Extraction Methodology Study 153
7.1.5 Study of Physical Design and Technology Scaling 154
7.2 Future Work 154
REFERENCES 156
PUBLICATIONS 162
VITA 165

LIST OF TABLES

1 Benchmark specifications 14

2 TSV alignment impact on on-chip stacked DDR3 20

3 Comparison of TSV and RDL options in Figure 9 20

4 Impact of dedicated TSVs and wire bonding 22

5 PDN sharing impact in stacked DDR3 23

6 Impact of intra-pair overlapping in stacked DDR3 for the cases in Figure 11 24

7 Impact of Memory state and I/O activity in off-chip stacked DDR3 25

8 Impact of architectural policy in stacked DDR3. Standard policy uses tRRD and tFAW. First-come-first-served and distributed-read are denoted as FCFS and DistR, respectively. 26

9 Case study for impact of IR drop on DRAM performance in off-chip stacked DDR3 design 27

10 Cost model summary for four benchmarks 29

11 Best options for four benchmarks (see Table 10 for the meaning of abbreviations). α is the weight factor. 30

12 Coupling S-parameter comparison between our model and 3D solver. TSV dimensions in μm and error in dB. 35

13 Inductance impacts on TSV nets 46

14 Primetime model comparison 47

15 Worst case and average case comparison 49

16 Design specifications 51

17 Average case and worst case comparison on total TSV net noise (V) 53

18 Silicon and E-field impacts on total TSV net noise (V), TSV-induced delay (ns) and power (μW) increase 55

19 Full-chip coupling optimization results of two design styles 57

20 Delay comparison between COMPX4 and BUFX4 62

21 Full-chip impact of differential TSVs 62

22 Full-chip analysis comparison with guard ring vs TSV shielding 63

23 Raphael extraction results of multi-ring structures. 68

24 Library comparison 74

25 Sample layout extraction results based on the single-TSV libraries. Capacitance is reported in fF. 77

26 Single-TSV extraction comparison with different libraries, where the total capacitance from Raphael simulation is 568fF. 79

27 Full-chip simulation runtime and memory space comparison. 79

28 Coupling capacitance between victim TSV to wires. 82

29 Sample layout extraction results based on the multi-TSV libraries. Capacitance is reported in fF. 83

30 Multi-TSV extraction comparison with different libraries, where the total capacitance from Raphael simulation is 423fF. 84

31 Full-chip impact of TSV-to-wire coupling on timing, power and noise. Both designs have 4.47pF total TSV MOS capacitance and 0.74pF total TSV-to-TSV coupling capacitance. 86

32 Keep-out-zone impact on FFT64 design up to M4. 89

33 Guard ring impact on full-chip designs. 92

34 Holistic extraction of F2F coupling capacitance. Capacitance valus is in fF. 117

35 Breakdown of coupling capacitance shown in Table 34 into intra-die vs. inter-die. 117

36 Die-by-die extraction of F2F coupling capacitance. Capacitance is in fF. . 118

37 Die-by-die extraction error analysis against holistic extraction. Capacitance is in fF. 118

38 In-context extraction of F2F coupling capacitance. We use top 2 metal layers for the interface. Capacitance is in fF. 119

39 In-context extraction error analysis against holistic extraction. Capacitance is in fF. 119

40 Comparison of interface-layer handling methods. Unit of total coupling capacitance of each layer is fF. 121

41 Impact of the interface-layer count on extraction accuracy. “In-C:N” denotes in-context extraction with N interface layers per die. Capacitance is in fF. 121

42 Full-chip comparison of die-by-die (D-D), holistic (Holi), and in-context (In-C) extraction with one interface layer per die. 123

43 Holistic extraction of FFT with orthogonal top metal layers. Capacitance is in fF. 133

44 In-context extraction errors. Number of interface layers is attached after the die. Capacitance is in fF. 133

45 Holistic extraction and in-context extraction of FFT shown in Figure 72. Capacitance is in fF. 134

46 Holistic extraction of FFT with bottom die in 45nm and top die in 28nm. Capacitance is in fF. 135

47 Technology nodes and F2F specs used in our study. Values are in µm. . . . 136

48 Inter-die coupling comparison of the two T2 designs shown in Figure 73. Capacitance and wirelength values are in pF and mm, respectively. 138

49 Impact of partitioning (LDPC design). Δ is with respect to min-cut partitioning. 139

50 Coupling capacitance breakdown for signal, clock, and power nets in T2 (holistic extraction used). 140

51 Impact of PDN shielding on signal net inter-die coupling. 141

52 Impact of die-by-die (DBD) vs. holistic extraction on various full-chip metrics for T2 designs shown in Figure 73. 143

53 Parasitic extraction comparison of the 45nm logic + 28nm memory design. Units are in pF. 147

54 Full-chip timing and power comparison. Power units are in mW. 148

55 Technology trends of inter-die coupling with values in pF. The specifications are shown in Table 47. 150

56 Technology trend summary. 150

57 Parasitic extraction comparison of the 45nm logic + 28nm memory design. Units are in pF. 151

LIST OF FIGURES

1 Coupling capacitance in (a) face-to-back (F2B) and (b) face-to-face (F2F) 3D ICs. 2

2 Default configurations of four 3D DRAM designs. (a) on-chip stacked DDR3, (b) off-chip stacked DDR3, (c) Wide I/O, and (d) HMC. 12

3 Our integrated architecture/CAD platform 14

4 Generated layouts and their R-Mesh models for T2 full-chip and stacked DDR3 15

5 On-chip vs. Off-chip Results 17

6 Validation of R-Mesh against Cadence EPS 17

7 I/O activity for DRAM operation 18

8 (a) C4-TSV alignment, and (b) TSV count and alignment impact in stacked DDR3 19

9 TSV locations in 3D DRAM vs. logic and their RDL needs. (a) edge (memory) + non-center (logic), (b) center + center, (c) edge + center + RDL, and (d) center + center + RDL 21

10 Wire-bonding cross-section view: (a) F2B, (b) F2F 22

11 Four cases of the two-bank interleaving read state 23

12 Performance results for the cases shown in Table 9 27

13 Traditional circuit model of 2-TSV coupling. 32

14 Transient coupling noise analysis result verification. 35

15 TSV MOS capacitance with substrate doping of (a) $10^{15}/cm^3$, (b) $10^{16}/cm^3$ 37

16 (a) 3-TSV test structure for multi-TSV coupling analysis. (b) Depletion region effects on TSV noise, delay and power. 37

17 Multi-TSV coupling model with depletion capacitance and body resistance 40

18 (a) TSV pitch impact with body resistance (b) Body resistance impact on delay, noise and power 41

19 (a) Two-TSV structure with grounded active layer (b) Grounded active layout impact on TSV coupling capacitance and resistance 41

20 Circuit model of 5-TSV case: (a) original, (b) E-field distribution-aware model 43

21 (a) E-field distribution of 5-TSV case. (b) Coupling S-parameter comparison. 44

22 Transmission S-parameter comparison. 46

23 Noise distribution comparison in full-chip level 47

24 Full-chip noise analysis flow 50

25 Transient analysis of victim voltage 51

26 Design layout. (a) and (b) are bottom and top die of irregular placement design, respectively, (c) and (d) are bottom and top die of regular placement design, respectively 52

27 (a) Guard ring model (b) Guard ring impact on substrate ground resistance . 56

28 Noise-optimized design layout. (a) and (b) are bottom dies of irregular and regular placement design, respectively, (c) and (d) are zoom-in shots 58

29 Signal transmission using TSV: (a) single-ended, (b) differential pair. 60

30 (a) 3-TSV coupling case (b) Signal skew impact on noise when the differential pair is aggressor 60

31 Hspice simulation of 3-TSV coupling case. 61

32 Digital comparator design: (a) schematic, (b) layout in 45nm technology. . 63

33 Design optimization with differential TSVs: (a) and (b) are layouts of irregular and regular design, respectively, (c) and (d) are zoom-in shots. . . . 64

34 TSV influence region results. (a) TSV height impact. (b) TSV-to-wire distance impact. 66

35 Multi-wire impact. (a) shows HFSS structure with a TSV and four rings. (b) shows the cross-section E-field around the TSV. 67

36 Corner segment impact. (a) Simulation structure with wire segments of 0.5μm in length. (b) Extraction results of each segment. 68

37 Impact of wire coverage around the TSV on coupling capacitance. 69

38 Our combined method. (a) shows the calculation of wire coverage, (b) shows the calculation of weighted average. 70

39 Test structures for library generation. (a) A line library structure. (b) A ring library structure. 73

40 Sample extraction layouts with a TSV and their surrounding wires. (a) and (b) are areas around TSV S1 and S2, respectively. Lengths are in μm. 77

41 Gate and TSV placement results of FFT64 design with a footprint size of 380μm $\times$380μm. (a) shows the bottom die, (b) shows the top die. 78

42 Full-chip Verification using combined method. (a) extraction result comparison, (b) error distribution. 78

43 HFSS structures. (a) Single-TSV, (b) Multi-TSV. 80

44 XY-plane E-field distribution comparison. (a) Single-TSV, (b) multi-TSV. . 80

45 Coupling capacitance extraction result of Figure 43. (a) Nearest wire, (b) middle wire, (c) furthest wire. 93

46 Library structure comparison. (a), (b), and (c) show the single-TSV line library, multi-TSV line library, and multi-TSV ring library, respectively. . . 94

47 Multi-TSV extraction verification. (a) shows correlation comparison between our pattern-matching algorithm and Raphael simulations, (b) shows error histograms of different libraries. 94

48 Top metal routing comparison. Only top dies are shown. (a) Design up to M4, (b) design up to M5. 94

49 Top die layout and zoom-in shots of FFT64 designs up to M4. (a) With 2.5µm KOZ, (b) with 0.5µm guard ring. 95

50 KOZ impact on wire length usage. (a) design up to M4, (b) design up to M5. 95

51 Guard ring capacitance model. (a) shows the simulated structure (b) shows the Raphael extraction result. 96

52 Sample multi-TSV line library structure with 1.5µm guard ring. 96

53 Top die layout and zoom-in shots of shielded FFT64 designs up to M4. Only M4 routing is shown. (a) 0.5µm guard ring, (b) 1.5µm guard ring. . . . 97

54 Comparison of F2F extraction methods. (a) die-by-die, (b) holistic, and (c) in-context extraction. 100

55 Technology configurations with 1µm F2F via thickness. (a) and (b) are homogeneous with 45nm and 28nm for both dies, respectively. (c) is a heterogeneous technology, where bottom die uses 45nm and top die uses 28nm. 101

56 Raphael structure for capacitance extraction. Both the top and bottom dies contain repeated layout patterns. D denotes the die-to-die distance while w, s, and t denote wire width, spacing, and thickness, respectively. 103

57 Die-to-die distance (= d in Figure 56) impact. (a) Single capacitor extraction, A to D are nets in Figure 56; (b) total capacitance extraction. 104

58 Wire-to-wire spacing (= s in Figure 56) impact. (a) Single capacitor extraction, A to D are nets in Figure 56, (b) total capacitance extraction. 105

59 CAD flow chart of our die-by-die extraction. 106

60 Sample interconnect technologies with four metal layers. (a) Die-by-die extraction, and (b) holistic extraction. 106

61 CAD flow chart of our holistic extraction. 107

62 3D holistic design generation. 109

63 CAD flow chart of our in-context extraction. 111

64 A sample in-context interconnect technology with four metal layers. 111

65 3D in-context design generation. 112

66 Double-counting capacitance in an in-context technology with four metal layers and two interface layers per die. 112

67 Correction weight for top in-context die in a 2-tier 3D IC with two interface layers per die. 115

68 Layouts of FFT64 benchmark using four metal layers. (a) holistic, (b) in-context with 1 metal layer from the other die for the interface. 126

69 Full-chip comparison of die-by-die (D-D) and in-context (In-C) against holistic extraction (Holi) on 3D nets, each of which is represented by one dot. (a) aggressor count, (b) wire capacitance. 127

70 Three cases of for in-context extraction with one interface layer, where (a) with connectivity information of the interface layer, (b) assumes signal nets, and (c) assumes ground nets. 130

71 Inter-die decoupling impact on 3D nets. (a) shows worst-case delay and (b) shows worst-case noise. 130

72 Layout shots of the FFT design, whose top die is in 28nm and bottom die in 45nm. (a) shows the placement and (b) shows the routing. 134

73 T2 core design flavors. (a) block-level design, (b) gate-level min-cut design. 137

74 F2F via options. (a) M6 wires are heavily blocked by F2F via pads, (b) M6 routing is not blocked because of the dedicated M7 for F2F via pads. 139

75 Clock tree of T2. (a) bottom die, (b) top die. 142

76 (a) M2-M4 routing of a memory block. (b) longest path delay calculation comparison. 147

77 Memory die layout comparison. (a) Memory die in GDS format. (b) Auto-generated context die in Encounter. 148

78 Block-level T2 layouts under various technology nodes. The footprint of 28nm, 14nm, and 7nm designs are $880 \times 880 \mu m$, $560 \times 560 \mu m$, and $340 \times 340 \mu m$. 149

79 Logic-memory design with 28nm memory die. (a) logic die in 45nm, (b) logic die in 14nm. 151

SUMMARY

As one of more-than-Moore technologies, 3D ICs enable next-generation systems with much higher device density without needs for technology scaling. However, designing reliable 3D IC systems with high performance and low power consumption is a challenging task. It is difficult to deliver power to all chips with a reduced footprint. And new parasitic elements in 3D ICs require accurate parasitic extraction and detailed design analysis in the full-chip level.

The objective of this research is to quantify power and signal integrity issues in 3D ICs, and develop CAD tools and methodologies to enable reliable 3D IC designs, as well as enhance physical design quality. This includes accurate parasitic extraction, timing, power analysis, and signal-power-thermal integrity analysis and optimization for both face-to-back and face-to-face bonded 3D ICs. To achieve this goal, CAD tools and methodologies for 3D IC design, analysis and optimization flows are implemented from multiple aspects of physical designs.

In this work, first, a holistic CAD platform is proposed to address the need for accurate modeling and analyzing IR drop issues in a 3D DRAM system with several optimization methods from design, packaging and architectural policy perspective. Also, accurate extraction methods are proposed for TSV-to-TSV coupling parasitic extraction by using multi-TSV model and pattern-matching algorithm. Then, several noise-protection methods are proposed to alleviate signal coupling in 3D ICs. Further, a holistic and an in-context methodology are proposed for extraction of inter-die coupling parasitics in F2F 3D ICs with accuracy and complexity tradeoff comparisons. Last, multiple impacts from physical design and technology scaling are studied with our tool flow demonstrated on extraction of the next generation heterogeneous F2F 3D ICs.

CHAPTER I

INTRODUCTION

1.1 Introduction to 3D ICs

By allowing higher transistor count on each chip, 3D IC is a promising solution to extend Moore's Law. Through-silicon-via (TSV) is widely used for vertical interconnections in 3D IC and provides wide connections between dies bonded in Face-to-Back (F2B) style. However, TSVs also introduce new parasitic elements to 3D ICs. Figure 1(a) shows a 3D IC structure where two dies are bonded in face-to-back. Unlike other small metal vias, TSVs are hundreds of times larger and they are buried inside the silicon substrate close to transistors. This makes them very sensitive to any noise coupled through silicon substrate. The TSV coupling not only is a threat to the signal integrity and the logic functionality, but also degrades the delay and power benefits since they introduce extra capacitances and inductances.

When only two dies are bonded, Face-to-Face (F2F) bonding structure can be used. Shown in Figure 1(b), though much smaller in size, F2F vias also contribute new parasitics into the 3D ICs and introduce observable impacts on the full-chip level. With technology scaling, future 3D ICs require much more densely routed designs and a much finer 3D via pitch. This requires a much smaller die-to-die distance as a result from difficulties in fabricating vias with large aspect ratio. With a closer neighbouring die, electric fields (E-fields) from metal wires on different dies couple to each other and introduce inter-die coupling impacts. Consequently, inter-die coupling parasitics contribute more into the total capacitance, especially in a future technology generation.

On the other hand, stacking of multiple dies requires a more robust power distribution network (PDN) and IR drop limits the 3D IC system since more transistors need power

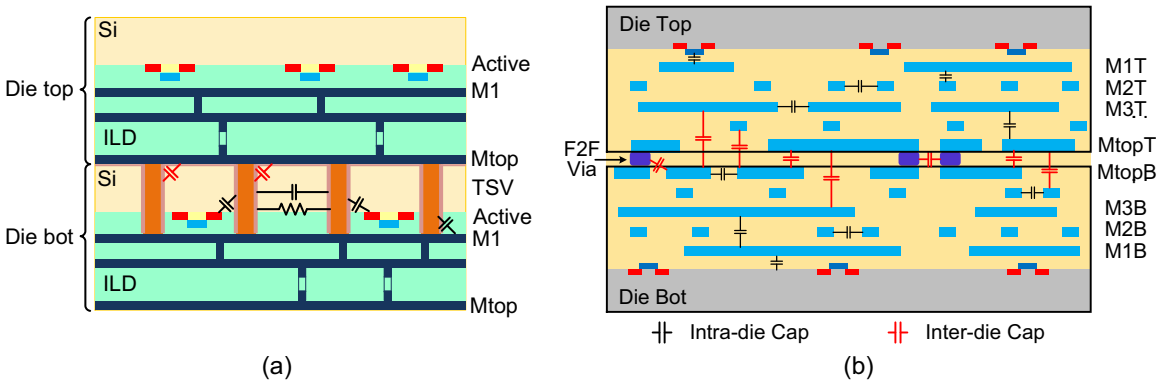

Figure 1: Coupling capacitance in (a) face-to-back (F2B) and (b) face-to-face (F2F) 3D ICs.

supply but the number of power pins are limited by the footprint, especially for 3D DRAMs which are sensitive to supply voltages for reliable functions. A simple solution is to limit the max frequency or number of parallel memory accesses so that max power is under control, but this leads to a lower system performance. Therefore, reducing IR drop while not limiting system performance is beneficial to the adaption of 3D IC designs.

1.2 Existing Work

1.2.1 3D Power Integrity Analysis and Optimization

To mitigate power delivery issues in 3D DRAM, several studies have proposed design [1] and packaging techniques. Edge TSVs are used in a stacked DDR3 design [2] to reduce power noise. Sub-bank partitioning with local decoupling capacitors is proposed in [3] to maintain DRAM regularity. Another study [4] found TSV alignment to be effective at reducing IR drop and current crowding. To achieve low power distribution network (PDN) impedance, a redistribution layer (RDL) is added between memory and logic die in [5]. From the memory controller perspective, the relationship between bank activity and IR drop in a hybrid memory cube (HMC) is characterized in [6], which proposed an optimized request scheduling policy that addresses the bank starvation problem. This policy is appropriate for designs with high vertical IR drop but has little impact on designs with many TSVs when horizontal IR drop dominates.

1.2.2 TSV-to-TSV Coupling Extraction

Some modern 3D IC designs such as a multi-die Wide I/O DRAM or a processor-memory system, group many TSVs inside a TSV farm to save area and avoid stress. For these designs, TSV-to-TSV coupling is relatively larger than other kinds of coupling as a result of close proximity between neighboring TSVs. A traditional pair-based TSV model calculates the coupling capacitance based on two cylindrical nodes. Kim *et al.* proposed analytical equations that accounted for the skin effect in high frequency [7]. Xu *et al.* proposed a compact RLCG model [8] with the depletion region. Ignoring E-field sharing from other TSVs, the pair-based model is accurate when no other conductor is next to the victim TSV. However, for multi-TSV structures, it overestimates coupling [9]. A distributed RLC-mesh model that handles a small area accurately [10], but the model is not applicable to the full-chip level with a long simulation time. Yao *et al.* employed this model into multi-ports for TSV arrays [11] and proposed a more general analytical model based on multi-conductor transmission lines to handle irregular TSV positions [12].

Since active components are buried inside the silicon substrate, both resistive and capacitive coupling paths connect TSVs with neighboring transistors. Compact modeling techniques proposed in [13] and distributed RLC circuits can be used to accurately model the electrical properties of the substrate [14]. A compact Π model that uses a resistor and capacitor pair to represent the coupling path is also proposed as a faster alternative solution [15] for TSV-to-device coupling extraction. Measurement results from [12, 16] show non-negligible TSV coupling noise captured by active devices.

1.2.3 TSV-to-wire Coupling Extraction

Previous studies on TSV-to-TSV and TSV-to-active coupling focus on a single die. Since TSVs penetrate the silicon substrate when multiple dies are stacked in 3D ICs, noise not only comes from the same die that the TSV is located on but also from a neighboring die through substrate coupling. Particularly in the full-chip level, TSV farms in the bottom die

may overlap with congested routing regions in the top die. For inter-die coupling, a parasitic element often ignored in current analyses is TSV-to-wire capacitance. Since TSVs are significantly larger and longer than regular vias, their E-fields cannot be captured with the traditional metal via handling. Because of the geometric complexity of a cylindrical TSV shape, the TSV extraction is more challenging than square vias. If a tapered TSV is used, their unevenly-distributed E-fields are even harder to extract. Moreover, unlike devices, TSVs penetrate the substrate, and their E-fields interact with many surrounding wires and components. Even for industry-standard extraction tools such as Calibre, only TSV-to-TSV coupling, not TSV-to-wire coupling, is extracted. Since active regions are connected to the power or ground supply, they form an E-field shield in the AC domain. Thus, only weak coupling exists between TSV-to-M1 in the same die. However, the other side of the substrate has no substrate connection shielding TSV E-fields, resulting in large coupling capacitance between TSVs and the top metal layer of the neighboring die. Therefore, TSV-to-wire coupling needs to be accurately extracted for timing, power, and noise analyses.

Although inter-die coupling is becoming more dominant with advanced technologies in which a thinner substrate is used to reduce the TSV dimension, few studies have addressed coupling issues from the standpoint of inter-die impact. Coupling issues affect the signal gain and resonance frequency in components such as TSV-based inductors [17] as well as performance and signal integrity in full-chip designs. Measurement results from [18] show that when the signal frequency is higher than 1GHz, inter-tier coupling is greater than intra-tier coupling. Liu *et al.* found that TSV-to-metal coupling impacts analog devices in [19] and showed that TSV-to-metal coupling has a non-negligible effect on the signal-to-noise ratio. However, extraction is limited to coupling between wires and TSV landing pads. TSV-to-wire parasitic extraction using a field solver is discussed in [20], and the author concludes that TSV-to-wire capacitance is not negligible for TSVs with a low aspect ratio. Assuming square-shaped TSVs, Kim *et al.* [21] derived an analytical solution and extended the empirical wire coupling model to handle TSV structures. Since this model is based on

closed-form formulas, calculations consume only negligible runtime. However, the model is not scalable because empirical equations handle certain fabrication technology, and curve fitting with various interconnect dimensions needs to be applied. As TSVs are fabricated in cylindrical shapes, the square-TSV assumption also introduces extra errors from geometric approximations.

Another general solution for capacitance extraction is based on a random-walk algorithm [22]. Commercial interconnect extraction engines such as Raphael NXT also use this algorithm for extraction of general structures. Random-walk-based extraction consumes shorter runtime than field solver-based extraction. However, achieving comparable accuracy requires many random walks and hops to capture the E-field distribution, especially with many conductors. As a result, a long extraction time is still needed on the full-chip level. Therefore, random-walk-based extraction is suitable for small designs for the purpose of sign-off verifications, but cannot be applied to fast parasitic extraction during placement and routing stages for timing and noise optimization purposes. Yu *et al.* proposed a random-walk-based TSV-to-wire extraction method with pre-calculated look-up tables that closely matched a field solver [23]. However, the performed extraction is based on randomly-generated layouts, so the full-chip impact from TSV-to-wire remained unknown.

1.2.4 Face-to-face Inter-die Coupling Extraction

F2F bonding with copper microbumps can achieve a die-to-die distance of 8.4μm [24]. This distance is comparable to the thickness of a regular redistribution layer (RDL) [25], which makes the coupling capacitance formed between dies observable. Advanced In-Au microbumps can reach a size of 1.6μm [26] and the gap between tiers can be reduced to 1.5μm [27], which increases inter-die parasitics significantly. Also, larger bonding pressure is required for better connection yields and lower resistance, which results in a even smaller die-to-die distance [28] and stronger inter-die coupling. Moreover, with a direct copper-to-copper bonding process [29], the thickness of the bonding interface layer and copper

pads can be reduced to less than 1μm [30]. This technology is commercialized by various foundries and packaging houses [31]. Such a close distance is similar to the thickness of inter-layer dielectrics (ILD) of the top metal layers, which makes parasitic extraction inaccurate without considering the electrical fields (E-fields) from the neighboring die.

There are many existing works on parasitic extraction methodologies for interconnects in traditional 2D ICs. These standard extraction techniques can be divided into deterministic methods, such as Finite Element Method (FEM) [32] and Boundary Element Method (BEM) [33, 34], as well as statistic methods such as Floating Random Walk (FRW) [22, 35]. Though techniques based on based on field solving or random walk can be accelerated further with an hierarchical approach with look-up tables and macro models [36], they require significant runtime on the full-chip level especially in advanced technology nodes with very fine pitch structures. Therefore, for efficiency reasons, pattern-matching based extraction are still widely used for large scale designs, with critical nets extracted using field solving.

Some recent works also demonstrate significant parasitic coupling in face-to-back bonded packages in both signal [18] and power distribution networks [37]. However, there are few existing work focusing on parasitic impacts on F2F bonded 3D IC designs, and all previous work assume a full knowledge of interconnection on both sides. The direct Cu-Cu bonding enables two dies to be tightly connected, thus the close die-to-die distance requires to consider both dies simultaneously for signal and power integrity issues [38]. To enable next generation of Heterogeneous F2F integration, it is also critical to define an interconnection interface to ensure designs from multiple sources can be integrated without violating signal integrity constraints. Though it is always possible to minimize the parasitic impacts by inserting large IO drivers with ESD protection circuits, only inter-die signal pins can be protected. For intra-die signal routing close to the die surface, parasitic components still have large impacts on its delay and noise. Moreover, for 3D ICs with many inter-die pins, large ESD cells introduce significant area and cost overhead. Therefore, one option is to well-control the driver circuit with a careful physical

design [39].

1.2.5 Signal Integrity Analysis and Optimization

Also, there are techniques to reduce the TSV-to-TSV coupling. One way is using grounded TSVs to block the coupling path between two direct-facing TSVs. Chang et al. insert 8 ground TSVs around the victim TSV [40] which is shown to be effective in noise reduction. But it introduces large area overhead because of the ground TSVs and therefore, it needs a re-placement and routing. Taigon et al. insert ground TSV into the empty spots in TSV farm to block the direct coupling [9]. Other methods include adding grounded blockages around TSVs to reduce coupling. Nauman et al. add ground plugs around TSV which are area efficient but hard to manufacture [41] because of the larger aspect ratio of these ground plugs. Jonghyun et al. use ground guard rings on device and metal layer to protect the victim TSV[42]. It is effective to reduce noise for a single TSV, however, its full-chip delay, power, and noise impacts are not studied. Differential signal is used for decades to improve signal transmission quality. Common-mode noise is rejected by the differential pair thus it provides better noise immunity. There are works that introduce differential TSVs in 3D ICs. Meng-Fan et al. use differential TSVs on a memory design[43]. A pair of TSV is studied in [44] with victim node on the substrate. Signal slew is shown to have large impact on the victim for differential pair transmission. There are also differential TSV models based on TSV-pair assumption[45, 46, 47]. However, up to 5 TSVs are studied in the model and they are only focusing on the noise impact without delay and power analysis. Also, the TSV-pair based model cannot handle many TSVs in a full-chip, thus the differential TSV impacts on full-chip level are still unknown.

Because ground TSVs also similarly impact TSV-to-wire coupling, they can also be used in TSV-to-wire coupling reduction techniques. Measurement results from [48] show that H-shaped TSVs provide better shielding than guard rings. However, these techniques

either consume large silicon areas and reduce placement utilization or require special fabrication technology and increase the chip cost.

1.3 Contributions of This Work

1.3.1 Power Integrity Analysis and Optimization for 3D DRAM

Most studies focusing on a single isolated solution are limited to face-to-back (F2B) bonding. Our goal is to conduct comprehensive research covering many key solutions from multiple domains. To accomplish this goal, we develop a cross-domain CAD platform that accurately models and evaluates DC power integrity in 3D DRAM. This work investigates the impact of logic/memory interaction, TSV and RDL optimization, wire bonding, face-to-face (F2F) bonding, and read scheduling policy on IR drop and performance. We use four modern 3D DRAM benchmarks: off-chip stacked DDR3, on-chip stacked DDR3, Wide I/O, and HMC shown in Figure 2. Our design, packaging, and architectural domain solutions are co-optimized to achieve the best solutions under IR drop, performance, and cost tradeoffs. To the best of our knowledge, this study is the first to comprehensively analyze and optimize the power integrity of modern 3D DRAMs across multiple domains.

1.3.2 TSV-to-TSV Coupling Extraction and Optimization

In this work, we proposed several full-chip level extraction and optimization techniques for TSV-to-TSV coupling: (1) We propose a new multi-TSV model that also considers the effects of silicon depletion region, silicon substrate, and E-field distribution with minimum components; (2) We propose two coupling analysis methods, for analyzing worst-case and average case TSV-to-TSV coupling, and perform a detailed extraction and analysis on the full-chip design using our multi-TSV model; (3) We perform an accurate full-chip coupling analysis considering all the silicon and field effects on two design-style, namely, regular placement design and irregular placement design showing TSV coupling impact; (4) We propose a guard-ring model and study the impact of guard-rings in full-chip level.

Based on our model, we show the impact of guard-ring on both regular and irregular placement design and show its effectiveness in noise reduction, delay, area, and design time; (5) We propose to use differential TSVs to alleviate coupling noise induced by TSV-to-TSV coupling and provide a comparison for multiple TSV-to-TSV coupling optimization techniques. With differential TSVs as signal transmission channel, the best tradeoff between silicon area and noise reduction is achieved.

1.3.3 TSV-to-Wire Coupling Extraction and Optimization

For full-chip TSV-to-wire parasitic extraction and optimization, the contributions of this work are as follows: (1) We investigate the overall impact of E-field sharing among multiple wires and TSVs in a holistic fashion; (2) We develop a fast and accurate pattern-matching algorithm that can extract hundreds of TSVs and their neighboring wires in seconds with small errors; (3) We show the full-chip timing, power, and noise impact of TSV-to-wire coupling; (4) We study the full-chip impact of two design optimization methods and show their effectiveness in reducing TSV-to-wire coupling.

1.3.4 Inter-die Coupling Extraction in Face-to-Face 3D ICs

In this work, we provide a comprehensive study on various extraction methodologies, runtime-accuracy tradeoffs, full-chip parasitic impacts for Heterogeneous F2F integration and define a practical interconnection interface to enable inter-die coupling consideration with intellectual property protection among collaborative companies. We start by introducing various methodologies for F2F inter-die coupling extraction and comparing their pros and cons using GDS-level full-chip benchmarks. Then we analyze the full-chip impacts from F2F inter-die coupling elements, with our pathfinding study into future Heterogeneous 3D ICs.

1.3.5 Impact of Physical Design and Technology Scaling on F2F Inter-die Coupling

In this work, we investigate the impact of design floorplan and partition, F2F via structure count, top metal routing direction, and technology scaling on F2F inter-die coupling. We study the impact of F2F coupling elements on both the PDN and clock-tree networks with large-scale designs. Further, we predict the trends in technology and impact of inter-die coupling as well as provide of design guidelines based on various extraction methods for both logic-logic and logic-memory designs.

CHAPTER II

POWER INTEGRITY ANALYSIS AND OPTIMIZATION FOR 3D DRAM

Modern computer systems require ever-increasing memory bandwidth and capacity. By stacking multiple DRAM dies and using through-silicon-vias (TSVs) as vertical connections, 3D DRAM becomes a promising solution that provides high memory bandwidth and capacity with low power consumption. One challenge in 3D DRAM is unreliable power delivery, the result of more devices requiring current while the number of bumps that can fit is smaller. In addition, DRAM dies are mounted on top of a processor, resulting in longer paths to the power supply.

2.1 DRAM Benchmarks

To provide wide coverage of various 3D DRAM applications, we choose stacked DDR3, Wide I/O, and HMC as benchmarks and assume that the stacked DDR3 can be configured as a separate chip (off-chip) or mounted on logic (on-chip). We use published designs references and scale power measurement results from Samsung and Micron into 20nm-class DRAM technology. To ensure that our study is both realistic and up-to-date, we obtain detailed DDR3 power maps through our industry collaborations from Samsung. Moreover, we use a full-chip OpenSPARC T2 processor in 28nm technology as the host chip. This ensures that our 3D system is complete and realistic to be fabricated. The design specifications of our benchmarks are listed in Table 1. We use stacked DDR3 as an example and results for all four benchmarks are provided in Section 2.7.

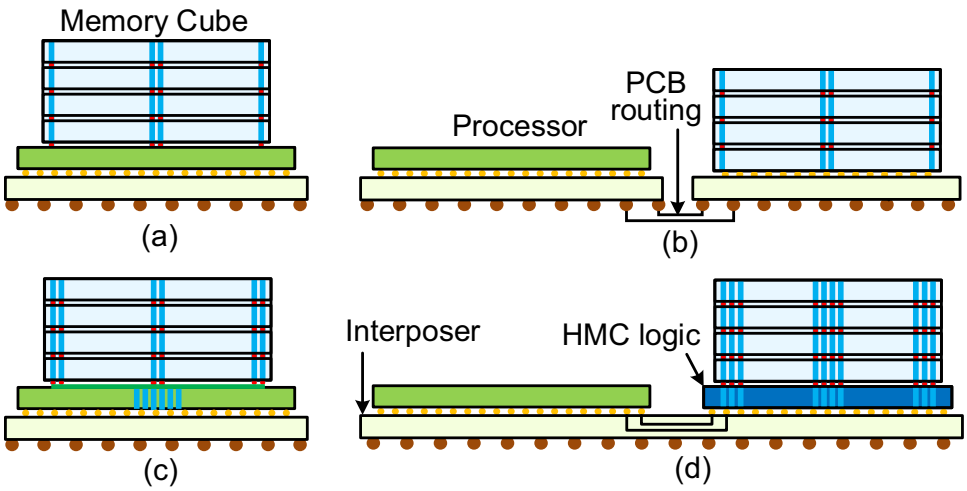

Figure 2: Default configurations of four 3D DRAM designs. (a) on-chip stacked DDR3, (b) off-chip stacked DDR3, (c) Wide I/O, and (d) HMC.

2.1.1 Stacked DDR3

Stacked DDR3, which follows the widely-accepted DDR3 specification, provides a low-cost and backward compatible solution. No re-design is needed for DDR3 memory controller thus the transition is easy. With no footprint increase, memory capacity can be easily extended. Each DRAM die remains cheap and total PCB area is saved. However, as the 3D stacking structure is not considered in DDR3 specification, the performance and power benefits from 3D IC are not fully utilized. This makes stacked DDR3 suitable for applications which requires larger memory spaces. We use Samsung's stacked DDR3 design presented in [2] as the reference. Each DRAM die includes 8 banks and form 1 rank in memory cube. Address and data buses are shared among dies and the memory cube forms a single channel. Power TSVs are located in both two edges and in the center. We assume the stacked DDR3 can be configured either as a separate chip (off-chip) or mounted on processor (on-chip).

2.1.2 Wide I/O

Wide I/O is a new JEDEC specification for 3D DRAM. With a large number of pins, Wide I/O is designed to be mounted on top of processors directly. The reduced wire load and operation frequency enable a much lower power consumption at similar bandwidth compared with DDR3 running at 1600MHz. These benefits makes it ideal for mobile applications which requires high memory bandwidth for graphics as well as long battery life. A Wide I/O prototype from Samsung [49] is used as our reference design. The memory cube is organized in 4 channels and 16 banks per die. The micro-bumps are located in the center correspond to JEDEC specification. To ensure IR drop is under control, RDL is used to supply power to the edges.

2.1.3 HMC

HMC is proposed by Micron [3] as the next generation high-performance memory solution. By arranging the memory cube into 16 volts where each volt consists of 2 banks per die, HMC dramatically increases parallelism in DRAM operation which leads to huge memory bandwidth. Due to its high performance, the memory cube is very power-hungry thus it is not suitable to be mounted on processor dies. Instead, the memory cube is mounted on top of its own logic die which handles communication with processor through silicon interposer. Because of these factors, HMC has its killer applications in GPU and servers.

2.2 CAD Platform for 3D DRAM Power Integrity

We implement an integrated CAD and architectural simulation platform shown in Figure 3. Our floorplan generator produces a block-level 3D DRAM floorplan based on the given design and architectural specifications. Then our PDN layout generator produces design files for PDN routing. Next, we perform special routes and produce a combined floorplan with both globally and locally-routed PDN using Cadence Encounter. Figure 4 presents two examples of our auto-generated layouts which are used for pre-design analysis of routing

Table 1: Benchmark specifications

Benchmark	Stacked DDR3 [2]	Wide I/O [49]	HMC [3]
Capacity	4Gb × 4 dies = 16Gb		
Stand-alone?	yes/no	no	yes
Stacked logic die	T2 (or none)	T2	HMC logic
Logic size (mm^2)	9.0×8.0	9.0×8.0	8.8×6.4
DRAM size (mm^2)	6.8×6.7	7.2×7.2	7.2×6.4
# banks per die	8	16	32
# channel	1	4	16
Speed (Mbps/pin)	1600	200	2500
Data width	8	512	512
3D IC benefit	capacity	low power	bandwidth
Target app	PC & laptop	mobile	GPU & server

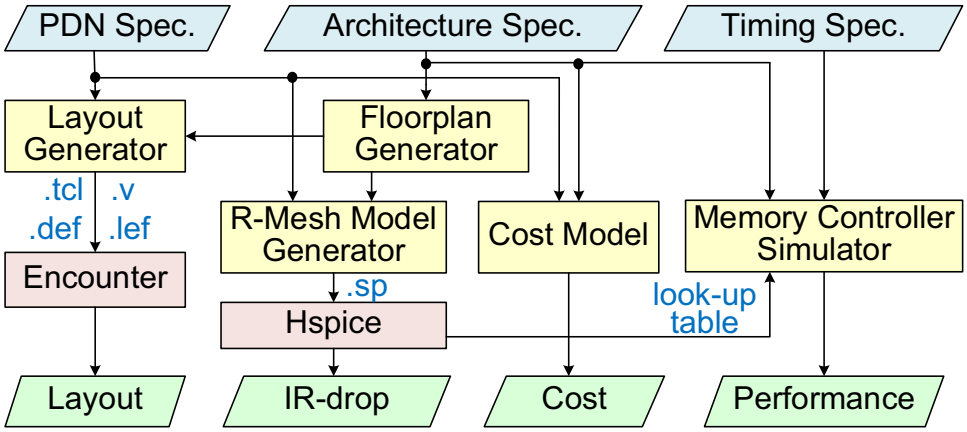

Figure 3: Our integrated architecture/CAD platform

congestion and early-stage routing planning. Lastly, we calculate the cost of design and packaging solutions that include metal usage, TSV count and location, RDL, and bonding style. We also use our memory controller simulator to obtain performance data.

2.2.1 R-Mesh Model and Validation

For IR drop calculation, we build a resistive mesh model (R-Mesh) for each metal layer based on design and technology information. PDN wire resistance is modeled depending on the metal layer usage which is defined as the area percentage of PDN on one layer. Local PDN supplies power within each block, while global PDN is used to connect them. The resistivity of each metal layer as well as its routing direction is read from the technology

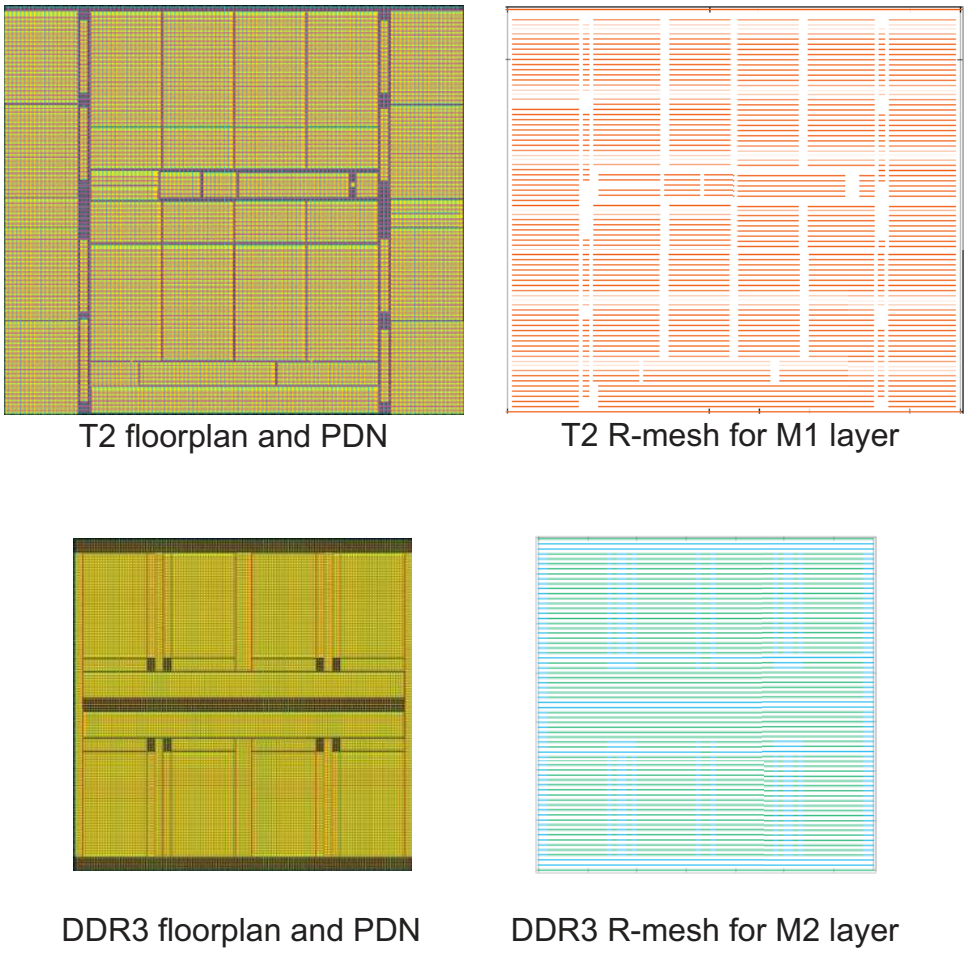

Figure 4: Generated layouts and their R-Mesh models for T2 full-chip and stacked DDR3

file. PG rings, vias, and inter-die connections are generated automatically. We use HSPICE to simulate it and calculate the IR drop. Figure 4 shows two R-Mesh model examples as well.

Since each row activation contains a write-back operation when the row is closed, we focus on read operations only. We generate a 2D DDR3 design using the aforementioned CAD method. For one bank operation, the max IR drop is 22.5mV for read and 22.4mV for write, and their IR drop distributions are similar. However, in 3D DRAM, the maximum IR drop depends on both single die operation and inter-die coupling. For naming convenience, the 3D DRAM memory state is represented as "R_1-R_2-R_3-R_4," where R_1 to R_4 are the number of active banks from the bottom DRAM die (DRAM1) to the top die (DRAM4).

The default state is 0-0-0-2 assuming zero-bubble interleaving read (IDD7) in our stacked DDR3.

To verify our R-Mesh model, we compare IR drop results with commercial tools shown in Figure 6. Using Encounter Power System (EPS) on the generated 2D DDR3 design, we perform IR drop simulation assuming that the left two banks are in the interleaving read mode. The max IR drops are 32.6mV and 32.2mV using EPS and R-Mesh, respectively. Our R-Mesh model shows only 1.3% error and achieves 517x speed up because it does not perform parasitic extraction from the layout and reduces the total resistor count. With three 3D DRAM benchmarks and proposed CAD platform, we are able to study DC power integrity in design, packaging and architectural level and obtain high quality solutions under noise, cost, and performance tradeoff.

2.2.2 Inter-die Impact

Figure 5 compares the on-chip stacked DDR3 where all power TSVs are attached to the logic die. As shown in on-chip case, the logic die IR drop is reflected on to the DRAM die through the connected TSVs and results in 1.178x larger IR drop. Also, DRAM die draws current from the logic die as well, which results in slightly increased logic die IR drop. However, since DRAM consumes much smaller power than logic, logic to DRAM reflection is the dominate factor in inter-die impact.

2.2.3 Memory Controller Simulator

For 3D DRAM operations, the IR drop constraint is a critical factor that affects the memory performance. In the standard JEDEC DDR3 specifications, two timing parameters, used to limit the maximum IR drop, are row to row delay (tRRD) and four active window (tFAW). Without considering detailed 3D stacking properties, these timing parameters limit the maximum number of banks that can be read in parallel. Thus, less parallelism reduces the maximum performance of the 3D DRAM.

