Robust Design Methodologies Under Performance Variations for More Than Moore Technologies: CNFETs & 3DICs

Satya Keerthi Vendra
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Part of the Electrical and Computer Engineering Commons

Let us know how access to this document benefits you.

Recommended Citation
https://doi.org/10.15760/etd.7783

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Robust Design Methodologies Under Performance Variations for
More Than Moore Technologies: CNFETs & 3DICs

by
Satya Keerthi Vendra

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical and Computer Engineering

Dissertation Committee:
Malgorzata Chrzanowska-Jeske, Chair
John M. Acken
Dan Hammerstrom
Bart Massey

Portland State University
2022
ABSTRACT

Various emerging technologies have shown great potential of supplementing silicon transistors as Moore’s law slows down. One such disruptive technology, Carbon Nanotube Field Effect Transistors (CNFET), among others, promises increased speed and integration with reduced power consumption. However, due to a limited controllability over the Carbon Nano-tube (CNT) growth process, CNFETs show large variations in their performance and behavior. It is therefore difficult to predict and model their behavior while design suggestions are also challenging. CNT variations are important for a realistic delay modelling. Due to the presence of CNT-specific variations, conventional CMOS evaluation techniques cannot be used for CNFETs.

This work focuses on predicting delay, power and functional yield of CNFET-based circuits under CNT variations, as accurately as possible using a statistical approach. Along with extensively studied CNT variations, we model CNT length variation for the first-time. Tube length variation is especially important if the same tube is used in a number of aligned transistors – called tube correlation. One of the well-known approaches to deal with variations is redundancy.

Redundancy, however, comes at the cost of increased power and area. To limit redundancy, we propose adding redundant tubes only to transistors on critical paths. The challenge with this approach is that with variations, critical paths may vary under CNT variations. Therefore, to consider all potential critical paths under assumed variations and add redundancy to all of them, we developed an efficient algorithm for fast identification of all paths that can become critical in the presence of variations.
This algorithm adds an optimized number of redundant tubes to critical-paths transistors only to minimize power increase at no additional area overhead and is much faster than time intensive Monte Carlo simulations. Our results on a set of ISCAS 85 benchmarks, show that, with our approach, we can achieve delay almost identical with delay without tube variations, >99.99% functional yield and less than 2% increase in circuit power.

Even with improved next-generation logic devices like CNFETs, system level performance will remain severely constrained by the growing interconnect performance bottleneck. To overcome this bottleneck, revolutionary digital system architectures with highly fine-grained integration of disparate technologies is required. Three dimensional (3D) integrated circuits (IC) are proposed as one way to address this problem among many other advancements in wafer level packaging techniques. While 3DIC technology has a lot to offer, lack of effective heat removal techniques continue to be a critical challenge for 3D IC circuit design. This is because up to millions of components produce a great quantity of heat in such a compact space of an integrated circuit, the temperature may sharply increase to significantly deteriorate the performance. In addition, early physical design stages like floor planning, being probabilistic, need thousands of runs to reach a desired stage of optimization. Including thermal analysis in the third dimension only further prolongs the solution evaluation in early design. We thus focus on developing fast methods for thermal goodness evaluation of 3DICs in early stages of physical design.

We propose a power based metric to quickly evaluate the relative thermal goodness of two given floorplans. The proposed algorithm is 29X faster on a grid size of 64x64x4 for GSRC benchmarks compared to a more accurate simulation based tool like Hotspot [1].
Moreover, each device layer in the 3D IC is at a different temperature and varying wire distribution. Ignoring the impact of temperature and interconnect density on 3D interconnect performance may lead to generating severely sub-optimal solution selection.

Experimental results on GSRC benchmarks show 40% underestimation in interconnect delay on average and 3.6X-4.5X variation in buffer count using room temperature parameters. An average difference of 19% in total delay in GSRC benchmarks using the proposed true thermal-aware and wire-density-aware interconnect performance evaluation when compared to only thermal-aware delay, emphasizes the need for a more realistic evaluation the 3D interconnect performance to avoid sub-optimal solution generation. Therefore, we incorporate thermal-delay aware floorplanning while also considering the effective wire density distribution to enable a more realistic evaluation of the interconnect performance to appropriately guide the 3D floorplan optimization.