To study DRAM performance, we build a 3D DRAM memory controller simulator that

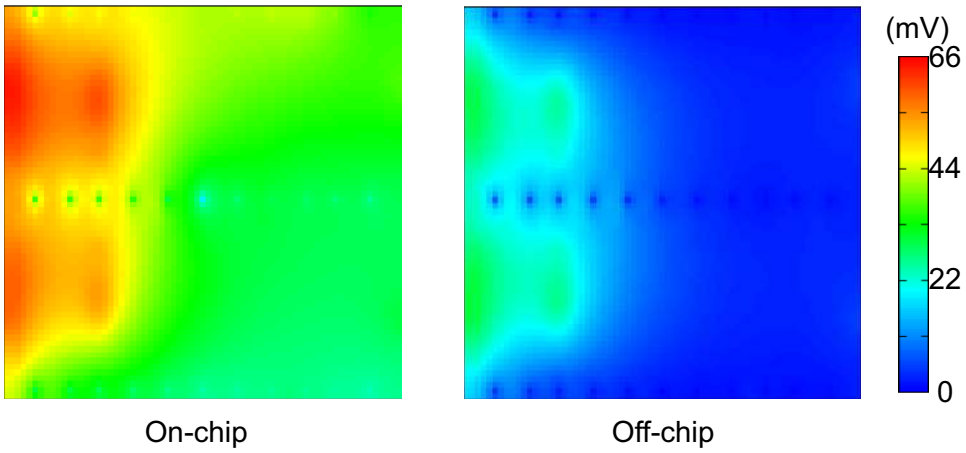

Figure 5: On-chip vs. Off-chip Results

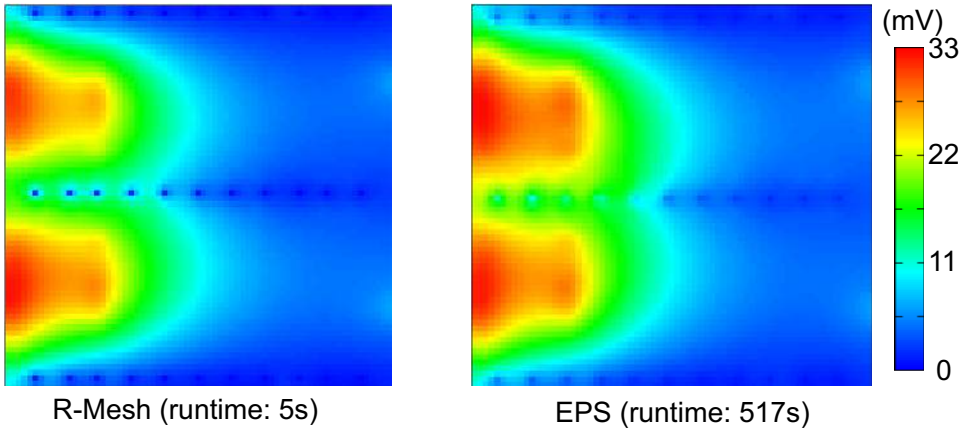

Figure 6: Validation of R-Mesh against Cadence EPS

performs cycle-by-cycle simulations for each DRAM bank and memory channel. Major DRAM read operation timing parameters such as tCL, tRCD, tRP, tRAS, and tCCD are modeled. If an active bank does not receive further read requests in a few cycles, the bank is closed to reduce IR drop. We generate 10,000 read requests with temporal and spacial locality under a row hit rate of 80%. For stacked DDR3, each read request arrives every five DRAM cycles with a burst length of eight, assuming a heavy work load. Figure 7 shows two-bank read interleaving timing diagram with tCL=8 and tCCD=4. Our memory controller has a priority queue of size 32 so that it can smartly schedule the requests for the best performance. Interleaving mode reads two banks per die in maximum to avoid current overdrawn from charge pump.

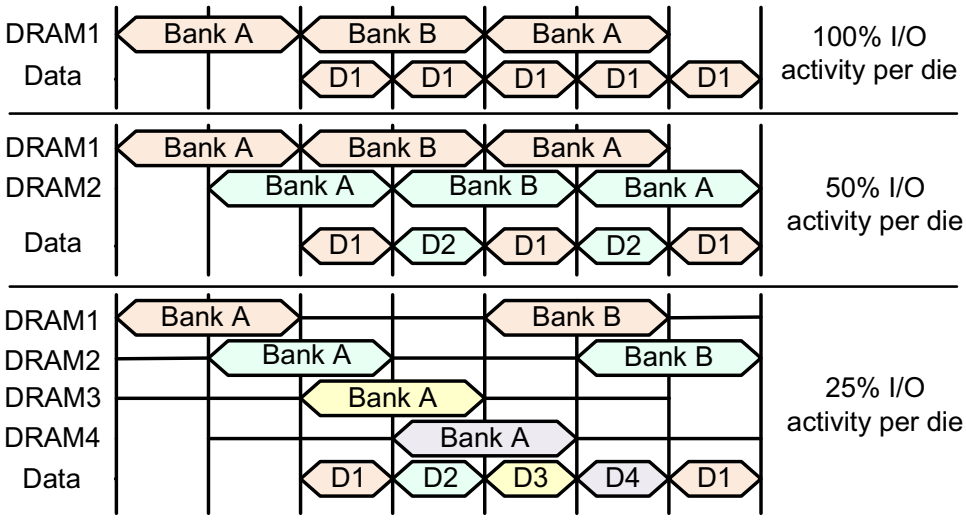

Figure 7: I/O activity for DRAM operation

2.3 Design Solutions

A traditional design technique for IR drop reduction is to increase metal usage which also applies to 3D IC. Assuming a 10% M2 usage and 20% M3 usage for VDD as baseline, with 2x PDN metal usage, IR drop is reduced more than 40% for stacked DDR3. However, the vertical IR drop becomes more significant in 3D IC. Thus, we explore unique design solutions in 3D ICs.

2.3.1 Stand-alone vs. Mounted on a Logic Die

3D DRAM can be mounted on logic (on-chip) or separated as a stand-alone chip (off-chip). For mounted memory, one solution to stabilize power supply is to add dedicated PG TSVs on the logic die. Assuming the same supply voltages of the logic and the DRAM die, power and ground nets from both dies can be connected together, thus their power noises are coupled. As results show, with a 50.05mV logic die power noise, the DRAM IR drop increases from 30.03mV in the off-chip stacked DDR3 design to 64.41mV in the on-chip design. Dedicated TSVs can be fabricated through via-last technology, which reduces TSV resistance and provides a clean power supply directly to memory dies. However, these dedicated TSVs penetrate the bottom die, occupy extra silicon area, and become routing

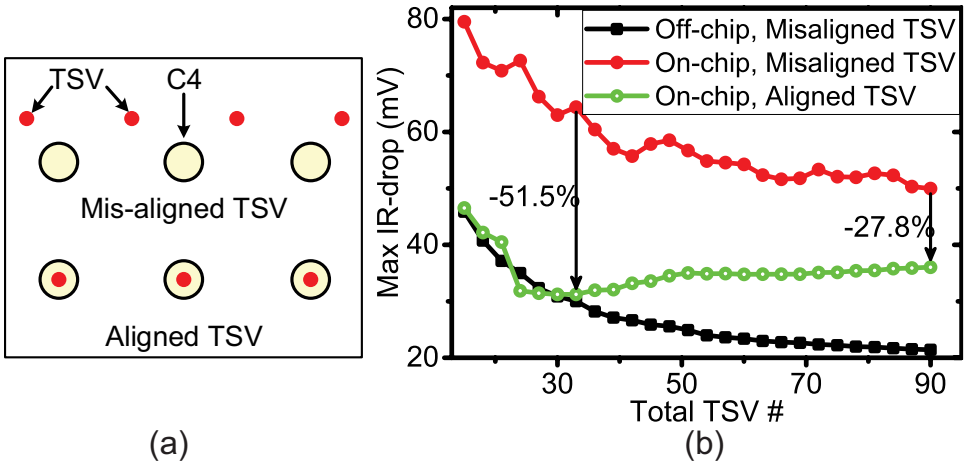

Figure 8: (a) C4-TSV alignment, and (b) TSV count and alignment impact in stacked DDR3

blockages on logic, increasing design complexity and logic die cost dramatically.

2.3.2 Impact of TSV Count and Alignment

Another intuitive design solution is to increase the PG TSV count. More PG TSVs reduce vertical IR drop and current crowding. However, if a uniform TSV pitch is assumed, not all TSVs can perfectly align with C4 bumps on the logic die. The misaligned TSV increases the inter-die coupling resulting in a higher IR drop on the DRAM die. Figure 8 compares the on- and off-chip designs with various TSV numbers. Table 2 provides comparisons between uniform but misaligned TSV and manually aligned TSV. The results show that using more TSVs reduces IR drop, but the reduction saturates with many TSVs. By carefully placing TSVs near C4 bumps on the logic die and reducing average C4-to-TSV distance, IR drop reduces by as much as 51.5% in on-chip stacked DDR3 while logic IR drop merely increases by 0.2%. More TSVs do not always guarantee a lower IR drop because of TSV misalignment, especially when the TSV count is small. For on-chip designs, increasing the TSV count leads to larger coupling from T2. Thus, the IR drop increases slightly on memory dies.

Table 2: TSV alignment impact on on-chip stacked DDR3

Total TSV#	27		33	
TSV aligned?	No	Yes	No	Yes
Avg C4-TSV distance (μm)	337.87	0	326.84	18.182
Max Logic IR drop (mV)	49.97	49.98	49.97	50.05
Max DRAM IR drop (mV)	66.28	31.46	64.41	31.24
Δ%	+2.9%	-51.2%	baseline	-51.5%

Table 3: Comparison of TSV and RDL options in Figure 9

Design option	(a)	(b)	(c)	(d)
Logic die cost	High	Low	Medium	Medium
DRAM die cost	High	Low	High	Medium
Overall cost	Highest	Lowest	High	Medium
IR drop(mV)	30.03	50.76	38.46	49.36

2.3.3 Impact of TSV Location and RDL

Various TSV design considerations affect the max IR drop. Edge TSVs can significantly reduce the IR drop by shortening the power supply path. However, dedicated edge TSVs introduce much higher cost to both logic and DRAM because large keep-out zones (KOZs) must be inserted around TSVs to avoid stress and noise issues. A low-cost solution called "center TSV" groups all TSVs into the center of the die and does not block routing on the logic die. To alleviate the high IR drop, the RDL can be added as a back-side routing layer. A RDL can be inserted only between logic and bottom DRAM die or on all dies. Figure 9 shows four design options, and Table 3 compares their tradeoffs between cost and IR drop. Center TSV without an RDL has the lowest cost but highest IR drop. Replacing edge TSVs with an RDL reduces cost but introduces higher power noise because of additional RDL resistance.

2.4 Packaging Solutions

2.4.1 Impact of Dedicated TSVs and Wire Bond

In addition to design techniques, advanced packaging solutions also help improve power integrity in 3D DRAM. To alleviate the inter-die impact shown in Section 2.3.1, dedicated

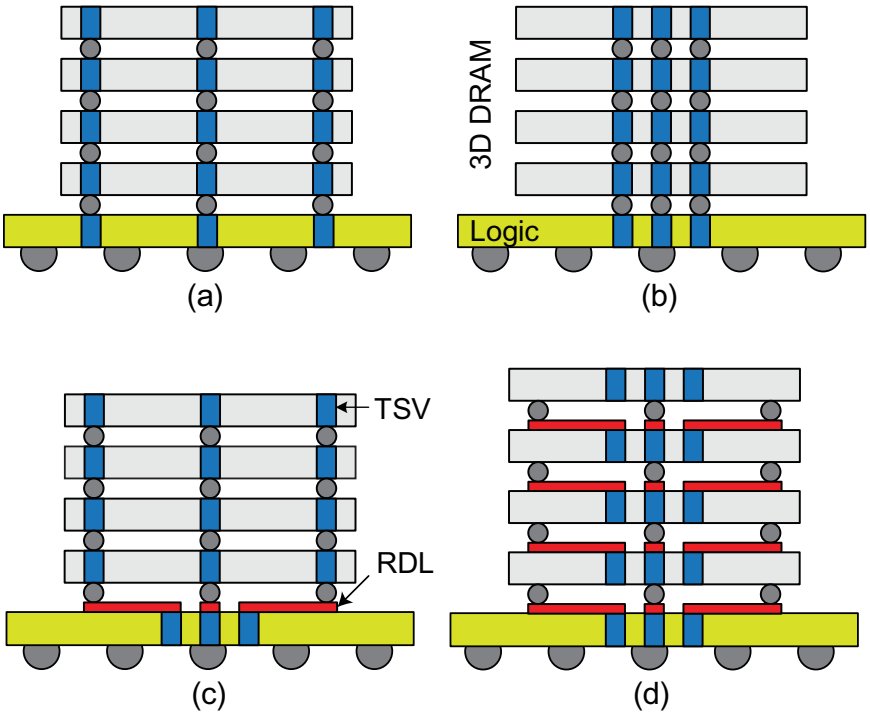

Figure 9: TSV locations in 3D DRAM vs. logic and their RDL needs. (a) edge (memory) + non-center (logic), (b) center + center, (c) edge + center + RDL, and (d) center + center + RDL

TSVs can be used to directly deliver power to the DRAM dies. With this packaging solution, the logic and the DRAM PDNs are fully decoupled, which results in an IR drop similar to that of the off-chip design.

In a 3D DRAM design, layouts of all DRAM dies are kept identical so that all memory dies share the same fabrication process, which improves the yield and cost. By taking advantage of the backside metallization process, additional metal pads for wire connections are formed on the backside. Figure 10 (a) shows the proposed packaging solution with wire bonding. Signal TSVs are used for low-power and high performance, and PG TSVs are used to supply power between memory dies. However, with backside wire bonding, an extra power delivery path is built from the top to the bottom die. With this method, the maximum IR drop reduces, and bonding wires can directly connect to large off-chip decoupling capacitors, which provide better AC power integrity. Table 4 summarizes impact of dedicated and wire bonding on the stacked DDR3 design. Both dedicated TSVs

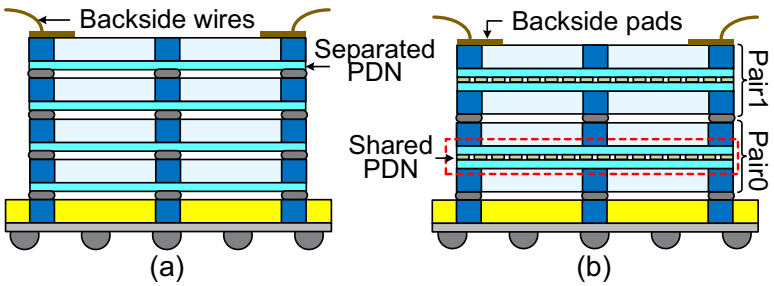

Figure 10: Wire-bonding cross-section view: (a) F2B, (b) F2F

Table 4: Impact of dedicated TSVs and wire bonding

Design	Dedicated TSV?	IR drop (mV)		
		Baseline	Wire-bonded	Δ%
On-chip	no	64.41	30.04	-53.4%
On-chip	yes	31.18	27.18	-12.8%
Off-chip	yes	30.03	27.10	-9.76%

and wire bonding reduce the IR drop as much as 50% for on-chip designs. However, since both wire bonding and dedicated TSVs provide direct power supply, a combination of both technologies provides only marginal additional benefits.

2.4.2 Impact of PDN Sharing with F2F Bonding

Another packaging technique also takes the advantage of layout regularities in 3D DRAM. Thus, by changing the die orientation of DRAM1 and DRAM3, F2F bonding can form between the two bottom dies and the two top dies. F2F vias can be placed almost everywhere, thus, PDNs of two F2F-bonded dies are tightly connected. In this way, a pair of F2F-bonded dies share their PDNs together. F2F bonding can also be used in combination with wire bonding, as shown in Figure 10 (b), and provides even larger IR drop benefits.

Unlike the F2B design, in which each DRAM die uses two metal layers for PDNs, a pair of DRAM dies in the F2F design can use four metal layers together. This feature, called PDN sharing, provides additional IR drop benefits. If one die in a pair is idle while another is active, the active die can use all four PDN layers. With PDN sharing, the IR drop of the idle die increases but leads to a significant IR drop reduction for the whole system. For example, under the 0-0-0-2 memory state, the overall maximum IR drop with

Table 5: PDN sharing impact in stacked DDR3

Die #	Pair #	Total banks being read	power (mW)	IR drop (mV) F2B	IR drop (mV) F2F+B2B
DRAM1	Pair 0	0	27	3.09	3.11
DRAM2		0	27	4.45	3.11
DRAM3	Pair 1	0	27	5.74	17.18
DRAM4		2	229.5	**30.03**	**17.18 (-42.8%)**

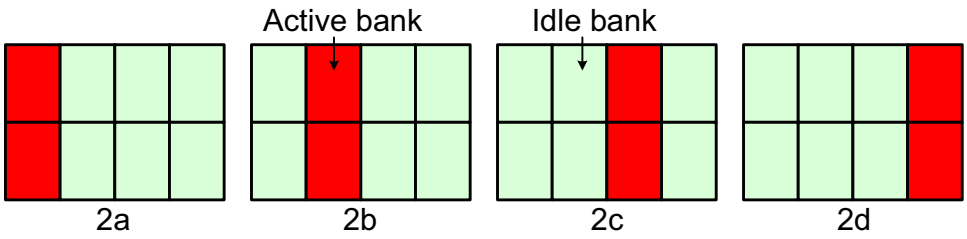

Figure 11: Four cases of the two-bank interleaving read state

F2F bonding decreases by 42.8% and 41.1% compared with F2B bonding in off-chip and on-chip stacked DDR3, respectively. Table 5 details this impact from PDN sharing.

2.4.3 Impact of Inter-Die Spatial Locality

The memory state has a large impact on F2F benefits as well. For example, Figure 11 shows four cases from the top-down view for the two-bank interleaving read mode, and Table 6 shows IR drop results. If two dies of a pair have active banks in the same location, it is called "intra-pair overlapping." With intra-pair overlapping, the current is congested in a small area, and both dies do not have extra PDN resources to share. Results also show that if the active regions on two dies are separated further, the IR drop reduction is larger with less current congestion. If active banks overlap in different pairs, the impact on the IR drop is small since PDNs between pairs are separated. Thus, F2F provides IR drop benefits over F2B, especially for designs with low bank activities and low probability of intra-pair overlapping. To avoid inter-pair overlapping, IR drop-aware read scheduling policies can rearrange bank activities so that the probability of inter-pair overlapping remains low.

Table 6: Impact of intra-pair overlapping in stacked DDR3 for the cases in Figure 11

Memory state	Intra-pair overlapping	Max IR drop (mV)		
		F2B	F2F+B2B	Δ%
0-0-2a-2a	yes	28.14	27.21	-3.3%
0-0-2b-2b		18.06	17.42	-3.5%
0-2a-0-2a	no	27.32	15.24	-44.2%
2a-0-0-2a		26.51	15.24	-42.5%
0-0-2b-2a	no	27.38	17.98	-34.3%
0-0-2c-2a		27.04	17.10	-36.8%
0-0-2d-2a		26.86	15.27	-43.1%

2.5 Architectural Solutions

2.5.1 Impact of Memory State and I/O Activity

If the IR drop is not considered during memory operations, the memory controller can activate as many banks as possible if there is no timing violation or bus conflict. However, parallel reading is always limited for power integrity concerns, especially in 3D DRAM. However, since the standard read policy is not aware of 3D stacking, simply limiting row activation pessimistically constrains parallel operations. Moreover, as balanced reads increase parallelism in 3D DRAM without IR drop overhead, distributing read requests evenly achieves the best tradeoff between the IR drop and performance.

Assuming zero-bubble reading, if more DRAM dies are activated, I/O activity per die decreases. Table 7 lists IR drop simulations for various cases. For the 0-0-0-2 state, 25% I/O activity reduces die power by 44.7%, which leads to 23.64% and 22.99% IR drop reductions for F2B and F2F+B2B designs, respectively. Moreover, if the read activity is balanced among dies (e.g., the 2-2-2-2 state), more banks can be activated in parallel, and the maximum IR drop of that state is even smaller than the 0-0-0-2 state with 100% I/O activity. In addition, worst IR drop cases for F2B and F2F differ. For F2F design with PDN sharing, the 0-0-0-2 state does not cause high IR drop. However, because of the intra-pair overlapping effect, the 0-0-2-2 state becomes the worst case. Compared with F2B, F2F reduces the worst-case IR drop by 9.4%.

Table 7: Impact of Memory state and I/O activity in off-chip stacked DDR3

Memory state	IO activity per die	Power (mW) active die	Power (mW) total	IR drop (mV) F2B	IR drop (mV) F2F+B2B
0-0-0-2	100%	229.5	310.5	30.03	17.18
2-0-0-0			310.5	26.26	14.61
0-0-0-2	50%	175.5	256.5	26.42	15.15
0-0-2-2			405.0	28.14	27.21
0-0-0-2	25%	126.9	207.9	22.93	13.23
2-2-2-2			507.6	24.82	23.57

2.6 Impact of the Read Scheduling Policy

From the perspective of performance, if the IR drop is not considered during memory operations, the memory controller can activate as many banks as possible if there is no timing violation or bus conflict. However, parallel reading is always limited for power integrity concerns, especially in 3D DRAM. However, since the standard read policy is not aware of 3D stacking, simply limiting row activation pessimistically constrains parallel operations. As shown in Section 2.5.1, impact of unique memory and I/O activity requires a detailed IR drop-aware policy for optimum performance. Moreover, as balanced reads increase parallelism in 3D DRAM without IR drop overhead, distributing read requests evenly achieves the best tradeoff between the IR drop and performance.

We propose IR drop-aware read policies based on a detailed look-up table. With our fast and accurate R-Mesh model, the max IR drops of each memory state with various I/O activities are saved in a look-up table read by the memory controller for read request scheduling. For each cycle, the memory controller checks all read requests in the priority queue and tries to send a request to each DRAM channel. Under a given IR drop constraint, the read request that can be sent to memory must satisfy all timing and IR drop constraints. This read policy is compared to JEDEC DDR3 standard policy with a tRRD of right and a tFAW of 32. Moreover, two request scheduling policies are implemented. One is called first-come-first-served (FCFS), and another is called distributed-read (DistR). For FCFS, the memory controller assigns a higher priority to the read request which comes in first.

Table 8: Impact of architectural policy in stacked DDR3. Standard policy uses tRRD and tFAW. First-come-first-served and distributed-read are denoted as FCFS and DistR, respectively.

IR drop policy	Standard	Our IR drop-aware policy	
Scheduling policy	FCFS	FCFS	DistR
IR drop constraint	none	24mV	24mV
Runtime (us)	109.3	84.68 (-22.6%)	75.85 (-30.6%)
Bandwidth (read/clk)	0.114	0.148 (+29.2%)	0.165 (+44.2%)
Max IR drop (mV)	30.03	23.98 (-20.2%)	23.98 (-20.2%)

For DistR, the memory controller tries to balance the read across multiple DRAM dies to increase die-level parallelism under the IR drop constraint. Thus, the read request, whose target die has the least number of active banks, has the highest priority.

Table 8 compares the performance of three read scheduling policies based on the F2B stacked DDR3 design. We set the IR drop constraint for our IR drop-aware policies to 24mV. With bank activation constraints, the standard policy results in a longer runtime and a lower average bandwidth. With a detailed IR drop look-up table, the memory performance improves by 22.6%. Furthermore, by taking advantage of DistR and balanced workloads, the performance improves by 30.63%. The maximum IR drop of our policy also decreases by 20.15% compared to the standard policy since memory states with high IR drops are avoided. Note that scheduling policy has a small impact if the IR drop constraint is high or the bank activity is low. In both cases, not the IR drop but single-bank performance becomes the system bottleneck.

2.6.1 Impact of IR drop on DRAM Performance

Since design and packaging optimizations reduce the IR drop, allowed memory states differ for various designs under the same IR drop constraint. Table 9 lists a few examples. With our memory simulator, impact of various IR drop optimization methods on performance is studied. Figure 2.6.1 shows runtime needed to finish all read requests. If the IR drop constraint is too tight, it allows no memory state. With a relaxed IR drop constraint, more states are allowed. Therefore, the memory controller can send more parallel read requests. As

Table 9: Case study for impact of IR drop on DRAM performance in off-chip stacked DDR3 design

Mounting style	off-chip			on-chip		
Case #	1	2	3	4	5	6
Bonding style	F2B	F2B	F2F	F2B	F2B	F2F
PDN metal usage	1x	1.5x	1x	1x	1x	1x
Wire bonding	no	no	no	no	yes	no
Max IR drop (mV)	30.03	22.15	17.18	64.41	30.04	65.43

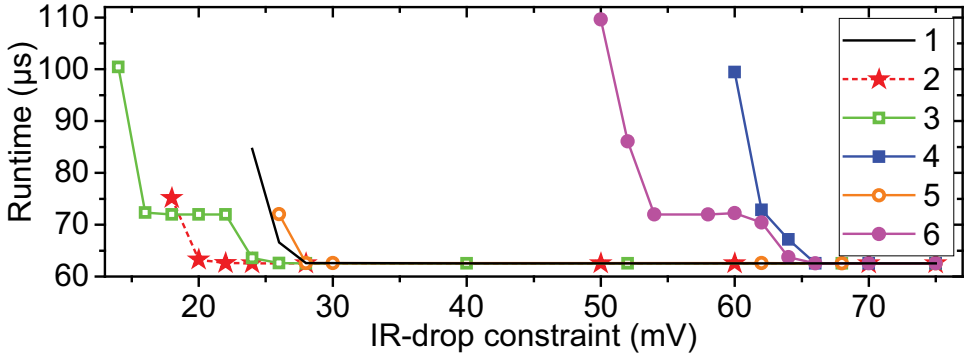

Figure 12: Performance results for the cases shown in Table 9

results show, all IR drop optimization methods are able to improve memory performance under a certain IR drop constraint. Interestingly, although the F2F design (Case 3) reduces the worst-case IR drop only by 9.4%, it outperforms the F2B design with 1.5x PDN (Case 2) with an IR drop constraint smaller than 18mV because PDN sharing shows larger benefits when bank activities are low. Therefore, F2F has a higher tolerance to low IR drop constraints.

2.7 Cross-Domain Co-optimization

2.7.1 Cost and IR drop Model

An intuitive way to lower the IR drop is using every solution available. However, this approach leads to a very expensive design with marginal IR drop benefits. Therefore, co-optimization of the IR drop, performance, and cost is critical to provide overall guidelines. We propose a cost estimation model with every technology parameter included as a cost term. Table 10 lists these cost terms. Except for the TSV count (TC), the cost of which is

calculated by a square root function, other terms are proportional to inputs. An input range ensures a realistic solution. For the Wide I/O design, the power TSV count is fixed at 160, which matches specifications. For stacked DDR3 and Wide I/O designs, only center and edge TSVs are options. For HMC, resulting from a high power consumption, PG TSVs are placed between banks. We call this TSV location style "distributed TSV." The minimum power TSV count is 160 for sufficient supply current.

To explore the whole design knobs in a short time for the purpose of finding the best PDN design, we built regression model with regular interval samples. We used Matlab stepwise function to build a general polynomial model, and additional error fitting with radial basis network. We set M2, M3 usage and Log scaled TSV count as inputs. Log scaling improves model accuracy to cover large TSV range with small samples. The reason why we chose 3 variables as inputs even though there exists other terms considered in Cost Model is that these terms such as TL, TD, BD, RL, WB are binary, which means they cannot be treated as continuous variables in regression model. Therefore, we made combination table of those terms to cover whole design knobs, and built a regression model group consists of each of regression models for each combination. Catching the global minimum point of each regression model is made by Matlab genetic algorithm, and we get the best PDN design by comparing with minimum points.

For technology co-optimization, brute-force searching for every combination in one benchmark takes 4637 hours on a four-core system. To reduce runtime, we choose a few sample cases for M2, M3, and TC, because they are continuous variables. For other optimization options, we search all valid combinations. After performing R-Mesh simulations on the sample cases, we use MATLAB regression analysis to obtain an IR drop model with a root mean square error (RMSE) of less than 0.135 and an R^2 of larger than 0.999. With the regression analysis, total runtime decreases to ten hours. Combined with total cost estimation, we define an IR-cost term by

$$IR\text{-}cost = IR\text{-}drop^{\alpha} \times Cost^{1-\alpha}, \tag{1}$$

Table 10: Cost model summary for four benchmarks

Solution	Abbreviation	Input Range	Cost Range
M2 VDD usage	M2	10%-20%	0.025-0.05
M3 VDD usage	M3	10%-40%	0.025-0.10
Power TSV #	TC	15-480	0.078-0.44
Dedicated TSV	TD	Yes/No	0.06/0
Bonding style	BD	F2B/F2F	0.045/0.06
RDL layer	RL	Yes/No	0.05/0
Wire bonding	WB	Yes/No	0.03/0
TSV location	TL	Center only (C)	0
		Edge and center (E)	0.5×TC
		Distributed (D)	TC

where $\alpha \in [0,1]$ is the weight factor. We perform MATLAB global optimization to obtain the best solutions. With α=0, we found the lowest cost solution, while α=1, the lowest IR drop solution.

2.7.2 Putting it Altogether: Best Solutions

Table 11 summarizes the best solutions for all four 3D DRAM designs. As expected, using no optimization option results in the lowest cost but the highest IR drop. By gradually increasing α, results show the priority of each optimization option. We achieve optimal tradeoff with α=0.3. Since packaging solutions such as wire bonding and F2F bonding are low-cost solutions but able to reduce IR drop significantly, they have higher priority. Because increasing the TSV count yields only a marginal gain but increases the cost significantly, placing more TSVs on a DRAM chip is unnecessary. The RDL is not a good option for the lowest IR drop. However, for Wide I/O design, since the specifications require that all PG pumps be located in the center, edge TSVs must be paired with RDL for interface connections. With edge TSVs, the IR drop can decline to below 20mV for the stacked DDR3 and the Wide I/O designs. However, only with distributed TSVs for HMC can the same IR drop be achieved. Because of the likelihood of inter-die overlapping, the F2F benefit declines in HMC. However, distributed TSVs are preferable for the stacked DDR3 and the Wide I/O designs.

Table 11: Best options for four benchmarks (see Table 10 for the meaning of abbreviations). α is the weight factor.

α	M2 (%)	M3 (%)	TC	TL	TD	BD	RL	WB	IR drop (mV) Matlab	IR drop (mV) R-Mesh	Cost
Stacked DDR3, off-chip											
0	10	10	15	C	Y	F2B	N	N	88.73	88.73	0.23
0.3	20	22	24	E		F2F	N	N	22.75	23.01	0.37
0.5	20	40	63	E		F2F	N	Y	11.3	11.07	0.54
1	20	40	360	E		F2F	N	Y	9.733	9.540	0.87
Baseline	10	20	33	E		F2B	N	N	30.03	30.03	0.35
Stacked DDR3, on-chip											
0	10	10	15	C	N	F2B	N	N	117.6	117.6	0.17
0.3	20	22	21	E	N	F2B	N	Y	25.51	27.09	0.32
0.5	20	40	60	E	Y	F2F	N	Y	11.61	11.36	0.53
1	20	40	420	E	Y	F2F	N	Y	9.864	9.843	0.92
Baseline	10	20	33	E	Y	F2B	N	N	31.18	31.18	0.35
Wide I/O											
0	10	10	160	C	N	F2B	N	N	110.1	110.2	0.35
0.3	20	40		E	Y	F2F	Y	Y	4.864	4.841	0.73
0.5	20	40		E	Y	F2F	Y	Y	4.864	4.841	0.73
1	20	40		E	Y	F2F	Y	Y	4.864	4.841	0.73
Baseline	10	20		E	Y	F2B	Y	N	13.56	13.62	0.62
HMC											
0	10	10	160	C	N	F2B	N	N	459.7	459.7	0.35
0.3	20	25	160	D	Y	F2B	N	Y	18.63	18.65	0.76
0.5	20	36	160	D	Y	F2B	N	Y	17.66	17.62	0.78
1	20	40	480	D	Y	F2B	N	Y	13.76	13.84	1.17
Baseline	10	20	384	E	Y	F2B	N	N	47.90	47.90	0.77

CHAPTER III

TSV-TO-TSV COUPLING AND OPTIMIZATION METHODOLOGIES

Through-Silicon-Via (TSV) is a popular choice to implement the vertical connections between dies in 3D ICs. However, TSVs also introduce new parasitic elements to 3D ICs. TSV is insulated from the substrate with a oxide liner which forms a MOS capacitor and it introduces a large capacitance coupling to other signals. Also, TSVs are hundreds of times larger in area compared with metal vias, which makes them as victims to many signal aggressors. The TSV coupling not only is a threat to the signal integrity and the logic functionality, but also degrades the delay and power benefits from 3D IC. Thus it is essential to have the coupling capacitance on TSVs extracted accurately for design verification, especially on critical 3D nets such as clock and power supplies. TSV-capacitance extraction should be a fast procedure so that during Place&Route stage, the TSV coupling information can be passed to the layout tools to perform 3D-aware delay and SI driven design optimization. Field solver tools can perform a detailed extraction on arbitrary structures, but the long simulation time and large memory requirement make it inappropriate for the full-chip extraction. Therefore, in this work, a fast and accurate extraction model is proposed for full-chip TSV-to-TSV coupling extraction.

3.1 Models for TSV-to-TSV Coupling

3.1.1 Two-TSV Model

The traditional 2-TSV model used in in [40] is based on a pair of parallel wires. Figure 13 shows the circuit components of this model. A pair of resistor and capacitor is

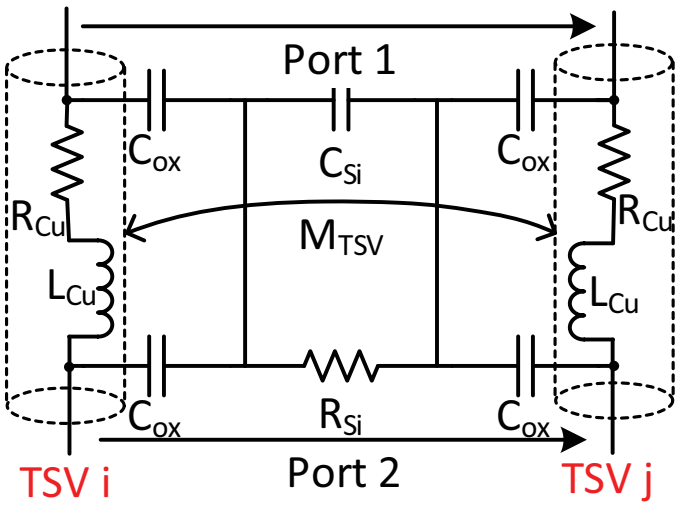

Figure 13: Traditional circuit model of 2-TSV coupling.

used to model the E-field coupling between two TSVs. On the full-chip level, the TSV-to-TSV coupling is calculated based on TSV pairs without considering impacts from neighbor TSVs.

This model assumes that there are no other object which either blocks the coupling path between TSVs or shares the E-field around. This assumption is satisfied only when there are two TSVs buried in the silicon substrate. The model is accurately matching with the measured results of a TSV-pair [50]. It is a good approximation even if the impact of devices and metal layers are considered since they affect only part of the E-field between TSVs. However, the TSV-pair assumption is no longer valid if there are other TSVs around. Those TSVs will share E-field all the way through silicon substrate. In real 3D designs, there are many TSVs around and all of them affect the E-field distribution within that area. Ignoring the multi-TSV effect, the traditional 2-TSV model overestimates the coupling between a TSV pair, and a new model which takes all the neighboring TSVs into account is desired for the full-chip analysis.

3.1.2 Multi-TSV Model

From the 2-TSV model, the total coupling capacitance on victim TSV increases linearly with number of aggressor TSVs if the coupling path distances are the same for all aggressors. However, from Raphael [51] simulation results, when there are 4 and 8 aggressor TSVs around the victim, the total coupling capacitance merely increases to 194% and 199% compared with 2-TSV case, respectively. As shown in [52], there is an upper bound of the coupling capacitance calculated by the capacitance of coaxial wire, given by the following formula:

$$C_{max} = \frac{2\pi\varepsilon_{si}L}{\ln(P/r)}, \tag{2}$$

where P and r is the outer and inner radius of the coaxial wire. In this case, r is the radius of the victim TSV and P is the minimum distance between other aggressors and the victim. According to (2), even when there are many aggressor TSVs, the total coupling capacitance cannot be more than 226% of the coupling between a TSV pair whose distance is P.

To model multiple TSVs, the multi-TSV model presented in [52] is used. By assuming the aggressor TSV array as the multi-conductor transmission line within silicon substrate and the victim TSV as the ground signal [53], the TSV-array inductance matrix $[L_{Si}]$ is computed by applying the following formula:

$$L_{Si,ij} = \begin{cases} \frac{\mu_{Si}L_{TSV}}{\pi}\ln\left[\frac{P_{i0}}{R_{TSV}+T_{ox}}\right] & \text{when } i = j \\ \frac{\mu_{Si}L_{TSV}}{2\pi}\ln\left[\frac{P_{i0}P_{j0}}{P_{ij}(R_{TSV}+T_{ox})}\right] & \text{when } i \neq j \end{cases}, \tag{3}$$

where P_{ij} is the distance between aggressor TSV i and j and the victim is labeled 0. Note in this formula, unlike the 2-TSV model, not only the distances between aggressor and the victim are considered, but also the distances between aggressor are considered. This makes it useful for any TSV placement style even when TSVs are not placed on a regular grid. By using the relation of homogeneous material between the capacitance matrix and the inductance matrix [54], the capacitance matrix for TSV array is calculated by:

$$[C_{Si}] = \mu_0\varepsilon_{Si}L_{TSV}^2\,[L_{Si}]^{-1}. \tag{4}$$

Since we focus on the coupling on victim TSV, only the coupling components between aggressor TSV i and the victim is used, which is given by:

$$C_{Si,i0} = \sum_{k=1}^{N} C_{Si,ik}. \tag{5}$$

Assuming a homogeneous substrate, the relationship between substrate coupling resistance and capacitance is given by:

$$R_{Si}C_{Si} = \varepsilon_{Si}/\sigma_{Si}. \tag{6}$$

Note that there are also coupling paths between aggressor TSVs, which is given by $C_{Si,ij}(i \neq j)$. However their impact on the victim TSV is small. This is because each aggressor is connected to a strong driving source with a full VDD swing, which is much larger than other coupling noise. The voltage waveforms of the aggressors are not affected much by the coupling. Previous work [53] used all of the coupling components between TSVs which is not a feasible solution in full-chip level. E.g., an array of 100 TSVs leads to more than 20000 RC components in the model. Therefore, in our work, the coupling paths between aggressors are ignored. To verify our model, many test cases are generated containing up to 8 TSVs and transient SPICE simulations are performed. 10 layouts are generated for each sample cases. Because of the large runtime and memory space required for field solver simulation, we cannot perform simulations with dozens of TSVs. However, since the test cases mainly contain TSVs which are facing directly to each other, they represent the main contributors of the coupling capacitance and noise in the full-chip level. Model calculated using our equations is compared with extraction results from field solver in frequency domain, and the maximum error on coupling S-parameter is reported in Table 12. We also perform a transient analysis in a 3-TSV case and the voltage waveform of a victim TSV is shown in Figure 14. The results show that for all tested layouts, the coupling parameter error of our extracted model is less than 0.02dB and we conclude our multi-TSV model accurately handles multi-TSV effects and is scalable with different TSV dimensions.

Table 12: Coupling S-parameter comparison between our model and 3D solver. TSV dimensions in μm and error in dB.

TSV radius	TSV height	TSV liner width	Max error
2	30	0.2	0.016
		0.5	0.011
2	60	0.2	0.017
		0.5	0.012
4	30	0.2	0.015
		0.5	0.014
4	60	0.2	0.018
		0.5	0.013

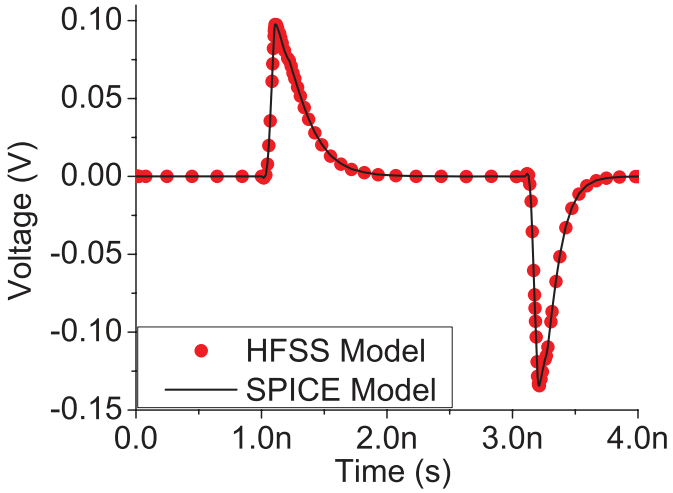

Figure 14: Transient coupling noise analysis result verification.

3.2 Silicon and E-Field Distribution Impacts

3.2.1 Impact of Silicon Depletion Region

In this section, the impact of silicon depletion region on TSV coupling is discussed. TSV, usually made of copper or tungsten, is insulated from the silicon substrate with a oxide liner, which together form a MOS structure. Due to the non-linearity of the MOS capacitance, many previous works [40, 52] ignore the depletion region around the TSV and assume that the oxide capacitance is the only part that contributes to the TSV capacitance. Also, the simulation tool is based on field solver (HFSS [55]) which does not take silicon semiconductor effects into consideration. To study the depletion region impacts, a TSV structure is built in device simulator Synopsys Sentaurus [56]. The TSV MOS capacitance extraction

result is shown in Figure 15 with different substrate doping concentration. Copper TSV and P type substrate are assumed for our simulation. The flat band voltage is calculated by the following formula:

$$V_{FB} = W_{Cu}/q - \varphi_{Si} - Q_s/C_{ox}, \tag{7}$$

where W_{Cu} (= 4.65eV), φ_{Si}, and Q_s are work function of copper, Fermi level of the silicon, and the charges inside oxide liner, respectively. Thus, for most digital systems, when the voltage on TSV is between 0V and VDD, a depletion region always exists around the TSV and it introduces a voltage dependent capacitance C_{dep}, calculated by:

$$C_{dep} = \frac{\pi \varepsilon_{Si} L_{TSV}}{\ln \frac{R_{TSV} + T_{ox} + W_{dep}}{R_{TSV} + T_{ox}}}. \tag{8}$$

While digital system usually has a clock running at several hundreds of MHz, it is safe to assume a complete depletion around TSV as in Figure 15. This is because the substrate is slightly doped and there are not enough carriers which can respond to such high signal frequency. Yang et al. proposed a simplified close form formula to calculate the depletion width in [57] :

$$W_{dep} = \frac{2\varepsilon_{Si}}{3\varepsilon_{ox}}(-T_{ox} + \sqrt{T_{ox}^2 + \frac{3\varepsilon_{ox}^2}{\varepsilon_{Si}} \frac{V_{TSV} - V_{FB}}{qN_a}}), \tag{9}$$

where V_{TSV} and N_a are TSV voltage and substrate doping concentration, respectively.

There are other works [58, 8] using numerical method to solve partial differential equations (PDEs) and get the depletion width. Direct solution from PDEs can be more accurate than the close-form formula with simplification, but they may lose in terms of the flexibility when TSV layout changes. Therefore, to avoid convergence issue in numeric solution, we use a close-form formula in our work. After considering the depletion region, the oxide thickness T_{ox} in (3) should be replaced by $(T_{ox} + W_{dep})$.

The following equation is used to calculate TSV MOS capacitance which is the serious combination of oxide capacitance (C_{ox}) and depletion capacitance (C_{dep}):

$$C_{MOS} = \frac{C_{ox} C_{dep}}{C_{ox} + C_{dep}} = \frac{\pi \varepsilon_{ox} L_{TSV}}{\ln \frac{R_{TSV} + T_{ox} + W_{dep}}{R_{TSV}}}. \tag{10}$$

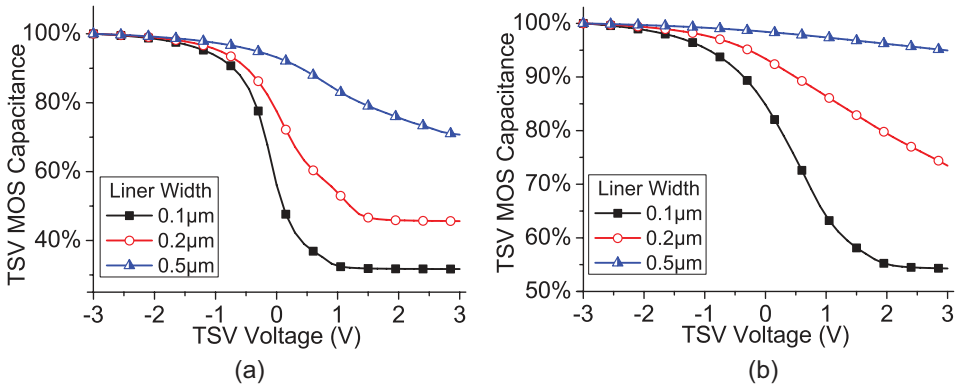

Figure 15: TSV MOS capacitance with substrate doping of (a) $10^{15}/cm^3$, (b) $10^{16}/cm^3$

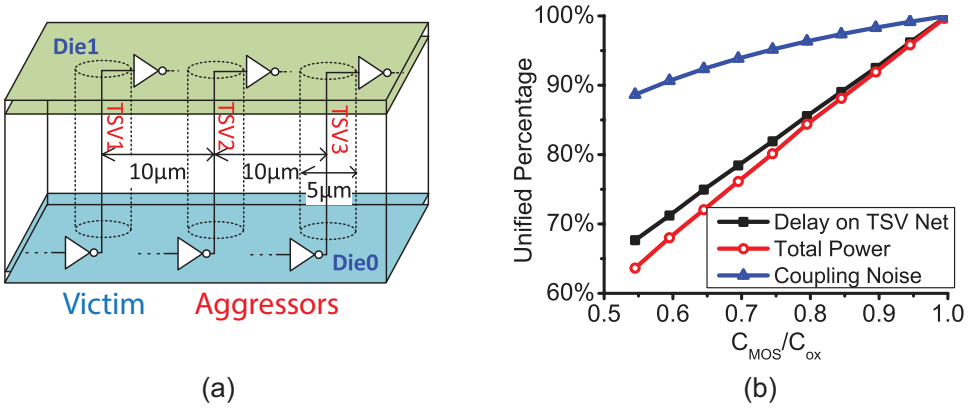

Figure 16: (a) 3-TSV test structure for multi-TSV coupling analysis. (b) Depletion region effects on TSV noise, delay and power.

The TSV MOS capacitance reaches its minimum value after the silicon is strongly inverted. The MOS capacitance depends on the substrate doping and liner thickness. With a thin liner and a lightly doped substrate, the depletion region impact is more significant, especially if TSV is scaled down in the future technology node. From Figure 15 the MOS capacitance can be as low as 36% of the oxide capacitance, thus it is overestimated if the depletion region is ignored. This results in a pessimistic estimation on TSV-induced delay and noise. Another observation is that the MOS capacitance becomes smallest when TSV voltage is tied at VDD while it reaches maximum value when TSV is grounded.

To find out how large is the depletion impact on TSV-induced delay, power and noise, a test structure with 3 TSVs is built. The structure is shown in Figure 16(a). TSVs have 5μm pitch. The substrate has $10^{15}/cm^3$ doping concentration. Each TSV is driven by an INVX4

and is driving an INVX4 as load. The victim TSV is driven at ground while the aggressors are switching. The switching delay and dynamic power on the aggressor TSV and the coupling noise on the victim TSV are measured in HSPICE. The HSPICE [59] simulation result on 3-TSV test structure is shown in Figure 16(b). Since the MOS capacitance has the largest capacitance value in the coupling model, any variation of the depletion region width has a large impact on the coupling noise and timing result. Also, as the TSV-induced delay and power are directly related to the load capacitance, they almost increase linearly with the MOS capacitance. The coupling noise, on the other hand, depends not only on the MOS capacitance, but also on the coupling and load capacitance as well as driving strength. It reduces by 13% if MOS capacitance is only half of the oxide capacitance. Highly doped substrate makes it difficult for the MOS capacitor to reach the strong inversion and the maximum depletion width. Therefore the depletion region impact is smaller and TSV-induce delay and noise are larger.

Though wide depletion region helps reducing TSV coupling noise and increasing performance, it increases the Keep-out-zone (KOZ) around TSV. Devices within depletion region are observed with a threshold voltage shift and performance difference. To prevent undesired side-effects introduced by TSVs, a 1μm region to avoid more than 10% performance variation, especially for smaller technology node [60].

3.2.2 Impact of Substrate Resistance

Since TSV is buried in doped silicon substrate, the substrate impact needs to be considered. Previous models used in [61, 52] assume the silicon substrate is a floating net. This assumption is not appropriate since most designs ground the substrate using substrate contacts. Even though each TSV has a KOZ, there is a finite impedance from substrate around the TSV to the ground node. Whenever a victim TSV is affected by the aggressor, charges will accumulate at silicon-oxide interface. With a finite silicon impedance, the MOS capacitance can be discharged through the discharging path of the substrate. Therefore, the

coupling noise on the victim TSV is reduced. Especially when the RC time constant of the discharging path is small and aggressor switching frequency is low, the accumulated charges can be quickly discharged even when the aggressor signal is still switching. Therefore, the peak noise voltage on the victim reduces due to fewer charges. The traditional model assumes a floating net at silicon substrate and therefore, overestimates the coupling noise on victim TSV since there is no discharging path. But it underestimates TSV-induce delay and power as the capacitance of the discharging path is also ignored. Therefore, the discharging path needs to be modeled using substrate resistors and capacitors. Figure 17 illustrates our proposed multi-TSV model with components to model silicon and E-field effects, where C_{Sig} and R_{Sig} represent the silicon substrate capacitance and resistance, respectively between TSV and the substrate contact to model the charging path.

To extract the substrate resistance and capacitance, a TSV structure with grounded substrate is built. The capacitance between TSV and substrate is extracted using Synopsys Raphael and the substrate resistance is evaluated using (6). Figure 18(a) shows the result comparison on the test structure with or without the silicon discharging path impact. The coupling noise value with $8\mu m$ TSV pitch is used as a reference. If the substrate is assumed to be floating, the coupling noise is largest for all different TSV pitches. Smaller body resistance makes the discharging path stronger and therefore reduces noise more. The substrate impact is more significant with larger TSV pitch. This is because if TSV-to-TSV distance is large, the coupling capacitance between TSVs is much smaller than the TSV MOS capacitance, so any E-field sharing between TSVs has large impact on the coupling and reduces more noise. Without considering substrate discharging path, the TSV pitch is found to have a small impact on the coupling noise [40]. However, the TSV distance becomes an important factor in TSV coupling with substrate impact considered and spreading the TSVs is more effective in noise reduction if the substrate is well grounded.

A 3-TSV test structure shown in Figure 16(a) is built to study the substrate impact on TSV-induced delay, power and noise. HSPICE simulation results are shown in Figure 18(b)

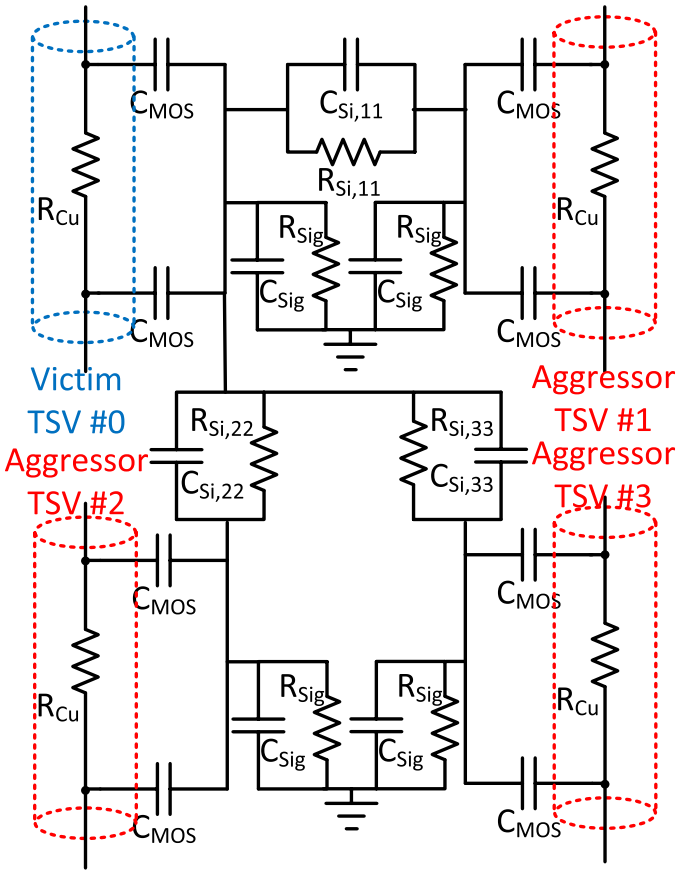

Figure 17: Multi-TSV coupling model with depletion capacitance and body resistance

and baseline is the simulation with 30KΩ body resistance. With smaller body resistance, according to (6), the substrate capacitance increases. Therefore, the delay and power on TSV noise increase while the coupling noise decreases.

Previous discussions only consider the substrate resistance impact on the E-field between TSV and the substrate. Furthermore, if the actual physical geometry of the grounded active region is considered, it has impact on the E-field between TSVs as well. E.g., a grounded active region is placed between two TSVs, it will reduce the coupling between two TSVs. This is because the active region shares some of the E-field, and part of the E-field between a TSV pair will be decoupled by the grounded region. This effect further reduces the crosstalk between TSVs. To study this impact, a structure with two TSVs is built. In this structure, a square grounded active region is placed between TSVs. Figure 19 shows the structure with TSV location held constant and the extraction results from

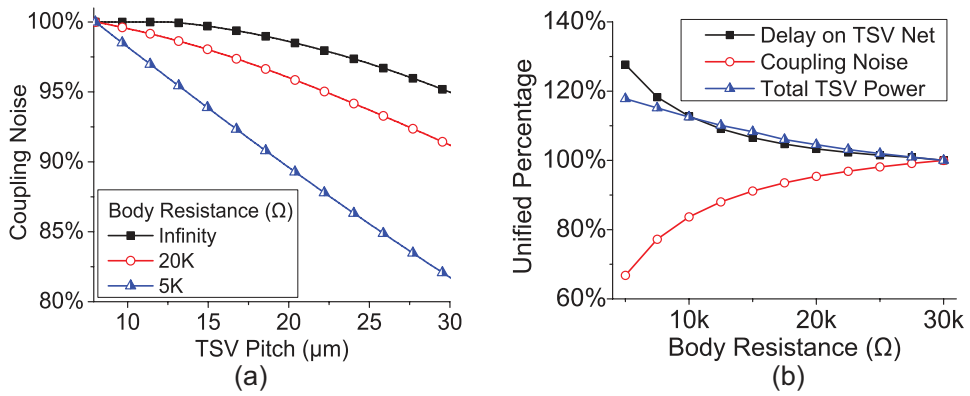

Figure 18: (a) TSV pitch impact with body resistance (b) Body resistance impact on delay, noise and power

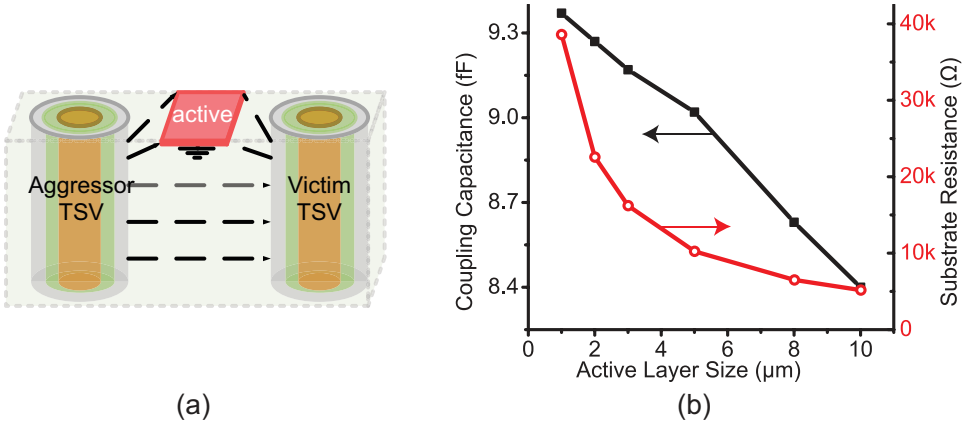

Figure 19: (a) Two-TSV structure with grounded active layer (b) Grounded active layout impact on TSV coupling capacitance and resistance

Raphael. Depending on the size of the grounded active layer and the distance between two TSVs, a maximum reduction of 9.6% and 87.1% exists in TSV coupling capacitance and body resistance, respectively. Smaller TSV coupling capacitance and larger substrate resistance indicate a weaker coupling path between TSVs. In this simulation, the TSV locations are kept the same, therefore the noise reduction comes from two aspects. Larger active region shares more E-field and leads to weaker coupling between TSV. Also, smaller distance between active layer and the TSV leads to a stronger discharging path to the ground. In general, if the victim is properly protected by the ground, it suffers less from the noise but more from the performance loss due to larger ground capacitance.

3.2.3 Impact of Electrical Field Distribution

In previous works, all of the coupling components connecting other TSVs share a single node around victim TSV which is connected to TSV net by the MOS capacitor. This model assumes that all sides of the TSV is electrically a same node. However, in real case, since the silicon substrate is not a perfect conductor, the electrical properties on different sides of the victim TSV are not the same. Moreover, the coupling between TSVs is mostly between two sides which is directly facing each other and there is few coupling on other sides. Especially in multi-TSV case, where each victim TSV is facing many aggressor TSVs in multiple directions, the E-field will be shared heavily. Consider a 5-TSV case which is shown in Figure 20, where there are 4 TSVs placed on each side of TSV V_1 and the E-field around the victim TSV is distributed among each aggressor. In this case, only neighbor TSVs are strongly coupled and there is only a weak coupling between TSV A_2 and TSV V_2 due to the E-field blocking effect of TSV V_1. Shown in Figure 20, the traditional model uses a common node P to connect all the coupling path from other aggressors. This creates a direct coupling path between TSVs which are weakly coupled. With the common node P, aggressor A_2 is directly coupling with victim V_2 through path B-P-D, which results pessimistically estimation in TSV-coupling. Figure 21(a) illustrates the HFSS simulation on E-field distribution of this structure. It is clearly seen from the plot that the coupling from each aggressor is mainly through one of the four sides of the TSV V_1, and there is fewer coupling between other sides of the victim and the aggressor because of the distributed E-field. In the traditional model, a single node is used for all sides of the TSV coupling, and it makes the coupling noise stronger since it assumes the coupling noise affects TSV on all sides at the same time. Therefore, it over estimates the coupling noise on the victim TSV.

To model the impact of the E-field distribution, 4 nodes around victim TSV are used to connect the coupling parameters to aggressors, shown in Figure 20(b). Regardless of

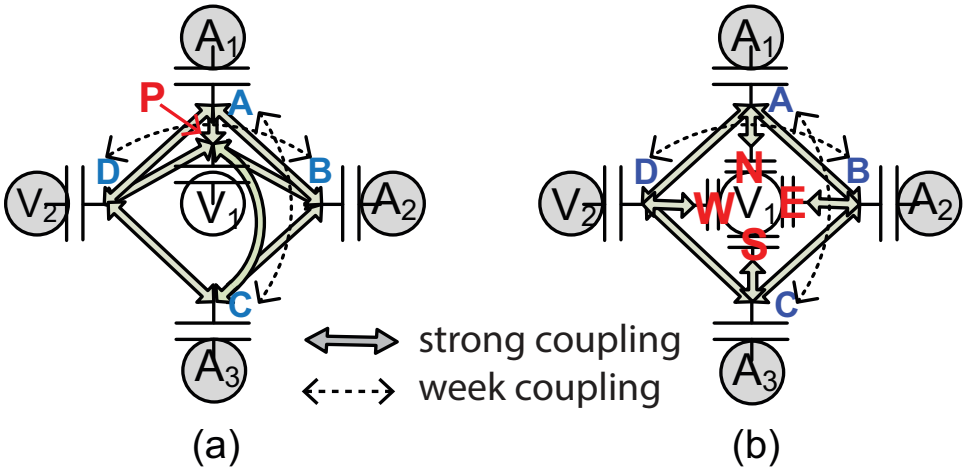

Figure 20: Circuit model of 5-TSV case: (a) original, (b) E-field distribution-aware model

the aggressor TSV number, 4 nodes are used to model the coupling E-field on their facing side of the area. Therefore, the connections of the aggressor will be attached to the corresponding node to consider E-field distribution around each TSV. Similar assumptions can be found in mesh-structure based TSV model [42, 44] where 4 nodes are used to consider the E-field distribution. Using more nodes is possible to consider more complicated E-field distribution, but the the conductance between TSV nodes needs to be considered as well. Depending on the relative location of aggressor TSVs, the coupling path will be connected using the facing node of the victim TSV. Therefore, the direct coupling path between weakly coupled TSV is eliminated in the new model. Figure 21(b) shows the coupling parameters of the circuit model compared with the results extracted using HFSS field solver. The result indicates overall both model match well with the field solver results on the coupling noise. But there is a 1.1dB over-estimation in coupling noise due to the direct path between TSVs in the original model. Our model shows smaller errors up to 15GHz not only in noise magnitude but also in noise phase compared with the original one. Therefore, we conclude that our model is more accurate to reflect the E-field distribution impact in TSV-to-TSV coupling.

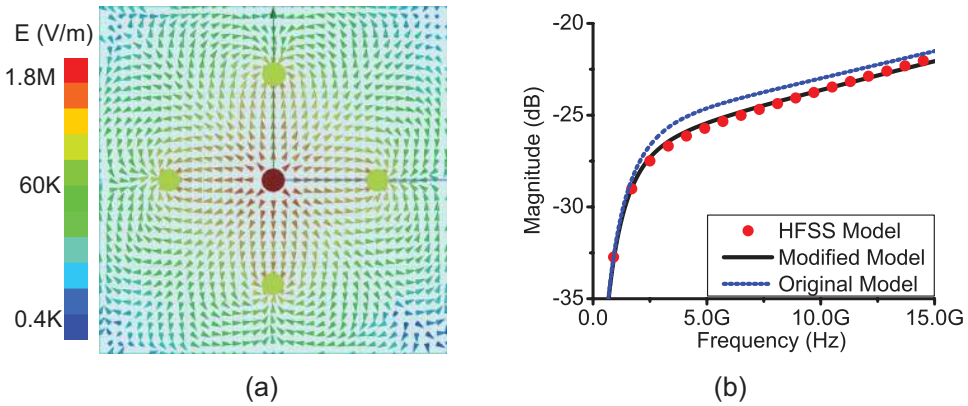

Figure 21: (a) E-field distribution of 5-TSV case. (b) Coupling S-parameter comparison.

3.3 Full-chip Analysis

3.3.1 Models Used for Full-chip Analysis

In the original multi-TSV model in [53], the number of components is too large to be simulated efficiently in circuit solver. Therefore, it is not a feasible solution in the full-chip level where simulation time and memory usage are big concerns. On the other hand, the widely used static timing analysis engines, such as Primetime [62], reject circuits with floating nets and inductors. Moreover, they cannot output a detailed voltage waveform and assume each net is driven at a certain logic level. To be able to perform full-chip analysis, we need to simplify the full circuit model of TSV-to-TSV coupling while still maintain the model accuracy.

First, the impact of TSV inductors is studied. To precisely model TSV-coupling, the inductors are included to model the magnetic field coupling between TSVs. In high-frequency range, ignoring the inductors lead to S-parameter discrepancy because the impedance of the inductors are comparable to the resistance of the TSVs. In the meantime, the mutual inductors contribute to the coupling between TSVs, and noise will be under-estimated if TSV mutual inductors are ignored. As shown in Figure 21(b), the SPICE model can be verified against field solver up to 15GHz which covers most analog circuit operation range. However, in a frequency range below 5GHz, like in most digital systems, the impact of the inductors are almost negligible in terms of noise, delay and

power. Within this range, the impact from capacitance and resistance dominates the coupling between TSVs. Table 13 lists the HSPICE simulation results on our 3-TSV test structure (shown in Figure 16). The results indicate that the inductors can be ignored while a good estimation on TSV-induced delay, power and noise is maintained. Therefore, to reduce simulation components, the multi-TSV model without TSV inductors is used in our full-chip analysis.

Second, a model which is compatible with the static timing analysis engine is proposed. Synopsys Primetime is used for full-chip timing and power analysis. There is a traditional TSV-to-TSV Primetime model used in [40, 52]. This model is derived from the SPICE model but it ignores the TSV MOS capacitors (C_{MOS}) so that the floating net between TSVs are eliminated. However, it under-estimates TSV-induced delay and power consumption since TSV MOS capacitor is much larger than coupling capacitor. Moreover, this model ignores the substrate impact and assumes a floating substrate. In our approach, a substrate net is added into the verilog netlist as the grounded substrate and the substrate capacitance is included. In addition, since substrate coupling capacitor is smaller by one-order magnitude compared to the TSV MOS capacitor (C_{MOS}). Therefore, it is ignored in full-chip analysis. Without the coupling capacitance, this model is not suitable for noise analysis especially in high-frequency regions where the capacitance dominates the coupling. However, it can be used for delay and power analysis as they are mainly affected by low frequency response. Figure 22 shows the transmission S-parameter comparison results up to 5GHz. Note that the transmission S-parameter is used instead of the coupling S-parameter as this model is not used for coupling noise analysis but for delay and power estimation using Primetime. HSPICE transient simulation result is shown in Table 14. Since the capacitance mainly affects high frequency range, both of the results show ignoring the substrate coupling capacitor gives a good estimation of the TSV coupling and compared with the original Primetime model, our modified model has a smaller error compared with the original model used in [40, 52].

Table 13: Inductance impacts on TSV nets

	wo/ TSV coupling	w/ inductor	wo/ inductor
Rise delay (ps)	22.63	168.05	168.06
Fall delay (ps)	11.92	108.88	108.96
Power (μW)	3.47	21.058	21.059
Peak noise (mV)	0	27.06	27.64

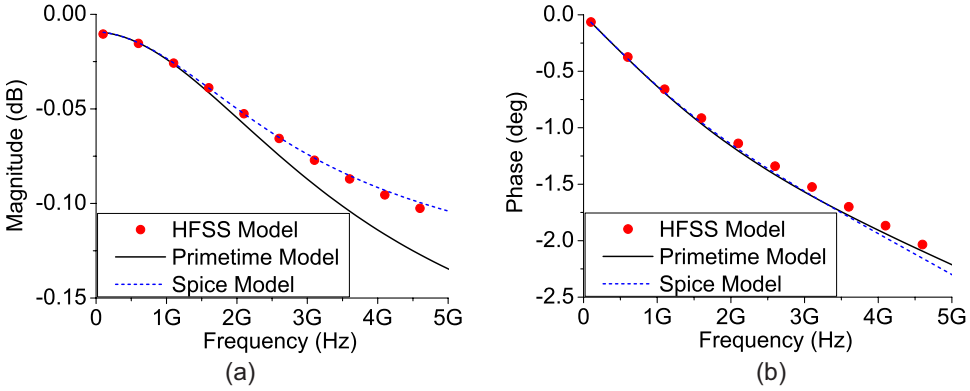

Figure 22: Transmission S-parameter comparison.

The comparison between the traditional 2-TSV model and our multi-TSV model is shown in full-chip level. The same number of aggressor TSVs is assigned around a victim TSV so that different models can be compared fairly. We also consider the E-field and silicon effects and compare the total coupling capacitance and resistance values. Figure 23 shows the noise distribution comparison between 2-TSV and multi-TSV model on a 3D design with 328 TSVs simulated. Since the 2D parasitics are the same for both models, the noise on 2D net is the same. As shown from the results, by using the 2-TSV model, the TSV net noise is much larger than that using multi-TSV model. One reason is the 2-TSV model overestimates the coupling capacitance between TSVs, and another is because it ignores the depletion, substrate and E-field distribution impact. Since our design is operating at 200MHz, TSV MOS capacitor dominates the coupling between TSVs within this range. However, using the 2-TSV model gives a total TSV net noise of 139.4V, which is 48.0% larger than total noise measured (94.2V) using our multi-TSV model.

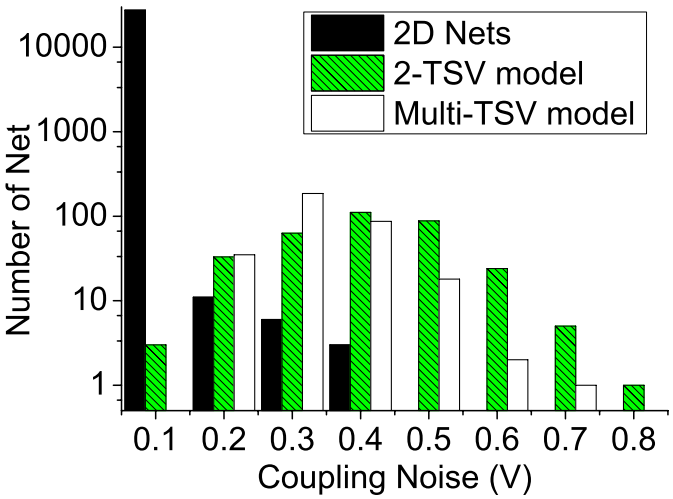

Figure 23: Noise distribution comparison in full-chip level

Table 14: Primetime model comparison

Body resistance (Ω)		0	1K	5K	10K
Multi-TSV model	Power (μW)	96.47	96.32	93.64	89.65
	Timing (ps)	54.0	45.5	40.0	39.1
Without C_{Si}	Power (μW)	96.47	93.64	93.67	89.87
	Timing (ps)	54.0	45.7	39.7	38.6
Without C_{MOS}	Power (μW)	70.24			
	Timing (ps)	37.7			

3.3.2 Full-chip Analysis Strategies and Flow

For full-chip analysis, we first extract TSV locations and 2D parasitics for each die separately from Cadence Encounter. Then a RC parasitic network is generated for all the TSVs using our multi-TSV model. The flow reported in [52] is updated, where TSV capacitance is calculated on one TSV after another. However, since the calculation of multi-TSV model gives the coupling capacitance between all TSV pairs, the runtime spent on TSV coupling capacitance calculation can be saved by using all the coupling information. In our flow, all TSVs are considered at the same time thus every coupling capacitor is computed in a single run. For our design with 330 TSVs, the original flow uses more than 13s, while our flow takes less than 2s on a XEON-E5 CPU. Note that if the number of TSV considered is the same, each calculation flow produces the same results for TSV coupling capacitance. Our

calculation flow has a great speedup compared with the traditional flow.

After the TSV coupling model is calculated, SPICE netlists as well as a top-level SPEF file are generated containing TSV parasitic information. For full-chip noise analysis, HSPICE simulation is performed and the coupling noises on victim nets are extracted. Different from the flow reported in [40] and [52], where the noises are measured at every nodes on a single net, and the coupling noise voltages are added all into the total noise. Thus, the total noise measured is several times larger than it should be. In our flow, only the maximum noise appears on a single net is measured so that the noise value is not counted many times. This procedure is performed on every TSV net in the design and the sum of maximum noise on all TSVs is used as the total noise. Figure 24 shows our noise analysis flow. Primetime is used to read the parasitic information for each die as well as TSV coupling information altogether and then perform full-chip static timing and power analysis.

Since TSV parasitics are depended on TSV voltage, and it is difficult to estimate the signal arriving time for all possible cases, different strategies are used for worst case and average case analysis. For worst case analysis, it is assumed that all the aggressive signals are arrived at the same time and they all have the same switching waveform from 0V to VDD. In this case, charges due to TSV coupling accumulate around the victim TSV and introduce a large voltage spike at the victim node. Then, the maximum voltage on the victim net is measured. Note it is only theoretically possible that all aggressors have the same waveform and the victim would see such a large noise, however, it is a good indicator of how severe is the coupling in the full-chip level and the result is only related to the design itself. We use TSV MOS capacitance measured when the TSV voltage is 0V since the depletion region width is minimum and TSVs are strongly coupled through the substrate.

For average case study, a time window is chosen which is no larger than the target clock period. We use the TSV MOS capacitance values measured at half of the VDD. Moreover, some aggressors may not even switch during the same clock cycle. Since not all

Table 15: Worst case and average case comparison

	Worst case	Average case
Time window	Clock period	$\leq$ Clock period
Start time	Fixed	Randomly chosen
Aggressor activity	1	0 to 1
Switching direction	Rise	Rise and fall
Noise definition	Maximum voltage	Peak-to-peak voltage

aggressor nets are switching at the same time, the arrive times of the aggressor signals are randomly located within the time window. A switching activity factor which is less than 1 is used to determine the possibility of signal switching. Note in worst case analysis, since all of the aggressors are switching, therefore, the switching factor is 1. Also, different from worst case analysis where all the signals are switching in the same direction, aggressor signals may rise or fall in our average case analysis. Therefore, after running HSPICE, the peak-to-peak voltage difference on the victim TSV net is measured as the noise value. Table 15 lists the comparisons between worst case and average case analysis, and Figure 25 shows the victim voltage waveform in different cases. The limitation of this method is, without static timing analysis on all possible input patterns and every timing path, this method cannot simulate the exact value of the noise under various input patterns. Instead, it provides an overview of the total TSV noise in full-chip scale. Therefore, if detailed signal integrity analysis is needed for each signal, combining static timing analysis engine with our multi-TSV model can solve this problem. The static timing analysis engine provides detailed signal waveform which includes arriving time and slew for every time path while our model computes the noise and delay on TSV net. Also, the detailed layout of the substrate contacts is not considered and a uniformed discharging path is assumed for the silicon substrate. This assumption is valid since the standard cell placement density is close to 60% everywhere in our design, therefore the substrate contact density is almost the same around each TSV. Pattern-matching algorithm may be used to extract the substrate parasitics and detailed discharging path can be built for each individual TSV.

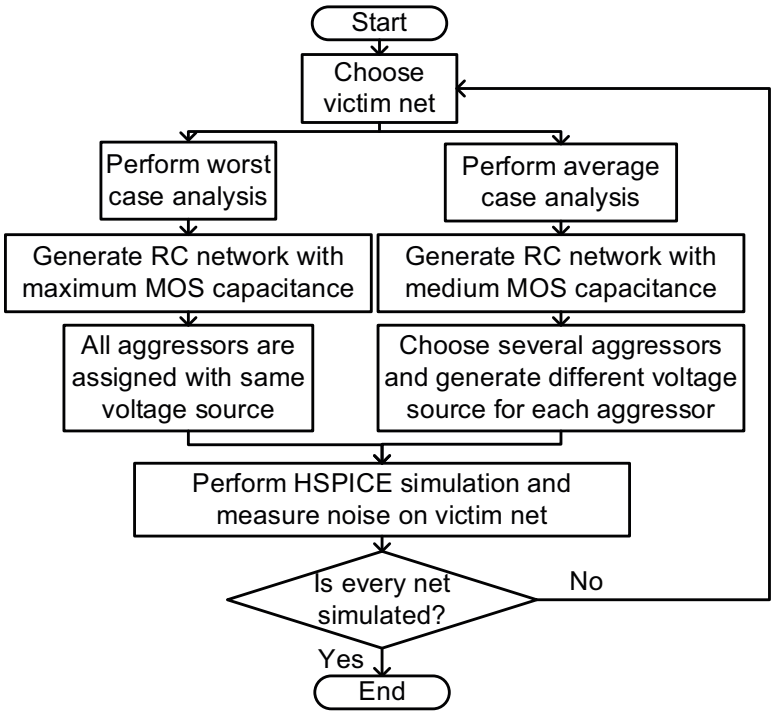

Figure 24: Full-chip noise analysis flow

3.3.3 Designs Specification

A 64 point FFT design is used to demonstrate the full-chip impact of TSV-to-TSV coupling. It has 47K gates and 330 TSVs. The target clock frequency is 200MHz. We implement this design on a 2-die 3D IC using 45nm technology with 5 metal layers. The TSV landing pad size is $5\mu m$ and TSV radius is $2\mu m$. The TSV liner thickness is $0.5\mu m$. Each TSV has a $1\mu m$ KOZ to ensure all the logic cells are outside of the TSV depletion region so that their threshold voltage and performance will not be affected by the depleted substrate. The total footprint area of the design is $380\mu m \times 380\mu m$, and the total TSV area is $16170\mu m^2$, which is 11.2% of the total area. Table 16 shows the detailed design information. An in-house 3D placer [63] is used to obtain the final placement and Cadence Encounter is used to refine placement and route the design. We apply different TSV placement strategies and obtain two kinds of designs. During regular placement, TSVs are placed on regular grid with a pitch of $20\mu m$. TSVs are distributed all over the design space and TSV placement density is about the same everywhere. For irregular placement, TSVs are treated the same way as

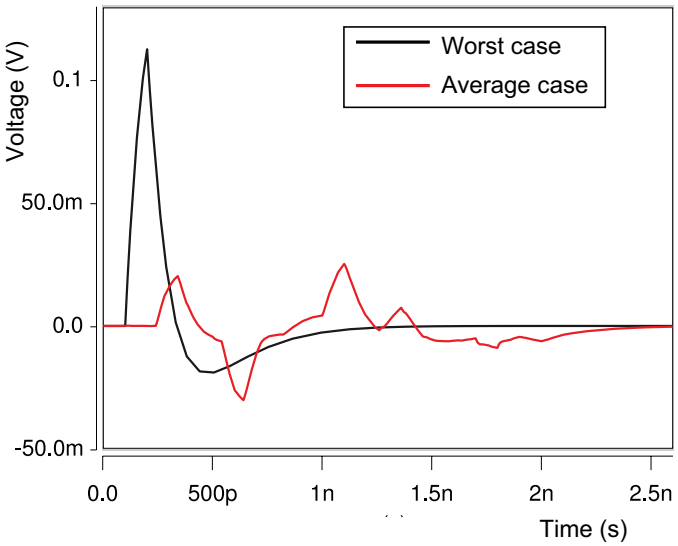

Figure 25: Transient analysis of victim voltage

Table 16: Design specifications

Placement style	Irregular	Regular
Minimum TSV pitch (μm)	12	19
Footprint (μm ^{2})	380×380	
TSV count	330	
TSV area (μm)	16170	

other logic cells and we try to minimize the total wirelength. The minimum TSV-to-TSV pitch with irregular placement is $11\mu m$ so that it can be manufactured. Figure 26 shows the die shots with TSV landing pads highlighted. Though a small digital design is used for full-chip analysis, the TSV placement density is similar to TSV farms in a large-scale design. For those designs, the layout can be partitioned into zones so that each zone can be extracted and simulated efficiently. Thus our method can be extended to 3D IC designs with large footprint without sacrificing the efficiently.

3.3.4 Worst case Analysis v.s. Average Case Analysis

From Figure 23, the largest coupling noise is measured on the TSV net rather than 2D nets and the average noise on TSVs is much larger than that of 2D nets. Also, compared with 2D nets, 3D TSV nets heavily suffer from coupling noise and delay. This is because

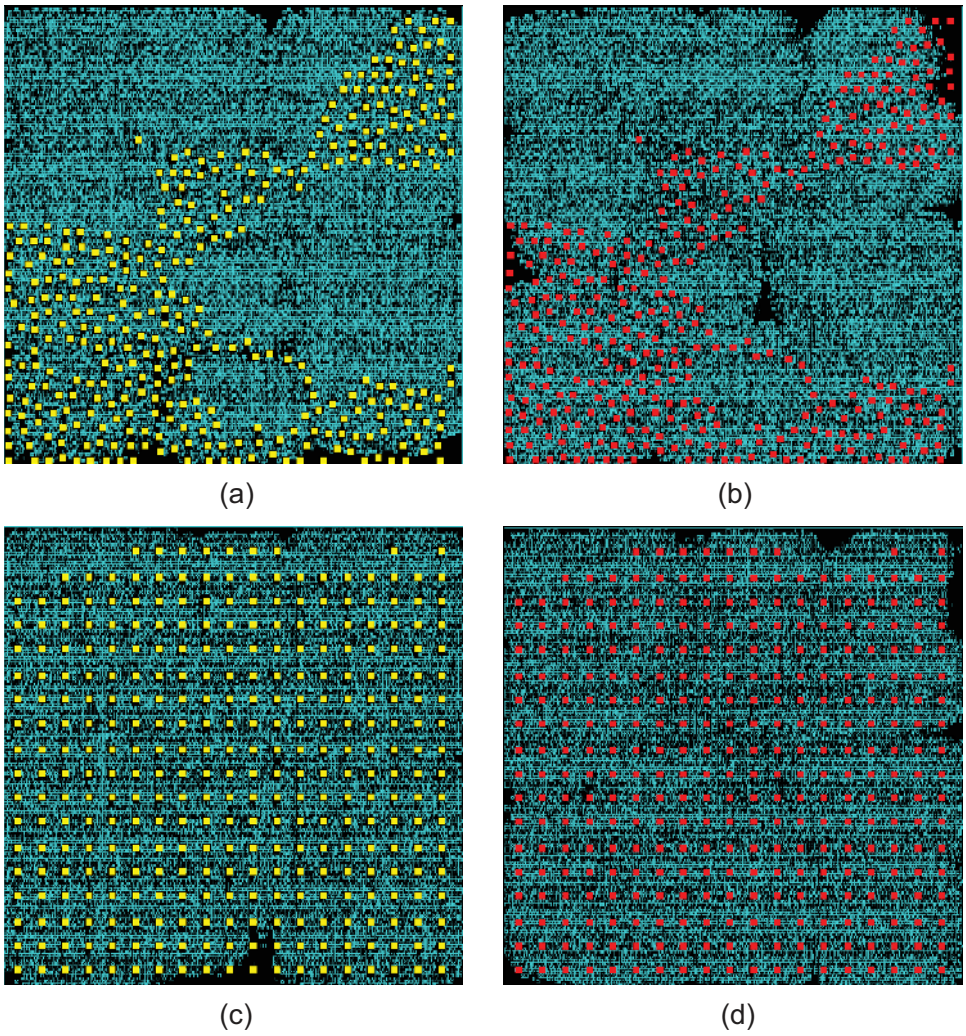

Figure 26: Design layout. (a) and (b) are bottom and top die of irregular placement design, respectively, (c) and (d) are bottom and top die of regular placement design, respectively

of the following reasons: (1) It is difficult for current technology to fabricate TSVs with very small dimensions and large aspect ratio. Therefore, TSV has large MOS capacitance due to its large length and radius; (2) In future technology nodes, more TSVs and higher TSV placement density are allowed to increase die-to-die bandwidth, therefore there will be larger coupling between TSVs; (3) The permittivity of the inter-layer dielectric (ILD) between 2D interconnections is very low if low-K material is used($2\sim3\varepsilon_0$). However, the silicon substrate that buries the TSV has a very high permittivity ($11.9\varepsilon_0$), which results in large TSV coupling capacitance. However, this can be alleviated by using Fully Depleted

Table 17: Average case and worst case comparison on total TSV net noise (V)

	Activity	Slew (ns)	Irregular	Regular
Average case	0.2	0.1	26.51	24.65
	0.5	0.1	39.61	35.37
Worst case	0.2	0.5	14.04	14.62
	1.0	0.1	139.01	132.44

Silicon On Insulator (FD-SOI) technology.

The average case analysis flow described in Section 3.3.2 is used for TSV-to-TSV coupling noise study. In average case, the victim TSV has much smaller peak-to-peak noise due to the following reasons: (1) Not all the aggressors switch in one clock period, and those switching aggressors do not start voltage transition at the same time. Smaller aggressor signal activity results in smaller coupling noise on victim TSV. (2) Due to the load capacitance, many aggressor nets have longer transition time, especially for nets with weak driver. Slower transition time on aggressor introduces fewer charges through the coupling path thus it reduces the coupling noise on victim TSV. Table 17 compares the two analysis in various metrics. The average case shows much smaller total TSV coupling noise than the worst case. The average case analysis provides an estimation on average noise level on TSV nets when multiple aggressors with different voltage waveforms are considered. The results show that both the switching activity and the signal slew have a large impact on the noise results on the TSV nets. Larger switching activity and smaller signal slew increase the TSV coupling noise significantly and they should be considered in noise analysis.

Moreover, compared with regular placement design, irregular placement design is showing 5% larger coupling noise. This is because, in irregular placement design, minimum distance between TSVs is smaller, and TSVs are placed with higher density. Therefore, irregular placement suffers more TSV coupling that results in a larger timing degradation. However, since the regular placement is a special case of irregular placement, it is possible to find a better irregular TSV placement which has smaller noise coupling.

3.3.5 Full-chip Substrate and Field Impact

To study the impact of field and substrate effects, we disable each field and silicon effect one by one while keeping other effects the same and perform noise analysis on the full-chip level. The worst case analysis flow is used because the average case analysis flow is random choice based and gives different results for each run. However, the worst case analysis result only depends on the circuit itself which makes it a fair comparison. Table 18 details chip-level E-field and silicon effects comparison. Without considering the depletion region, TSV MOS capacitance is overestimated, especially when TSV liner thickness is thin and the substrate doping concentration is low. Since our design runs at 200MHz, full depletion around TSV is assumed. If the depletion region is ignored, the result show a 10.5% and 10.2% increase in total TSV net noise for irregular and regular design, respectively. This is because the MOS capacitance is overestimated by 17%. Moreover, ignoring substrate resistors and capacitors is also a pessimistic estimation on coupling noise. The discharging path through a substrate is critical to limit the peak noise on the victim and it also affects delay and power consumption. Also, without considering the electrical field distribution, the noise is over-estimated because every aggressor sees the whole TSV MOS capacitance around victim TSV, even though it only faces to one side of the victim TSV. Since the electrical field distribution effect does not change any capacitance value, the calculated delay and power is the same using Primetime. Overall, the depletion region impact has the largest impact on full-chip metrics as the MOS capacitance is the dominating component in TSV-to-TSV coupling.

3.4 TSV-to-TSV Coupling Noise Reduction Using Guard Ring

3.4.1 Guard Ring Model

Since the silicon substrate provides a discharging path to the ground, it can be used to reduce the coupling noise on TSVs by making the discharging easier and reducing substrate-to-ground resistors (R_{Sig}). We use a grounded guard ring proposed in [42] in the active

Table 18: Silicon and E-field impacts on total TSV net noise (V), TSV-induced delay (ns) and power (μW) increase

Irregular TSV	Total TSV noise	TSV-induced Delay	TSV Net Power
no depletion region	153.7 (+10.5%)	0.85 (+7.6%)	13.53 (+6.7%)
no body resistance	144.9 (+4.2%)	0.78 (-1.2%)	12.54 (-1.1%)
no E-field distribution	146.3 (+5.2%)	0.79 (0%)	12.68 (0%)
all-effects-included	139	0.79	12.68
Regular TSV	Total TSV noise	TSV-induced Delay	TSV Net Power
no depletion region	145.9 (+10.2%)	0.98 (+7.7%)	13.66 (+7.0%)
no body resistance	138.9 (+4.9%)	0.90 (-1.1%)	12.63 (-1.1%)
no E-field distribution	138.9 (+4.9%)	0.91 (0%)	12.77 (0%)
all-effects-included	132.4	0.91	12.77

layer with P+ doping to build a short discharging path for the victim TSV. The ring is connected with grounded rings on Metal1. Therefore, the TSV is protected by ground ring in active layer and landing pad is protected by ring on Metal1. In [42], the guard ring is divided into many cells, and each cell contains 6 to 12 components. This model uses too many components which makes it unsuitable for full-chip analysis. To reduce the model complexity, we propose a new guard ring model with few added components to multi-TSV model. The proposed guard ring structure is shown in Figure 27(a). The discharging path through the grounded ring contains two components C_{Sig} and R_{Sig}, and we use Synopsys Raphael to extract the substrate capacitance to the ground. Detailed extraction results are listed in Figure 27(b), with various edge-to-edge distance and guard ring width. Small ground resistance leads to a strong connection between the substrate and the ground net, thus it can help shielding coupling noise introduced by TSV-to-TSV coupling. The ring width shows a large impact on the ground resistance. Thus, the coupling noise reduces further if the width of the guard ring is increased. However, the distance between TSVs and the guard ring does not affect much on the ground resistance. Longer edge-to-edge distance between TSV and guard ring results in a larger guard ring but the coupling E-field strength is reduced. The drawbacks of this method include a slight timing degradation on TSV nets due to the increased ground capacitance and a small area overhead. Wider guard ring shows larger noise reduction but they introduce longer delay. On other hand, while

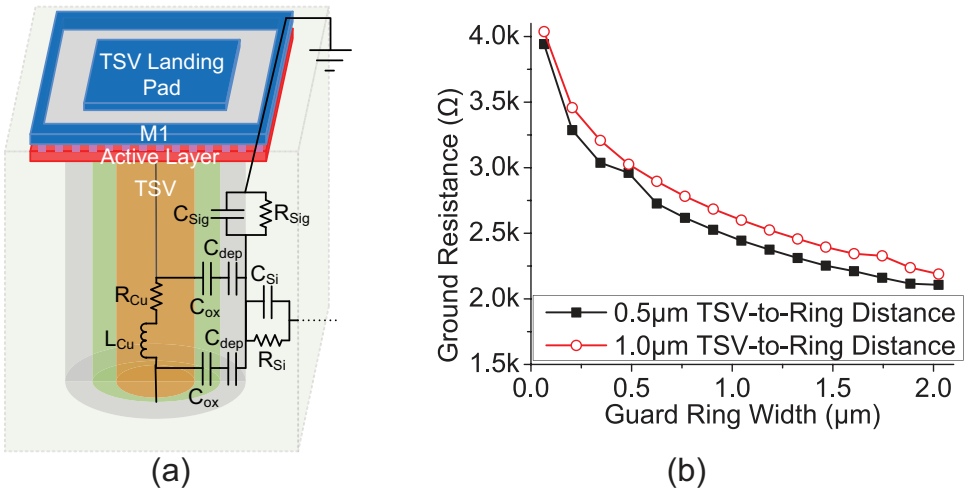

Figure 27: (a) Guard ring model (b) Guard ring impact on substrate ground resistance

the silicon around TSV is depleted and cannot be used for devices, the guard ring in the active area can make use of this area and help reducing noise. This makes the guard ring more appropriate for designs with large KOZ and increase the silicon utilization. Transient analysis is performed on the 3-TSV test structure with our multi-TSV model and the guard ring shows 47.5% noise reduction on victim TSV net.

3.4.2 Optimization Flow and Results

In [40], the authors proposed a TSV shielding technique. The coupling path impedance between TSVs is used to select which TSV should be protected. However, the coupling path impedance is not a good indication of coupling noise because of the following reasons: (1) Not only neighbor TSVs, but also the 2D nets are aggressors for a victim TSV. Using only coupling between TSVs cannot reflect the coupling from 2D aggressors. (2) TSV coupling path impedance and the coupling noise is not in a linear relationship. (3) The number of coupling neighbors also affects the noise value. Therefore, to efficiently find TSVs which need noise protection, the following strategy is utilized to perform the noise optimization. First, a worst case noise analysis is performed on the full-chip design and obtain the noise levels on each TSV. Then, the TSVs are sorted according to the noise levels and guard rings with different widths are added around TSVs. To minimize the area overhead, a minimum

Table 19: Full-chip coupling optimization results of two design styles

Placement style	Irregular	Regular
Total TSV noise without guard ring (V)	139.0	132.4
Total TSV noise with guard ring (V)	101.1	96.5
Noise reduction	27.3%	27.1%
TSV-induced delay (ns)	0.81	0.93
TSV-induced power (μW)	12.75	12.86

noise threshold is used below which no guard ring will be added. Above the threshold, TSVs that suffer larger coupling noise are protected with a wider guard ring and vice versa. Worst case analysis is used here as it is not random-seed dependent. Figure 28 shows the layout with TSV and guard ring highlighted after the optimization is performed on our regular placement and irregular placement designs. Shown in the layout, TSVs with large coupling noise are mostly located in the center of the die where TSVs are surrounded by more aggressor TSVs as well as standard cells.

After the guard rings are added to the design, the overlapping in the layout is fixed using incremental placement and routing and then perform worst case analysis on the new layout. Table 19 shows the noise optimization results. There is a 27.3% reduction in total TSV net noise with only 7.65% area overhead from guard rings. The delay of the design also increases a little due to the increased substrate ground capacitance. Our results show that guard ring protection is very effective in TSV noise reduction with minimum area overhead.

3.5 TSV-to-TSV Coupling Noise Reduction Using Differential TSV pair

3.5.1 Differential TSV Impact on Modeling

Another method to enhance the signal transmission reliability is using differential TSV pairs. Figure 29 compares the single-end TSV and the differential TSV transmission. In differential TSV transmission case, voltages on a pair of TSVs are compared and the difference is used to determine the output level. There are differential TSV models proposed in [45, 46, 57]. These models analyze a pair of differential TSV with grounded TSV nearby.

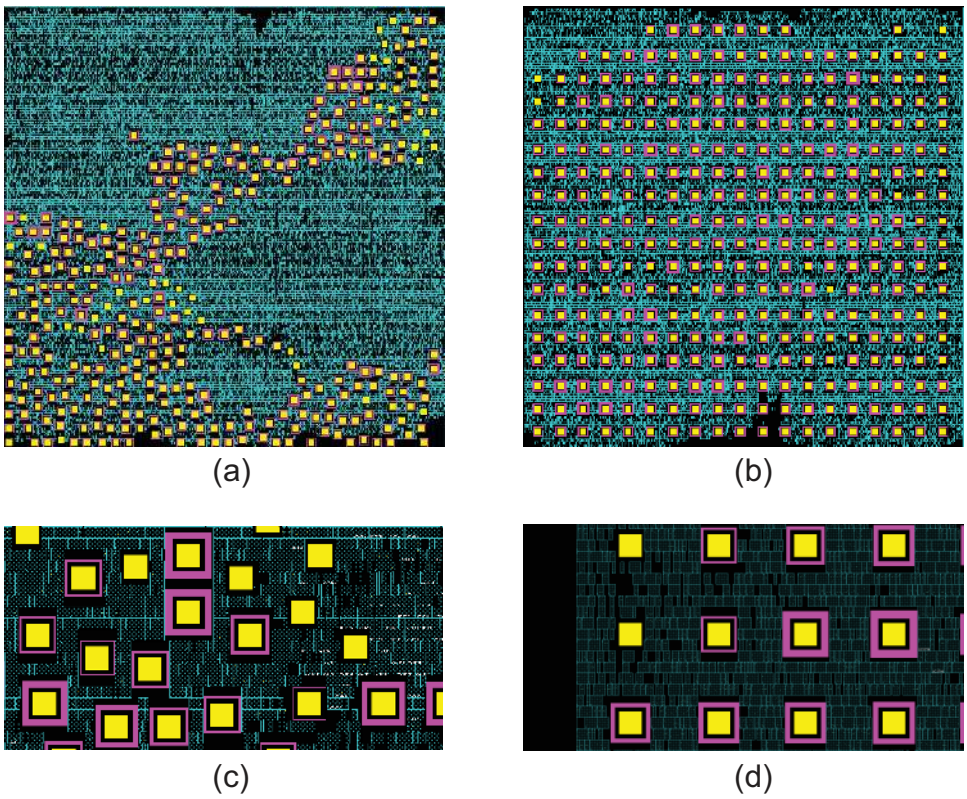

Figure 28: Noise-optimized design layout. (a) and (b) are bottom dies of irregular and regular placement design, respectively, (c) and (d) are zoom-in shots

The model matches the measurement result, however, they ignore the coupling from other TSVs and the E-field distribution. Also, in many 3D ICs, signal TSVs are often placed in TSV farms where there is no power/ground TSV around. Therefore, to analyze the full-chip impact of differential TSV, our multi-TSV model is used in the following discussion.