This work is thus a step towards generating both thermally optimal 3D solutions and variation-tolerant reliable CNFET circuits.
DEDICATION

This thesis is dedicated to all the people who did not give up on me.

TO MY BELOVED PARENTS

for raising me to believe that anything is possible. I’m sure there wasn’t a parenting book in the world that could have prepared you for my eccentricity.

TO MY HUSBAND, Srikanth Kadali

for making everything possible with his endless love and continual, unrelenting support. His hours of work in loving our son enabled my hours of research and contemplation needed for finishing this thesis. My eternal gratitude.

TO MY SON, Vijith Sahas

for constantly reminding me that sun always shines again after the storm. For being my constant source of joy and for tolerating my time away from him.

TO MY BROTHER, Venkat Kalyan Vendra

for teaching me that it’s not always that easy, but that’s life. Without whom, I would not have picked up the courage to do this study.

iv
ACKNOWLEDGMENTS

This thesis becomes a reality with the kind support and help of many. I would like to extend my sincere thanks to all of them.

Foremost, I am truly grateful to my thesis advisor, Professor Malgorzata Chrzanowska-Jeske, for her constant guidance, help, and support throughout my PhD study. Without her, none of this work would have been possible. Her patience, motivation, enthusiasm, and immense knowledge have deeply inspired me. It is a privilege and honor to work and study under her invaluable guidance. I would also like to extend my heartfelt thanks to her for her empathy, understanding and for being there during my highs and lows throughout this 5-year journey, personally and professionally.

I cannot express enough thanks to my committee members, Professor John M. Acken, Professor Dan Hammerstrom and Professor Bart Massey for their continued support, encouragement, insightful comments and valuable suggestions to better this study. My special thanks to Professor. John Acken for his selflessness and willingness to hear me out every time I stopped by. I truly appreciate you and your time you spent helping me in many occasions.

Great deal of appreciation goes to the invaluable help I received from my friends and lab-mates, Rehman Ashraf, Sucheta Mohapatra, Shayla Bhuiya and Renju Rajeev, from fruitful research discussions to the experimental validations and data collection, contributed directly to parts of this thesis. I would also like to extend my thanks to all my friends at Portland State who made my life here enjoyable and made me feel home away from home.
I appreciate the continuing support from the department staff of the Electrical and Computer Engineering department. The financial support from the Maseeh College of Engineering and Computer Science Fellowship and Laurels Scholarship from Portland State University are gratefully acknowledged.

Finally, I owe my deepest gratitude to the support and love of my family- my parents, Mahalakshmi and Murthy and my brother, Venkat for encouraging me, and giving me love and care throughout my career. You all kept me going. My deepest love and indebtedness must go to my husband, Srikanth, for his rock-solid support, endless love and patience to complete my academic journey at Portland State University. Thank you for pushing me to always do my best.
# Table of Contents

Abstract .......................... i  
Dedication .......................... iv  
Acknowledgements .................. v  
List of Tables ......................... xi  
List of Figures ....................... xiii  
Glossary ............................. xvi  
List of Symbols ....................... xvii  

## Chapter 1  Introduction  1  
1.1 Need for planar CMOS alternatives. ........................................... 1  
1.1.1 Emerging logic device: Carbon Nanotube FETs. .......................... 3  
1.1.2 Vertical Integration: 3D ICs with TSVs. .................................... 4  
1.2 Technology Challenges .......................................................... 5  
1.2.1 CNT-specific Fabrication Imperfection .................................... 6  
1.2.2 Thermal management in 3D ICs. ............................................ 11  
1.3 Motivation and Significance .................................................... 14  
1.4 Thesis Contributions ............................................................. 16  
1.4.1 *Circuit-level design methodology for reliable and variation-tolerant CNFET circuits under CNT variations (PartA)* ............................... 16  
1.4.2 *CAD approaches for fast thermal goodness evaluation and thermal management in 3D ICs (Part B)* ............................................. 17  
1.5 Organization of the dissertation. ............................................... 19  

## Chapter 2  Previous Work  21  
2.1 Carbon Nanotube FETs ............................................................. 21  
2.2 Vertical Integration: 3D ICs with Through Silicon Vias ................. 22  