A test case with 3-TSV is shown in Figure 30(a). TSV A and B form a differential pair and TSV C is a single-end TSV. Each TSV is driven with a signal slew of 0.1ns and an INVX4 as load. If TSV A and B are aggressors and TSV C is a victim, when one of the aggressors is switching, the noise voltage on victim is 0.16V. And if both aggressors are switching with the same waveform, the noise on victim C is almost doubled to 0.31V. However, if TSV A and TSV B form a differential pair and their signals are perfectly symmetric in ideal case, there is no noise on the victim since the aggressive signals cancel each other. Even in real cases when there are unsymmetrical factors due to signal skew and

process variation, and the differential signals are not perfectly synchronized, the noise can still be smaller than single-end TSV coupling. The signal skew impact is shown in Figure 30(b) when the differential pair is the aggressor, and the noise voltages are measured in peak-to-peak swing. For differential signal, once the signal skew is larger than the input transition time (0.1ns), there is no benefit on the victim noise level since the aggressive signals can be treated as two individual signals in those cases. The unsymmetrical location between the victim and the differential pair does not heavily affect the coupling noise on the victim. This is because according to (6), the RC time constant between TSVs is the same. Thus the signal arrive time from TSV A and TSV B will be similar if the signal input skew is small. In our 45nm technology, the signal skew between a INVX4 and BUFX4 is 16.4ps without load capacitance and is 4.9ps with 50fF load capacitance. Therefore, differential TSV can effectively reduce the TSV coupling noise.

On the other hand, the differential TSV transmission improves the noise immunity and reliability. Consider the case when TSV C is the aggressor, and TSV A and B are victims. Since the voltage is compared at the end of the differential pair and the common-mode noise is assumed to be perfectly rejected by the comparator, the absolute value of the voltage subtraction is used as the noise. HSPICE simulation is performed with our multi-TSV model. Figure 31 shows the voltage waveform. Even though each TSV still sees a 0.16V voltage noise on its waveform, the subtraction voltage perfectly rejects the coupling noise. Also, the subtraction voltage has a swing of two times of VDD, which gives more room for signal detection. Note that TSV is used as the aggressor in this analysis, same strategy can be applied for noise reduction when the aggressor is a 2D net.

3.5.2 Full-chip Optimization Flow and Analysis With Differential TSVs

For full-chip implementation, a simple digital comparator proposed in [64] is used. Figure 32 shows the circuit and the layout. Our comparator (COMPX4) is designed using the same footprint and output transistor width as a BUFX4. Table 20 compares the delay of

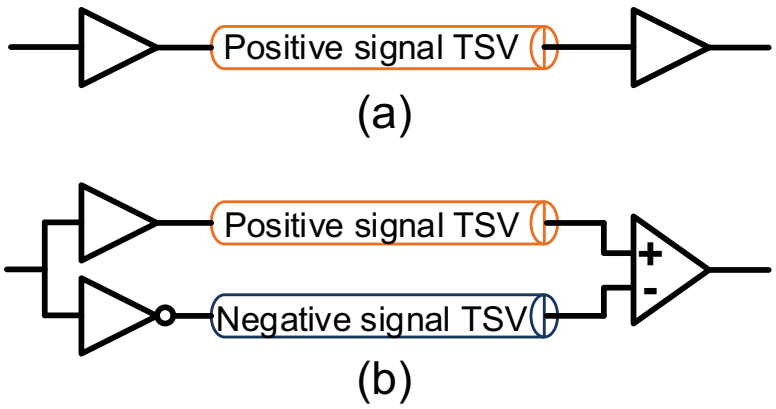

Figure 29: Signal transmission using TSV: (a) single-ended, (b) differential pair.

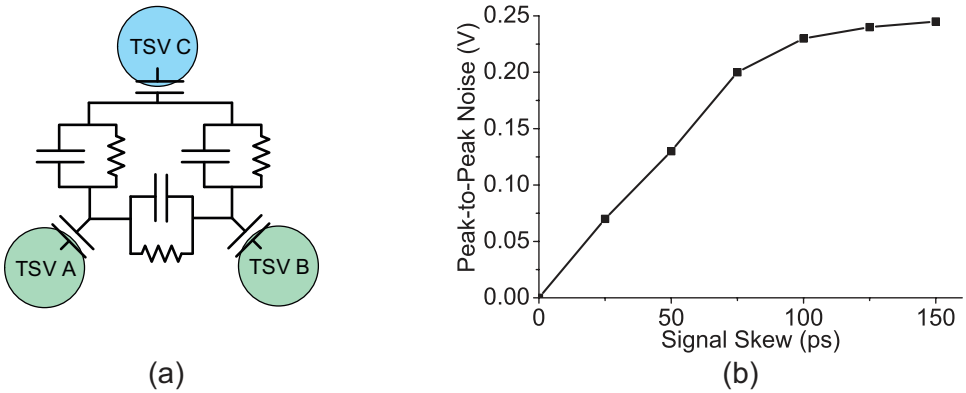

Figure 30: (a) 3-TSV coupling case (b) Signal skew impact on noise when the differential pair is aggressor

these two cells. For regular designs, similar as in Section 3.4, the worst case noise analysis results from original design is used to set a noise threshold. TSVs with noise above the threshold will be replaced by differential TSV pairs. However, for irregular design, since TSVs are placed closer, it is possible that when a single TSV is replaced by a differential pair, the inserted TSV overlaps with existing TSVs. Therefore, for irregular design, starting from the TSV with largest noise, we try to replace TSVs with differential pair, unless the new inserted TSV will cause overlapping in the layout. Compared with regular design, a slightly lower noise threshold is used if same number of TSVs are protected. After the differential TSV insertion, a refine placement is performed to fix TSV overlapping with

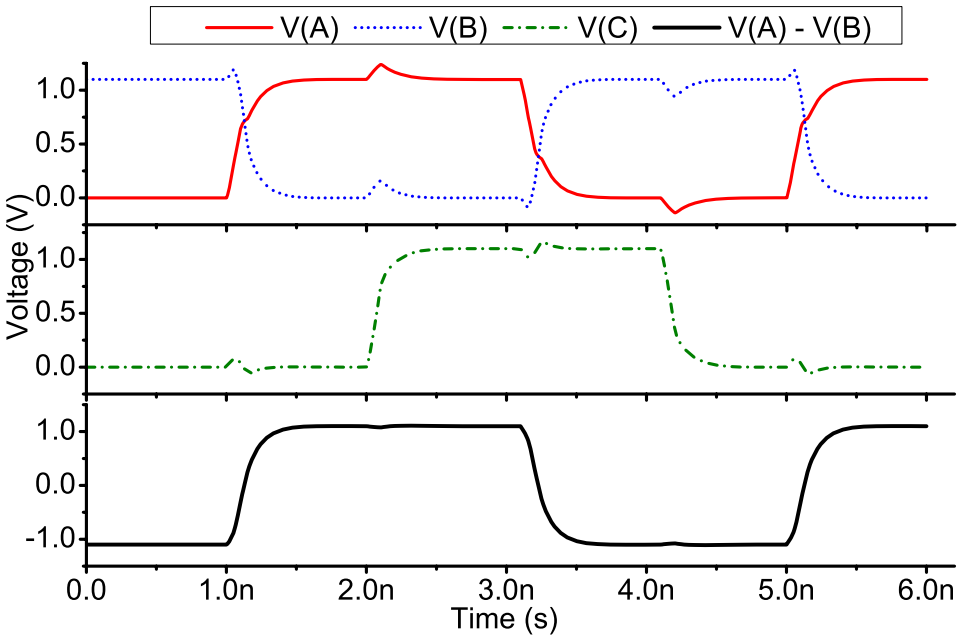

Figure 31: Hspice simulation of 3-TSV coupling case.

standard cells and to insert new cells such as comparators. Then incremental routing is performed so that no major re-design is needed when applying the differential TSV insertion. Note the differential TSV pair impact comes from two aspects. Once a victim TSV is replaced by a differential pair, it has better noise immunity. On the other hand, when the other TSV is considered as the victim, the coupling noise from each member of the differential pair cancels each other and it results in a smaller noise on the victim TSV. To consider both effects, full-chip analysis flow needs to be modified for differential-TSV-awareness.

To perform the full-chip noise analysis, the worst case flow in Section 3.3.2 is modified to consider the differential TSV impact. First, differential pair is divided into positive TSV and negative TSV where pthe ositive TSV has the same voltage switching direction as other aggressors and the negative TSV has the opposite switching direction. Moreover, for differential TSV pair, noises are compared at the both TSVs and the absolute value of voltage subtraction is taken as the noise of a differential pair of TSVs rather than the peak-to-peak voltage in single-end TSV case. Layouts of designs with differential TSV pairs are shown in Figure 33 and full-chip analysis results are shown in Table 21. For the regular

Table 20: Delay comparison between COMPX4 and BUFX4

Input slew (ns)	Cell	Load Capacitance (fF)		
		10	100	300
0.1	BUFX4	35.8	67	148
0.1	COMPX4	50	189	523
0.5	BUFX4	92.7	129.3	209.2
0.5	COMPX4	126.6	266.4	577.2

Table 21: Full-chip impact of differential TSVs

Design style	Irregular		Regular	
With differential TSV?	no	yes	no	yes
Protected TSV#	0	100	0	100
Area increase	-	3.4%	-	3.4%
LPD (ns)	4.62	4.64	4.36	5.02
Total TSV noise (V)	139.0	76.5	132.4	83.1
TSV noise reduction	-	44.9%	-	37.2%
Total TSV coupling cap (pF)	4.32	8.10	3.27	6.33
Total TSV MOS cap (pF)	21.9	28.5	21.9	28.5

design, the TSV on the critical path is replaced by a differential TSV pair so there is a small increase in the longest path delay due to slower comparators and longer signal transition time. However, for the irregular design, such TSV is not protected. Therefore, only minor change exists on the longest path delay. From the results, we conclude that differential TSV transmission is very efficient in TSV coupling noise reduction with a small overhead in timing and area.

3.6 TSV Noise Optimization Method Comparison

In this section, we compare different full-chip TSV noise optimization method including ground TSV insertion(TSV shielding [40]), guard ring protection and differential TSV pair insertion. Table 22 shows the detailed comparison. TSV shielding method uses a FIR design while our methods use an FFT design. TSV shielding is very effective in TSV-to-TSV noise reduction, but there are major drawbacks for this technique: (1) It requires large additional area for ground TSVs, and for every TSV protected, eight additional TSVs are

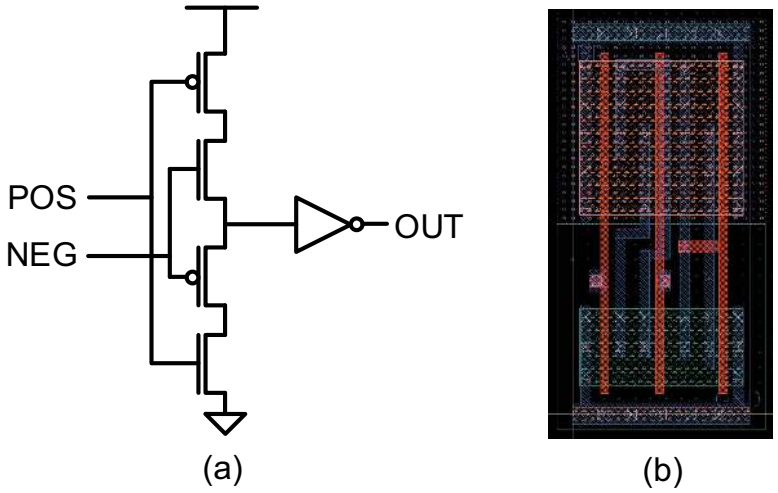

Figure 32: Digital comparator design: (a) schematic, (b) layout in 45nm technology.

Table 22: Full-chip analysis comparison with guard ring vs TSV shielding

	TSV shielding [40]	Guard ring	Differential TSV
Base design	FIR	FFT	FFT
Protected TSV #	118	298	110
Initial TSV size (µm)	49	49	49
Protected TSV size (µm)	361	68.89 ~ 121	105
Initial footprint (μm^2)	402×402	380×380	380×380
Final footprint (μm^2)	421×421	380×380	380×380
Noise reduction	42.04%	27.3%	49.2%
Area overhead (µm)	42598 (26.4%)	11053 (7.65%)	4900 (3.9%)

inserted which results in a large area overhead. (2) TSV shielding needs to enlarge the footprint area and perform a redesign to achieve good noise reduction. Thus it requires more design time compared with guard rings which is easier to implement. (3) The ground TSVs also introduce a large capacitance to the victim TSVs, which will cause delay increase on paths through the protected TSV. As the worst case noise is used to find out TSVs which are heavily affected by the coupling, the coupling direction is not considered. Thus we assume the victim TSV needs to be protected on all sides. As shown in [52], the authors use fewer grounded TSVs inside the TSV farm. This leads to smaller impact on timing and power in the full-chip level, but the noise reduction is also compromised. On the other hand, guard ring protection introduces smallest overhead to the design since no additional TSVs are required and minimum changes are needed for full-chip optimization. Thus, it is

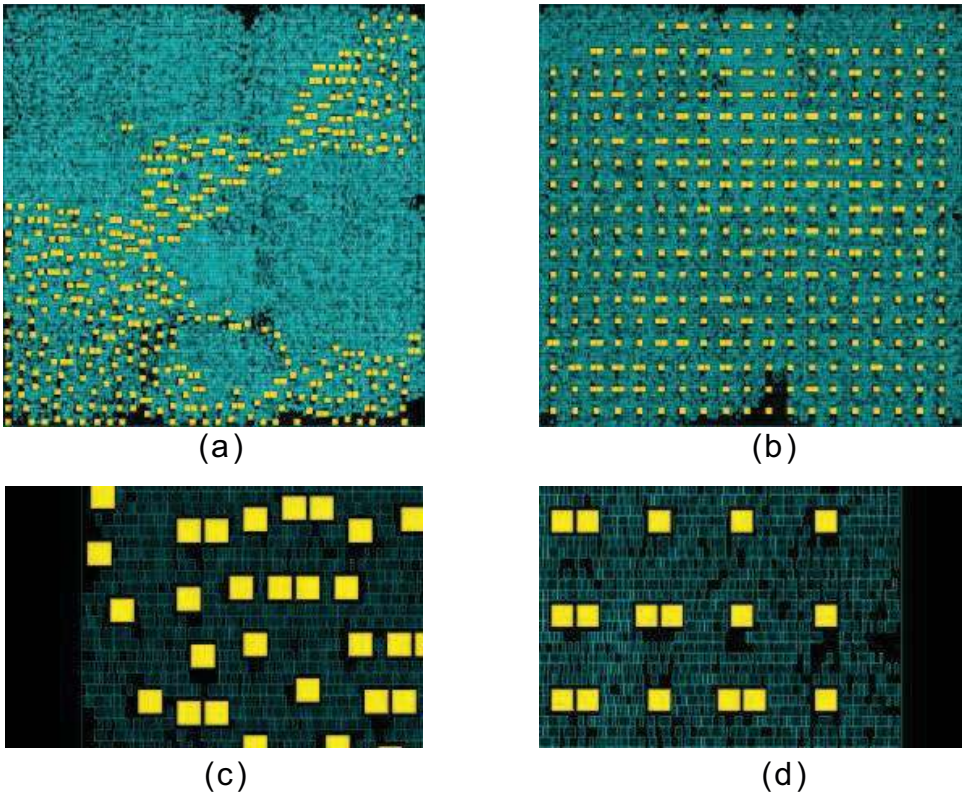

Figure 33: Design optimization with differential TSVs: (a) and (b) are layouts of irregular and regular design, respectively, (c) and (d) are zoom-in shots.

a cost-effective method. However, the noise reduction percentage is also smallest among all of the techniques. The differential TSV insertion introduces small area increase but relatively larger longest path delay increase due to the comparator. Their noise reduction percentage is also large thanks to the differential signal transmission. One benefit from guard ring protection and differential pair TSV insertion is that no re-floorplan is needed if the placement is not heavily congested, which saves a lot of design time and efforts. Overall, our conclusion is, for TSVs on the critical path, guard ring protection is the best solution with minimum delay overhead. For other TSVs, differential TSV is a good choice to minimize area overhead and TSV shielding can be applied on TSVs which needs full protection on every side.

CHAPTER IV

TSV-TO-WIRE COUPLING EXTRACTION AND OPTIMIZATION METHODOLOGIES

Because of increased wire and TSV density, parasitic components between TSVs and wires become important contributors to signal coupling in 3D ICs. One way to avoid heavy TSV-to-wire coupling is to leave a large keep-out zone (KOZ) around the TSV or provide additional shielding around critical signals. However, these techniques are not cost efficient because the area and wirelength increase dramatically. A smarter choice is to carefully extract coupling elements from TSVs based on their physical sizes as well as silicon substrate effects and perform signal integrity analysis to ensure that timing and noise are under control. This process is particularly critical for advanced technologies and mobile applications, in which the supply voltage is low and the signal swing is reduced for low power operation, to obtain a good signal-to-noise ratio (SNR) and a low bit error rate (BER).

4.1 E-field Sharing Impact

4.1.1 TSV Influence Region

Since capacitance is geometry dependent, the interconnect dimension significantly affects coupling capacitance. A large TSV results in stronger coupling as its E-field affects more neighbor conductors. We define a TSV influence region for TSV-to-wire extraction, and only wires within the influence region have their coupling capacitance extracted. We build a special structure, in which a TSV is surrounded by a wire ring, to study the TSV influence region. The TSV radius is 2.5μm and height is 15μm. The ring has the same width (0.14μm) and thickness (0.28μm) as wires in M4 to M6 layers of a 45nm technology. Low-K materials are used in the inter-layer dielectric (ILD) layer with a relative permittivity of

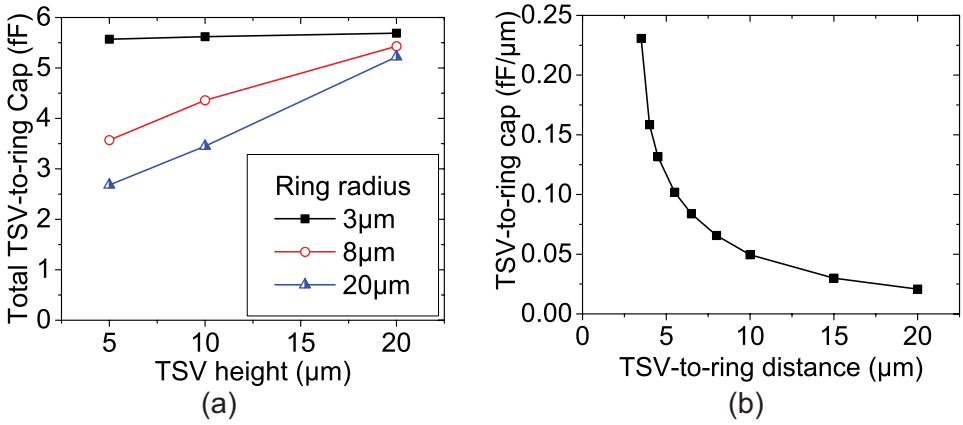

Figure 34: TSV influence region results. (a) TSV height impact. (b) TSV-to-wire distance impact.

2.2. This symmetrical structure is used so that TSV-to-wire distances are the same for all parts of a wire ring.

Single-ring-extraction results are shown in Figure 34(a) with various TSV dimensions. A short TSV does not heavily affect faraway wires, and most of its E-field is restricted within a 10μm range. However, a tall TSV affects wires as far as 20μm. We use bump-less 3D IC technology [65] and directly bond the TSV pads on the bottom die to the top metal layer landing pads on the top die [66], because this technology provides much higher TSV density. Figure 34(b) shows the coupling strength measured by the unit length capacitance, which is calculated by dividing total ring capacitance by the ring circumference. TSV-to-wire coupling is majorly within a 10μm influence region, and wires located farther than 20μm from the TSV show negligible coupling capacitance.

4.1.2 Multi-Wire Impact

The traditional empirical TSV-to-wire model considers a TSV and wire pair at one time [21] and ignores E-field sharing from other interconnect components. Though careful curve fitting can accurately model simple structures, extraction errors on a complicated structure can be large. This is because multiple wires share the E-field around the TSV. We build a structure with four metal rings in HFSS shown in Figure 35(a), and extract their E-field

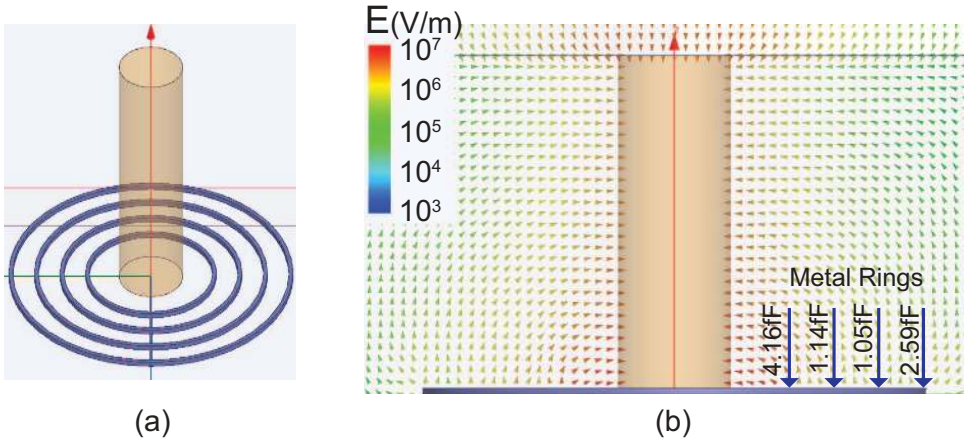

Figure 35: Multi-wire impact. (a) shows HFSS structure with a TSV and four rings. (b) shows the cross-section E-field around the TSV.

interactions with the TSV. Figure 35(b) shows the cross-section E-field distribution map simulated with TSV-to-ring coupling capacitance extracted. As results shown, the strongest coupling E-field forms between the TSV and the nearest wire, and their coupling capacitance is the largest. The outer-most wire also shows large capacitance because no outside neighbor conductor shares the coupling E-field. However, for middle rings having neighbor conductors on both inner and outer sides, only small coupling capacitance is formed as a result from strong E-field sharing. As results shown, without considering E-field sharing, using a formula based on a TSV and wire pair to extract all wire capacitance results in large overestimation. It is also difficult to come up with a compact model for various complicated geometries.

Another observation from the multi-ring structure is that if the ring pitch is small, coupling capacitance of all middle rings is close because of a similar E-field distribution in this region. When more rings are simulated (*e.g.*, from five rings to nine rings), coupling capacitance differences are less than 5% for middle rings. Table 23 shows total capacitance results based on various multi-ring structures. Therefore, if the ring pitch is small enough, we can use fewer rings to estimate the cases with more rings and significantly reduce the library generation time. This condition is often satisfied: If many wires locate inside a TSV influence region, the wire pitch decreases which results in a similar coupling E-field for all

Table 23: Raphael extraction results of multi-ring structures.

Ring count	Ring radius (μm)	Total Ring Capacitance (fF)		
		Nearest ring	Middle rings	Furthest ring
3	5~9	4.21	1.29	3.21
4	5~11	4.16	1.05~1.41	2.59
5	5~13	4.14	0.86~1.05	2.11
5	6~7.12	3.50	2.08~2.09	3.16
9	6~8.24	3.12	1.99~2.02	3.12

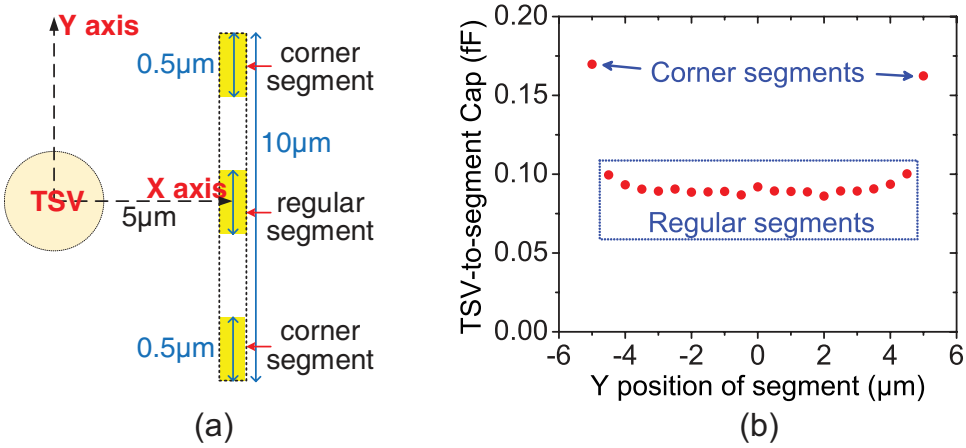

Figure 36: Corner segment impact. (a) Simulation structure with wire segments of 0.5μm in length. (b) Extraction results of each segment.

wires. In our study, we use up to five wires for TSV-to-wire library generation. A larger library with more wires improves accuracy at the cost of a longer library generation time.

Moreover, the E-field sharing effect is also observed even for a single wire. If a wire is divided into several segments, a regular segment has neighbours on both sides while a corner segment has only one neighbour. We build a single wire structure which is divided into 0.5μm segments and extract the capacitance of each segment using Raphael. Figure 36 shows all regular segments have similar coupling capacitances to the TSV but corner segments show 80% larger capacitances even though they are located further from TSV. This is because sidewalls of corner segments also contribute to the fringe capacitance and there is no outside neighbour which shares the coupling E-field.

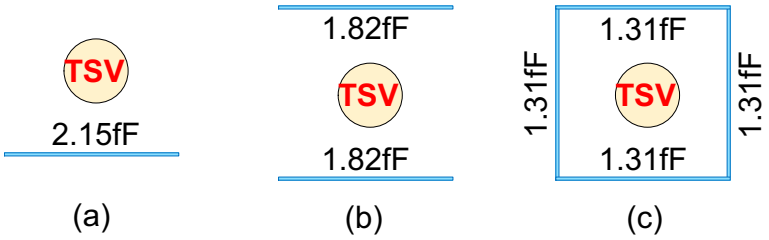

Figure 37: Impact of wire coverage around the TSV on coupling capacitance.

4.1.3 Wire Coverage Impact

If multiple wires surround a TSV, another E-field sharing impact is observed. As shown in Figure 37(a), if the TSV is only facing wires on one side with little E-field sharing, total coupling capacitance for a single wire is 2.15fF. However, if the TSV is facing to wires in more directions as shown in Figure 37(b) and (c), the single wire capacitance decreases to 1.31fF. This is because TSV-to-wire coupling is evenly distributed to all four neighbors. We use a wire coverage factor to represent how much a TSV is surrounded by wires. A wire coverage calculation example is shown in Figure 38(a) and wire coverage factors for structures in Figure 37(a) to (c) are 25%, 50% and 100%, respectively. Larger wire coverage results in stronger E-field sharing and smaller capacitance per unit length. However, since more conductors are around the TSV total, TSV-to-wire capacitance increases. Therefore, for accurate TSV-to-wire capacitance extraction, the wire coverage effect needs to be considered carefully, especially when routing is congested in the full-chip design.

4.2 TSV-to-Wire Extraction Technique

4.2.1 Pattern Matching Technique

To handle 3D full-chip TSV-to-wire extraction, we propose a pattern-matching technique. This technique is similar to traditional 2D full-chip extraction tools which correlate closely with silicon measurements. But our technique accounts for every special TSV-related impact which traditional tools cannot handle. The first stage of a pattern-matching extraction

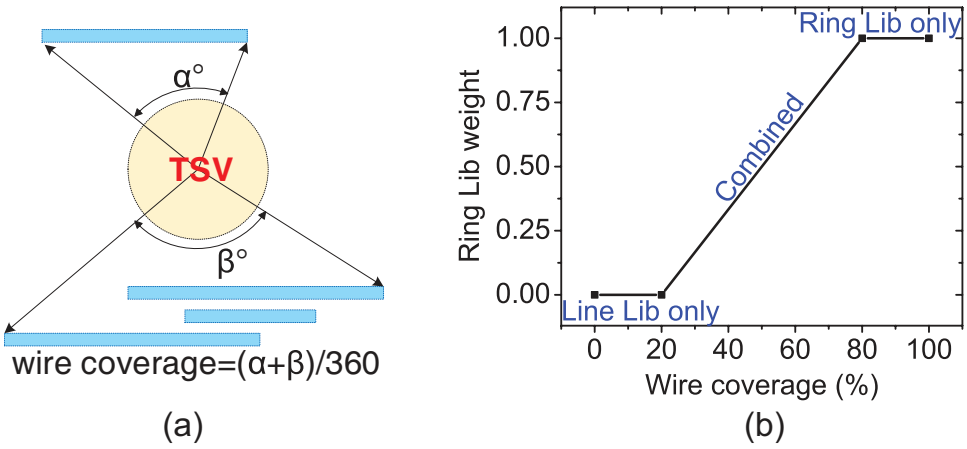

Figure 38: Our combined method. (a) shows the calculation of wire coverage, (b) shows the calculation of weighted average.

is using a general extraction engine such as a field solver, to perform extraction on various pre-defined structures. Results are saved into a database called library. Then during extraction stage, full-chip interconnects are compared to the library and extraction results of pre-calculated structures closest to the layout are used for the capacitance calculation. Modern extraction engines such as Calibre xRC are able to generate a series of extraction rules based on library results. Curve-fitted equations and interpolation methods are used during structure matching to provide a more accurate estimate.

Though generating the library and extraction rules takes a long time as thousands of layout geometries need to be simulated, a common library or a set of rules can be used for certain technology on various designs. Therefore, these extraction files are provided in the process design kit (PDK) by the foundry. Since only library look-up and math calculations are performed during extraction, pattern-matching extraction can extract parasitics of a large-scale circuit within minutes, and they are suitable for extraction of next generation 3D ICs with billions of transistors and thousands of TSVs. Also, the pattern-matching method is also a promising solution for parasitic extraction of next generation monolithic 3D ICs [67].

However, traditional pattern-matching engines can only handle 2D designs, where interconnect coupling is limited to several neighboring wires. Vias in 2D ICs does not have

large parasitics as their sizes are small and the via coupling capacitance is often ignored by the extraction engine by default. Unlike metal vias, TSVs are hundreds of times larger and they interact with many surrounding neighbors as a result of their large influence regions. E-field sharing from multiple conductors also introduces new challenges which must be accounted for during extraction. Therefore, we focus on the special impact from TSVs and propose a first-of-its-kind 3D extraction method. Since it is compatible with the pattern-matching-based 2D extraction tools, our method can be easily integrated into current CAD flow and provide a smooth transition to the next generation of 3D IC designs. Also, it can be easily parallelized and has a great potential for runtime improvement on a multi-core system.

To handle all aforementioned effects, we build three special libraries for TSV-to-wire coupling extraction. Two libraries (*i.e.*, a line library and a ring library) are used for regular segments, while a third corner library is used specifically for corner segments. These libraries enable detailed consideration of E-field sharing among wires. In our libraries, the TSV radius, TSV height, wire thickness, and wire width are used as library indexes. This enable extraction of TSVs and wires with various dimensions. To handle relative location between a TSV and a wire, the nearest TSV-to-wire distance, wire pitch, and wire location angle are included as indexes as well. To handle multiple wires, we build libraries containing various numbers of wire, and include the wire count as another library index. During extraction, wires are divided into segments and their capacitance is calculated for each segment. The geometry information of the segment and its context is used to match patterns in the library, and a linear interpolation of closest structures is used when no pattern exactly matches the segment. Our libraries contain thousands of structures covering a wide range of possible scenarios based on 45nm technology. TSV has a height of 15μm and a radius of 2.5μm with a minimum placement KOZ of 0.5μm. Wire dimensions are based on technology files.

Since Raphael does not handle the frequency-dependent silicon substrate, we use a

dielectric material with a relative permittivity of 11.9 in our TSV-to-wire extraction. The silicon conductivity is ignored because the top metal layer is not directly connected to the substrate of the neighboring die and we assume a lightly-doped substrate on the backside. If a highly-doped substrate is used, the substrate resistance can be calculated based on the RC relationship of homogeneous materials[7]. Moreover, as shown in Figure 1(b), the active regions are located near the M1 layer of the bottom die. Thus, their E-field sharing only affects the coupling capacitance between a TSV and its neighboring top metal wires on the top die. If a silicon effect-aware field solver is used to handle these properties around TSVs, it can provide more accurate extraction results. However, the semiconducting electrical properties of the silicon substrate and the E-field sharing from active layers affect TSV-to-TSV coupling capacitance. These are major E-fields inside the substrate. Therefore, the substrate resistive path and the E-field sharing in the active layers cannot be ignored in TSV-to-TSV coupling extraction. Thus, in our TSV-to-TSV coupling extraction, we model the silicon depletion regions, substrate resistance, and E-field sharing from active regions to improve the accuracy.

4.2.2 Line Library

We build the line library for TSVs with a low wire coverage. As shown in Figure 39(a), the line library is built by placing straight wires on only one side of the TSV. All wires are segmented and a single structure is able to produce results for many segments with various locations. This increases extraction parallelism and reduces the library generation time. The length of each wire segment depends on its relative location to the TSV. Each segment always has a facing angle of 5° to the TSV and wire segments far from the TSV are longer. This is because that wires far from the TSV has weaker coupling and smaller capacitance per unit length. A finer grid provides more accurate results at the cost of longer runtime. Our segmenting method takes advantage of cylindrical shapes of TSVs so that capacitance of each wire segment is in a similar range to prevent accumulations of small errors. Similar

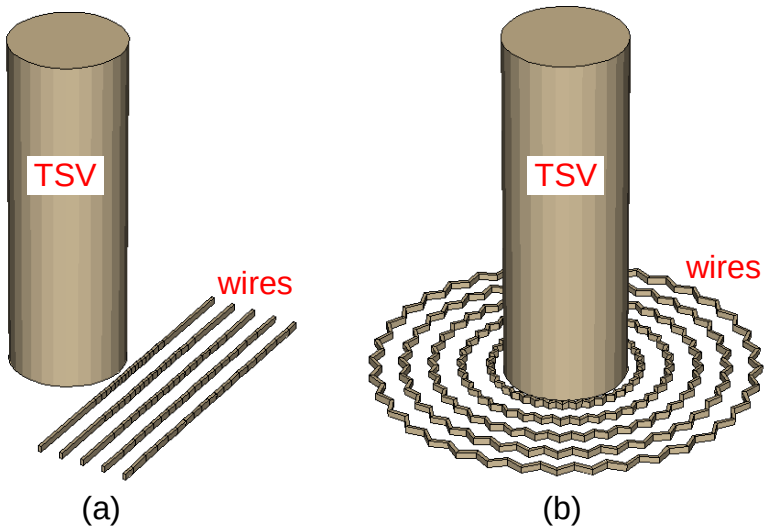

Figure 39: Test structures for library generation. (a) A line library structure. (b) A ring library structure.

to the finite element analysis (FEA), finer segmenting is used on areas where the E-field is strongest and rapidly changing while coarser segmenting is used on less critical areas. This enables a best tradeoff between the simulation time and accuracy.

The line library assumes that only one side of a TSV is surrounded by metal wires and only weak E-field sharing exists around the TSV, thus unit length capacitance of a wire is high. Therefore, this library is suitable for layouts where TSVs are covered by a few wires around. For a general case where the TSV is surrounded by wires on multiple sides, the line library gives overestimated capacitance since the line library always assumes a weak E-field sharing. In terms of the library generation time, since the line library consists of less complicated geometry structures such as straight wires, it is faster to generate.

4.2.3 Ring Library

To handle layouts where TSVs are surround by many wires on all sides, another ring library is built. As shown in Figure 39(b), we duplicate wire segments with various locations to form a ring around the TSV. In this structure, as the E-field of the TSV is evenly distributed in all directions, we extract the total capacitance of the ring and divide it by the total number of ring segments. We place various numbers of rings around the TSV to simulate multiple

Table 24: Library comparison

Library	Ring	Line	Corner
Target segment	regular	regular	corner
E-field sharing	strong	weak	weak
Unit length capacitance	small	medium	large
Geometry complexity	high	low	low
Generation time	long	short	short

wires. As the wire coverage for a ring structure is 100%, E-field sharing around the TSV is high while unit length capacitance in the ring library is small. Unlike the line library, the ring library always assumes strong E-field sharing around TSVs, thus they are suitable for designs with congested routing wires. For a general case where the TSV is surrounded by few wires, and wire coverage is low, the ring library underestimates TSV-to-wire capacitance. Thus, the ring library is complementary to the line library to provide accurate extraction for general cases. However, as the ring structure is built with many segments, the complicated geometry needs a longer extraction time for field solving.

4.2.4 Corner Library

As in previous discussions, wire segments with a single neighbor have larger coupling capacitance due to sidewall capacitance and less E-field sharing from neighbors. Therefore, based on the line library, a special corner library is built to extract corner segment capacitance at various locations. The corner library structure is similar to that of the line library. However, only capacitance of the corner segment is extracted and saved. Compared to line and ring libraries, the unit length capacitance of surrounding wires is the highest and geometry complexity for the corner library is low. However, since there are not many corner segments in the full-chip level, especially for top metal layers, its impact on system performance and noise metrics is small. On the other hand, for short wires, the extraction error is significantly reduced with corner segment effects resolved. Comparisons of all three libraries are listed in Table 24.

4.3 Pattern Matching Algorithm

Once all libraries are built, we divide surrounding wires into segments and choose closest library structures to obtain TSV-to-wire coupling capacitance. We develop an algorithm shown in Algorithm 1 for pattern-matching-based TSV-to-wire coupling capacitance extraction. Extraction is performed on each TSV. Areas around the TSV is divided into 72 circular sectors, each with 5° in central angle and the same radius as the TSV influence region. These sectors are numbered clockwise. In this case, wires closer to the victim TSV have finer segments and only segments within the TSV influence region are handled. Similar to the line library structure, wires are segmented at the sector boundary and all wire segments in the same sector are gathered into a list. The wire dimension and location, number of wires, average pitch of wires are used as indexes to search through the library. The lookup procedure takes place on each list and compares the layout structure to the pre-generated libraries. Linear interpolation is used when the library structure does not exactly match the extraction structure.

For corner segments, results from the corner library is used. For regular segments, we combine both the line library and the ring library based on wire coverage around the TSV. As shown in Figure 38(b), if wire coverage is above 80%, we only use the ring library because coupling capacitance per unit length is small. On the other hand, if coverage is below 20%, we only use the line library assuming weak E-field sharing. Otherwise, results from both libraries are combined and a weighted average is calculated depending on wire coverage. This enables wire coverage consideration during full-chip extraction. After all lists are parsed, TSV-to-wire parasitics are exported into a standard parasitic exchange format (SPEF) file which can be integrated into the standard full-chip CAD flow for further timing and noise analyses.

Algorithm 1: Pattern-matching extraction algorithm

Input : Ring, Line, and Corner libraries; Routed layout
Output: TSV-to-wire capacitance

```
foreach TSV i do
    foreach Wire j within the influence region of TSV i do
        Divide j into segments;
        foreach Segment k within the influence region of TSV i do
            d←sector index;
            Append k to list S[d];
    foreach Sector d do
        foreach Segment k inside d do
            t← nearest wire distance;
            p← average wire pitch in S[d];
            if k is a regular segment then
                LookUp(d, k, S[d], t, p) in the line library;
                LookUp(d, k, S[d], t, p) in the ring library;
                Calculate the combined value based on wire coverage;
            else
                LookUp(d, k, S[d], t, p) in the corner library;
Export capacitance in SPEF format;
```

4.3.1 Single-TSV Validation

For library comparison and verification, we perform extraction on sample layouts with two TSVs and four wires shown in Figure 40. Table 25 compares our extraction results based on single-TSV libraries with Raphael results. Using the line library is accurate when wire coverage is low, while using the ring library is accurate when wire coverage is high. But a combined method accounted for wire coverage and E-field sharing, always extracts capacitance more accurately. With single-TSV libraries, the maximum error is 0.17fF and the average error is 0.05fF.

To validate our extraction method in the full-chip level, we implement a two-die 64-point fast Fourier transform (FFT64) design and apply our method to all TSVs. The placement result of this design is shown in Figure 41. After reading the routing results, for each TSV, we build a Raphael structure exactly as the layout around it. We set the TSV influence region as 10μm to save Raphael simulation time and capture most coupling around

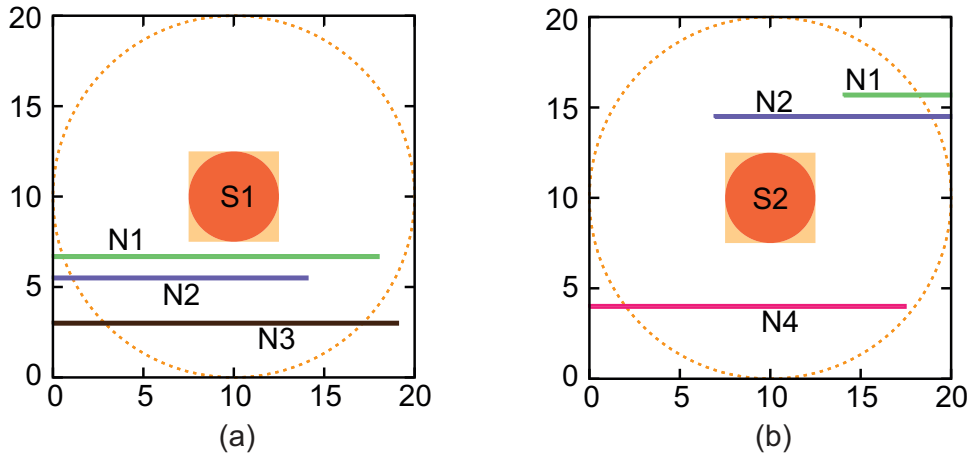

Figure 40: Sample extraction layouts with a TSV and their surrounding wires. (a) and (b) are areas around TSV S1 and S2, respectively. Lengths are in μm.

Table 25: Sample layout extraction results based on the single-TSV libraries. Capacitance is reported in fF.

TSV	Wire	Raphael	Our method		
		Single-TSV	Ring Lib	Line Lib	Combined
S1	N1	1.76	1.49 (-15%)	2.07 (+17%)	1.93 (+9.3%)
S1	N2	0.76	0.68 (-10%)	0.78 (+2.5%)	0.76 (-0.7%)
S1	N3	0.81	0.86 (+6.2%)	0.79 (-2.8%)	0.81 (-0.6%)
S2	N1	0.31	0.29 (-7.9%)	0.34 (+6.9%)	0.31 (-0.3%)
S2	N2	1.38	1.28 (-6.6%)	1.37 (-0.7%)	1.33 (-3.6%)
S2	N4	1.62	1.49 (-7.8%)	1.57 (-2.9%)	1.53 (-5.3%)

the TSV. Extraction results from the field solver and our pattern-matching algorithm are compared in Figure 42(a), where each dot represents a coupling capacitor between a TSV and a neighboring wire. The error histogram is shown in Figure 42(b) for all extracted capacitors compared with Raphael. Results show that our pattern-matching extraction is highly accurate in the full-chip level.

Table 26 compares extraction results using different libraries. Without resolving wire coverage impact, using a single ring library gives 8.3% underestimated total capacitance, while using a single line library gives 5.3% overestimated total capacitance. However, if results from both libraries are combined, the total capacitance error is only 1.9% and

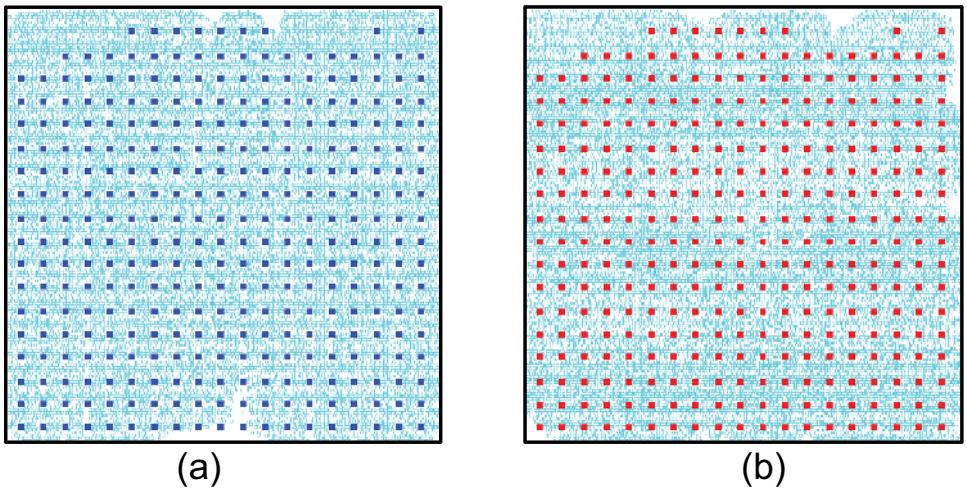

Figure 41: Gate and TSV placement results of FFT64 design with a footprint size of 380μm ×380μm. (a) shows the bottom die, (b) shows the top die.

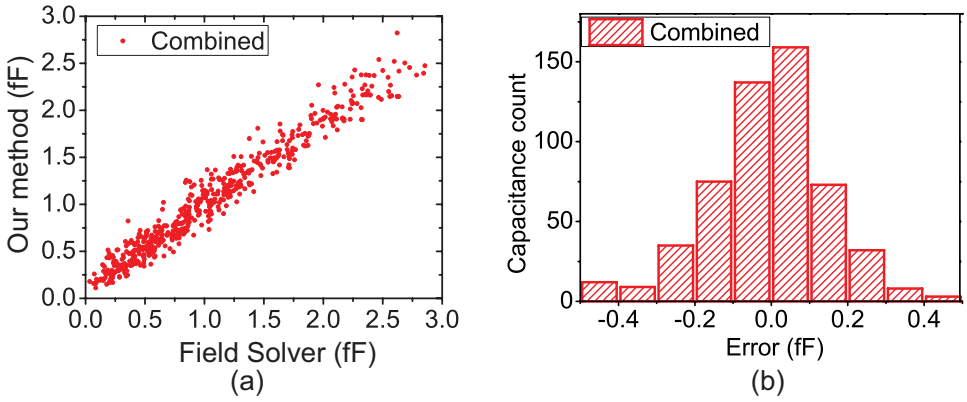

Figure 42: Full-chip Verification using combined method. (a) extraction result comparison, (b) error distribution.

the average error decreases to only 0.112fF. Compared with Raphael, which needs significant runtime and memory, our pattern-matching method achieves 11250 times speedup and 29.29 times smaller memory space as shown in Table 27. Therefore, it is a practical solution even for a large-scale 3D IC with many TSVs. Therefore, we conclude that our pattern-matching method, which handles E-field sharing impact with ring, line and corner libraries, is highly fast and accurate for full-chip TSV-to-wire extraction.

Table 26: Single-TSV extraction comparison with different libraries, where the total capacitance from Raphael simulation is 568fF.

	Ring Lib	Line Lib	Combined
Total Cap (fF)	538	618	579
Total Cap error	-8.3%	+5.3%	-1.9%
Correlation coefficient	0.971	0.966	0.981
Average error (fF)	0.171	0.163	0.112

Table 27: Full-chip simulation runtime and memory space comparison.

Extraction method	Raphael	Pattern-matching	Improvement
Runtime	7.5h	2.4s	11250x
Memory space	615MB	21MB	29.29x

4.4 Extraction With Multi-TSV

4.4.1 E-Field Sharing With Multi-TSV

Previous studies [21] are based on TSV-to-wire extraction with a single TSV. Because of fabrication yield and cost issues, TSVs are usually placed regularly where a TSV can only locate at a pre-defined grid point. However, modern TSV fabrication technology allows much denser TSV placement, where multiple TSVs are placed close to each other and their E-fields interact with each other. A full-chip level study has shown that ignoring multi-TSV impact results in an overestimation on TSV-to-TSV coupling [9]. Therefore, we need to handle the E-field interaction with multi-TSV for accurate TSV-to-wire extraction.

Unlike TSV-to-wire coupling, TSVs even far away from each other have non-negligible E-field interaction. As a result, even though a TSV is located beyond the TSV-to-wire influence region of another TSV, it still affects extraction results. We build a sample structure in HFSS to illustrate multi-TSV impact. The single-TSV structure is shown in Figure 43(a), where three wires are placed around a victim TSV. The multi-TSV structure is shown in Figure 43(b) with two nearest neighboring TSVs placed around the victim TSV. All TSVs and wires have the same dimensions as those in Section 4.1.1 with a TSV pitch of 18μm. Since our extraction is performed on each victim TSV, the victim TSV is numbered at 0 while two neighboring TSVs are numbered at 1 and 2. The origin of the coordinate system

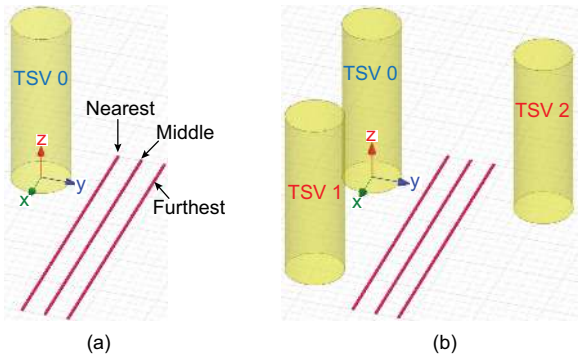

Figure 43: HFSS structures. (a) Single-TSV, (b) Multi-TSV.

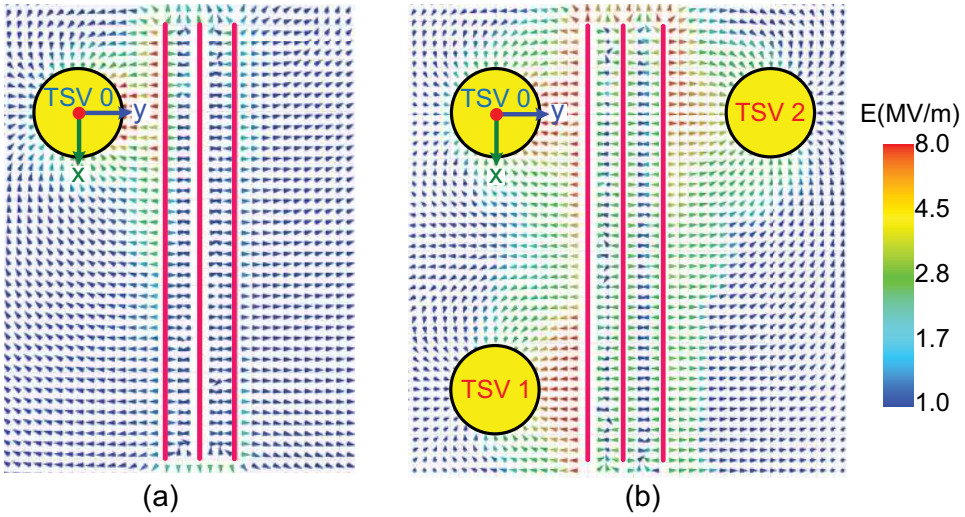

Figure 44: XY-plane E-field distribution comparison. (a) Single-TSV, (b) multi-TSV.

is located at the center of TSV 0. The XY-plane E-field distribution comparison between single-TSV and multi-TSV, shown in Figure 44, is based on HFSS simulations.

With the single-TSV structure, only areas where the victim TSV is close to wires have strong coupling E-fields. This results in large capacitance between the TSV and the nearest wire. Also, as there is no other neighbor conductors, all coupling fields of the furthest wire go to the victim TSV as well. With multi-TSV structure, while the E-field distribution around the victim TSV remains the same to single-TSV case, the coupling E-field changes significantly around the neighboring TSVs. The neighboring TSV not only increases the total E-field strength but also alters the E-field direction around wires. E-fields from wires are heavily shared by neighboring TSVs thus the coupling between wires and the victim

TSV is reduced. For the furthest wire to the victim TSV 0, where most of its coupling goes to the right neighboring TSV 2, its coupling to the victim TSV decreases significantly. For the nearest wire, E-field sharing from the TSV 1 mostly affects areas that are beyond the influence region of the victim TSV, and the major portion of its coupling to the victim TSV remains the same. For the middle wire, its coupling to the victim TSV is also smaller compared to single-TSV model because of E-field sharing from both TSV 1 and 2.

To study of the multi-TSV impact on parasitic components, we divide neighboring wires into small segments and perform extraction with Raphael. Figure 45 shows the coupling distribution comparison between single-TSV and multi-TSV models and Table 28 summaries extraction results. For all wires, the coupling capacitance to the victim TSV is reduced due to E-field sharing. For the nearest wire which has the largest coupling capacitance to the victim TSV, its total coupling capacitance to the victim TSV is reduced by 14.8% compared with the single-TSV model since most E-field sharing affects areas where the coupling to the victim TSV is weak. E-field sharing from both TSVs results in a 24.3% reduction for the middle wire, and a 35.1% reduction for the furthest wire. In addition, with the single-TSV model, total coupling capacitance of the furthest wire is larger than the middle wire. However, it becomes the smallest with the multi-TSV model. From victim TSV perspective, total TSV-to-wire capacitance is 1.68fF with the single-TSV model but it is only 1.30fF with the multi-TSV model. From results we conclude that, for layouts where multiple neighboring TSVs are located around, ignoring the E-field sharing from multi-TSV results in an overestimation of TSV-to-wire coupling capacitance. Therefore, for accurate TSV-to-wire coupling extraction, E-field sharing from multi-TSV needs to handled carefully.

4.4.2 Multi-TSV Libraries

To handle multi-TSV impact, we extend our patter-matching algorithm with multi-TSV structures. To avoid a long library generation time, four nearest neighboring TSVs are

Table 28: Coupling capacitance between victim TSV to wires.

Target wire	Coupling capacitance (fF)		
	Single-TSV	Multi-TSV	Δ
Nearest wire	0.765	0.652	-0.113 (-14.8%)
Middle wire	0.448	0.339	-0.103 (-24.3%)
Furthest wire	0.476	0.309	-0.167 (-35.1%)

added into multi-TSV library structures because they have largest impact around the victim TSV and shield E-fields from further TSVs. An illustrative comparison of different library structures is shown in Figure 46. With additional TSVs, the library construction time is increased with more conductors, but E-field interactions among TSVs are captured. All three libraries are constructed with neighboring TSVs and are used in the full-chip extraction. Compared with single-TSV libraries, coupling capacitance is smaller, especially for wire segments which are far from the victim TSV. However, since multi-TSV libraries can only handle layouts with the same TSV pitch, we build our multi-TSV libraries with a TSV pitch of 18μm which is the same in our design layouts. This limitation is usually not a concern because most TSV technologies such as the one used in [68] require a regular TSV placement. For other technologies which allow irregular TSV placement, multi-TSV libraries with various TSV pitches are needed to provide more accurate extraction. Also, if there is only a few TSVs which are placed far away from each other, single-TSV extraction can still provide accurate results in the full-chip level.

4.4.3 Multi-TSV Validation

To verify our multi-TSV extraction algorithm, we first applied this method to sample layouts in Figure 40. Instead of comparing our extraction results with those extracted from Raphael structures with a single TSV, we expanded the simulation window to cover influence regions of four nearest neighboring TSVs as well. Thus, the Raphael simulation captures E-field interactions from all five TSVs. Extraction results are shown in Figure 29. With all E-field interactions from neighboring TSVs captured in our multi-TSV libraries,

Table 29: Sample layout extraction results based on the multi-TSV libraries. Capacitance is reported in fF.

TSV	Wire	Raphael Multi-TSV	Our method Ring Lib	Our method Line Lib	Our method Combined
S1	N1	1.41	1.23 (-13%)	1.61 (+14%)	1.50 (+6.1%)
S1	N2	0.57	0.50 (-12%)	0.58 (+2.3%)	0.56 (-1.9%)
S1	N3	0.49	0.48 (-1.6%)	0.49 (+1.6%)	0.49 (-0.6%)
S2	N1	0.17	0.16 (-3.0%)	0.19 (+13%)	0.18 (+4.8%)
S2	N2	1.00	0.94 (-5.5%)	1.04 (+4.6%)	0.98 (-1.2%)
S2	N4	1.03	0.97 (-6.3%)	1.08 (+4.9%)	1.01 (-1.9%)

the extraction accuracy improves. The maximum error is only 0.09fF with an average error of 0.023fF. We observe that, with single-TSV extraction, coupling capacitance between TSV S1 and wire N3 is larger than that between TSV S1 and wire N2. This is because wire N3 is the furthest wire around the TSV and E-field sharing from its outside neighboring TSV is not captured. With our multi-TSV libraries, this inaccuracy is corrected and extraction results match well with Raphael simulations. Compared with single-TSV extraction, total capacitance decreases by 29.2% from 6.67fF to 4.72fF.

For full-chip verification, the same flow described in Section 4.3.1 is used. For each TSV, Raphael structures with four neighboring TSVs are used for comparison. Extraction results with multi-TSV libraries are compared with field solver extraction in Figure 47(a), and the error histogram is shown in Figure 47(b). Table 30 compares extraction results using different libraries. The line library still overestimates coupling capacitance, and the ring library underestimates it, but a combined method gives the most accurate results: The total capacitance error is only -0.9% and the average error is only 0.063fF. Both numbers are significantly improved compared with results using single-TSV libraries. Except for input library changes, the pattern-matching algorithm remains the same, thus the performance speedup and memory reduction are still valid for multi-TSV extraction.

Table 30: Multi-TSV extraction comparison with different libraries, where the total capacitance from Raphael simulation is 423fF.

	Ring Lib	Line Lib	Combined
Total Cap (fF)	386	459	419
Total Cap error	-8.7%	+8.5%	-0.9%
Correlation coefficient	0.986	0.988	0.989
Average error (fF)	0.100	0.087	0.063

4.5 Full-Chip TSV-to-Wire Impact

4.5.1 Design Specification and Analysis Flow

As a case study of TSV-to-wire coupling impact in the full-chip level, we use our FFT64 design. There are 47K gates in this design, which is partitioned into two dies. TSVs are 15μm in height and 2.5μm in radius with a landing pad size of 5μm. There are 330 signal TSVs in total which connect the M1 of the bottom die with back end of line (BEOL) of the top die. TSVs are placed regularly with a pitch of 18μm. To provide a fair comparison, we use the same 18μm as the TSV influence region for extraction with both single-TSV and multi-TSV libraries. The supply voltage is 1.1V as in our 45nm technology. In this work, we focus on extraction between TSVs and the top metal layer. This is because E-fields of other metal layers are usually blocked by PDNs, and signal routings on the top metal layer.

In our technology, metal dimensions are the same from M4 to M6. Therefore, the same library can be used to handle designs with top metal layer from M4 to M6. To provide wide coverage for our study, we implement two design variants. One uses up to M4 and the other uses up to M5, but both of them shares the same placement of gates and TSVs. The latter design has more routing resources than the other one. Thus, the router can better choose routing tracks on multiple metal layers to avoid heavy coupling between routed wires. As a result, the routing congestion on the top metal layer and the longest path delay (LPD) of the design up to M5 decrease, compared with the other design. Figure 48 compares top metal layer routing between these designs.

We apply our pattern matching method to FFT designs for TSV-to-wire extraction. For

TSV-to-TSV coupling extraction, a silicon-effect-aware multi-TSV coupling model is used, and 2D parasitics are extracted using Encounter. The same full-chip analysis flow is used where the full-chip static timing and power analysis is performed with Primetime, and the worst case noise analysis is performed with Hspice to measure TSV noise with an accurate multi-TSV model.

4.5.2 Full-Chip TSV-to-Wire Impact

Since the target of our design partition is to minimize the TSV count, and system performance is not taken into consideration, critical paths of both designs are a same 3D path. It starts from a register in the top die, goes through the bottom die by TSV89 and TSV274, and ends on another register of the top die. As a result, both TSV-related parasitics and top-metal parasitics affect full-chip timing results. We apply the pattern matching technique to both designs with both single-TSV and multi-TSV libraries. Table 31 summarizes full-chip TSV-to-wire impact on timing, power, and noise. Note that the LPD change only comes from TSV nets since we assume the clock network is ideal. If a real clock tree network is included, then the LPD is further affected since the clock signal needs to be delivered to the top die using TSVs as well, and TSV-to-wire parasitics affects the clock skew. Therefore, it also calls for fast and accurate TSV-to-wire coupling extraction for high quality clock tree synthesis.

For the design up to M4, if TSV-to-wire coupling capacitance is ignored, timing and noise analyses are inaccurate. From the result, the LPD is only 4.48ns without TSV-to-wire coupling. This is underestimated since interconnect capacitance is not fully captured. After TSV-to-wire capacitance is annotated from the SPEF file, the LPD increases to 4.83ns because of increased capacitance mostly on TSV89 and TSV274. Many other 3D nets are affected by TSV-to-wire coupling as well. Even if the critical path of the original design is not a 3D path, with TSV-to-wire coupling extracted, the critical path may change and timing impact is noticeable. Note that since 2D routers and timing optimization engines are not

Table 31: Full-chip impact of TSV-to-wire coupling on timing, power and noise. Both designs have 4.47pF total TSV MOS capacitance and 0.74pF total TSV-to-TSV coupling capacitance.

FFT64 design up to M4			
TSV-to-wire extraction method	none	single-TSV	multi-TSV
Total TSV-to-wire Cap (pF)	0	2.01	1.32
Longest path delay (ns)	4.48	5.08 (+13.4%)	4.83 (+7.81%)
Total TSV net power (mW)	0.303	0.356 (+17.6%)	0.335 (+10.6%)
Total net power (mW)	2.42	2.50 (+3.31%)	2.46 (+1.65%)
Total power (mW)	22.9	23.0 (+0.44%)	23.0 (+0.44%)
Average TSV noise (mV)	98.5	237 (+104%)	185 (+88.3%)
FFT64 design up to M5			
TSV-to-wire extraction method	none	single-TSV	multi-TSV
Total TSV-to-wire Cap (pF)	0	0.579	0.419
Longest path delay (ns)	4.43	4.58 (+3.39%)	4.50 (+1.58%)
Total TSV net power (mW)	0.302	0.316 (+4.64%)	0.310 (+2.65%)
Total net power (mW)	2.42	2.44 (+0.83%)	2.43 (+0.42%)
Total power (mW)	22.9	22.9 (+0%)	22.9 (+0%)
Average TSV noise (mV)	90.3	130 (+44.0%)	112 (+24.5%)

aware of the TSV-to-wire capacitance, not enough buffers are inserted to TSVs and wires on the top metal layer. This results in a large delay increase after TSV-to-wire extraction is applied. Pattern-matching-based extraction is preferred in the full-chip level because it can support fast and incremental estimation for timing and routing optimization. With correct TSV-to-wire parasitic information, timing paths with large interconnect capacitance can be effectively buffered so that timing violations can be addressed.

Also, ignoring the TSV-to-wire coupling results in a significant underestimation in TSV net noise. This is because traditional 2D extraction only extract parasitics from TSV landing pads, thus TSV coupling capacitance is heavily underestimated. Moreover, the influence region of a TSV landing pad is significantly smaller than that of a TSV. As a result, many aggressors are ignored with 2D extraction simply because wires are too far away. Even though TSV-to-TSV coupling contributes to TSV noise significantly, its impact is large if the TSV is tall and TSVs are placed closely. Without extracting TSV-to-wire coupling, average TSV noise is underestimated. From full-chip analyses, it increases to 185mV

with multi-TSV extraction. In terms of power, though there is a large increase in TSV net power, we only observe a negligible increase in total power resulting from TSV-to-wire coupling. This is because the major portion of total power is consumed by transistors, and 3D nets are only a small portion of all nets. Note that the FFT64 design is a small circuit, if a design has larger footprint with more TSVs and longer wirelength, the TSV-to-wire impact on the power will increase.

With shorter wirelength on the top metal layer, the FFT64 design up to M5 shows much smaller impact from TSV-to-wire coupling. Unlike the other design where the TSV-to-wire coupling capacitance is larger than TSV-to-TSV coupling capacitance, total TSV-to-wire coupling capacitance is reduced to 0.419pF with multi-TSV libraries. However, TSV-to-wire still has noticeable impact on timing as well as average TSV noise results. With less routing congestion, the design up to M5 has better performance than the other design. This provides an example of design and cost tradeoff by changing number of metal layers. With more metal layers and a higher cost, wires are less congested, parasitic components are reduced, and timing-aware routers can easily find better tracks to allocate signal wires.

As discussed in Section 4.4, ignoring E-field sharing among multiple TSVs and using single-TSV library overestimate TSV-to-wire coupling. By using the single-TSV library, total TSV-to-wire capacitance is 2.01pF, but it is reduced to 1.32pF using our multi-TSV libraries. This 34.3% reduction in coupling capacitance results in smaller TSV-to-wire impact on full-chip timing, power, and noise. For the design up to M5, TSV-to-wire coupling capacitance decreases by 36.0% with multi-TSV libraries as well. Therefore, for accurate TSV-to-wire coupling analyses, multi-TSV libraries are needed for full-chip analyses. If only a few TSVs are placed far away, single-TSV library can still provide accurate results.

4.6 Coupling Minimization

To alleviate TSV-to-wire coupling impact on timing and noise in the full-chip level, we propose two physical design approaches, *i.e.*, increasing Keep-Out-Zone and guard ring

protection. In this work, all TSV-to-wire coupling results are based on multi-TSV extraction which accounts for four nearest neighboring TSVs.

4.6.1 Increasing Keep-Out-Zone

Since TSV-to-wire coupling is majorally between TSVs and the nearest metal layer. Therefore, a simple technique for TSV-to-wire coupling reduction is increasing the minimum distance between the TSV and its nearest routing wire. To implement this method, we place a routing blockage on top metal layer around each TSV. This blockage region is the routing KOZ. To study impact of various KOZ sizes, we implement two designs with KOZ sizes of 2.5µm and 5µm, respectively. Figure 49(a) shows design layouts with KOZs. The original design has KOZs of 0.5µm. With a larger KOZ, capacitance between a TSV and its nearest wire further decreases because of a weaker coupling E-field. In addition, a larger KOZ reduces routing resources available on the top metal layer. This results in a reduction in the top metal layer wirelength and the number of aggressors around the TSV. Note that increasing the routing KOZ does not have any silicon area overhead. Therefore, the placement is the same as the original design placement, and only incremental routing is performed to correct any routing violations. Thus, layouts have minimum changes, and it is a fair comparison among all designs.

However, one drawback for increasing the KOZ is that we observe heavier routing congestion on other layer as a result of the reduced number of routing tracks on the top metal layer. Potentially, this may result in a degradation of design quality and a increase in coupling noise between 2D wires. Figure 50 compares wirelength distribution with different KOZ sizes. For the FFT64 design up to M4, since top metal wires are reduced by 43.7% with 5µm KOZs, wire congestion on other metal layers is more severe. Wirelength on M2 increases by 29.9% and total wirelength increases from 373.9mm to 376.3mm. On the contrary, since the design up to M5 has enough routing resources and its top metal wirelength is small, increasing the KOZ only has slight impact. Therefore, for some designs which

Table 32: Keep-out-zone impact on FFT64 design up to M4.

KOZ size (UM)	0.5	2.5	5
M4 wirelength (mm)	82.5	70.5 (-14.5%)	46.5 (-43.6%)
Total wirelength (mm)	374	375 (+0.19%)	376 (+0.64%)
Longest path delay (ns)	4.83	4.77 (-1.2%)	4.64 (-3.9%)
Total TSV net power (mW)	0.335	0.326 (-2.7%)	0.316 (-5.7%)
Total net power (mW)	2.46	2.45 (-0.4%)	2.43 (-1.2%)
Average TSV noise (mV)	185	165 (-11.1%)	139 (-25.0%)

have limited routing resources, increasing the KOZ size may not be beneficial because of increased routing congestion.

We perform the full-chip analysis on designs with routing KOZs, and results are summarized in Table 32. Since the placement is the same, TSV-to-TSV coupling elements are unchanged. With larger KOZs, top metal layer wirelength is reduced, and their coupling capacitance is reduced. Therefore, TSV-to-wire coupling also shows smaller impact on full-chip timing and noise. Compared with the original design, the LPD decreases by 1.2% and 3.9% for 2.5μm and 5μm KOZs, respectively. In terms of signal integrity, a larger KOZ also lowers the TSV net noise by reducing the aggressor count and TSV-to-wire coupling capacitance. The average worst-case noise on TSV nets can be reduced by 11.1% and 25.1% for 2.5μm and 5μm KOZs, respectively. On the other hand, KOZ impact on the full-chip power result is much smaller since TSV-to-wire capacitance decreases but wire-to-wire capacitance increases on other layers. Overall, we conclude that increase top metal layer KOZ is effective in reducing TSV-to-wire coupling at the cost of higher routing congestion.

4.6.2 Guard Ring Protection

Another way to protect the victim TSV is to provide E-field shielding using grounded conductors around TSVs. Similar technologies are widely used to increase the SNR in communication applications. In this work, we propose a physical design optimization technique specifically designed to reduce TSV-to-wire coupling. Unlike the previous work where

grounded guard rings in active layer are added around TSV [10], grounded wire guard rings on the top metal layer are inserted around TSV. Therefore, there is no overhead on silicon areas and standard fabrication technology is used. However, the guard ring consumes some routing tracks on the top metal layer and needs additional routing to connect the ring to the ground. The grounded guard ring shields some E-fields around TSV, and introduces a ground capacitor to the TSV net. As a result, there are small delay and power overheads on TSV nets, but it reduces TSV coupling noise. Moreover, the guard ring now becomes the nearest wire around the TSV, thus all other wires have neighbors on both sides. Therefore, coupling capacitance between the victim TSV and signal wires is reduced and the TSV is better shielded with coupling noise.

To model guard ring impact, we build a guard ring model shown in Figure 51(a), where C_{TW}, C_{TG}, and C_{WG} represent TSV-to-wire, TSV-to-ring, and wire-to-ring capacitance, respectively. The ring is assumed to connect with ground ideally, and a 10μm long wire on M4 is located 8μm far from the center of the TSV. We perform Raphael simulations on guard ring structures with various guard ring widths, and results are shown in Figure 51(b). With a wider guard ring, the TSV is better shielded from TSV-to-wire coupling, thus TSV-to-wire coupling capacitance is smaller. However, a wider guard ring increases ground capacitance on both the TSV and the wire, which leads to small delay and power increases. Therefore, it is better to protect TSVs which are not located on the critical path so that additional ground capacitance has no impact on design performance. In this work, we insert wire guard rings to every TSV so that we can observe performance impact from guard rings.

To study full-chip impact of wire guard rings, we build three libraries with multi-shielded TSVs, *i.e.*, line library, ring library, and corner library, in which TSVs are surrounded by grounded guard rings. To study the guard ring width impact, libraries are built with both 0.5μm and 1.5μm guard rings. Figure 52 shows a sample structure from the line library, where each TSV is surrounded by a 1.5μm guard ring. In addition, we implement

FFT64 designs with 0.5μm and 1.5μm guard rings on the top metal layer. Figure 49(b) shows die layouts with guard rings. These two designs are based on our previous FFT64 design with 2.5μm KOZs. We insert the guard ring into the KOZ so that placement and signal routing are kept the same. Similarly to KOZ insertion, guard rings also consume routing resources on the top metal layer. However, they provide better protections to TSVs.

We perform TSV-to-wire extraction using these libraries with multi-shielded TSVs and full-chip analysis results are summarized in Table 33 for the design up to M4. Compared with the original design, the LPD increases by 0.21% and 1.45% for designs with 0.5μm and 1.5μm guard rings, respectively. This impact comes from two aspects: Ground capacitance on TSV nets increases but TSV-to-wire capacitance decreases. If they are compared with the design with 2.5μm KOZs, delay overheads from larger ground capacitance are shown clearly: The LPD increases 1.4% and 2.7% for designs with 0.5μm and 1.5μm guard rings, respectively. Total capacitance on TSV nets always increases with wider guard rings since E-fields around TSVs are stronger, and we observe a slight timing overhead from increased capacitance. Guard ring impact on power is negligible, since TSV MOS capacitance is the major load on TSV nets and it is not changed. From results, we find that the ground guard ring is very effective in TSV net noise reduction. Compared with the design with 2.5μm KOZs, the average TSV net noise decreases by 11.6% and 16.4% with 0.5μm and 1.5μm guard ring, respectively. Compared to the original design, the average TSV net noise decreases by 21.4% and 25.6% with 0.5μm and 1.5μm guard ring, respectively. Meanwhile, for the design up to M5, we only observe a noticeable noise reduction and there is a negligible timing and power overhead, since TSV-to-wire coupling is much weaker on this design. Overall, we conclude from our full-chip results that both increasing KOZ and inserting guard rings are very effective for TSV-to-wire coupling noise reduction with minimum overheads on design qualities.

Table 33: Guard ring impact on full-chip designs.

Guard ring width (μm)	0	0.5	1.5
KOZ size (μm)	0.5	2.5	2.5
Longest path delay (ns)	4.83	4.84 (+0.21%)	4.90 (+1.45%)
Total TSV net power (mW)	0.335	0.340 (+1.49%)	0.349 (+4.18%)
Total net power (mW)	2.46	2.46 (+0%)	2.47 (+0.41%)
Average TSV noise (V)	185	146 (-21.4%)	138 (-25.6%)

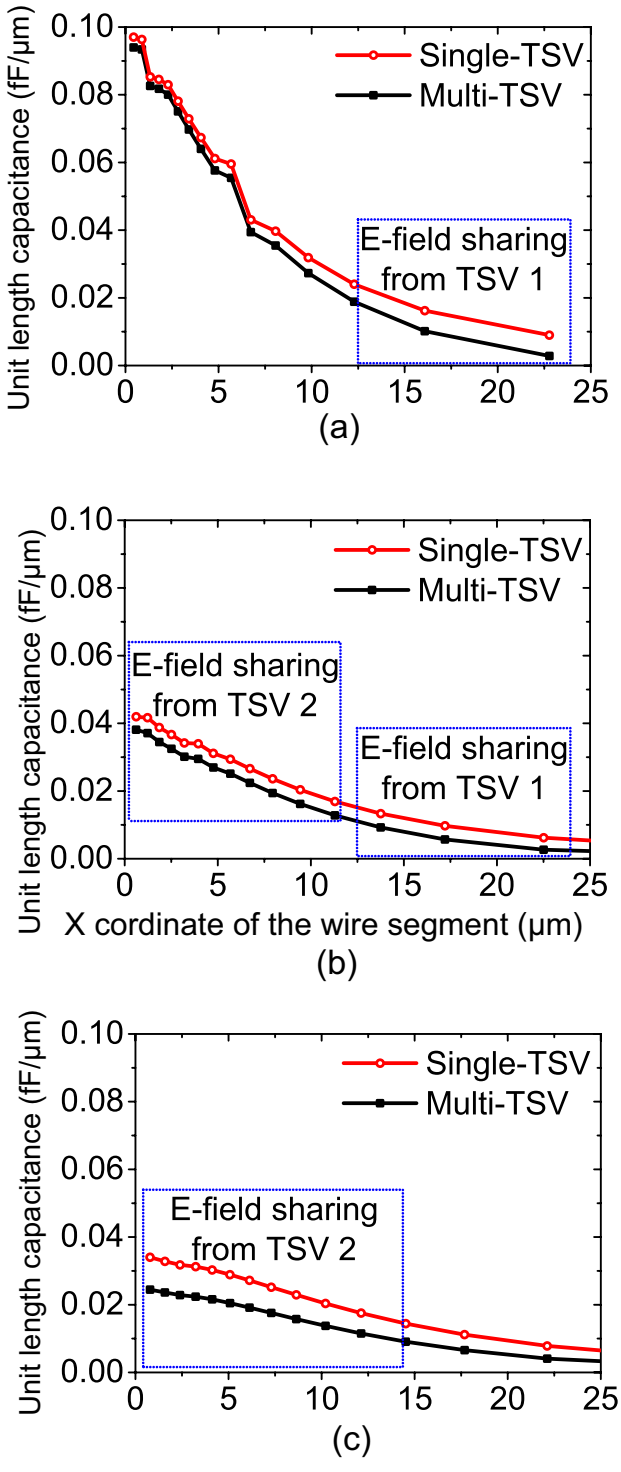

Figure 45: Coupling capacitance extraction result of Figure 43. (a) Nearest wire, (b) middle wire, (c) furthest wire.

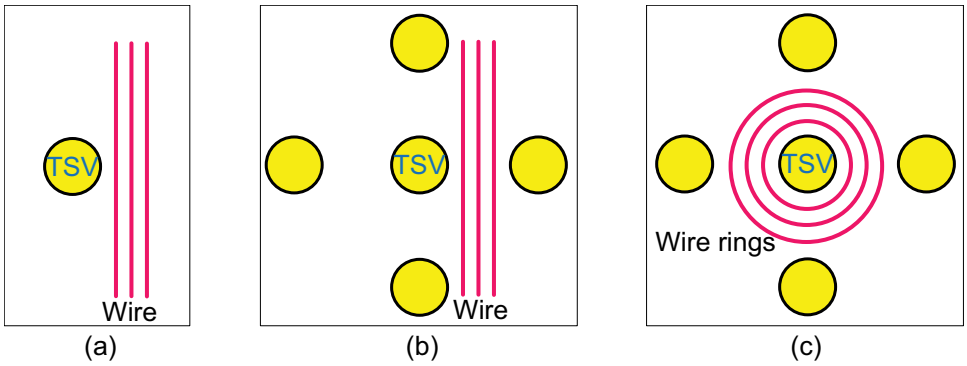

Figure 46: Library structure comparison. (a), (b), and (c) show the single-TSV line library, multi-TSV line library, and multi-TSV ring library, respectively.

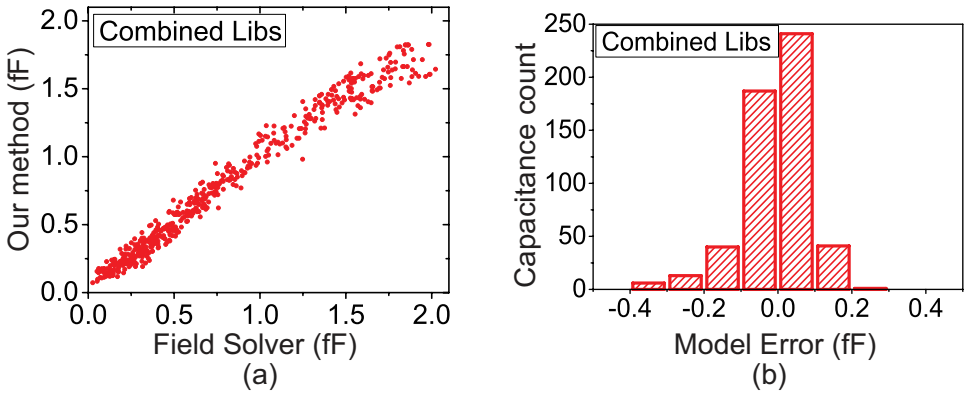

Figure 47: Multi-TSV extraction verification. (a) shows correlation comparison between our pattern-matching algorithm and Raphael simulations, (b) shows error histograms of different libraries.

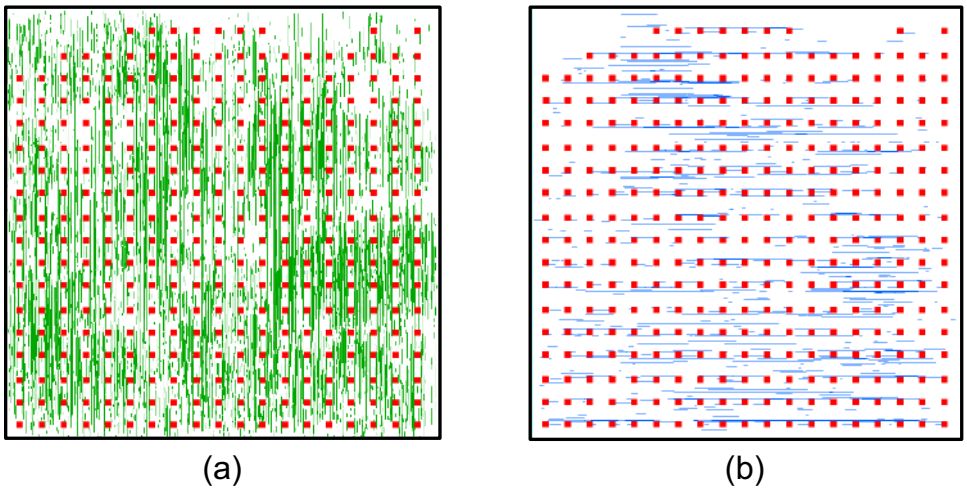

Figure 48: Top metal routing comparison. Only top dies are shown. (a) Design up to M4, (b) design up to M5.

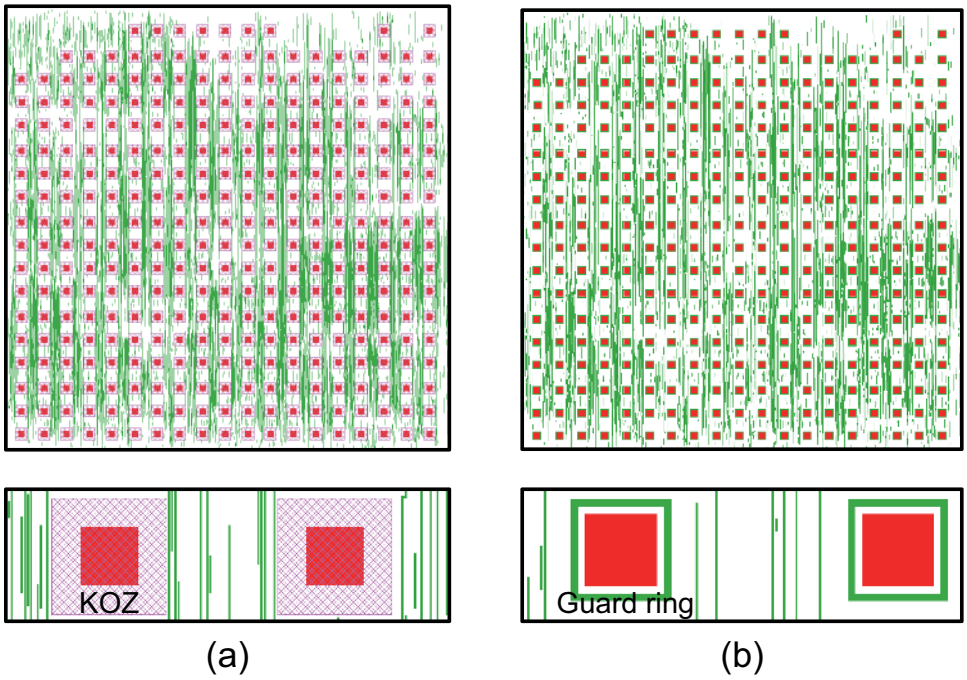

Figure 49: Top die layout and zoom-in shots of FFT64 designs up to M4. (a) With 2.5µm KOZ, (b) with 0.5µm guard ring.

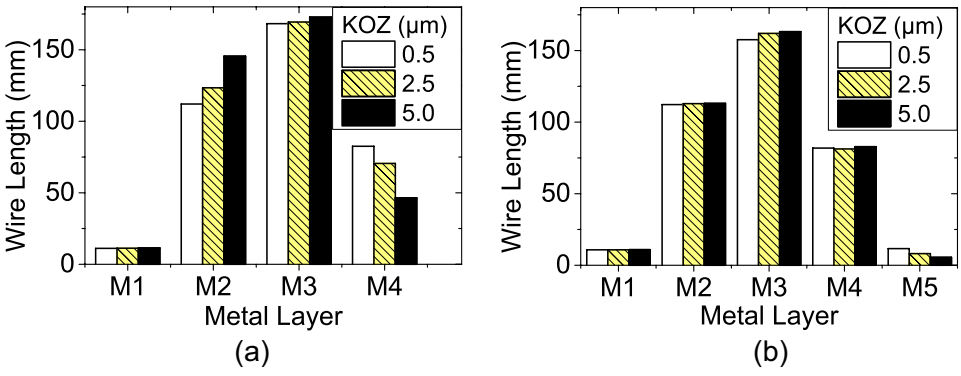

Figure 50: KOZ impact on wire length usage. (a) design up to M4, (b) design up to M5.

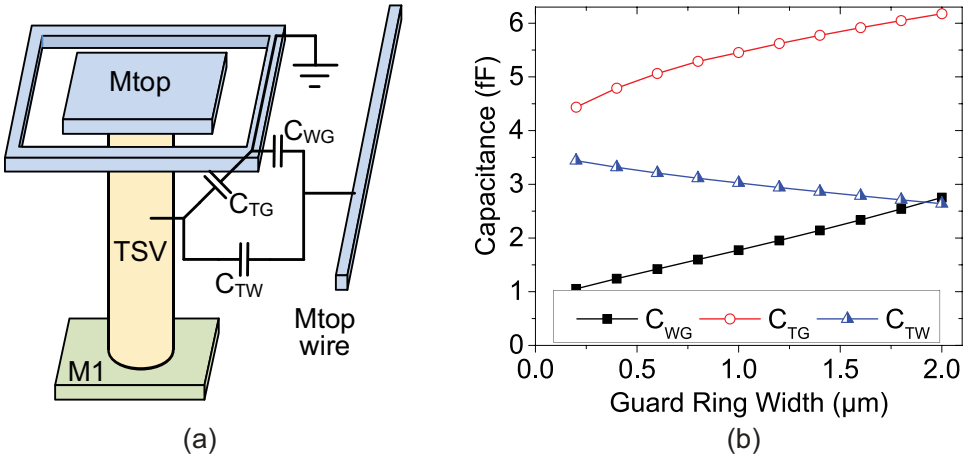

Figure 51: Guard ring capacitance model. (a) shows the simulated structure (b) shows the Raphael extraction result.

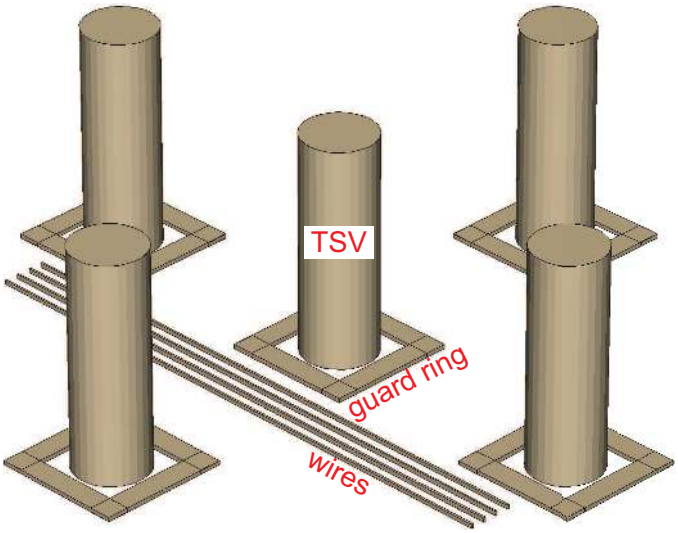

Figure 52: Sample multi-TSV line library structure with 1.5μm guard ring.

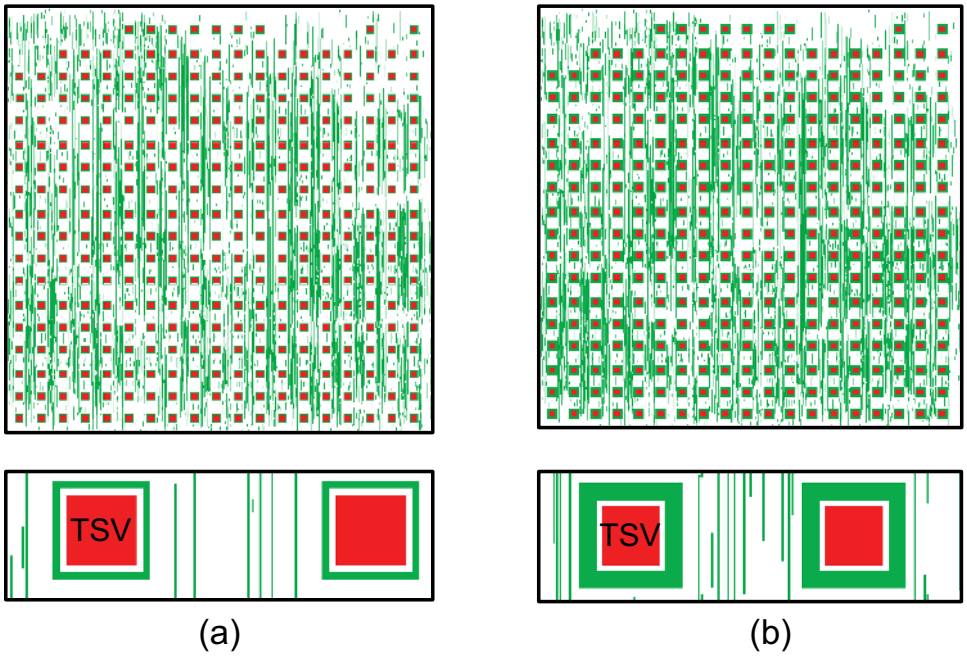

Figure 53: Top die layout and zoom-in shots of shielded FFT64 designs up to M4. Only M4 routing is shown. (a) 0.5μm guard ring, (b) 1.5μm guard ring.

CHAPTER V

INTER-DIE COUPLING EXTRACTION METHODOLOGIES IN FACE-TO-FACE 3D ICS

Traditional technology scaling in sub-20nm nodes is expensive. To lower cost, reduce power consumption, and increase signal bandwidth on a smaller footprint, 3D ICs are promising solutions to extend Moore's Law. A common 3D IC technology uses face-to-back (F2B) bonding, which builds through-silicon vias (TSVs) in the silicon substrate as vertical interconnects. With this technology, however, increasing 3D via density is difficult because TSVs penetrate a thick silicon substrate, and fabricating TSVs with high aspect ratio is prohibitively expensive and complex. Unlike F2B bonding, in which the vertical interconnection density is limited by the TSV size, face-to-face (F2F) bonding technology connects top metal layers from both dies with F2F vias [69]. F2F designs achieve much higher three-dimensional (3D) connection density with F2F vias in a few microns [70].

5.1 F2F Extraction Methodologies

5.1.1 Die-by-die Extraction

In order to handle various F2F technologies and configurations, we propose and exam three extraction methods in this work. First, the die-by-die extraction extracts the bottom and top dies individually similarly to current 2D IC extraction, as shown in Figure 54(a). It ignores coupling capacitance between dies and can be implemented easily using traditional 2D extraction engines such as Calibre xACT [71]. Presuming extractions for each die, the only requirement is a method that can stitch together these individual die netlists with parasitics. The die-by-die extraction is accurate as long as the die-to-die distance is large and the E-fields from both dies do not couple to each other. It can also be applied if the top

metals from both dies are perfectly shielded and decoupled by, e.g., PDN layers. However, even in such cases, the second ground plane from the neighboring die should also be considered during technology characterization for an accurate extraction, since the ground capacitance of the top metal layers increases. On the other hand, the die-by-die extraction is also considered as "LVS-friendly", since LVS can be done without knowing any geometries from the neighbor die. Since any sign-off parasitic extraction needs to be performed after LVS check and all the geometry are properly netlisted, the die-by-die extraction completely decouples the designs of each dies allowing for a faster time-to-market and easier industrial collaboration which are critical to allow parasitic extraction of heterogeneous 3D ICs. Therefore, die-by-die extraction is currently used in commercialized technology, and demonstrated F2F 3D ICs are all based on this extraction technique, since with current technology the die-to-die distance is still much larger than any wire thickness.