### Part A. Circuit-level design methodology for reliable and variation-tolerant CNFET circuits under CNT variations  25  

## Chapter 3  Statistical Evaluation of Critical Path Delay for Functional Yield Enhancement  26  
3.1 Introduction .............................................................. 26  
3.2 Methodology- Statistical Critical Path Delay Evaluation .............. 34
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.1. CNT Diameter and Count Variations</td>
<td>92</td>
</tr>
<tr>
<td>6.2.2. Carbon Nanotube Correlation</td>
<td>93</td>
</tr>
<tr>
<td>6.3 Methodology</td>
<td>96</td>
</tr>
<tr>
<td>6.3.1 CNT Length Variation</td>
<td>96</td>
</tr>
<tr>
<td>6.3.2 Cases Analyzed</td>
<td>97</td>
</tr>
<tr>
<td>6.4 Experiment Results</td>
<td>100</td>
</tr>
<tr>
<td>6.5 Chapter Conclusions</td>
<td>104</td>
</tr>
<tr>
<td>Part A.2. Significance of work and Major Conclusion</td>
<td>106</td>
</tr>
<tr>
<td>Part B. <em>CAD approaches for fast thermal goodness evaluation and thermal management in 3D ICs</em></td>
<td>107</td>
</tr>
<tr>
<td>B.1 Introduction</td>
<td>108</td>
</tr>
<tr>
<td>B.2 3D floorplanning and problem formulation</td>
<td>110</td>
</tr>
<tr>
<td>B.3 General Assumptions</td>
<td>111</td>
</tr>
<tr>
<td>B.4 Benchmark Specifications</td>
<td>113</td>
</tr>
<tr>
<td>Chapter 7 <em>Buffered Interconnect Power and Performance with Temperature profile</em></td>
<td>115</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>115</td>
</tr>
<tr>
<td>7.2 Assumptions</td>
<td>117</td>
</tr>
<tr>
<td>7.3 Background</td>
<td>119</td>
</tr>
<tr>
<td>7.3.1 3D Interconnect delay</td>
<td>119</td>
</tr>
<tr>
<td>7.3.2 3D Interconnect power</td>
<td>120</td>
</tr>
<tr>
<td>7.4 Interconnect delay dependence on temperature</td>
<td>121</td>
</tr>
<tr>
<td>7.5 Interconnect power dependence on temperature</td>
<td>122</td>
</tr>
<tr>
<td>7.6 Experiment Results</td>
<td>124</td>
</tr>
<tr>
<td>7.7 Chapter Conclusions</td>
<td>128</td>
</tr>
<tr>
<td>Chapter 8 *Fast thermal goodness evaluation of a 3D IC Floorplan</td>
<td>130</td>
</tr>
<tr>
<td>8.1 Introduction</td>
<td>130</td>
</tr>
<tr>
<td>8.2 Chapter contributions</td>
<td>132</td>
</tr>
<tr>
<td>8.3 Background</td>
<td>133</td>
</tr>
<tr>
<td>8.4 Fast thermal goodness evaluation</td>
<td>134</td>
</tr>
<tr>
<td>8.4.1 Floorplan to grid translation</td>
<td>136</td>
</tr>
<tr>
<td>8.4.2 Thermal Goodness Value (TGV)</td>
<td>139</td>
</tr>
<tr>
<td>8.5 Experiment Results</td>
<td>141</td>
</tr>
<tr>
<td>8.6 Chapter Conclusions</td>
<td>148</td>
</tr>
</tbody>
</table>
Chapter 9  Delay Optimization during 3D-IC Floorplanning using Thermal-aware Evaluation 150
9.1 Introduction .................................................. 150
9.2 Previous works ............................................. 152
9.3 Chapter Contributions ..................................... 155
9.4 Preliminaries ................................................ 156
  9.4.1 3D IC Floorplanning .................................... 156
  9.4.2 Correlation Coefficient ................................. 158
  9.4.3 Buffer Insertion Length (BIL) .......................... 158
9.5 Fast Fusion Model for improved temperature correlation .......... 161
  9.5.1 Simplified Vertical Thermal Model ..................... 161
  9.5.2 Thermal Goodness Evaluation of 3D IC .................. 163
  9.5.3 Proposed Fusion of VM and TGV ...................... 166
9.6 Accurate 3D interconnect delay evaluation ........................ 168
  9.6.1 Influence of temperature on wire resistance .......... 169
  9.6.2 Influence of interconnect density on wire capacitance . 170
  9.6.3 Impact of \( R_w(T) \) and \( C_{w_{\text{ID}}} \) on buffer insertion length . 170
  9.6.4 3D Floorplan Evaluation ............................... 172
9.7 Experiment Results ......................................... 175
  9.7.1 TGV vs VM correlation and runtime comparison .... 176
  9.7.2 Proposed fusion model correlation with Hotspot tool ... 178
  9.7.3 3D Floorplanning with \( V_M T_G \) ....................... 181
  9.7.4 Influence of interconnect density and temperature on interconnect delay and buffer count ............................... 185
9.8 Chapter Conclusions ........................................ 189