5.1.2 Holistic Extraction

The second method is the holistic extraction, where all layers from both dies are taken into account during technology calibration and parasitic extraction. As shown in Figure 54(b), this extraction requires a full knowledge of both dies, from device layer all the way to the top metal layer. By performing a holistic LVS of all dies, the geometry connectivity can be fully netlisted. It can achieve maximum accuracy by capturing all E-fields from both dies, therefore, this work uses holistic extraction as a reference in our F2F extraction, and compare other extraction methods to it. However, holistic extraction is extremely challenging computationally both during pre-calibration and runtime. First, considering all layers requires many more library structures to be built and there are more combination of possible structures. This can significantly increase calibration time. Ideally, it is upon system designer to choose vendors for each components.

For heterogeneous integration, it is difficult to consider all possible combination of different technologies from multiple foundries beforehand, especially with various metal

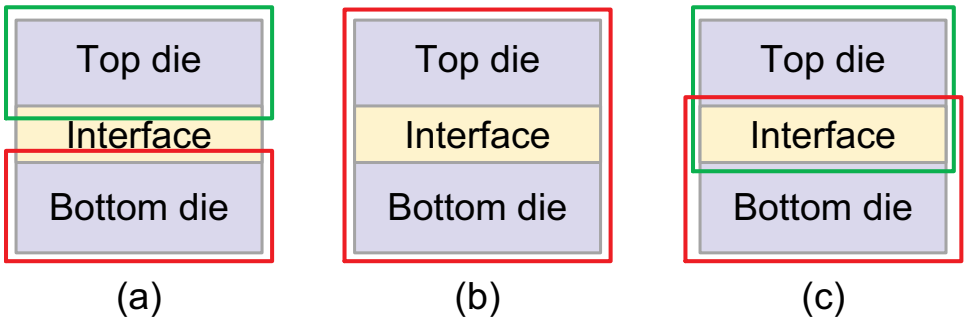

Figure 54: Comparison of F2F extraction methods. (a) die-by-die, (b) holistic, and (c) in-context extraction.

stack configurations. Moreover, it requires to code holistic LVS and extraction rule decks that can properly recognize all the devices, connect two dies stacked on top of each other, and perform extraction for all layers. As dies may from different technologies, foundries need to share all information of their technologies including critical layers such as devices and local interconnects which are needed for holistic rule decks. For homogeneous 3D ICs, where both dies are from the same foundry, it is not impossible but it takes time to regenerate rules for both dies and carefully resolve any layer conflicts. However, if multiple foundry technologies are used, it requires foundries to share their critical trade secretes to the public and their competitors. Either one foundry is responsible to incorporate layers from the neighboring die, and maintain the holistic rule deck, or a third party, likely a packaging house for F2F bonding, is required to combine rule decks, which are generally encrypted they are sent to design houses. Not only it is time consuming to resolve conflicts and combine LVS and extraction rules, but also it threats intellectual property protection of design houses. Designers for both dies need to reveal all of their layouts and netlists, which opens doors to back-engineering. Therefore, though holistic extraction may be possible with homogeneous 3D ICs, it may not be efficient and realistic for commercial use, especially with Heterogeneous 3D ICs.

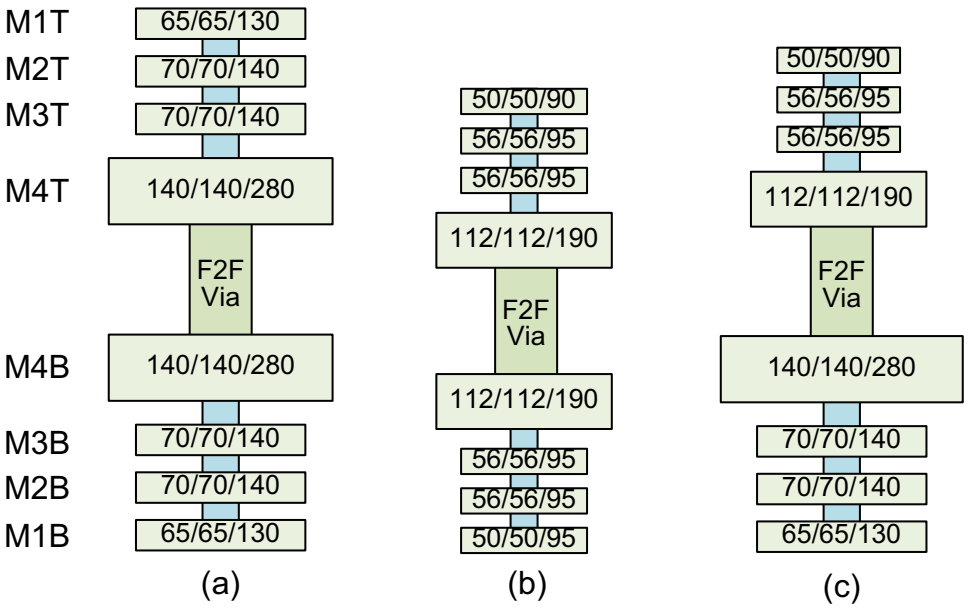

Figure 55: Technology configurations with 1μm F2F via thickness. (a) and (b) are homogeneous with 45nm and 28nm for both dies, respectively. (c) is a heterogeneous technology, where bottom die uses 45nm and top die uses 28nm.

5.1.3 In-context Extraction

To improve parasitic extraction accuracy without imposing the need for detailed information from the neighboring die, in-context extraction is proposed to take advantages of die-by-die extraction without losing much accuracy compared with holistic extraction. Previous study has shown that most of the coupling E-field are formed within limited depth into the other die [72], we define this as the "coupling depth." It often can only reach one or two layers, thus only coupling capacitances between neighboring layers are significant enough to affect any full-chip analysis. Therefore, to efficiently perform extraction without sacrificing accuracy, in-context extraction only takes a few layers, called "interface layers," from the neighboring die into account during both technology calibration and parasitic extraction. As shown in Figure 54(c), similar to die-by-die extraction, the top and bottom dies are extracted separately but both are extracted with the knowledge of interface layers. Dies with interface layers from the neighbor are called "in-context dies." Though still need extra layers, in-context extraction significantly reduces number of pre-calibrated structures

required since structures with small dimensions, such as devices and local metals, are not included. Also, for advanced technologies nodes, thousands of design rules strictly applying to those small structures can also be avoided and it is much easier to handle top metals with larger dimensions. On the other hand, in-context extraction still remains LVS-friendly, because only a few top metal layers are needed to code an LVS deck for in-context dies. And new rule decks can be calibrated incrementally by reusing of existing rule decks. Note that revealing the non-critical properties, such metal dimensions and dielectric properties are not critical issues, calibration of heterogeneous 3D IC technologies needs only extend the existing 2D rule files. Therefore, this approach reduces the complexity of handling all layers simultaneously and can be carried out independent of device fabrication process. Previous work implemented the in-context extraction with homogeneous technology, in this work, we are focused on heterogeneous 3D IC integration.

5.2 Field Sharing Analysis

To find out how two E-fields from both dies interact with each other, we build a test structure shown in Figure 56. The ground planes are located 3μm away from wires, and wire width (w) and thickness (t) are fixed as 0.8μm and 1.2μm, which are the same as top metal layer dimensions in a 45nm technology. We duplicate patterns of Net A and B on the top die, and Net C and D on the bottom die. In this structure, the coupling capacitance can be divided into three groups: intra-die coupling capacitance, inter-die overlapping capacitance, and inter-die fringe capacitance. The repeated patterns ensure that any capacitors of the same kind have the same value. Therefore, intra-die coupling, inter-die overlapping, and inter-die fringe capacitance can be represented by Cap AB (or Cap CD), Cap AC, and Cap AD (or Cap BC), respectively. Note that because of the symmetric structure, total intra-die capacitance can be represented by 2x Cap AB while total inter-die capacitance can be represented by 1x Cap AC plus 2x Cap AD. Capacitance is extracted assuming an infinite wire length with a 2D extraction with a unit of fF/μm.

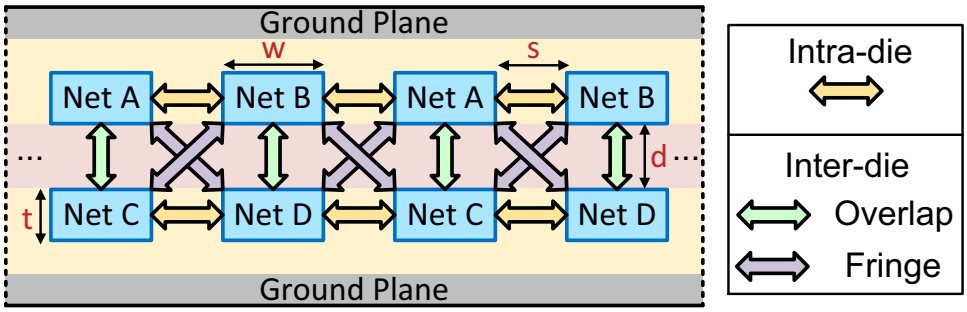

Figure 56: Raphael structure for capacitance extraction. Both the top and bottom dies contain repeated layout patterns. D denotes the die-to-die distance while w, s, and t denote wire width, spacing, and thickness, respectively.

5.2.1 Die-to-die Distance Impact

First, we vary the die-to-die distance (d) from 0.5μm to 8μm and find out its impact on the coupling capacitance. Field solver extraction results are shown in Figure 57, where capacitance values are taken by measuring the average of ten wires on each die. The wire spacing (s) is fixed as 0.9μm, which is the minimum spacing of M4 to M6 in the target technology. With a closer die-to-die distance, inter-die coupling capacitance (represented by Cap AC) increases significantly, while inter-die fringe capacitance (represented by Cap AD) increases slightly. Also, because of the E-field sharing from the neighbor die, with a closer die-to-die distance, intra-die coupling capacitance decreases. It only changes slightly when dies are far from each other, but it decreases significantly when die-to-die distance is less than 5μm since the E-field sharing from the other die is much stronger. And with a die-to-die distance smaller than 1μm, inter-die coupling capacitance becomes comparable to intra-die coupling capacitance even with minimum wire spacing. Therefore, inter-die coupling can no longer be ignored with a close die-to-die distance. Shown in [73], if the die-to-die distance is similar to the top metal dimensions, the inter-die coupling becomes comparable to the intra-die coupling of the top wires. Note that the total capacitance always increases with a closer die-to-die distance, and the portion of inter-die coupling keeps increasing as well. Therefore, die-by-die extraction, which is unaware of the neighboring die and ignores the inter-die E-field sharing, cannot extract the inter-die coupling capacitance

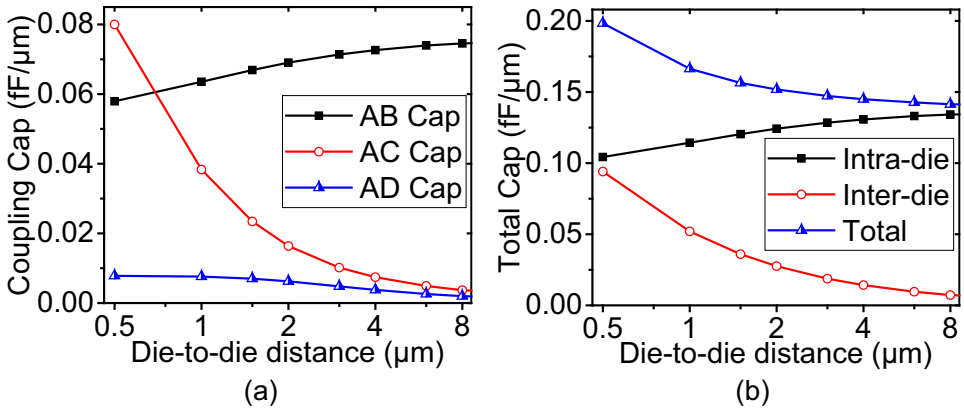

Figure 57: Die-to-die distance (= d in Figure 56) impact. (a) Single capacitor extraction, A to D are nets in Figure 56; (b) total capacitance extraction.

accurately when die-to-die distance is smaller than 5μm in this technology.

5.2.2 Wire Spacing Impact

Then, we vary the wire spacing while keep the die-to-die distance to be 1μm. Raphael extraction results are shown in Figure 58. With a large wire spacing, both intra-die coupling and total coupling capacitance decrease. However, the inter-die coupling capacitance percentage increases with a larger wire pitch. Also, E-field sharing from neighboring wires within the same die is weaker, thus stronger coupling is formed between overlapped wires across dies, which is the major portion in the inter-die capacitance. As a result, total inter-die capacitance increases with a wire spacing up to 3μm. However, if wire spacing increases further, intra-die E-field sharing is very weak, the increase of overlap capacitance (Cap AC) saturates. Therefore, total inter-die capacitance slightly decreases with a smaller fringe capacitance (Cap AD). Overall, inter-die capacitance becomes comparable to intra-die capacitance with a wire spacing larger than 1μm. From these results, inter-die coupling cannot be ignored in designs with sparsely-routed top metal layers, while intra-die E-field sharing cannot be ignored with densely routed wires during parasitic extraction.

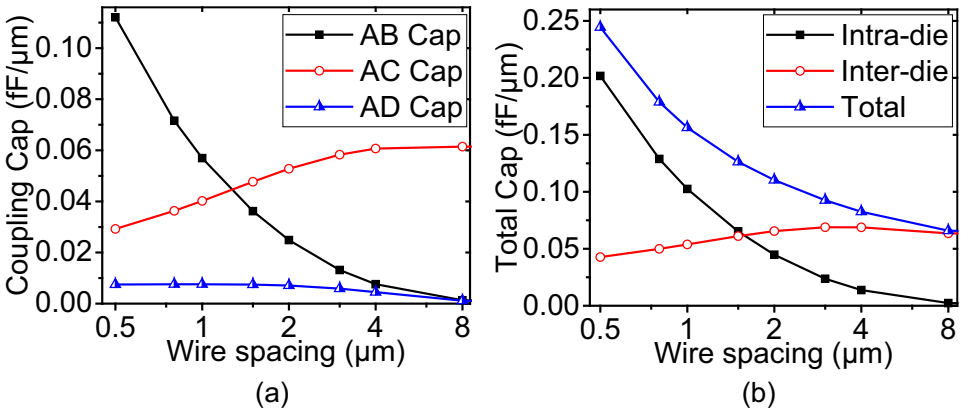

Figure 58: Wire-to-wire spacing (= s in Figure 56) impact. (a) Single capacitor extraction, A to D are nets in Figure 56, (b) total capacitance extraction.

5.3 Full-Chip Extraction Flows

In this section, we demonstrate our CAD flows of all three extraction methods discussed in Section 5.1. Our flows can be easily ported to any full-chip extraction extraction engines such as Calibre xACT or xRC.

5.3.1 Die-by-die Extraction

The CAD flow of homogeneous die-by-die extraction is shown in Figure 59. If a heterogeneous technology is used, two sets of extraction rules can be calibrated independently. A sample technology with four metal layers is shown in Figure 60(a) for die-by-die extraction, where the same 2D technology calibration can be used. Since currently no commercial design tool is able to handle timing and power optimization of 3D designs, commercial 3D ICs have their dies designed separately only with pre-defined 3D vias as interface to the neighboring dies, which reduces CAD complexity and accelerates the design process. LVS can be done similarly as a 2D design to match the layouts or extract a netlist for parasitic extraction. And after parasitics are obtained from both top and bottom dies, designers need to include a top-level netlist, which describes 3D connections and I/O interfaces between dies, as well as a top-level parasitic file which includes capacitance of 3D vias. The full-chip analysis can be easily performed by merging of all the parasitics with a connected

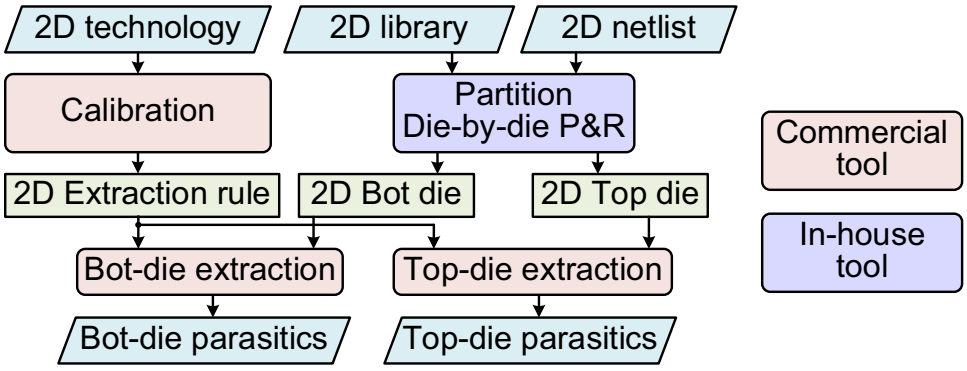

Figure 59: CAD flow chart of our die-by-die extraction.

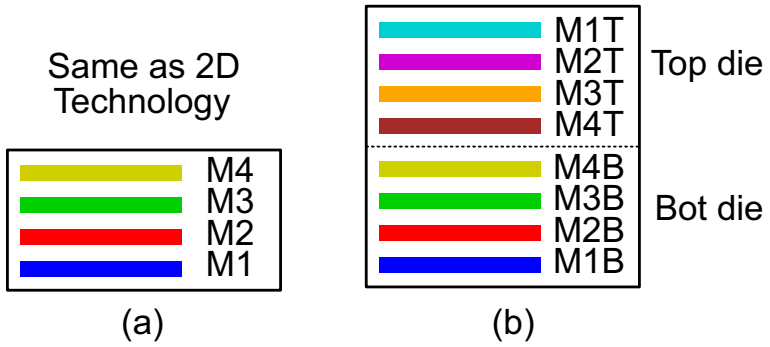

Figure 60: Sample interconnect technologies with four metal layers. (a) Die-by-die extraction, and (b) holistic extraction.

netlist for the whole system. Ignoring the inter-die capacitance, this flow is widely adopted for both F2F and F2B designs, and it is the fastest approach and the only feasible way nowadays.

5.3.2 Holistic Extraction

Compared with the die-by-die approach, holistic extraction requires to consider all layers simultaneously as shown in Figure 60(b). The metal layers located in the bottom die are denoted with a postfix of "B" while the metal layers in the top die are with "T". With F2F bonding, top metal layers from both dies are heavily coupled. Especially when only a few metal layers are used, the inter-die coupling capacitance consumes a large portion of total coupling capacitance. However, there is currently no commercial full-chip extraction engine which is able to handle two device layers simultaneously. Therefore, we implement

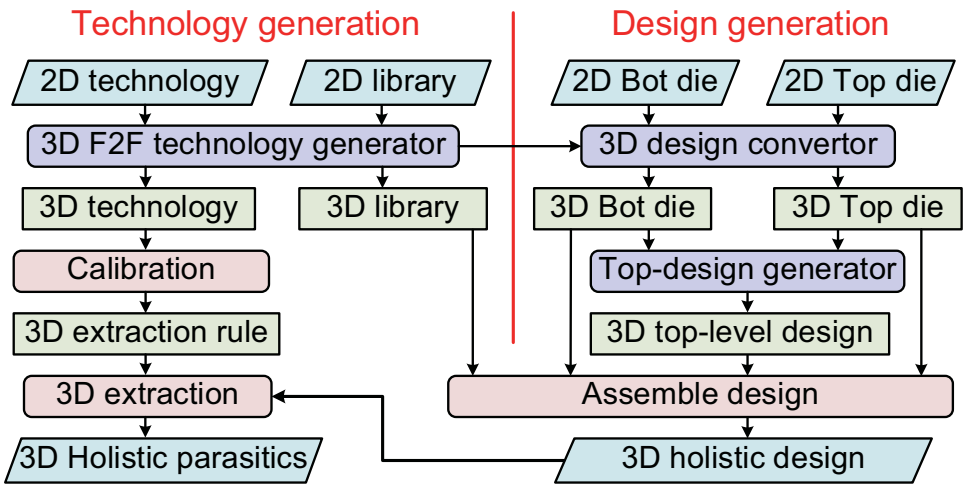

Figure 61: CAD flow chart of our holistic extraction.

the holistic extraction flow shown in Figure 61 by considering the top die device layer as a conducting layer. This will introduce some errors mostly on the M1T layer in holistic extraction. However, it still gives reasonable results since parasitics inside standard cells should be extracted separately and included in the post-layout cell netlist, and its timing and power impacts should be considered by cell characterization. Since most of M1 areas are used for intra-cell connections, we are performing a full-chip top-level extraction with very few M1 wires for inter-cell connections, thus only a small portion of coupling capacitance is formed on M1 layers of both dies.

Our holistic extraction flow contains both in-house tools and commercial tools for design automation. We are reusing commercial extraction tools to provide silicon-validated results for our study. First, to create a holistic technology, a technology generator reads the 2D technology and library, and duplicates metal layers and cells in the F2F fashion as shown in Figure 60(b). In order to avoid naming conflict, the cells located in the top and bottom dies are renamed with different postfixes and their pin layers are renamed accordingly to indicate which die they are from. The generated 3D technology and library contain all metal layers as well as the bottom die substrate and device layer. Note that it is also possible to apply holistic extraction to heterogeneous extraction as long as the extraction

tool is able to handle wires with various dimensions and metal pitches from different technologies. For F2F bonding layer, we adopt the design method used in [74], in which F2F connections from both dies are combined into F2F vias between top metal layers of both dies. This is to make the 3D technology similar to a 2D metal stack where a via layer is between any two neighboring metal layers. With all layers calibrated, holistic extraction is able to fully cover all E-field interactions inside the F2F bonding layer as well as any E-field sharing impacts from metal layers. Compared with die-by-die extraction, longer calibration time is required for solving E-field and building 3D technology libraries, however, once these extraction rules are generated, the runtime of full-chip holistic extraction is still comparable with die-by-die extraction, since all layers are extracted in a single run.

Since current physical design flow implements each die in 3D ICs separately, we implement a CAD flow to generate the the 3D holistic design from die-by-die designs. First, a 3D design convertor takes in both designs and converts each design according to the output of the 3D technology generator. Cells and layers are renamed according to their host dies so that they are compatible with our holistic technology. Then, by taking the LEFs of both dies, our top-design generator creates a top-level layout which has the same footprint as the 3D chip but only contains dies and F2F via connections. In this design, the top die and bottom die are overlapped design blocks in the floorplan with F2F vias as block pins. Note that for a heterogeneous design with different top and bottom die sizes, the top-level design has to align the F2F vias from both dies to ensure a valid connection. With all three designs ready (*i.e.*, the top-level design as well as two converted 3D dies), there are two ways of performing holistic extraction. If the extraction engine supports hierarchical extraction, such as Calibre xACT, simply supplying the design files of the top-level floorplan as well as both dies is enough. Without cell and layer conflicts, parasitic extraction can be performed as if dies are sub-blocks in the top-level design. Another way is to use the assemble design feature from physical design tools. It reads the sub-block layouts and flatten the top-level

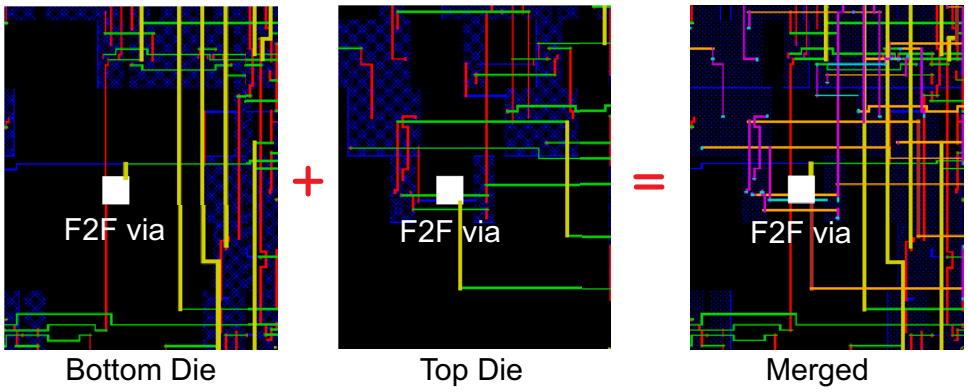

Figure 62: 3D holistic design generation.

design which can then be supplied to extraction tools. Or it can be done by merging individual die GDS files, and perform extraction with the holistic GDS. Note that cells from both dies will overlap on the floorplan, thus it cannot pass the geometry check. Luckily, all major extraction tools are able to handle this layouts without problem. Figure 62 illustrates this design merging process.

5.4 In-Context Extraction

We also implement our first-of-its-kind in-context extraction flow combining both commercial engines with our in-house tools. Our goal is to use a similar flow as traditional die-by-die extraction but with an inter-die extraction accuracy similar to that of the holistic design. There are two intuitive way to extend dies with extra layers. Either growing extra layers from a single die or excluding unnecessary layers from the holistic design. We take the latter method as the holistic design contains connectivity information in the netlist to avoid touching the design netlist.

5.4.1 Technology and Design Generation

Unlike holistic extraction, which handles multiple substrate and device layers simultaneously, in-context extraction does not require to create new extraction engines for multiple dies, thus our flow is fully compatible with all major CAD tool vendors. For naming convenience, we use “In-C:N” to denote in-context extraction with N interface layers per die.

Note that holistic extraction can be considered as a special case of in-context extraction, where all metal layers become interface layers. Also, our flow is able to handle heterogeneous 3D ICs even with mismatched die footprint or unsymmetrical F2F bonded designs in which number of metal layers or interface layers from top and bottom dies are not the same.

To enable such extraction, for each in-context die, we must include enough data about connectivity and geometries from its neighboring die. Our in-context flow for homogeneous 3D ICs is shown in Figure 63. If a heterogeneous design is used, in-context extraction rules for bottom and top in-context dies require to be calibrated separately. For technology generation, we simply extend the basic 2D technology and library to create in-context technology files so that there is minimum changes to the technology description files. Also, incremental calibration can be applied to reuse existing rule decks and ensure the silicon-validated 2D extraction rules are unchanged.

An example with four metal layers and one interface layer per die is shown in Figure 64. For the bottom die, we need to add the top die interface layer, which is recognized as M5 by CAD tools. Similarly for the top die, the M4B layer is recognized as the new M5 layer. Note that if both in-context dies have exact the same layer stack, only one technology calibration is required and generated rule decks can be reused. We call the outmost metal layer in our in-context technology "surface layer", though no metal layer is physically located at the surface in real F2F technology. For example, with the metal stack (In-C:1) shown in Figure 64, M4B and M4T layers are surface layers of top and bottom in-context dies, respectively. Similarly for In-C:2 extraction, M3B and M3T become surfaces layers of top and bottom dies. The surface layer is special since it has one missing neighbor layer in the in-context technology. Since each in-context technology only includes one substrate and device layer, it can be calibrated similarly as a traditional 2D technology.

Similar to holistic extraction, our generator takes in design files from both dies and renames the cells and layers. However, only interface layers are included during design

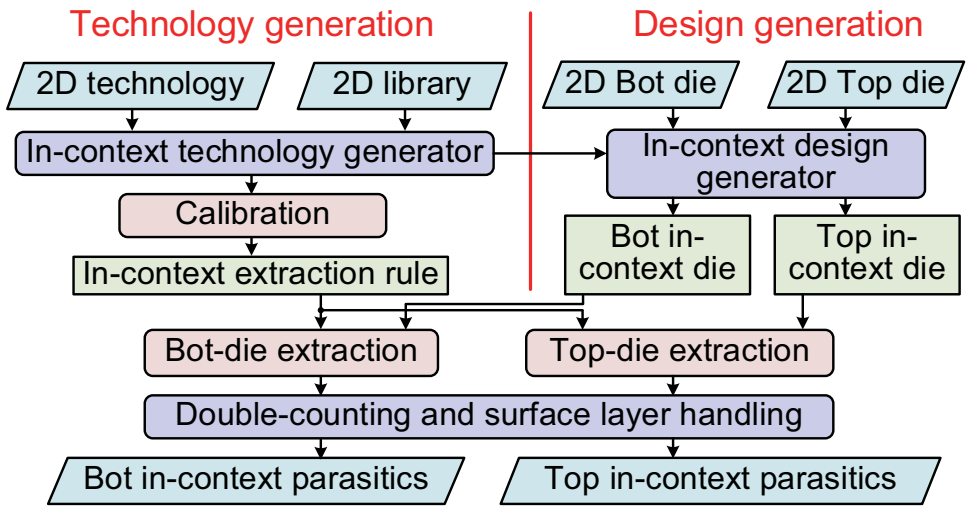

Figure 63: CAD flow chart of our in-context extraction.

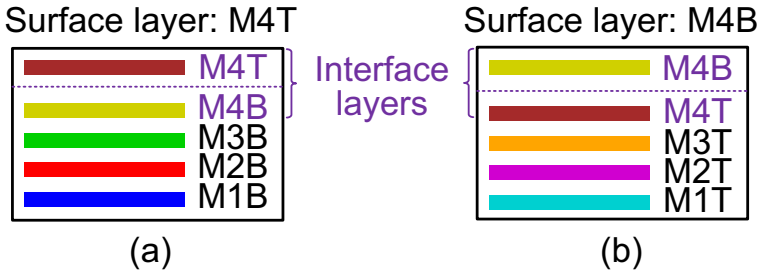

Figure 64: A sample in-context interconnect technology with four metal layers.

layout merging, and other layers as well as cells from the neighboring die are discarded. Therefore, bottom and top in-context designs are generated separately matching with the previously generated in-context technology. Figure 65 illustrates an in-context design generation process for technology shown in Figure 64. After in-context designs are generated, they are extracted similarly as the die-by-die flow. Another advantage with in-context extraction over the holistic extraction is that, without cell overlapping, the in-context die is able to pass all geometry and connectivity check performed by the physical design tools, which makes it much easier to catch any design mistakes. Since most of the inter-die E-fields are formed within neighbor layers, in-context extraction provides a close-to-optimum solution with easy implementation. Also, it is much easier to avoid intellectual property issues with heterogeneous designs.

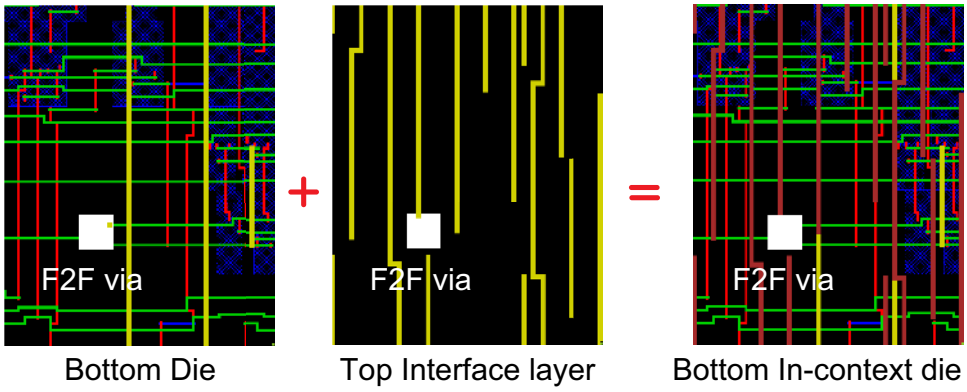

Figure 65: 3D in-context design generation.

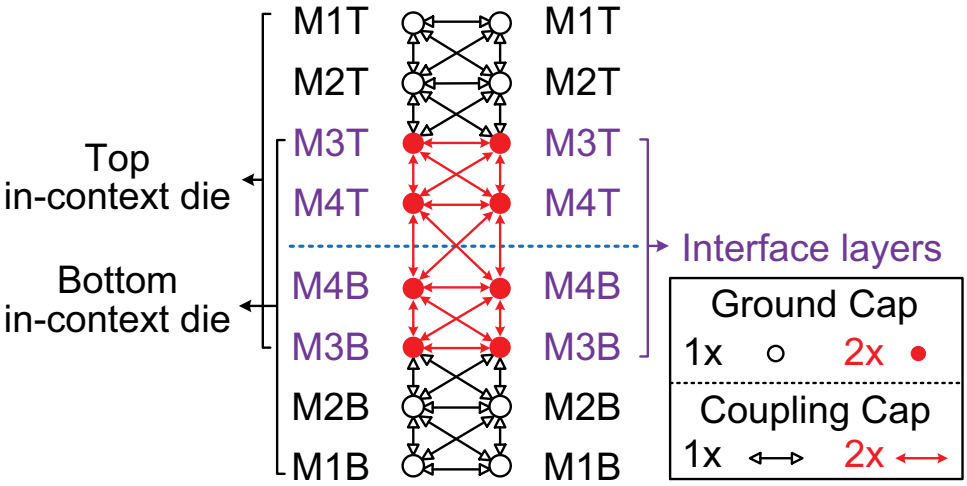

Figure 66: Double-counting capacitance in an in-context technology with four metal layers and two interface layers per die.

5.4.2 Double Counting Correction

Though in-context extraction provides a way to stay compatible with current CAD tools and extract dies separately, the interface layers need to be handled with extra care to avoid any inaccuracy. First is to avoid double counting the capacitance formed between the interface layers. If we directly add parasitics from both dies together, the capacitance will be significantly overestimated since interface layers are extracted both in the top and bottom in-context designs. As an example shown in Figure 66, ground capacitance on M4B and coupling capacitance between M3T and M4B are calculated twice, but the coupling between M4T and M2T is not.

To solve the double counting problem, we extract capacitance with their geometry information annotated into the SPEF file. Then we implement an SPEF analyzer, which reads the extended SPEF file and look up the capacitance layer connection one-by-one. An intuitive way to solve the double counting is to divide every double-counted capacitance by half. It is effectively calculating the average value between top and bottom in-context parasitics. We call this method as "In-C halved" and the method simply merging both in-context parasitics as "In-C original." With an In-C halved extraction, overestimation of inter-die coupling can be corrected. However, this is still not full accurate, since the overestimated capacitance is not exactly twice as large as their correct value. Neither bottom die or top die has the full information of the whole design, and even for the same capacitor, its value is different in two dies, because the extraction environment is not the same in both dies.

5.4.3 Surface Layer Correction

Another issue which also affects the in-context extraction accuracy is the surface layer handling. Shown in Figure 64, surface layers of both in-context dies are the outmost metal layers missing one neighbor layer in the metal stack. As discussed in Section 5.2, E-field sharing in the F2F design significantly affects coupling capacitance. However, with in-context designs, E-field sharing impacts are not fully considered since a few metal layers are missing during the technology calibration. Most of the E-field interactions are between neighboring metal layers, and surface layers are mostly affected by inaccurate extraction. Unlike other metal layers where E-field sharing from both sides are taken care of, the capacitance extracted on the surface layer only considers the E-field sharing from one of its neighboring layer. The In-C halved method is able to correct the double-counting but unable to fix the inaccurate surface layer capacitance.

To solve this issue, we propose an "In-C weighted" method. The motivation is simple, as we observe that a surface layer in one in-context die is not the surface layer in the other in-context die. For example, as in Figure 64, ground capacitance on M4T can not

be extracted accurately with bottom in-context die, because layers from M1T to M3T are missing. However, it is accurate in the top die, where M4T is not the surface layer and has both its neighboring layers. Therefore, when stitching together capacitance of both dies, imbalanced weights should be used depending on how close a layer is to the surface.

To implement this, we use a parameter D for each metal layer as the distance to surface. In any in-context technology, the surface layer has a D value of zero, while D increments by one for each metal layer starting from the surface layer. For example, in Figure 64, D value of M2B is three in the bottom in-context technology, while D value of M3T is two in the top in-context die. Generally, with a larger distance to surface, more E-field sharing can be considered for that layer. We define an R ratio for each interface layer as the ratio between its D values in the bottom in-context die and the top in-context die. It is used as a weight to merge capacitance extracted from both dies. To combine calculation of both ground capacitance and coupling capacitance, we define the R ratio of the ground layer as 1:1.

Then, we can calculate the capacitance from interface layers based on a weighted average from both dies. Note that we do not need to handle capacitance which are not double-counted. As long as the total weight of both dies is equal to one, there is no overestimation in inter-die coupling. Therefore, for a double-counted capacitor connecting two layers, we normalize the product of R ratios of these layers to 1, and use it as the weight between the bottom in-context die and the top in-context die. Figure 67 illustrates an example technology with four metal layers and two interface layers. Our in-context extraction algorithm gives more weights to layers far from the surface so that the inaccuracy from E-field sharing impact is mitigated. As in the example, larger weight is given to ground capacitance in M3T in the top die, but M3B in the bottom die. Also, we use half from bottom die and half from top die for coupling between M4T and M4B.

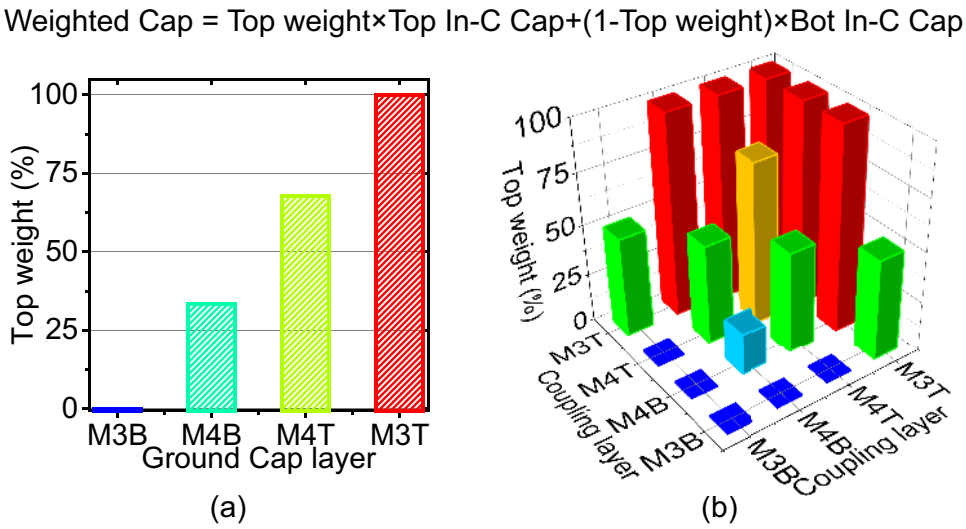

Figure 67: Correction weight for top in-context die in a 2-tier 3D IC with two interface layers per die.

5.5 Full-chip Extraction Results

In this section, we build a 64 point FFT (FFT64) circuit in a 45nm technology shown in Figure 55(a) and apply all three extraction methods on it for comparison. The F2F via has a size of 1µm × 1µm, and the F2F bonding layer is 1µm in thickness and filled by SiO_2 with a relative permittivity of 3.9. We implement the flows described in Section 5.3 and generate design layouts in all three styles: die-by-die, holistic, and in-context. Figure 68 shows FFT64 design shots. This design is routed up to M4 and has a footprint of 380µm × 380µm with 38K gates, which is similar to a digital block in a modern system. The F2F via resistance is assumed as 1Ω connecting between M4B and M4T.

5.5.1 Inter-die vs. Intra-die Breakdown

First, we analyze how much coupling in a F2F design is contributed by inter-die coupling using holistic extraction shown in Table 34. In our extraction, both ground capacitance and coupling capacitance are extracted. We keep as much coupling capacitance as possible, since most extraction tools have decoupling algorithms to simplify the extracted netlist by dividing a coupling capacitor into two ground capacitors so that the generated SPEF file

can be simplified and reduced in size. The table is symmetric thus only the lower triangle is shown. It can be divided into three parts: intra-bottom-die coupling, intra-top-die coupling, and inter-die coupling. Table 35 summarizes the capacitance breakdown for each metal layer. As results shown, intra-die coupling is still the most dominate portion in total capacitance. and most inter-die coupling is mostly between top metal layers of both dies. The M4B-to-M4T coupling contributes to 83% of all inter-die coupling. The inter-die coupling contributes to 34% of the total coupling capacitance on M4B and 39% of the total coupling capacitance on M4T layer. We also observe a noticeable contribution from inter-die coupling on total coupling capacitance of second-topmost layers (8.4% and 9.1% for M3B and M3T, respectively). For lower metal layers, the contribution from inter-die coupling is negligible. Overall, inter-die coupling contributes to 23% in total coupling capacitance in the F2F-bonded FFT64 design. If more metal layers are used for design implementation, the inter-die coupling percentage will decrease relatively, so that its impacts on full-chip timing and power results will be smaller. However, the absolute value of coupling capacitance will still be significant, especially for top metal layers.

The results validate two of our motivations: 1. Inter-die coupling is not negligible especially for the top metal layers, therefore, die-by-die extraction is not sufficient for accurate extraction of F2F designs; 2. Inter-die coupling E-fields are mostly limited between a few metal layers because of E-field shielding from metal wires. In this configuration, we conclude that the coupling depth is around two metal layers. Therefore, it is safe to ignore a few metal layers in our in-context extraction, which still captures most of inter-die coupling E-fields. From the results, we conclude that our holistic extraction is highly accurate to capture all E-field interactions inside F2F designs.

5.5.2 Die-by-die vs. Holistic Extraction

Then, we analyze how much error is introduced with die-by-die extraction. The total extracted ground capacitance is very similar between die-by-die extraction (39476fF) and

Table 34: Holistic extraction of F2F coupling capacitance. Capacitance valus is in fF.

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
M1B	5.76							
M2B	3.03	381						
M3B	17.1	147	1261					
M4B	0.13	396	231	1826				
M4T	0.03	18.6	9.9	**1184**	1311			
M3T	0.14	0.69	140	18.6	196	1226		
M2T	0.00	2.58	0.72	46.7	369	148	442	
M1T	0.00	0.01	0.28	0.12	0.28	25.3	4.63	7.54

Table 35: Breakdown of coupling capacitance shown in Table 34 into intra-die vs. inter-die.

	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Intra	26.0	927	1,656	2,454	1,876	1,595	963	37.8	9,536
Inter	0.18	21.9	151	1,249	1,212	160	50.0	0.42	2,845
Inter %	0.7%	2.3%	**8.4%**	**34%**	**39%**	**9.1%**	4.9%	1.1%	23%

holistic extraction (39247fF) with only a 0.58% difference. This is because the substrate, which serves as the ground plane, is far from the inter-die interface layers. Most differences between these two method come from coupling capacitance. Die-by-die extraction results are shown in Figure 36. Note that all inter-die coupling is zero with this die-by-die extraction which leads to a significant underestimation in total coupling capacitance, especially for top metal layers of each die. Die-by-die extraction also ignores the F2F bonding layer, as there is no F2F via connections. As shown in Table 37, die-by-die extraction underestimates total coupling capacitance by 35% compared with holistic extraction. Though with more metal layers in each die, percentage difference between die-by-die and holistic extraction will be smaller, but accurate extraction is still essential for critical nets on the top metal layer. Therefore, we conclude that die-by-die extraction cannot accurately capture all coupling capacitance and E-field interactions inside the F2F designs, especially for technologies with a close die-to-die distance.

Table 36: Die-by-die extraction of F2F coupling capacitance. Capacitance is in fF.

Layer	M1B	M2B	M3B	M4B	Layer	M4T	M3T	M2T	M1T
M1B	5.33	2.36	12.3	0.09	M4T	905	203	305	0.16
M2B	2.36	337	139	377	M3T	203	1055	127	13.6
M3B	12.3	139	1216	253	M2T	305	127	313	2.46
M4B	0.09	377	253	1325	M1T	0.16	13.6	2.46	4.97

Table 37: Die-by-die extraction error analysis against holistic extraction. Capacitance is in fF.

	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	26.2	949	1,808	3,703	3,089	1,755	1,013	38.2	12,381
D-D	20.1	856	1,620	1,955	1,413	1,399	747	21.2	8,032
Err	-6.06	-93.4	-187	-1,747	-1,676	-356	-266	-17.0	-4,349
Err %	-23%	-9.8%	-10%	-47%	-54%	-20%	-26%	-45%	-35%

5.5.3 In-Context vs. Holistic Extraction

To validate our in-context extraction, we compare extraction results with holistic extraction, which is assumed as our golden model. Note that since holistic extraction cannot handle the top die substrate and device layer, M1T layer parasitics extracted with holistic extraction are less reliable. Targeting a coupling depth of two layers, Table 38 shows extraction results of in-context extraction with two interface layers per die (In-C:2). Since M1 and M2 are not interface layers, inter-die coupling capacitance on those layers is zero with in-context extraction. But the in-context extraction still remains as highly accuracy since the inter-die coupling contributions from M1 and M2 are small, and negligible errors are introduced. If higher accuracy is desired, more interface layers can be added into in-context extraction, and LVS complexity is still much lower than holistic extraction, since adding a few interconnect layer with large dimensions is still much easier than analyzing multiple device layers or local interconnection layers.

Table 39 summarizes the extraction comparison between in-context and holistic extraction. As results shown, for all layers, our in-context extraction is highly accurate in both ground capacitance and coupling capacitance. Since our in-context extraction ignores a few inter-die coupling elements on M1 and M2, total capacitance extracted with our in-context

Table 38: In-context extraction of F2F coupling capacitance. We use top 2 metal layers for the interface. Capacitance is in fF.

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
M1B	5.76							
M2B	3.02	380						
M3B	17.2	148	1265					
M4B	0.13	399	235	1818				
M4T	0.03	18.9	9.88	**1165**	1303			
M3T	0.14	0.54	127	17.8	195	1218		
M2T	0	0	0.48	43.6	365	149	438	
M1T	0	0	0.19	0.09	0.25	25.6	4.63	7.27

Table 39: In-context extraction error analysis against holistic extraction. Capacitance is in fF.

	Ground capacitance								
	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	1,136	6,588	9,240	3,878	2,664	8,320	6,306	1,117	39,247
In-C	1,137	6,583	9,249	4,159	2,639	8,183	5,986	949	38,886
Err	1.10	-4.20	9.00	281	-24.9	-136	-319	-168	-361
Err%	0.1%	-0.1%	0.1%	7.2%	-0.9%	-1.6%	-5.1%	-15%	-0.9%
	Coupling capacitance								
	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	26.2	949	1,808	3,703	3,089	1,755	1,013	38.2	12,381
In-C	26.3	950	1,803	3,679	3,058	1,734	1,001	38.0	12,287
Err	0.15	0.81	-5.15	-24	-31.0	-21.3	-12.3	-0.22	-93.3
Err%	0.6%	0.1%	-0.3%	-0.7%	-1.0%	-1.2%	-1.2%	-0.6%	-0.8%

flow is underestimated slightly. As results show, total ground capacitance is underestimated only by 0.9%, and total coupling capacitance is underestimated only by 0.8%. Note that coupling capacitance errors on M4B and M4T are only 0.7% and 1.0%, respectively. Since these two inter-die coupling elements are largest in absolute value, indicating that almost all inter-die coupling capacitors are captured with our in-context extraction. Therefore, we can conclude that our in-context extraction is highly accurate and efficient to capture most E-field interactions inside the F2F designs without adding too much CAD complexity.

5.5.4 Impact of Interface Layer Handling

Previous results are extracted based with the In-C weighted method which corrects both double counting and surface layer errors. We compare various interface layer handling methods discussed in Section 5.4.3 for accuracy. Note that the runtime required for post-extraction handling is very small compared to extraction runtime. For example, Calibre xACT requires 6 minutes to generate results for in-context designs, while parsing the extended SPEF file and handling interface layers only takes about 10 seconds. Therefore, the runtime difference from interface handling is negligible. Table 40 summarizes full-chip extraction results with three handling methods on M3B and M3T. As results shown, interface layer handling significantly affects extraction accuracy. If the coupling capacitance is simply added up from both dies, the In-C original method overestimates coupling capacitance in the interface layer significantly. The total coupling capacitance errors for M3B and M3T are 77% and 112%, respectively. Total coupling capacitance is also overestimated for M4B and M4T as well. Note that even for the same capacitor, its capacitance value is different when extracted with bottom and top in-context dies, because its context and the E-shield sharing from neighbor layers differ.

By dividing every capacitance value by half, extraction errors are significantly reduced to -12% and -5.8% for M3B and M3T, respectively. However, the extraction accuracy is still not high enough because E-field sharing impacts are not handled well for surface layers as discussed in Section 5.4.3. With our proposed method using a weighted average, our in-context extraction is highly accurate compared to holistic extraction. Total coupling capacitance errors for M3B and M3T are reduced to -0.3% and -1.4%, respectively, which is almost negligible for full-chip analyses. Our interface layer capacitance handling does not affect the number of coupling capacitance, thus the number of aggressors is the same, but it affects the coupling strength of the aggressors. Overall, we can conclude that our in-context extraction algorithm using weighted average to handle interface layers is highly effective and accurate.

Table 40: Comparison of interface-layer handling methods. Unit of total coupling capacitance of each layer is fF.

Layer	Method	M3B	M4B	M4T	M3T	Total	Err	Err%
M3B	Holistic	1261	231	9.9	140	1642	-	-
	original	2220	413	16.4	255	2904	1262	77%
	halved	1110	206	8.2	127	1452	-190	-12%
	weighted	1265	235	9.9	127	1637	-5.27	**-0.3%**
M3T	Holistic	140	18.6	196	1226	1581	-	-
	original	255	32.9	377	2682	3347	1766	112%
	halved	127	16.4	188	1341	1673	92.3	5.8%
	weighted	127	17.8	195	1218	1559	-22.4	**-1.4%**

Table 41: Impact of the interface-layer count on extraction accuracy. "In-C:N" denotes in-context extraction with N interface layers per die. Capacitance is in fF.

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	26.2	949	1808	3703	3089	1755	1013	38.2	12,381
In-C:1	26.1	953	1701	3708	2994	1604	994	37.8	12,018
In-C:2	26.3	950	1803	3679	3058	1734	1001	38.0	12,287
In-C:3	26.2	949	1794	3671	3057	1745	1012	38.2	12,292

Previous in-context extraction results are based on two interface layers per die. However, we also study the in-context extraction accuracy with various numbers of interface layers. Table 41 summarizes these results. Interestingly, even with only one interface layer per die, in-context extraction is quite accurate. Total coupling capacitance only has a 2.9% error compared with holistic extraction, which can actually be regarded as In-C:4 for a technology with four metal layers. With more interface layers, accuracy increases. Total coupling capacitance errors of In-C:2 and In-C:3 are -0.76% and -0.68%, respectively, compared with holistic extraction. Note that since in-context extraction still ignores some inter-die coupling, thus it generally extracts less coupling capacitance than holistic extraction. From these results, we conclude that most of inter-die coupling capacitance can be extracted even with one interface layer from each die. If higher accuracy is required, more interface layers can be included into the in-context extraction to provide detailed consideration of the neighboring die and metal layers.

5.6 Full-chip Power, Performance, and Noise Analysis

In this section, we present our full-chip timing, power, and signal integrity analysis results of our FFT64 benchmark using Primetime. After SPEF files are generated from our extraction flows, we stitch all parasitics together and use TCL scripts for design analysis.

5.6.1 Impact of Inter-die Coupling on 3D Nets

Since inter-die coupling are mostly between top metal layers of both dies, we focus on the 3D nets which connect between bottom and top dies. Except for the clock net, which is assumed to be an ideal network, all other 329 F2F vias are measured in detail. Other 2D nets have fewer routing wires on the top metal layers, and are less affected by inter-die coupling. The results are shown in Figure 69, where each dot represents one 3D net, and its X value is the result with holistic extraction. As results show, using die-by-die extraction, number of aggressors is significantly underestimated for 3D nets, because aggressors from the neighbor die are ignored. However, with our in-context extraction, most aggressors are correctly captured even with one interface layer per die. With multiple interface layers included for extraction, more aggressors are captured. Similarly, wire capacitance of each 3D net is underestimated with die-by-die extraction as well, though the error is smaller, since ground capacitance is the major portion in the wire capacitance. This results in a underestimated delay and power consumption.

5.6.2 Full-Chip Power, Performance, and Noise

To find out how large inter-die coupling impacts have on the full-chip metrics, we compare full-chip analysis results run with all three extraction methods as shown in Table 42. The longest path reported by Primetime is a 3D path which starts from a register in the top die, goes to the bottom die through a F2F via, and ends on another register in the top die. Since the parasitics of inter-die coupling mainly affect wires on the top metal layer, 3D paths are more affected by inter-die coupling. As results show, without inter-die coupling, die-by-die

Table 42: Full-chip comparison of die-by-die (D-D), holistic (Holi), and in-context (In-C) extraction with one interface layer per die.

metric	Holi	D-D	Err%	In-C	Err%
Longest path delay (ns)	3.90	3.66	-6.2%	3.83	-1.8%
3D nets switching power (mW)	1.05	1.01	-3.5%	1.04	-0.4%
Total switching power (mW)	12.1	11.9	-1.7%	12.0	-0.8%
Total coupling cap on 3D nets (fF)	4.37	2.96	-32%	4.21	-3.7%
Total wire cap on 3D nets (fF)	10.8	9.35	-13%	10.7	-1.1%
Average aggressor # on 3D nets	285	200	-30%	253	-11%
Max noise on 3D nets (mV)	41.3	30.40	-26%	38.8	-6.1%

extraction underestimates the longest path delay by 6.2%. Also, total wire capacitance on 3D nets is underestimated by 13%. Therefore, die-by-die extraction is not enough for accurate full-chip analysis. Note that though inter-die coupling capacitance is a large portion of total coupling capacitance, ground capacitance and pin capacitance are major contributors to the capacitive load of a net. Therefore, inter-die coupling only affects slightly on the switching power consumption of F2F designs. From our results, ignoring inter-die coupling and the F2F bonding interface layers, die-by-die extraction underestimates 3.5% of total switching power on 3D nets, while we only observe 1.7% underestimation on the switching power.

However, in terms of signal integrity, inter-die coupling shows much larger impact, especially on top metal layer wires. Total coupling capacitance reported on 3D nets is underestimated significantly by 32%. Similarly, average number of aggressors for 3D nets is also underestimated by 30%. Because of fewer aggressors and a weaker coupling, the maximum noise on 3D nets is underestimated by 26% with die-by-die extraction as well. Therefore, for sign-off verification and post silicon analysis, where highly accurate parasitic extraction is required, the die-by-die extraction introduces significant errors and inter-die coupling needs to be handled carefully.

With our in-context extraction, most of the inter-die coupling and E-field interaction

is captured accurately. As results show, the timing error is only 1.8% even using our in-context extraction with one interface layer per die, and total switching power is underestimated by only 0.8%. For signal integrity analyses, in-context extraction is also able to capture most of coupling aggressors. For 3D nets, only 3.7% and 1.1% underestimation is observed on total coupling capacitance and total wire cap, respectively. And the max noise underestimation is only 6.1% with in-context extraction. Note that only one interface layer per die is included, and more coupling aggressors will be captured using in-context extraction with more interface layers. However, their coupling strengths are relatively weak thus their impacts are much smaller.

5.6.3 Summary of Various Methodologies

In general, die-by-die extraction is the most time- and cost-efficient parasitic extraction that does not require new CAD tools. It is accurate on designs with thick die interface layers and small inter-die coupling capacitance. Holistic extraction, by contrast, is the most complex and time-consuming procedure but provides the highest accuracy across various technologies. It is more suitable for homogeneous integration or designs in which information about both designs is provided beforehand. However, it requires updating a current CAD infrastructure with multiple-die handling, which will take some time before it is widely adopted.

Compared with holistic extraction, in-context extraction entails fewer layers, so the technology calibration time decreases as much as 42.8%. Similarly, the extraction time for each in-context die decreases as much as 30.7%, if parallel extraction is carried out on each die. Moreover, in-context extraction does not require the simultaneous extraction of two device layers, which introduces significant difficulties for LVS checking. By treating each die separately but remaining aware of the neighboring die, in-context extraction resolves the issue to code a complicated LVS deck by mixing two technologies, and requires only a simple extension of current CAD methodologies. Since foundries need to reveal

their interface layers, but they do not need to share important device fabrication details, in-context extraction can also accelerate the commercialization of 3D ICs. For future commercial products, we propose establishing an industry standard that includes connectivity, layout geometries and technology configurations for at least two metal layers. In-context extraction can also accelerate the commercial adoption with heterogeneous F2F 3D ICs.

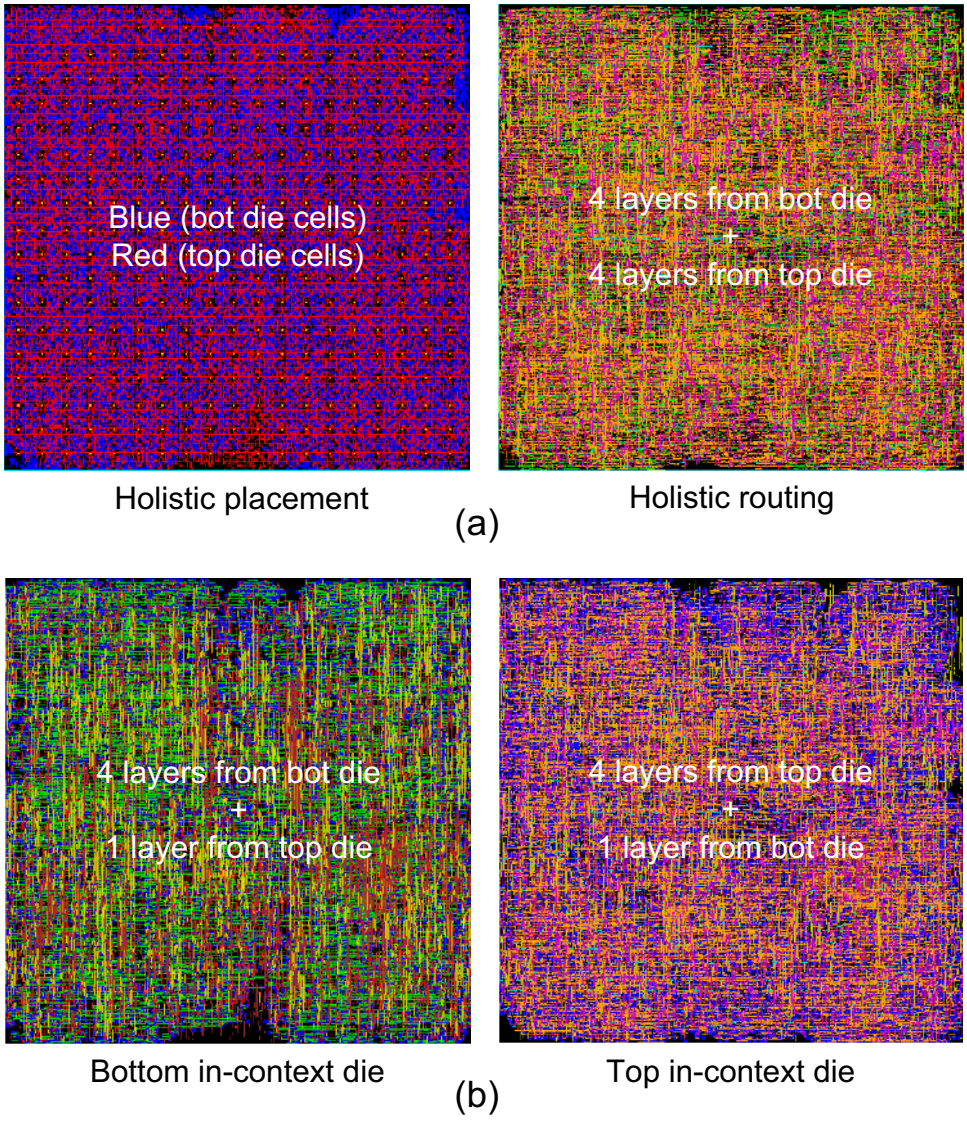

Figure 68: Layouts of FFT64 benchmark using four metal layers. (a) holistic, (b) in-context with 1 metal layer from the other die for the interface.

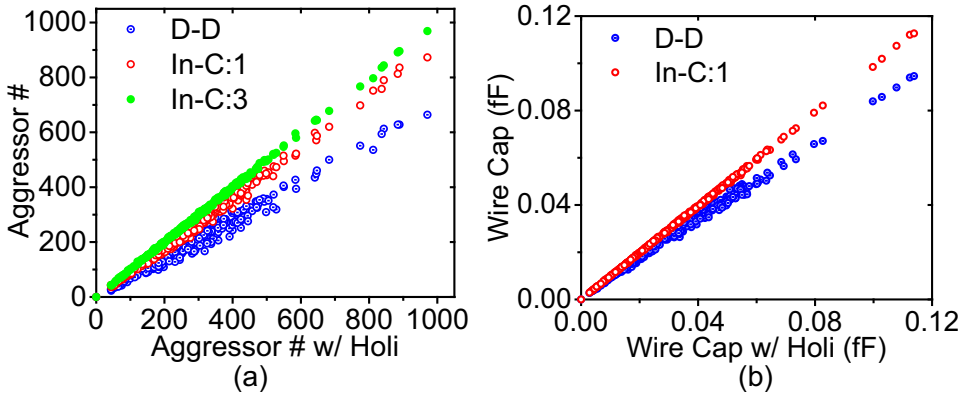

Figure 69: Full-chip comparison of die-by-die (D-D) and in-context (In-C) against holistic extraction (Holi) on 3D nets, each of which is represented by one dot. (a) aggressor count, (b) wire capacitance.

CHAPTER VI

TOWARDS FUTURE TECHNOLOGY

6.1 Extraction for Heterogeneous 3D ICs

Previous design is still based on the homogeneous technology where fabrication processes of both bottom and top dies are the same and designers have a full knowledge of the connectivity and geometry of the system. As discussed in Section 5.1, though in-context extraction provides a fast and accurate approximation and is easier for implementation, holistic extraction is still the most accurate solution and can be implemented without problem. Once the CAD tools are completely migrated to handle multiple dies, holistic extraction provides a straightforward solution. However, when multiple vendors are responsible for design and fabricating different dies, in-context extraction is preferred to protect intellectual property and decoupled the design with multiple companies. In this section, we discuss several issues in heterogeneous integration and the tradeoffs with in-context extraction. We also implemented a heterogeneous design and perform full-chip extraction to validate our in-context flow.

6.1.1 Methodology

For accurate parasitic extraction, the connectivity (or netlists) of both dies are required. However, with heterogeneous integration, it may not be possible because of intellectual property protection. This will result in tradeoffs between extraction accuracy and CAD complexity. An example is shown in Figure 70 with two nets. Net A is in the top die and net B is in the bottom die. Both nets span across two layers with multiple wire segments. For an in-context extraction with one interface layer, various handling methods can be applied for heterogeneous integration. If the extraction engine has a full knowledge of the connectivity, as shown in case (a), the extraction can be performed with correct E-field distribution, and

all extracted capacitance can be netlisted correctly. In this case, capacitance C1 and C2 can be further reduced into one. However, if only the layout geometry is known, as shown in case (b) and (c), there are two ways of handling the interface layer. Note that current analysis engine generally ignores floating nets, so either it can assume all wires on the neighboring die are independent signal nets or they are grounded wires.

However, both methods have to sacrifice the extraction accuracy. In case (b), since wire A1 and A3 belongs to different nets, it introduces an extra coupling capacitance C4 between them. Because of the E-field sharing represented by C4, some of the E-fields are redistributed to coupling between wire A1 and A3. This results in all capacitance C1 to C3 to become smaller in values. On the other hand, wire A1 and A3 become two independent signal nets, which also differ from case (a). As of case (c), all the capacitance can be extracted as ground capacitance but parasitics between two dies are completed decoupled. This results in some errors in noise and delay analyses as well. If net B is a victim, since both wire A1 and A3 are aggressors in case (a), they generates some noise through capacitor C1 to C3 when switching. However, these capacitors become grounded in case (c). Note only the inter-die aggressors are missing, but also the total ground capacitance on net B increases, which makes net B more difficult to switch. Therefore, the coupling noise on net B is underestimated. On the other hand, the impact on the timing comes from Miller Effects. In case (a), the worse-case delay is when net A and net B are switching to the opposite direction. Because of the Miller capacitor C1 to C3, the delay of both nets are larger. However, case (c) can only provide an average estimation for the delay after inter-die capacitance are decoupled.

Since Primetime does not consider Miller effects on timing and power, we rebuild the environment of each 3D nets and perform Hspice simulation one by one for worst-case timing and noise analysis. All aggressors are assumed to have the same waveform switching in the opposite direction to the victim nets, and we measure the delay and noise on each victim net between coupled capacitance as in case (a) and decoupled capacitance as in case

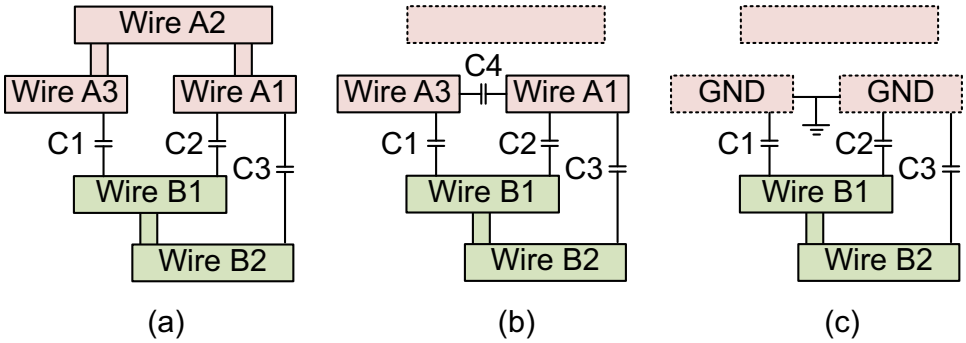

Figure 70: Three cases of for in-context extraction with one interface layer, where (a) with connectivity information of the interface layer, (b) assumes signal nets, and (c) assumes ground nets.

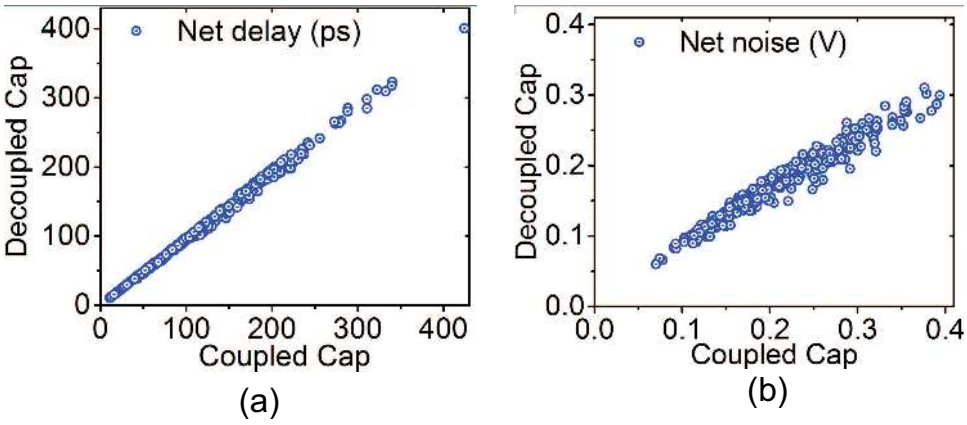

Figure 71: Inter-die decoupling impact on 3D nets. (a) shows worst-case delay and (b) shows worst-case noise.

(c). The results are shown in Figure 71. With decoupled capacitance, the worst-case delay and noise are underestimated by 4.7% and 17.3% in average. Note that for a full timing path, the difference is small since most 2D nets are not affected much. However, if signal integrity is critical, designers need to provide both layout geometries as well as netlist connectivity for the interface layer to allow maximum accuracy with in-context extraction. This can be done by providing an annotated GDS file for the interface layers, where wire geometries are labeled with their connectivity information.

6.1.2 Routing Direction Impact

Another issue with heterogeneous integration lies in the routing directions of metal layers. If wires on the neighboring layers are routed in the same direction, it is more likely that several wires are routed along in a long range. This will significantly increase the coupling between wires on the neighboring layers. Therefore, in common modern designs, wires on the neighboring layers are routed in orthogonal directions to avoid large coupling capacitance, except for M1 which may be routed in the same direction of its neighboring layer for manufacturing alignment issues. Previous design assumes a homogeneous technology such that metal stack configuration of both dies are the same. Therefore, the coupling capacitance is mainly formed between two top metal layers which are routed in the same direction. This helps the in-context extraction to achieve very good accuracy when only one interface layer is included.

However, in a heterogeneous design, the designer and manufacturer of both dies are different and dies are designed separately, routing directions of top metal layers are likely to be orthogonal. This significantly changes the inter-die coupling E-field distribution in the interface layers. Intuitively, inter-die coupling may reduce because smaller coupling capacitance is formed between top layers of both dies. However, non-neighboring interface layers are routed with the same direction which significantly increases the inter-die coupling between them. Take the metal stack shown in Figure 60 (b) as an example, if M4B and M4T are routed in the orthogonal direction, the coupling between them will reduce. However, the coupling between M4B and M3T as well as the coupling between M3B and M4T increases since they become in parallel routing direction. Therefore, if top metal layers are changed from parallel routing direction to orthogonal routing direction, its impact on inter-die coupling depends on the technology configuration such as metal dimensions and dielectric properties, as well as design layouts which determine the wirelength distribution of each layers. And inter-die coupling may increase or decrease depending on E-field distribution.

To illustrate this, we design our FFT circuit with top metal layers routed in orthogonal directions for comparison. To avoid changing the wirelength distribution, we redesign the top die by keeping its cell placement and F2F via locations the same, while rotate the routing directions of all its layers by 90 degree. Then we perform an incremental routing on the top die to fix any design violations. After the designs are generated, we perform holistic and in-context extraction on the new design and compare it to the original one. However, since we are focusing on the heterogeneous designs which does not know its neighboring die before bonding, in-context extraction results are divided into two parts, one for bottom die and one for top.

Table 43 shows the holistic extraction of the redesigned FFT. As results shown, unlike the original design where the maximum inter-die capacitance is between M4B and M4T, in this design with orthogonal top metal layers, the maximum inter-die coupling is between non-neighboring layers. The inter-die coupling between M4 layers significantly decreases to 214fF because of the routing direction change. Therefore, the coupling depth of this design increases to around two metal layers. This also changes the in-context extraction accuracy, as shown in Table 44. As results indicate, because the inter-die coupling increases significantly, in-context extraction on each individual die with only one interface layers is no longer accurate enough. The coupling depth is not fully covered by one interface layer, so adding more interface layers are necessary. By including two interface layers, it is guaranteed that at least one layer with horizontal routing direction and one layer with vertical routing direction will be included. The extraction error is significantly decreased. Further, the benefits of including three interface layers are small since it is out of the coupling depth. Therefore, we conclude that in-context extraction with heterogeneous designs need to include at least enough interface layers covering the coupling depth. Most likely, one interface layer if top layers of both die are routed in with parallel direction, and two layers if routed in orthogonal directions. Note that orthogonal top layer routing is not a problem if designers have a full knowledge to both dies including layouts and connectivity, as in

Table 43: Holistic extraction of FFT with orthogonal top metal layers. Capacitance is in fF.

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
M1B	5.76							
M2B	3.02	380						
M3B	17.0	146	1268					
M4B	0.13	396	234	1824				
M4T	0.24	1.36	343	51.3	1278			
M3T	0.02	12.9	4.69	492	214	1681		
M2T	0.02	0.15	9.63	1.76	243	128	377	
M1T	0.00	0.02	0.02	0.33	0.14	5.97	5.02	7.10

Table 44: In-context extraction errors. Number of interface layers is attached after the die. Capacitance is in fF.

Die	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
bot:1	-0.03	-13.11	-12.30	-487	9.04			
top:1				-76.11	-348	-89.92	-10.09	-0.13
bot:2	-0.01	0.56	-8.93	24.17	1.46	26.26		
top:2			-62.67	-35.7	-62.53	-47.79	-2.69	-0.10
bot:3	0.00	0.30	-2.57	8.95	-3.07	11.50	-1.37	
top:3		-3.24	-29.65	-17.25	-30.8	-18.03	-1.51	-0.07

homogeneous designs, because the weighted interface layer handling methods are able to correct the extraction error by combining both dies.

6.1.3 Full-chip Extraction of Heterogeneous Technologies

With heterogeneous integration, it is possible that top and bottom dies are designed and fabricated in different technology nodes. To illustrate this, we redesigned our FFT circuits with heterogeneous integration shown in Figure 55(c). The top die is designed in a 28nm technology and the die footprint size is measured at 300μm square. As shown in Figure 72, the bottom die is still in a 45nm node and the cell placement is the same as previous designs with a footprint size of 380μm. However, in order to fit the F2F vias into the top die footprint, the F2F vias densities are shrunk by using a one-one mapping method while the F2F via dimensions are the same. The bottom and top dies are still bonded with a 1μm dielectric layer in between. We perform holistic extraction and in-context extraction with

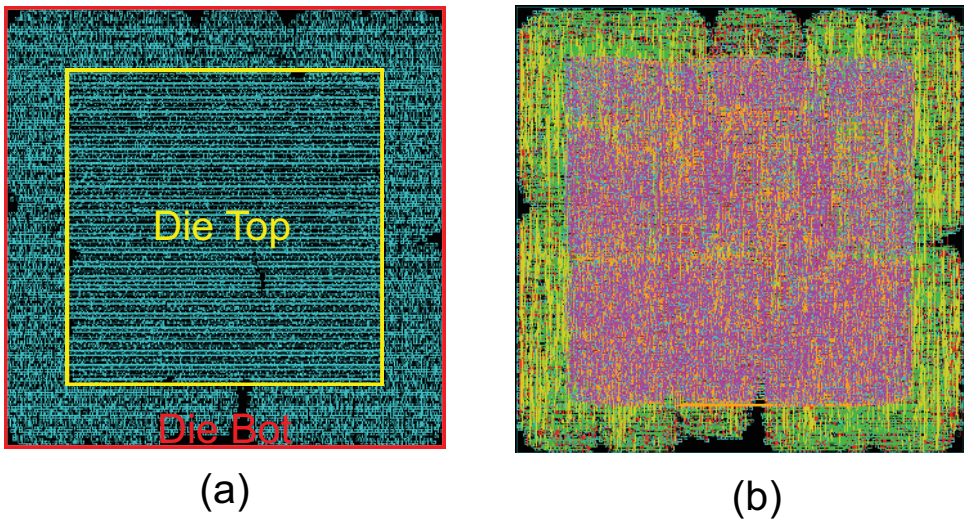

Figure 72: Layout shots of the FFT design, whose top die is in 28nm and bottom die in 45nm. (a) shows the placement and (b) shows the routing.

Table 45: Holistic extraction and in-context extraction of FFT shown in Figure 72. Capacitance is in fF.

	Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
Holi	intra-die	32.63	1056	1865	2602	1768	2161	1651	56.84
	inter-die	0.20	15.46	134	781	677	146	105	1.40
INC	intra-die	32.98	1081	1876	2626	1752	2145	1623	56.29
	inter-die	0.21	11.03	118	764	669	130	93.39	1.13

two interface layers on this design and results are shown in Table 45. For the bottom die, the coupling capacitance is smaller for its top layers since the top die is shrunk which leaves an empty region to its boundary. This also results in a reduction in total inter-die coupling since the die-to-die distance is unchanged. However, if bonding technology improvement is considered which requires a thinner inter-die dielectric layer, the inter-die coupling is still comparable to previous designs. As results shown, our in-context extraction is still accurate for designs in heterogeneous technologies while keeps its CAD simplicity. However, if extraction of each die is conducted independently, including layers at least covering inter-die coupling depth is required for high accuracy.

Table 46: Holistic extraction of FFT with bottom die in 45nm and top die in 28nm. Capacitance is in fF.

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
M1B	6.22							
M2B	4.20	459						
M3B	22.0	185	1368					
M4B	0.18	408	289	1905				
M4T	0.03	10.3	7.02	660	1175			
M3T	0.16	1.45	124	20.2	108	1883		
M2T	0.01	3.69	1.53	100.2	484	138	1017	
M1T	0.00	0.04	0.92	0.43	0.80	32.4	11.60	12.03

6.2 Physical Design Impact

We select two benchmarks to study the impact of full-chip inter-die coupling with logic-logic stacking. We use a low-density parity-check (LDPC) design that is a widely used encryption engine and an OpenSPARC T2 processor core. The LDPC design is a pin-dominated design with 4105 IO pins while T2 is a cell-dominated design with 401k gates. These benchmarks enable us to cover a wide range of applications with realistic layouts. Current designs are much more complicated, so they require careful PDN and clock tree analysis for reliable performance and design yields, especially with advanced technologies in which mask expenses are so high that ensuring a high probability of first-time success is crucial. Since PDNs and clock nets are global nets that are usually routed with upper metal layers, they are more likely to be affected by the inter-die coupling and any other coupling elements in the F2F stack.

6.2.1 F2F Bonding Technology Settings

To conduct the study of technology trends, we use three technology nodes in this work: A commercial FD-SOI 28nm technology, an open source 14nm FinFET technology [75], and a 7nm FinFET technology from an industry IP vendor. We choose these three nodes since they cover a wide range of designs, and they have one node between them, which provides a thorough examination of four-year trends in technology according to Moore's

Table 47: Technology nodes and F2F specs used in our study. Values are in μm.

Node	28nm	14nm	7nm
Fin Pitch	-	0.04	0.036
Poly and M1 Pitch	0.116	0.064	0.054
M6 spacing/width	0.05/0.05	0.036/0.028	0.02/0.02
F2F via size/spacing	0.5/2.0	0.25/1.0	0.13/0.5
Pessimistic D2D distance	1.0	0.7	0.5
Optimistic D2D distance	0.7	0.5	0.35

Law. With every two-node technology step, the interconnect dimension shrinks by roughly 0.5x, and cell density increases by roughly 3.5x. To ensure a realistic and representative study, we also compare the results of the interconnect dimension and cell density with those of commercial foundries and IDMs to ensure that our design matches state-of-the-art designs. Details regarding interconnect technologies are listed in Table 47.

6.2.2 Design Hierarchy Choice

Since no standard design flow exists for 3D ICs, designers may choose various CAD tools and flows for design partitions, floorplan and placement, which leads to significant variation in final design metrics. Also, depending on design implementation, inter-die coupling also varies significantly, especially for large-scale designs with detailed architectural hierarchies. We use T2 core to study the impact of the design floorplan on wirelength and inter-die coupling. The traditional gate-level design flow flattens the whole design and uses min-cut as the partition scheme. However, unaware of the design hierarchy, the partitioner divides standard cells that belong to the same block into several dies. Such partitioning results in more 3D vias as well as longer overall wirelength.

As T2 core consists of several blocks, a careful partition and floorplan should take hierarchical information into consideration. As shown in Figure 73(a), while the gate-level design uses a partitioner to obtain a heuristic min-cut solution based on the flattened netlist, the block-level design uses the manual partition based on the block hierarchy. The wirelength and coupling capacitance are compared in Table 48. As results show, block-level

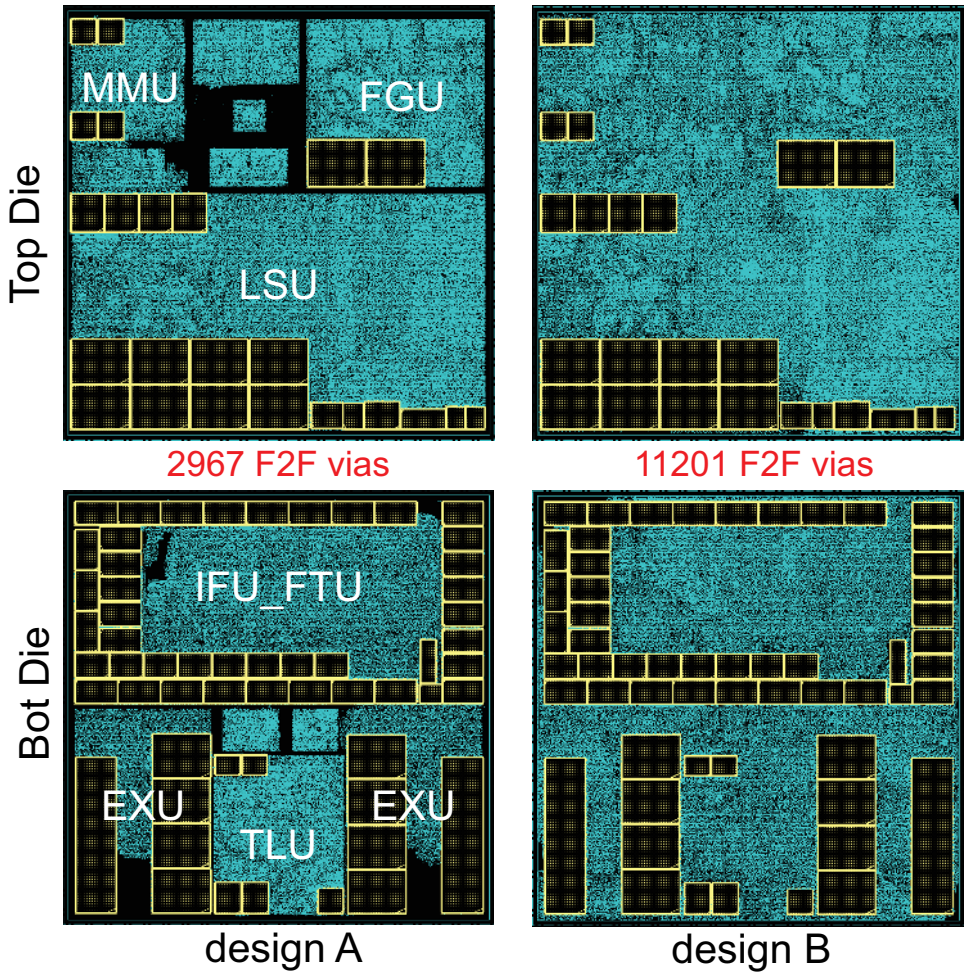

Figure 73: T2 core design flavors. (a) block-level design, (b) gate-level min-cut design.

design significantly reduces the total wirelength by 28.1%, which leads to a significant reduction of 27.5% in all coupling capacitance, especially for inter-die coupling capacitance on the top metal layers. Note that unlike the block level flow used in [74], our flow is still based on the flattened netlist, allowing for design tools that further optimize across block boundaries. Traditional block-level flow only performs optimization within each block and then on top-level separately. With our flattened design with hierarchy awareness, tools can take advantage of every cell information and perform optimization onto the entire design. Therefore, for the best design quality and inter-die coupling reduction, hierarchy-aware design partition and floorplan are needed.

Table 48: Inter-die coupling comparison of the two T2 designs shown in Figure 73. Capacitance and wirelength values are in *pF* and *mm*, respectively.

Block-level	M5B	M6B	M7	M6T	M5T	Other	Total
Wirelength	1429	1260	0	1434	1860	8411	14394
Intra-die	40.36	51.51	0.12	58.16	55.99	283.1	489.3
Inter-die	0.77	2.93	0.14	2.94	0.78	0.65	8.19
Gate-level	M5B	M6B	M7	M6T	M5T	Other	Total
Wirelength	2742	2166	0	1806	2490	10806	20009
Intra-die	90.51	87.3	0.52	65.4	76.86	354.8	675.4
Inter-die	1.18	4.59	0.53	4.52	1.16	0.27	12.31

6.2.3 Routing Blockages by F2F Vias

Another effect comes from the routing blockages caused by F2F vias. To analyze how much inter-die coupling capacitance is contributed by F2F vias, we build a T2 design that only routes up to M6 but uses M7 purely for F2F via landing pads. Removing top layer routing significantly reduces the inter-die coupling from 18.9pF to 8.19pF. The holistic extraction results are shown in Table 49. Most of the coupling capacitance comes from M6 while only a small percentage comes from M7. Therefore, we conclude that the F2F vias do not contribute much to the total inter-die coupling capacitance by itself.

However, with more F2F vias, connecting these vias requires more routing on the top metal layer. As a result, longer wirelength is routed on the top metal layer, which leads to larger inter-die coupling capacitance. With more routing on the top metal layer and larger caps, inter-die coupling increases with more F2F vias, which are also routing blockages. If too many F2F vias are introduced into the top metal layer, their landing pads heavily block the routing tracks. As an example, we build a similar design with a max-cut partition in which we maximize the use of the F2F via. As shown in Figure 74, because of heavy routing blockage on the top metal layer, the wirelength on the top metal layer significantly decreases.

To illustrate the impact of F2F vias, we build three variants of the LPDC designs. All

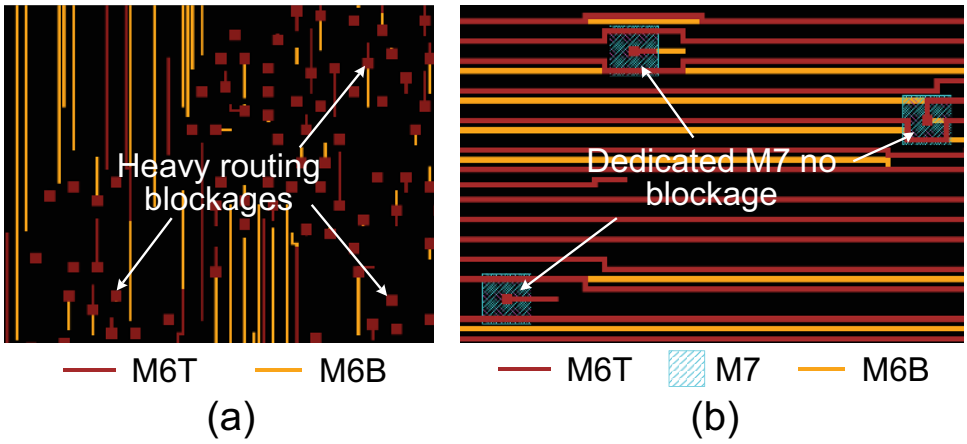

Figure 74: F2F via options. (a) M6 wires are heavily blocked by F2F via pads, (b) M6 routing is not blocked because of the dedicated M7 for F2F via pads.

Table 49: Impact of partitioning (LDPC design). Δ is with respect to min-cut partitioning.

Partition	Wirelength (mm)		F2F Via		M6-to-M6 Cap (fF)	
	Both M6	Δ	F2F#	Δ	Cap	Δ
Min-cut	392	-	3,866	-	792	-
Mid-cut	523	33.5%	6,878	77.9%	1,162	46.6%
Max-cut	451	15.1%	19,798	412%	1,038	31.0%

three designs are made with the same flow but different partition schemes: min-cut, max-cut, and mid-cut. Table 49 lists the holistic extraction results. Both min-cut and max-cut have a shorter top routing wirelength than the mid-cut. Compared with the inter-die coupling with the min-cut partition design, that with max-cut increases by 31.0% because of its 15.1% longer M6 wires. However, for the mid-cut option, into which 6787 F2F vias are inserted, inter-die coupling is the strongest because of its 33.5% longer M6 wires. Therefore, the inter-die coupling cap maximizes with long wires on the top metal layers. Therefore, the impact of the F2F via on inter-die coupling does not directly result from the F2F count, but more because of the related wires on the top metal layer that form the major coupling between dies in an F2F 3D IC. As for design guidelines, to reduce inter-die coupling, fewer top metal wires and dedicated F2F via layers would be helpful.

Table 50: Coupling capacitance breakdown for signal, clock, and power nets in T2 (holistic extraction used).

Net	Signal		Clock		Power	
Layer	intra-die	inter-die	intra-die	inter-die	intra-die	inter-die
M1B	1154	0.17	9.9	0.01	90.4	0.01
M2B	14250	3.29	981	2.61	17042	54.3
M3B	35885	29.9	1606	2.22	2921	0.12
M4B	52742	276.9	1788	28.5	14818	191.5
M5B	49547	1050	3668	110.8	2448	132.8
M6B	45186	6473	3810	727.7	5972	351.6
M6T	61491	6611	4791	748.3	5990	457.5
M5T	76271	1049	5736	108.9	3425	125.9
M4T	71715	157.4	2499	9.29	18021	90.3
M3T	58139	13.8	2679	1.22	4615	0.11
M2T	21314	1.16	1711	4.16	27342	56.6
M1T	1473	0.23	10.2	0	103.9	0
Total	489166	15667	29288	1744	102788	1461
%	96.60%	3.37%	94.40%	5.62%	98.60%	1.40%

6.2.4 Coupling Impact on Power Net

Unlike other signal nets, power and ground nets are mostly routed on the top metal layers to minimize wire resistance. To analyze the inter-die coupling on PDNs, we generate T2 designs with PDNs routed from M4 to M6. We use 10%, 15%, and 20% of the total area for PDN routing from M4 to M6, respectively, and M1 to M3 are used only for signal nets. The results in Table 50 show that PDN coupling capacitance consumes a large portion of total inter-die coupling, since they are mostly routed in the top metal layers. Thus, a thorough understanding of dynamic power integrity necessitates a careful analysis of inter-die coupling. However, since the PDNs are treated as DC signals and instead of generating a coupling capacitance, most extraction tools generate ground capacitance instead, so they do not generate any noise. In addition, PDNs can share an E-field between wires, so they reduce the coupling field between other signals. Therefore, to minimize inter-die coupling, using more PDN wires on the top metal layers to shield the coupling E-field can help reduce any coupling noise from the neighboring die.

Though PDN significantly affects the parasitic extraction, it also provides E-field

Table 51: Impact of PDN shielding on signal net inter-die coupling.

Top layer	PDN	M5B	M6B	M6T	M5T	Other	Total
M6	M4-M6	1.29	7.55	7.82	1.28	0.92	18.87
M7	M4-M6	0.77	2.93	2.94	0.78	0.79	8.2
	M4-M7	0.55	2.46	2.52	0.56	0.39	6.5

shielding for other signal nets. In addition, more PDN wires reduce top metal layer wirelength since the PDNs also occupy additional spaces and reduce number of available routing tracks for signal wires. Therefore, it provides a perfect way of inter-die coupling reduction so that aggressive noises from the neighbouring die can be minimized. On the other hand, with additional PDN wires, the overall cost increases since those wires are routing blockages, and may limit the possible F2F via locations resulting in longer wirelength. In this section, we provide detailed analysis by using PDN as protection wires for inter-die E-field shielding.

To demonstrate this, we insert an additional M7 on top of the 28nm T2 design, while keeping the same F2F via location. The extraction results are shown in Table 51. Because of the additional D2D spacing, the total inter-die signal coupling significantly reduces by 56.5%. Then, we insert additional PDN wires on the empty space of M7. The PDN occupies 20% of the total M7 area and the rest space is used for F2F via connection. As results shown, the total inter-die coupling on signal wires further reduced by 20.9% with additional PDN routing. Note that, the inter-die coupling from PDN themselves increases with additional M7 PDN wires, however, it is generally beneficial to have larger capacitance on PDNs themselves, as these parasitics can act as decoupling capacitors for reduce dynamic voltage droop. For example, compared with the original design with M6, the total inter-die capacitance on PDN wires increases from 1.46pF to 3.6pF. From the results, we conclude that adding an extra PDN layer can significantly reduce inter-die coupling on signal wires.

Figure 75: Clock tree of T2. (a) bottom die, (b) top die.

6.2.5 Coupling Impact on Clock Net

Similar to power nets, the clock network is also routed heavily on the top metal layers. Figure 75 shows a clock tree network of a 28nm T2 design. As most of these clock routes are above M4, they are more sensitive to inter-die coupling. If die-by-die extraction is used, the clock delay, skew, and transition time will be underestimated. Since any timing change on the clock net affects all timing paths connected to it, it is critical to analyze the impact of inter-die coupling on clock networks.

To illustrate the impact on clock nets, we use a 28nm T2 design, which has many memory macros with a significant amount of flip-flops and requires many clock wires. We route clock trees in both block- and gate-level designs (see Figure 73) with a target clock period of 1.5ns. Note that currently no standard tools provide a 3D clock tree synthesis. Therefore, we use only one clock TSV for the clock tree and a 2D clock tree synthesis with Encounter. This results in a large clock skew across dies. The full-chip timing and power analysis results are shown in Table 52. As the results indicate, if die-by-die extraction is used on the clock tree, the max delay and clock transition are significantly underestimated. Note that the impact of inter-die coupling capacitance on signal net timing is relatively

Table 52: Impact of die-by-die (DBD) vs. holistic extraction on various full-chip metrics for T2 designs shown in Figure 73.

	Block-level partition			Gate-level partition		
	DBD	Holi	Δ%	DBD	Holi	Δ%
Clock delay (ns)	1.02	1.16	**13.7%**	1.08	1.21	**12.0%**
Clock transition (ns)	0.83	0.96	**15.7%**	1.06	1.25	**17.9%**
Clock skew (ns)	0.54	0.59	**9.3%**	0.55	0.64	**16.4%**
Switching power (W)	0.17	0.17	0.4%	0.17	0.17	0.2%
Total power (W)	0.33	0.33	0.2%	0.34	0.34	0.1%
Worst-case noise (V)	0.48	0.47	-2.1%	0.51	0.53	3.9%
WNS (ns)	-0.07	-0.05		-0.06	-0.10	

small because of the large pin cap; however, these small delay increases accumulate on a clock tree with more than five levels of clock buffers and clock gates. Also, the signal skew also increases up to 16.4%, because of the clock net delay changes. Therefore, the clock nets observe a much larger impact from inter-die coupling and increase in delay and clock tree synthesis for 3D IC needs a detailed inter-die coupling-aware parasitic extraction.

Another trend, shown in Table 50, is that the clock network has the smallest total coupling capacitance compared with signal nets and power nets. However, their inter-die coupling capacitance portion is the largest. Both power and clock networks are routed in the top level. With same PDNs for both dies, all power wires on the top metal layer is overlapping with wires of the same net, this results in a smallest inter-die coupling. However, unlike power wires, clock routes in both dies are significantly different. Therefore, clock routes are likely to interact with other nets than itself, which leads to a large inter-die coupling portion.