Chapter 10  Thermal Management in 3D Design using Graphene based TIM to counteract thermal impact on 3D interconnect performance 191
10.1 Introduction ................................................ 191
10.2 Graphene/Graphite/Copper Intermediate Layer Properties .... 193
10.3 3D IC layer configuration test cases ........................ 195
10.4 Experiment Results ......................................... 197
10.5 Chapter Conclusions ....................................... 200

Chapter 11  Conclusions and Future Scope 202
11.1 Thesis Conclusions ....................................... 203
11.2 Future Work ............................................... 214

Bibliography .................................................... 216
List of Tables

Table 1.  Improvement in critical path delay evaluation and functional yield for ISCAS benchmark circuits with statistical delay calculation approach with 10 tubes in the channel region and diameter range of 1nm-2nm. 1.3X Ideal case is the nominal delay allowed for the circuit to be functional ........................................ 40

Table 2.  ISCAS’85 logic benchmark circuits’ worst-case delay, ideal case delay and delay degradation tolerance limits in the range of 5%-30% ........................................ 50

Table 3.  Improvement in Mean Critical Path Delay and Functional Yield of selected ISCAS’85 benchmark circuits with delay degradation tolerance in the range of 5%-30% when proposed minimum tube redundancy is added . . . . . . . . . . . . . . 51

Table 4.  Nomenclature/Values used in ROCP approach ........................................ 55

Table 5.  Critical Path Statistics for ISCAS’85 C432 ........................................ 69

Table 6.  Reduction in Gate Failure Probability in C432 Circuit due to Increasing Transistor Correlation when P_m=10% ........................................ 75

Table 7.  POWER-DELAY PRODUCT OF ROCP vs. RAF ........................................ 83

Table 8.  Critical Path Delay and Delay-Limited Yield in ISCAS ’85 Benchmarks With Declining Tolerance Limit, DT (1.3X-1.05X) ........................................ 84

Table 9.  Average Delay-limited Yield compared by Correlation Style In Investigated ISCAS ’85 circuits (P_m=10%). ........................................ 86

Table 10.  Probability of Failure (P_F) of Logic Gates When CNFETs Are Uncorrelated vs. Correlated in 10,000 Instances of C432 ISCAS’85 Benchmark Circuit with P_m=10% and 40nm Wide CNFETs ........................................ 95

Table 11.  Probability of Failure (P_F) of Correlated CNFETs under various Cases of CNT Length Variation ........................................ 103

Table 12.  Impact of CNT Length Selection on Variation Of Total Delay In 1M Transistors For 1000 MC Simulations with P_m=10% ........................................ 103

Table B1.  Specifications of GSRC benchmarks used in our experiments ........................................ 113

Table 13.  45nm Technology parameters ........................................ 118

Table 14.  Evaluated values at given temperatures ........................................ 121

Table 15.  Leakage power of an 8X buffer with increasing temperature ........................................ 124

Table 16.  Total delay of buffered interconnect for n100, n200 and n300 benchmark circuits ........................................ 127

Table 17.  Interconnect dynamic power (P_{wire\_dyn}, P_{buf\_dyn}), leakage power (P_{buf\_leakage}) and the total interconnect power (P_{total}) for n100 and n300 ........................................ 128

Table 18.  Runtime speedup achieved with TGV ........................................ 142

Table 19.  Technology Specifications Used in our Experiments ........................................ 159

Table 20.  Nomenclature ........................................ 160

Table 21.  Comparison of Evaluation Parameters for n100 Benchmark with γ = 40, 45 and 50 ........................................ 174

Table 22.  Modified GSRC benchmarks for buffer insertion ........................................ 175

Table 23.  Interconnect delay comparison at room temperature vs device layer temperature ........................................ 181
Table 24. Comparison of Total Interconnect delay using CF1 and CF2 In GSRC Benchmarks. $T_{SV_{dia}} = 3\mu M$, $T_{SV_{AR}} = 10$, $T_{SVp} = 6\mu m$, $WS = 5\%$.

Table 25. Total Delay and Buffer Count Comparison in Expanded GSRC Benchmarks using CF1 and CF2. $T_{SV_{dia}} = 3\mu M$, $T_{SV_{AR}} = 10$, $T_{SVp} = 6\mu m$.

Table 26. Material thermal properties in 3D-IC.

Table 27. Simulated test cases using the HotSpot tool.
List of Figures

Figure 1. Evolution of advanced packaging/integration styles .................................................. 2
Figure 2. Innovations in CMOS technology scaling until 22nm showcase the need for beyond CMOS nano-scale devices .......................................................... 2
Figure 3. CNFET side view and top view with an array of 3 semiconducting CNTs and one metallic CNT .............................................................. 3
Figure 4. (a) Homogeneous 3DIC with through silicon vias(TSVs) (b) Heterogeneous 3DIC .............................................................. 4
Figure 5. Contribution of individual CNT specific variations to ON current variation in CNFETs .............................................................. 6
Figure 6. Gaussian distribution for CNT diameter variation with $\mu =1.5\text{nm}$ and $3\sigma=0.5\text{nm}$ .............................................................. 8
Figure 7. Ideal case and actual case in CNFET correlation under CNT length variation and fractured CNTs. Fractured CNTs are CNTs damaged during the transfer process of CNTs from growth substrate to target substrate .............................................................. 9
Figure 8. Heat dissipation problem in 3D ICs due to increased power density compared to 2D .............................................................. 12
Figure 9. (a) Test circuits – planar and 3D (b) Thermal evaluation using 3D-ICE tool (c) Thermal evaluation using HOTSPOT .............................................................. 13
Figure 10. CNFET layout of inverter with parallel tube configuration ........................................ 26
Figure 11. ISCAS 85 C17 Benchmark circuit .............................................................. 29
Figure 12. Diameter distribution of CNTs diameter with respect to $\mu$ and $3\sigma$ diameter of 1.5nm and 0.5nm respectively .............................................................. 30
Figure 13. CNT ON current distribution for diameter distribution given in Figure 12 of 10,000 CNT population .............................................................. 31
Figure 14. $I_{ON}$ variation of 1X Inverter at 32nm technology node of 1XCNFET Inverter showing increase in the ON current variation with increase in initial presence of metallic tubes .............................................................. 32
Figure 15. Critical path delay distribution of ISCAS 85 C432 benchmark circuit showing the delay spread for $P_m=0\%$ .............................................................. 38
Figure 16. Gate delay distributions of 10,000 instances of 1X inverters for 20 runs, using 40, 35 and 32 CNTs/CNFET (left to right) showing the improvement in variation tolerance (narrow distributions) and average mean gate delay reduction with redundancy .............................................................. 45
Figure 17. $Y_f$ achieved with minimum increase in circuit area when minimum tube redundancy is added vs a given redundancy in c432 benchmark circuit .............................................................. 48
Figure 18. Statistical critical path delay reduction achieved in c432 benchmark with and without redundancy when $P_m=5\%$ and 10% .............................................................. 49
Figure 19. Comparison of $Y_f$ in c432 benchmark for various delay degradation acceptance when tube redundancy is introduced .............................................................. 50
Figure 20. (a) Oblique view and (b) Top view of a CNFET .............................................................. 54
Figure 21. Correlated CNFETs sharing the same set of tubes and uncorrelated CNFETs with individual sets of tubes .............................................................. 61
Figure 22. Two instances of C17 circuit showing the change in critical path delay and the Critical path itself under CNT diameter and tube count variations. .......................... 65
Figure 23. ROCP design methodology with adaptive optimization ................................. 67
Figure 24. Statistical critical path count in C432 benchmark obtained using MC simulations ................................................................. 71
Figure 25. (a) XOR2 gate in conventional CMOS-type layout with PU and PD networks separately correlated (b) XOR2 gate in staggered layout style with PU and PD transistors correlated as a group. .................................................. 73