6.3 Logic-Memory Extraction

6.3.1 Context Creation Methodology

Though both holistic and in-context extraction accurately handle F2F designs during sign-off verification stage, they require a LVS-clean design to generate the interface layers with their electrical connections annotated to the layout structures. If the netlist is not clean or

the connection information is not provided, wires can only be treated as floating or ground, which decreases the accuracy by applying approximations. However, when heterogeneous 3D ICs are designed, bottom die and top die of the same chip may come from different vendors, and can be fabricated by different foundries. To save design-to-market time, each die of the 3D ICs may be designed in parallel and it is difficult to exchange detailed interface layouts before sign-off stage.

Therefore, during initial design stage, for procedures such as floorplaning, placement and routing, designers may not have LVS-clean interface layers from the neighboring die for extraction. But if the one die is designed unaware of the other, inaccurate parasitics lead to miscalculated timing, power and noise results. This increases the risks of redesigning the whole chip after two dies are bonded. Traditionally, to solve the issue, designers of individual dies have to leave a lot of design margins and consider for the worst case. It requires to insert lots of large sized buffers for the IO interface, which increases area cost and power. Even if all F2F via nets are buffered, inter-die coupling still affects single die performance, since 2D nets which are routed on the top metal layer are also affected by the neighboring die. Therefore, the E-field sharing from the neighboring die needs to be considered even during early stage designs.

As discussed in Section 5.4, accurate extraction can be achieved by creating an extraction context for a single die. To handle early stage designs, we propose an effective way of creating the extraction context by taking advantage of the regularity of the top layer metal geometry. If the top layers of the neighboring die are following certain layout patterns, only a small amount of information is needed to rebuilt the extraction environment. This is a very common situation, since logic chips usually have their top layers covered by PDNs in a regular fashion, while memory chips usually have regular layouts for both signal and power nets.

6.3.2 Extraction of Logic-Memory Design

We demonstrate our context creation method with a heterogeneous logic-cache partitioned 3D IC design routed up to M4, where the bottom die is a 45nm signal processor unit and the top die is a 28nm L2 cache die. As shown in Figure 76(a), the memory die has a highly regular layouts in layers from M2 to M4, and top 2 layers are mostly used for PDN. Therefore, we only need the memory floorplan, metal pitch and spacing of each layer to rebuild the extraction context. These parameters can be determined even before the memory die design stage. To demonstrate this, we build a floorplan generator which takes these information and automatically rebuild memory flooplan with all blocks by using power and ground wires. Since the M1 of the memory die consists many non-manhattan routing, the floorplan does not contain M1 layer geometries. However, this does not degrade in-context extraction accuracy since the impact from M1 to the bottom die is small. As shown in Figure 77, the auto generated layouts accurately mimic the original design which is in the GDS format.

With the auto-generated context die with M2 to M4, we apply the in-context extraction on the logic die assuming the top die metals are floating. We compare the extraction results of single die, in-context die and holistic extraction with full GDS. The results are shown in Table 53. Without the context, the extraction of the logic die is inaccurate. The ground capacitance is underestimated by 2.46%, since the inter-die coupling between M4 and the memory PDN is ignored. The coupling capacitance is overestimated by 2.51% since the E-field sharing of the top die is ignored. With our context creation method, the extraction errors significantly reduce to less than 0.39% and 0.41% for ground and coupling capacitance, respectively. The context creation method is highly accurate by taking advantage of the regular top layer routing. Though still in early stage, with accurate extraction, designers are able to perform accurate static timing analysis, which helps improve physical design and optimization quality.

Since the inter-die coupling mostly affects wires on top metal layers, only part of the

nets are affected. However, as we observed in Section 6.2.5, the delay calculation error propagates along the path, and even if only one node has incorrect load capacitance, timing calculation becomes incorrect for all following nodes. This is because the delay and power calculation depend on not only the load capacitance of a node itself, but also the input transition time and signal arrival time. If only node has underestimated capacitance load, both the delay and output transition time are reduced. This results in a faster input transition time at next logic level, and delays of all following fan-out nets are further underestimated even if their load capacitance is correct. Therefore, though only a part of nets have routing on the top layer, the delay miscalculation propagates through the whole chip, and amplifies along the timing path.

We perform Primetime timing and power analysis, and the critical path delay compared in Figure 76(b). Without the extraction context, the longest path delay is underestimated by 14.1%, and results clearly show delay error prorogation after a logic depth of 5, even though not all nets have incorrect load capacitance. But with the auto generated neighboring die, timing error is reduced significantly to only 0.13%. In terms of power, the inter-die coupling shows much smaller impacts. As the power is generally dominated by the pin capacitance and the cell internal power, inter-die coupling impacts are relatively small, but still noticeable. As results show in Table 54, with the created context die, the error of net switching power is reduced from 6.76% to 1.35%.

6.4 Technology Scaling Impact

6.4.1 Logic-Logic Design

In this section, we discuss the impact of future technology scaling on inter-die coupling and full-chip metrics. We design LDPC and T2 cores in all three nodes: 28nm, 14nm, and 7nm to provide a comprehensive analysis. All designs are routed up to M6 without dedicated F2F via layers. As we do not have memory compiler for FinFET nodes, memory macros are scaled accordingly. A comparison of T2 core layout is shown in Figure 78.

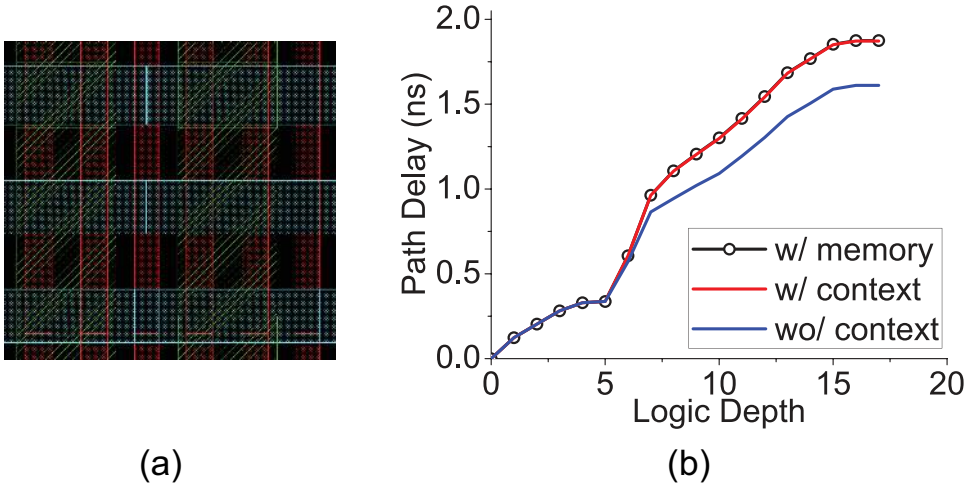

Figure 76: (a) M2-M4 routing of a memory block. (b) longest path delay calculation comparison.

Table 53: Parasitic extraction comparison of the 45nm logic + 28nm memory design. Units are in pF.

Logic die + memory GDS						
Layer	M1B	M2B	M3B	M4B	Total	Err%
GCap	18.2	126.6	221.5	122.8	489.1	-
CCap	1.23	28.6	71.4	92.7	193.9	-
Logic die only						
GCap	18.2	126.6	218.5	113.7	477.1	-2.46%
CCap	1.24	28.9	72.7	95.9	198.8	2.51%
Logic die + context die						
GCap	18.2	126.7	220.9	125.2	491.0	0.39%
CCap	1.23	28.6	71.3	92.0	193.1	-0.41%

If dies are fabricated with the same technology, one impact we observe from the previous discussion is that the average distance between intra-die wires decreases while the average inter-die wire distance remains about the same. This significantly reduces the inter-die coupling cap portion in the advanced technology node. For example, a comparison of LDPC in 14nm vs 7nm is shown in Table 55. With much smaller wire dimensions, the inter-die coupling capacitance decreases in 7nm with a D2D distance of 0.5µm, resulting in a smaller impact when using the extraction of die-by-die vs. holistic extraction. Also, a general trend with the advanced technology node is that more metal layers are needed to complete routing. Therefore, the intra-die portion is likely to increase further because more

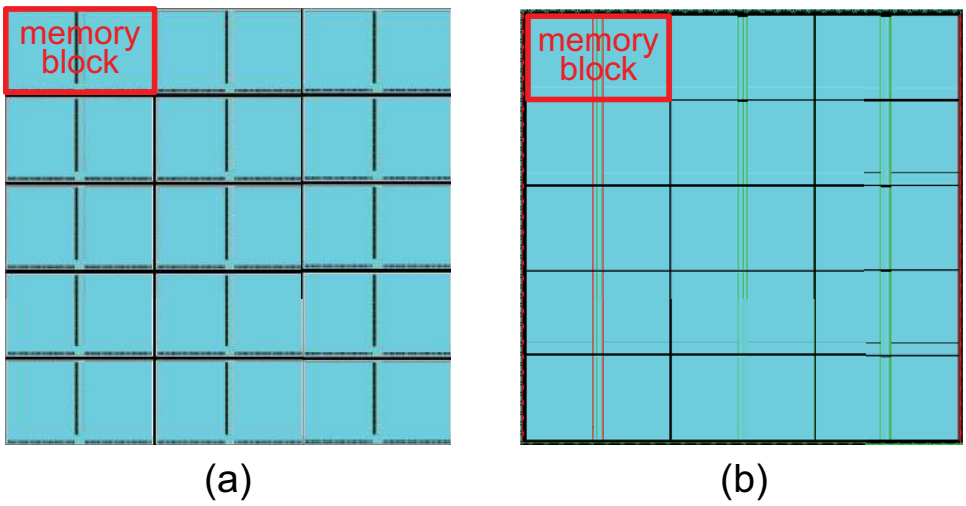

Figure 77: Memory die layout comparison. (a) Memory die in GDS format. (b) Auto-generated context die in Encounter.

Table 54: Full-chip timing and power comparison. Power units are in *mW*.

Design	w/ GDS	wo/ context	Err%	w/ context	Err%
LPD (ns)	1.875	1.611	-14.1%	1.872	-0.16%
Net power	135.6	128.8	-5.01%	137.8	1.62%
Cell power	798.0	797.2	-0.10%	798.4	0.05%
Leakage	6.85	6.85	0%	6.85	0%
Total power	940.5	932.9	-0.81%	943.0	0.27%

coupling capacitors are formed within each die.

Another impact with advanced technology comes from bonding scaling. Without D2D distance scaling and F2F via dimension scaling, it will be difficult to design a complicated 3D chip with most of the top metal layer fully occupied by the F2F pads. Therefore, along with the technology node scaling, because of D2D distance shrinking, the inter-die coupling capacitance increases. For example, when we compare the LDPC in 7nm, we observe that inter-die coupling significantly increases by 45% with a 0.7x closer D2D distance. Also, intra-die coupling capacitance decreases slightly as a result of the impact of E-field sharing. If the D2D distance shrinks further with future technologies such as monolithic 3D ICs, inter-die coupling will play a more important role since the D2D distance shrinks to less than 100nm. A full summary of both T2 and LDPC design is listed in Table 56. As results show, if the D2D distance is kept the same, inter-die coupling portion declines.

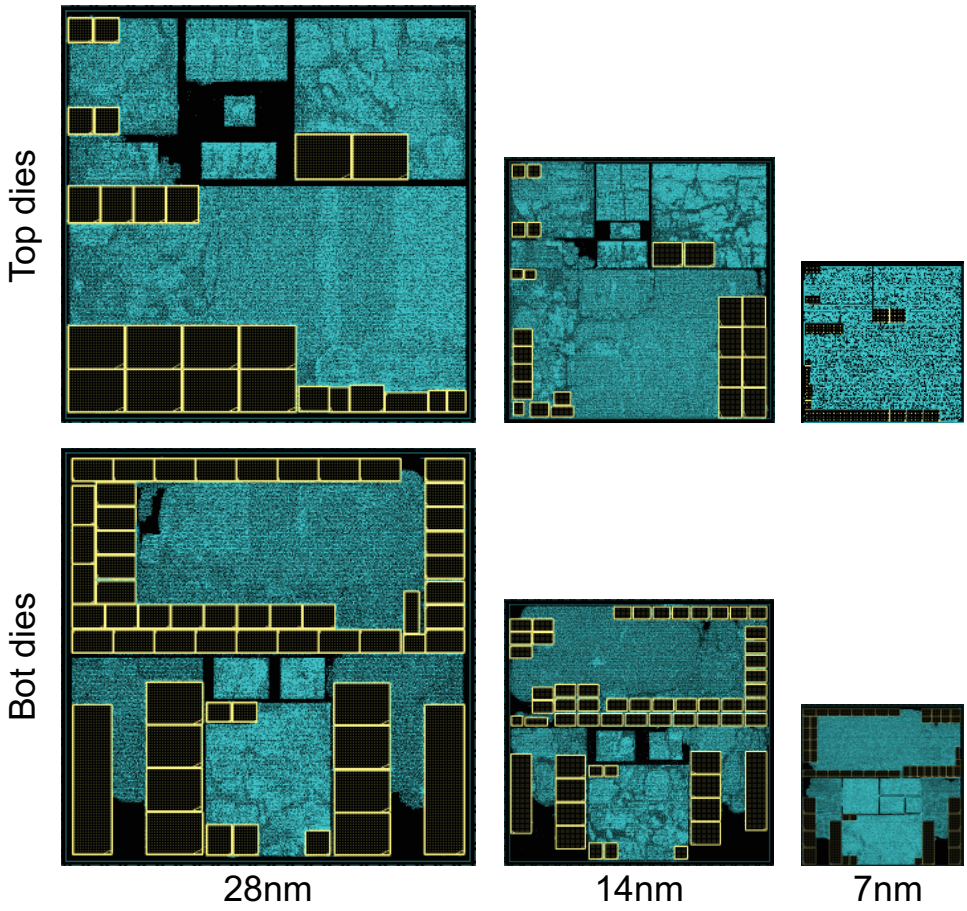

Figure 78: Block-level T2 layouts under various technology nodes. The footprint of 28nm, 14nm, and 7nm designs are $880 \times 880 \mu m$, $560 \times 560 \mu m$, and $340 \times 340 \mu m$.

With both the technology node and advanced bonding technology, a similar portion of inter-die coupling remains. Therefore, we conclude that the impact of inter-die coupling still needs to be carefully extracted and analyzed even with future technologies and a high metal density.

6.4.2 Logic-Memory Design

To verify our context creation method across technology node, we also implement the logic-memory design in advanced nodes. In this new design, the logic die is shrunk to 14nm FinFET node, which results in more than two times performance increases. The layouts of our logic-memory designs are shown in Figure 79. Though the wire dimension shrunk in advanced node, compared with the logic die in 45nm, the inter-die coupling impact

Table 55: Technology trends of inter-die coupling with values in pF. The specifications are shown in Table 47.

Node	Die gap	Layer	M4B	M5B	M5T	M4T	All	%
28nm	1.0μm	intra-die	22.2	20.6	18.42	21.49	208.3	96.3%
		inter-die	0.24	3.75	3.72	0.25	8.11	3.74%
	0.7μm	intra-die	22.2	20.2	18.03	21.42	207.3	95.0%
		inter-die	0.28	5.13	5.10	0.30	10.97	5.03%
14nm	0.7μm	intra-die	11.9	12.6	2.01	8.13	59.5	97.7%
		inter-die	0.07	0.65	0.61	0.07	1.42	2.34%
	0.5μm	intra-die	11.9	12.5	8.97	8.10	59.3	96.8%
		inter-die	0.08	0.91	0.87	0.09	1.99	3.25%
7nm	0.5μm	intra-die	5.09	4.31	3.69	4.18	37.6	97.4%
		inter-die	0.05	0.45	0.45	0.05	1.00	2.58%
	0.35μm	intra-die	5.06	4.20	3.66	4.17	37.4	96.3%
		inter-die	0.06	0.66	0.66	0.06	1.45	3.73%

Table 56: Technology trend summary.

	28nm	14nm	7nm
Die-to-die distance (μm)	1.00	0.50	0.35
LDPC inter-die coupling (pF)	208.3	59.3	37.4
LDPC intra-die coupling (pF)	8.10	1.99	1.45
LDPC intra-die coupling %	3.74%	3.25%	3.73%
T2 inter-die coupling (pF)	621.2	256.7	191.0
T2 intra-die coupling (pF)	18.9	14.9	5.55
T2 intra-die coupling %	2.95%	5.49%	2.82%

increases in 14nm and results in large error for single die extraction without the context. This comes from two reasons. First, D2D distances shrinks from 45nm to 28nm node, as the bonding distance is determined by the older node of the die pair. Second, the logic die dimension shrinks from a square of 1.4mm to 0.5mm. In 45-28nm node, the memory die only covers only 50% of the logic die in the center, while the memory die covers the whole logic die in 14nm node. This is different from previous designs in Section 6.4.1 with both die scaling. Therefore, the inter-die coupling impact area increases. As shown in Table 57, our context creation method is still highly effective to reduce extraction error in advanced nodes.

Table 57: Parasitic extraction comparison of the 45nm logic + 28nm memory design. Units are in pF.

Logic die + memory GDS						
Layer	M1B	M2B	M3B	M4B	Total	Err%
GCap	0.75	60.4	94.9	54.9	210.9	-
CCap	0.00	8.59	35.0	37.2	80.8	-
Logic die only						
GCap	0.75	60.6	93.0	50.1	204.4	-3.08%
CCap	0.00	8.67	35.4	39.4	83.5	3.38%
Logic die + context die						
GCap	0.75	60.5	94.3	53.6	209.1	-0.87%
CCap	0.00	8.61	35.0	37.1	80.7	-0.13%

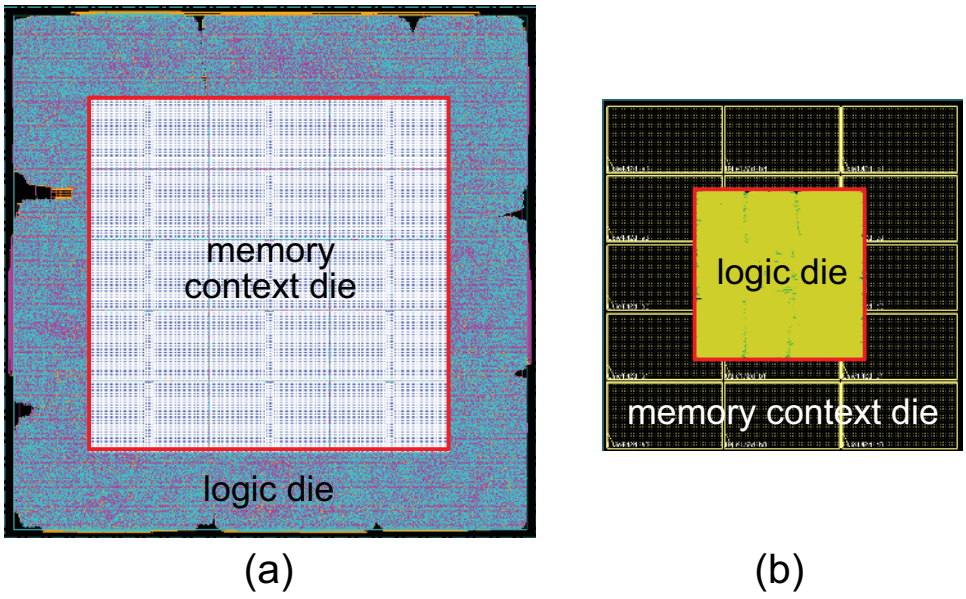

Figure 79: Logic-memory design with 28nm memory die. (a) logic die in 45nm, (b) logic die in 14nm.

CHAPTER VII

SUMMARY AND FUTURE WORK

7.1 Summary and Conclusions

7.1.1 Power Integrity Analysis and Optimization for 3D DRAM

In this work, we investigated impact of various design, packaging, and architectural policy options on 3D DRAM DC power integrity. Based on our CAD/architectural platform and four 3D DRAM benchmarks, results showed that inter-die coupling, the TSV count, location, and alignment strongly affected the IR drop. We used the RDL to replace edge TSVs at the cost of a higher IR drop. Packaging solutions such as backside wire bonding and F2F bonding reduced the IR drop significantly with low cost overhead. With regard to performance, our IR drop-aware policies optimized performance as much as 30.6%. Distributing activity to multiple DRAM dies reduced the IR drop and increased performance under a tight IR drop constraint. Based on the regression analysis, we proposed best co-optimization solutions for the stacked DDR3, Wide I/O, and HMC designs.

7.1.2 TSV-to-TSV Coupling Extraction and Optimization

In this work, we studied the TSV-to-TSV coupling and its impact on 3D IC. We proposed a compact multi-TSV model that can be applied to full-chip TSV-to-TSV coupling analysis that considers E-field and substrate effects. Our multi-TSV model was shown to be highly accurate compared with 3D field solver. Depletion region, substrate impedance, and E-field distribution effects were found to be critical in TSV modeling. We proposed worst case and average case analysis methods and full-chip analysis showed that TSV-to-TSV coupling has large impact on full-chip timing and noise. To alleviate the TSV-to-TSV coupling noise, we proposed a novel guard-ring model and an optimization method to protect the victim TSVs by grounded active region. Our analysis results showed that this optimization method can

reduce the coupling noise up to 27.3% with the maximum area overhead by only 7.65%. Also, with differential TSV insertion, the total TSV noise can reduce up to 49.6% with only 3.9% area overhead. Results showed that our optimization method is very effective, easy to implement and area efficient.

7.1.3 TSV-to-Wire Coupling Extraction and Optimization

In this work, we studied various factors affecting the TSV influence region and TSV-to-wire coupling capacitance. For fast and accurate full-chip TSV-to-wire capacitance extraction, we built three libraries based on multi-TSV structures. We proposed a pattern-matching algorithm which accounted for various E-field sharing impact, *i.e.*, multiple wire impact, corner segment impact, wire coverage impact, and multi-TSV impact. We verified our method using a two-die 3D FFT64 design against field solver simulations in the full-chip level. We also studied multi-TSV impact on TSV-to-wire coupling. Results showed that ignoring E-field sharing and using a single-TSV model on multiple TSVs lead to an over-estimation on coupling capacitance. Applying our pattern-matching algorithm, we studied full-chip TSV-to-wire impact on timing, power, and noise. Increasing metal layer usage reduced impact of both top metal layer signal routing and TSV-to-wire coupling. Analysis results showed that TSV-to-wire coupling was none-negligible and had large impact on full-chip delay and TSV net noise. To alleviate TSV-to-wire coupling, we proposed two physical design solutions, *i.e.*, increasing the KOZ around TSV in top routing layer and adding a ground guard ring. We showed that both methods were very effective in TSV net noise reduction with small overheads on design qualities.

7.1.4 Inter-die Coupling Extraction Methodology Study

In this work, we compared three extraction methods in F2F 3D ICs. We implemented a holistic extraction method for homogeneous integration and found that it is the most accurate at capturing all inter-die coupling. We also proposed an in-context extraction method for heterogeneous integration that is compatible with traditional CAD tools but

includes interface layers from a neighboring die during extraction. We demonstrated the impact of E-field sharing and determined that inter-die coupling cannot be ignored in F2F-bonded 3D ICs. While die-by-die extraction underestimates total coupling capacitance, holistic extraction more accurately estimates coupling capacitance by capturing all inter-die coupling but with higher complexity. Our in-context extraction is highly accurate and captures most E-field interactions across dies. In addition, as it is LVS-friendly, it can easily be implemented to simplify collaboration across multiple companies.

7.1.5 Study of Physical Design and Technology Scaling

In this work, we analyzed inter-die coupling impact on full-chip 3D F2F designs from perspectives of extraction methodology, physical design, and future technology scaling. Though small in value, the impact of inter-die coupling significantly affects full-chip performance and noise. Physical design choices determine the inter-die coupling, and both the PDN and the clock network are significantly affected. Moreover, with advanced technology, the inter-die coupling portion decreases with thinner and denser wires. However, with advanced bonding technologies, inter-die coupling still remains in a similar portion and cannot be ignored.

To alleviate inter-die coupling and improve the quality of the physical design, hierarchy-aware floorplan and partition reduce the total wirelength by 28.1% and inter-die coupling by 27.5%. Reducing the F2F via and the top metal wirelength is critical to reducing inter-die coupling. Depending on the generation of technology, using orthogonal routing on the top metal layers reduces coupling of the neighbor layer at the cost of increasing coupling of the non-neighbor layer. For maximum reduction of inter-die coupling, more PDN areas on the top metal layer and a dedicated layer for F2F via pads can be used.

7.2 Future Work

As the power integrity of DRAM contains two parts: DC power integrity and AC power integrity. Therefore, it is critical to consider the dynamic behaviour of the PDN in 3D

DRAMs. To achieve this, capacitive and inductance components needs to be extracted and simulated accurately within acceptable runtime. The dynamic switching activity and power consumption of DRAM PDNs need to be resolved as well.

In this work, only capacitive parasitics are extracted and analyzed. One missing parasitic components which are often ignored in lower frequency is the parasitic inductance. However, with a high clock frequency, the inductance components are also critical to the signal integrity analysis in 3D ICs as well as 2.5D ICs. In some cases, long signal wires on top of the die will generate strong magnetic field that couples with wires on the package. These inductive coupling may result in a strong noise between die and package and it is critical to estimate their impact on timing, power and noise. Therefore, a fast and accurate inductive extraction engine is needed to resolve this issue. We will continue working on these research topics and further improve the reliability of 2.5D and 3D ICs.

REFERENCES

[1] S. K. Samal *et al.*, "Full chip impact study of power delivery network designs in monolithic 3D ICs," in *IEEE Int. Conf. on Computer-Aided Design*, Nov 2014, pp. 565–572.

[2] U. Kang *et al.*, "8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology," *Journal of Solid-State Circuits*, vol. 45, no. 1, pp. 111–119, Jan 2010.

[3] Q. Wu and T. Zhang, "Design Techniques to Facilitate Processor Power Delivery in 3-D Processor-DRAM Integrated Systems," *Transactions on Very Large Scale Integration Systems*, vol. 19, no. 9, pp. 1655–1666, Sept 2011.

[4] X. Zhao, M. Scheuermann, and S. K. Lim, "Analysis and Modeling of DC Current Crowding for TSV-Based 3-D Connections and Power Integrity," *Transactions on Components, Packaging and Manufacturing Technology*, vol. 4, no. 1, pp. 123–133, Jan 2014.

[5] W. Beyene *et al.*, "Signal and power integrity analysis of a 256-GB/s double-sided IC package with a memory controller and 3D stacked DRAM," in *Electronic Components and Technology Conference*, May 2013, pp. 13–21.

[6] M. Shevgoor *et al.*, "Quantifying the Relationship Between the Power Delivery Network and Architectural Policies in a 3D-stacked Memory Device," in *International Symposium on Microarchitecture*, ser. MICRO-46, 2013, pp. 198–209.

[7] J. Kim *et al.*, "High-Frequency Scalable Electrical Model and Analysis of a Through Silicon Via (TSV)," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, pp. 181–195, 2011.

[8] C. Xu *et al.*, "Compact AC Modeling and Performance Analysis of Through-Silicon Vias in 3-D ICs," *IEEE Trans. on Electron Devices*, vol. 57, no. 12, pp. 3405–3417, Dec 2010.

[9] T. Song *et al.*, "Full-chip multiple TSV-to-TSV coupling extraction and optimization in 3D ICs," in *ACM Design Automation Conf.*, May 2013, pp. 1–7.

[10] J. Cho *et al.*, "Modeling and Analysis of Through-Silicon Via (TSV) Noise Coupling and Suppression Using a Guard Ring," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, pp. 220–233, 2011.

[11] W. Yao *et al.*, "Modeling and Application of Multi-Port TSV Networks in 3-D IC," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 4, pp. 487–496, April 2013.

[12] J.-E. Lorival *et al.*, "An efficient and simple compact modeling approach for 3-D interconnects with IC's stack global electrical context consideration," *Microelectronics J.*, vol. 46, no. 2, pp. 153 – 165, 2015.

[13] C. Xu, R. Suaya, and K. Banerjee, "Compact modeling and analysis of coupling noise induced by through-Si-vias in 3-D ICs," in *IEEE Int. Electron Devices Meeting*, Dec 2010, pp. 8.1.1–8.1.4.

[14] H. Wang, M. H. Asgari, and E. Salman, "Compact model to efficiently characterize TSV-to-transistor noise coupling in 3D ICs," *Integration, the VLSI J.*, vol. 47, no. 3, pp. 296 – 306, 2014.

[15] X. Gu and K. Jenkins, "Mitigation of TSV-substrate noise coupling in 3-D CMOS SOI technology," in *IEEE Conference on Electrical Performance of Electronic Packaging and Systems*, Oct 2013, pp. 73–76.

[16] C. Bermond *et al.*, "RF characterization of the substrate coupling noise between TSV and active devices in 3D integrated circuits," *Microelectronic Engineering*, vol. 130, no. 0, pp. 74 – 81, 2014.

[17] U. R. Tida, C. Zhuo, and Y. Shi, "Novel Through-Silicon-Via Inductor-Based On-Chip DC-DC Converter Designs in 3D ICs," *ACM J. on Emerging Technologies in Computing Systems*, vol. 11, no. 2, pp. 16:1–16:14, Nov. 2014.

[18] Y. Araga *et al.*, "Measurements and Analysis of Substrate Noise Coupling in TSV-Based 3-D Integrated Circuits," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, vol. 4, no. 6, pp. 1026–1037, June 2014.

[19] W. Liu *et al.*, "Design methodologies for 3D mixed signal integrated circuits: A practical 12-bit SAR ADC design case," in *ACM Design Automation Conf.*, June 2014, pp. 1–6.

[20] K. Salah, "Analysis of coupling capacitance between TSVs and metal interconnects in 3D-ICs," in *IEEE Int. Conf. on Electronics, Circuits and Systems*, 2012, pp. 745–748.

[21] D. H. Kim, S. Mukhopadhyay, and S. K. Lim, "Fast and Accurate Analytical Modeling of Through-Silicon-Via Capacitive Coupling," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, vol. 1, no. 2, pp. 168–180, Feb 2011.

[22] W. Yu *et al.*, "RWCap: A Floating Random Walk Solver for 3-D Capacitance Extraction of Very-Large-Scale Integration Interconnects," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 3, pp. 353–366, March 2013.

[23] ——, "Random walk based capacitance extraction for 3D ICs with cylindrical inter-tier-vias," in *IEEE Int. Conf. on Computer-Aided Design*, Nov 2014, pp. 702–709.

[24] Y. Chang *et al.*, "Electrical characterization and reliability investigations of Cu TSVs with wafer-level Cu/Sn-BCB hybrid bonding," in *Symp. on VLSI Technology*, April 2012, pp. 1–2.

[25] T. Lacrevaz *et al.*, "Electrical Broadband Characterization Method of Dielectric Molding in 3-D IC and Results," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, vol. 4, no. 9, pp. 1515–1522, Sept 2014.

[26] M. Murugesan *et al.*, "High density 3D LSI technology using W/Cu hybrid TSVs," in *IEEE Int. Electron Devices Meeting*, Dec 2011, pp. 6.6.1–6.6.4.

[27] M. Motoyoshi *et al.*, "Stacked SOI pixel detector using versatile fine pitch-bump technology," in *IEEE International 3D Systems Integration Conference*, Jan 2012, pp. 1–4.

[28] H.-G. Lee *et al.*, "Wafer-Level Packages Using B-Stage Nonconductive Films for Cu Pillar/Sn-Ag Microbump Interconnection," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, vol. 5, no. 11, pp. 1567–1572, Nov 2015.

[29] L. Peng *et al.*, "Ultrafine Pitch (6 μm) of Recessed and Bonded Cu-Cu Interconnects by Three-Dimensional Wafer Stacking," *IEEE Trans. on Electron Devices*, vol. 33, no. 12, pp. 1747–1749, Dec 2012.

[30] L. Benaissa *et al.*, "A vertical power device conductive assembly at wafer level using direct bonding technology," in *International Symposium on Power Semiconductor Devices and ICs*, June 2012, pp. 77–80.

[31] "Tezzaron Semiconductor," http://www.tezzaron.com/.

[32] G. Chen, H. Zhu, T. Cui, Z. Chen, X. Zeng, and W. Cai, "ParAFEMCap: A Parallel Adaptive Finite-Element Method for 3-D VLSI Interconnect Capacitance Extraction," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 2, pp. 218–231, Feb 2012. [Online]. Available: https://www.mentor.com/products/ic_nanometer_design/verification-signoff/

[33] W. Shi and F. Yu, "A divide-and-conquer algorithm for 3-D capacitance extraction," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 23, no. 8, pp. 1157–1163, Aug 2004.

[34] T. El-Moselhy, I. M. Elfadel, and L. Daniel, "A Markov Chain Based Hierarchical Algorithm for Fabric-Aware Capacitance Extraction," *IEEE Trans. on Advanced Packaging*, vol. 33, no. 4, pp. 818–827, Nov 2010.

[35] W. Yu *et al.*, "Utilizing macromodels in floating random walk based capacitance extraction," in *Design, Automation and Test in Europe*, March 2016, pp. 1225–1230.

[36] Y. Zhou *et al.*, "Macro Model of Advanced Devices for Parasitic Extraction," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[37] G. Kumar *et al.*, "Design and Demonstration of Power Delivery Networks With Effective Resonance Suppression in Double-Sided 3-D Glass Interposer Packages," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, vol. 6, no. 1, pp. 87–99, Jan 2016.

[38] Z. Li, Y. Li, and J. Xie, "Design and package technology development of Face-to-Face die stacking as a low cost alternative for 3D IC integration," in *IEEE Electronic Components and Technology Conf.*, May 2014, pp. 338–341.

[39] "ESD in 3D IC Packages," Feb 2015. [Online]. Available: http://www.3dincites.com/2015/02/esd-3d-ic-packages/

[40] C. Liu *et al.*, "Full-chip TSV-to-TSV coupling analysis and optimization in 3D IC," in *Design Automation Conference, 2011 48th ACM/EDAC/IEEE*, 2011, pp. 783–788.

[41] N. H. Khan, S. M. Alam, and S. Hassoun, "GND Plugs: A Superior Technology to Mitigate TSV-Induced Substrate Noise," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 3, no. 5, pp. 849–857, 2013.

[42] J. Cho *et al.*, "Modeling and Analysis of Through-Silicon Via (TSV) Noise Coupling and Suppression Using a Guard Ring," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 1, no. 2, pp. 220–233, 2011.

[43] M.-F. Chang *et al.*, "A High Layer Scalability TSV-Based 3D-SRAM With Semi-Master-Slave Structure and Self-Timed Differential-TSV for High-Performance Universal-Memory-Capacity-Platforms," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 6, pp. 1521–1529, June 2013.

[44] H. Wang, M. H. Asgari, and E. Salman, "Compact model to efficiently characterize TSV-to-transistor noise coupling in 3D ICs ," *Integration, the VLSI Journal*, vol. 47, no. 3, pp. 296 – 306, 2014, special issue: VLSI for the new era.

[45] J. Kim *et al.*, "Modeling and analysis of differential signal Through Silicon Via (TSV) in 3D IC," in *CPMT Symposium Japan, 2010 IEEE*, Aug 2010, pp. 1–4.

[46] K.-C. Lu *et al.*, "Wideband and scalable equivalent-circuit model for differential through silicon vias with measurement verification," in *Electronic Components and Technology Conference, 2013 IEEE 63rd*, May 2013, pp. 1186–1189.

[47] K.-C. Lu and T.-S. Horng, "Comparative modelling of differential through-silicon vias up to 40 GHz," *Electronics Letters*, vol. 49, no. 23, pp. 1483–1484, Nov 2013.

[48] S. Uemura *et al.*, "Isolation Techniques Against Substrate Noise Coupling Utilizing Through Silicon Via (TSV) Process for RF/Mixed-Signal SoCs," *IEEE J. of Solid-State Circuits*, vol. 47, no. 4, pp. 810–816, April 2012.

[49] J.-S. Kim *et al.*, "A 1.2 V 12.8 GB/s 2 Gb Mobile Wide I/O DRAM With 4 $\times$ 128 I/Os Using TSV Based Stacking," *Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 107–116, Jan 2012.

[50] J. Cho *et al.*, "Through-silicon via (TSV) depletion effect," in *Electrical Performance of Electronic Packaging and Systems, 2011 IEEE 20th Conference on*, 2011, pp. 101–104.

[51] Synopsys, "Raphael." [Online]. Available: http://www.synopsys.com/Tools/silicon/tcad/interconnect-simulation/Pages/raphael.aspx

[52] T. Song *et al.*, "Full-chip multiple TSV-to-TSV coupling extraction and optimization in 3D ICs," in *Design Automation Conference, 2013 50th ACM / EDAC / IEEE*, 2013, pp. 1–7.

[53] Y.-J. Chang *et al.*, "Novel crosstalk modeling for multiple through-silicon-vias (TSV) on 3-D IC: Experimental validation and application to Faraday cage design," in *Electrical Performance of Electronic Packaging and Systems, 2012 IEEE 21st Conference on*, 2012, pp. 232–235.

[54] C. R. Paul, *Analysis of multiconductor transmission lines*. Lexington, KY: John Wiley and Sons, 1994.

[55] ANSYS, "HFSS." [Online]. Available: http://www.ansys.com/Products/Electronics/ANSYS-HFSS

[56] Synopsys, "Sentaurus." [Online]. Available: https://www.synopsys.com/Tools/silicon/tcad/Pages/default.aspx

[57] Y. Yi and Y. Zhou, "Differential Through-Silicon-Vias modeling and design optimization to benefit 3D IC performance," in *Electrical Performance of Electronic Packaging and Systems, 2013 IEEE 22nd Conference on*, Oct 2013, pp. 195–198.

[58] G. Katti *et al.*, "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs," *Electron Devices, IEEE Transactions on*, vol. 57, no. 1, pp. 256–262, Jan 2010.

[59] Synopsys, "HSPICE." [Online]. Available: https://www.synopsys.com/tools/Verification/AMSVerification/CircuitSimulation/HSPICE/Pages/default.aspx

[60] W. Guo *et al.*, "Copper through silicon via induced keep out zone for 10nm node bulk FinFET CMOS technology," in *Electron Devices Meeting, 2013 IEEE International*, Dec 2013, pp. 12.8.1–12.8.4.

[61] J. Kim, *et al.*, "High-Frequency Scalable Electrical Model and Analysis of a Through Silicon Via (TSV)," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 1, no. 2, pp. 181 –195, feb. 2011.

[62] Synopsys, "Primetime." [Online]. Available: http://www.synopsys.com/Tools/Implementation/SignOff/Pages/PrimeTime.aspx

[63] K. Athikulwongse *et al.*, "Stress-driven 3D-IC placement with TSV keep-out zone and regularity study," in *Computer-Aided Design, 2010 IEEE/ACM International Conference on*, 2010, pp. 669–674.

[64] R. Anglada and A. Rubio, "A digital differential-line receiver for CMOS VLSI currents," *Circuits and Systems, IEEE Transactions on*, vol. 38, no. 6, pp. 673–675, Jun 1991.

[65] C. S. Tan and G. Y. Chong, "High throughput Cu-Cu bonding by non-thermo-compression method," in *IEEE Electronic Components and Technology Conf.*, May 2013, pp. 1158–1164.

[66] A.-Y. Park *et al.*, "Thermo-mechanical simulations of a copper-to-copper direct bonded 3D TSV chip package interaction test vehicle," in *IEEE Electronic Components and Technology Conf.*, May 2013, pp. 2228–2234.

[67] S. Panth *et al.*, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 4, pp. 540–553, April 2015.

[68] D. H. Kim *et al.*, "3D-MAPS: 3D Massively parallel processor with stacked memory," in *IEEE Int. Solid-State Circuits Conf.*, Feb 2012, pp. 188–190.

[69] ——, "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)," *IEEE Trans. on Computers*, vol. 64, no. 1, pp. 112–125, Jan 2015.

[70] C. S. Tan *et al.*, "Three-Dimensional Wafer Stacking Using Cu-Cu Bonding for Simultaneous Formation of Electrical, Mechanical, and Hermetic Bonds," *IEEE Trans. on Device and Materials Reliability*, vol. 12, no. 2, pp. 194–200, June 2012.

[71] Mentor Graphics, "Calibre." [Online]. Available: https://www.mentor.com/products/ic_nanometer_design/verification-signoff/

[72] T. Song and S. K. Lim, "Die-to-Die Parasitic Extraction Targeting Face-to-Face Bonded 3D ICs," in *J. of Information and Communication Convergence Engineering*, vol. 13, no. 3, Sep 2015, pp. 172–179.

[73] T. Song *et al.*, "Coupling capacitance in face-to-face (F2F) bonded 3D ICs: Trends and implications," in *IEEE Electronic Components and Technology Conf.*, May 2015, pp. 529–536.

[74] M. Jung *et al.*, "On enhancing power benefits in 3D ICs: Block folding and bonding styles perspective," in *ACM Design Automation Conf.*, June 2014, pp. 1–6.

[75] M. Martins *et al.*, "Open Cell Library in 15Nm FreePDK Technology," in *Int. Symp. on Physical Design*, 2015, pp. 171–178.

PUBLICATIONS

This dissertation is based on and/or related to the works and results presented in the following publications in print:

[1] **Yarui Peng**, Taigon Song, Dusan Petranovic, and Sung Kyu Lim, "Silicon Effect-aware Full-chip Extraction and Mitigation of TSV-to-TSV Coupling," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 33, no. 12, pp.1900-1913, Dec. 2014

[2] **Yarui Peng**, Dusan Petranovic, and Sung Kyu Lim, "Multi-TSV and E-Field Sharing Aware Full-chip Extraction and Mitigation of TSV-to-wire Coupling," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

[3] Taigon Song, Chang Liu, **Yarui Peng**, and Sung Kyu Lim, "Full-Chip Signal Integrity Analysis and Optimization of 3D ICs," IEEE Transactions on Very Large Scale Integration Systems.

[4] Taigon Song, Chang Liu, **Yarui Peng**, and Sung Kyu Lim, "Full-Chip Multiple TSV-to-TSV Coupling Extraction and Optimization in 3D ICs," ACM Design Automation Conference, 2013.

[5] Taigon Song, Chang Liu, **Yarui Peng**, and Sung Kyu Lim, "Full-Chip Multiple TSV-to-TSV Coupling Extraction and Optimization in 3D ICs," SRC TECHCON Conference, 2013.

[6] **Yarui Peng**, Taigon Song, Dusan Petranovic, and Sung Kyu Lim, "On Accurate Full-Chip Extraction and Optimization of TSV-to-TSV Coupling Elements in 3D ICs," IEEE International Conference on Computer-Aided Design, 2013.

[7] **Yarui Peng**, Dusan Petranovic, and Sung Kyu Lim, "Fast and Accurate Full-chip Extraction and Optimization of TSV-to-Wire Coupling," SRC TECHCON Conference, 2014. Best in Session Award.

[8] **Yarui Peng**, Dusan Petranovic, and Sung Kyu Lim, "Fast and Accurate Full-chip Extraction and Optimization of TSV-to-Wire Coupling," ACM Design Automation Conference, 2014.

[9] **Yarui Peng**, Bon Woong Ku, Younsik Park, Kwang-Il Park, Seong-Jin Jang, Joo Sun Choi, and Sung Kyu Lim, "Design, Packaging, and Architectural Policy Co-Optimization for DC Power Integrity in 3D DRAM," ACM Design Automation Conference, 2015.

[10] **Yarui Peng**, Taigon Song, Dusan Petranovic, and Sung Kyu Lim, "Full-chip Inter-die Parasitic Extraction in Face-to-Face-Bonded 3D ICs," IEEE International Conference on Computer-Aided Design, 2015

In addition, the author has completed works unrelated to this dissertation presented in the following publications in print:

[1] Sandeep Samal, **Yarui Peng**, Mohit Pathak, and Sung Kyu Lim, "Ultra-Low Power Circuit Design with Sub/Near-Threshold 3D IC Technologies," IEEE Transactions on Components, Packaging, and Manufacturing Technology, vol.5, no.7, pp.980-990, July 2015

[2] Moongon Jung, Taigon Song, **Yarui Peng**, and Sung Kyu Lim, "Fine-Grained 3D IC Partitioning Study with A Multi-core Processor," IEEE Transactions on Components, Packaging, and Manufacturing Technology, vol.5, no.10, pp.1393-1401, Oct. 2015

[3] Sandeep Samal, **Yarui Peng**, Yang Zhang, and Sung Kyu Lim, "Design and Analysis of Ultra Low Power Processors Using Sub/Near-Threshold 3D Stacked ICs," International Symposium on Low Power Electronics and Design, 2013.

[4] Sandeep Samal, **Yarui Peng**, and Sung Kyu Lim, "Design and Analysis of Ultra Low Power Processors Using Sub/Near-Threshold 3D Stacked ICs," SRC TECHCON Conference, 2014.

[5] Moongon Jung, Taigon Song, Yang Wan, **Yarui Peng**, and Sung Kyu Lim, "On Enhancing Power Benefits in 3D ICs: Block Folding and Bonding Styles Perspective," ACM Design Automation Conference, 2014.

[6] **Yarui Peng**, Moongon Jung, Taigon Song, Yang Wan, and Sung Kyu Lim, "Thermal Impact Study of Block Folding and Face-to-Face Bonding in 3D IC," IEEE International Interconnect Technology Conference, 2015

[7] Taigon Song, Moongon Jung, Yang Wan, **Yarui Peng**, and Sung Kyu Lim, "3D IC Power Benefit Study Under Practical Design Considerations," IEEE International Interconnect Technology Conference, 2015.

[8] Can Rao, **Yarui Peng**, Tongqing Wang, Sung Kyu Lim, and Xinchun Lu, "Investigation of Post-annealing Stress and Pop-out in TSV Front-side CMP," International conference on planarizationCMP technology, 2016, Best student paper award

VITA

Yarui Peng was born in Changsha, China in 1990. He received the B.S. degree from Tsinghua University, Beijing, China in 2012 and M.S. degree from Georgia Institute of Technology, Atlanta, USA in 2014. He is currently working toward the Ph.D. degree under the supervision of Dr. Sung Kyu Lim in the School of Electrical and Computer Engineering, at Georgia Institute of Technology.

His research interests are in physical design, analysis and optimization for 3D ICs, including parasitic extraction and optimization for signal integrity, and alleviating reliability issues in thermal and power delivery. He is the recipient of best-in-session award in SRC TECHCON 14 and best student paper award in ICPT 2016.