Figure 26. Statistical critical path delay distribution for C432 when Pm=10%. (a) No redundancy (b) RAF (c) ROCP ............................................................... 79
Figure 27. Total circuit static power vs. delay in C432 benchmark for RAF and ROCP. 81
Figure 28. Power savings obtained with ROCP compared to RAF. Power minimization of (a) up to 6.7% when Pm = 5% and (b) up to 8.3% when Pm = 10% Pm = 10%. 82
Figure 29. Statistical critical path delay distribution using (a) Uncorrelated (b) CMOS-type Correlated, and (c) Correlated-staggered style CNFETs in C432. .................. 87
Figure 30. Average critical path delay variation in ISCAS’85 benchmarks with and without redundancy compared by increasing correlation with P_m = 10%. .......................... 88
Figure 31. (a) Ideal (Case(i)): All CNTs are equally long and cover channels of all the CNFETs in the direction of CNT growth (b) Actual (Case ii): CNFETs with varying CNT lengths in the direction of CNT growth. ................................. 91
Figure 32. CNFET 1 and CNFET 2 are correlated in direction of CNT growth(x). CNFET 1 and CNFET 3 are uncorrelated in the direction perpendicular to CNT growth(y) ................................................................. 94
Figure 33. (a) Minimum CNT length from the CNT length distribution is chosen to decide the maximum number of FETs in a row and improve reliability. (b) The CNT length chosen is equal to (µ-3σ) of the CNT length distribution. FETs with degraded performance may be formed in this case. ................................. 98
Figure 34. (a) Ideal case (b) CNT length variation (c) Min. CNT Length (d) (µ-3σ) Length ................................................................. 100
Figure 35. Variation (σ/µ) in gate delay due to individual CNT length variation(Case(ii)), and together with P_m% = 10% and CNT diameter variation(Case(ii)). CNT length variation contributes nearly 30% to ON current, I_ON variation in CNFETs. ................................................................. 101
Figure B2. General flow diagram of EA-based 3D floorplanning as used in this work and proposed evaluation tuning as highlighted. ........................................... 114
Figure 36. The temperature values used on each die with heat sink close to the bottom die and a rising vertical temperature profile in a 3DIC ........................................... 116
Figure 37. TSV enable 3D interconnect spanning multiple dies. The wirelength will be reduced to 2D on each die ignoring the TSVs ........................................... 119
Figure 38. Wire dynamic power compared with total dynamic power consumption for n200 GSRC benchmark. Minimal increase in wire dynamic power and major contribution by buffer dynamic power to be observed ................................... 126
Figure 39. Comparison of buffer leakage and total (dynamic + leakage) interconnect power consumption in n200 GSRC benchmark at room temperature and with vertical temperature profiles 1 and 2 ........................................... 126
Figure 40. 4-layer TSV-based 3D IC showing layer thicknesses and vertical temperature distribution. ............................................ 131
Figure 41. Proposed fast thermal goodness evaluation ............................................. 135
Figure 42. Grid cell and module coordinate notations ............................................. 136
Figure 43. (a) Layout to grid translation (b) Total grid cell power calculation (c) Module m to grid cell g-overlap area calculation............................................. 137
Figure 44. 3D thermal goodness evaluation model ................................................. 139
Figure 45. Positive correlation between TGV and Hotspot for n100 benchmark ....... 143
Figure 46. Positive correlation indicated by the dotted trend line between TGV and Hotspot for GSRC benchmarks using 60 floorplan samples ...................... 144
Figure 47. Sample 2D floorplan powers and associated thermal map comparing TGV and peak temperature .................................................. 144
Figure 48. Thermal map of 4-layer n100 benchmark comparing peak temperature and TGV ................................................................. 145
Figure 49. TGV and peak temperature with increasing area .................................... 146
Figure 50. Success rate analysis on 10 samples of GSRC benchmarks .................... 147
Figure 51. Trapped heat between stacked layers in 3D ICs due to increased power density compared to 2D heat dissipation path ............................... 151
Figure 52. (a) 3D tile stack array (b) Single tile stack analysis .............................. 162
Figure 53. 3D thermal goodness evaluation model .............................................. 164
Figure 54. Flow of proposed fusion methodology $V_{MTG}$: 1-dimensional heat flow analysis on TGV model ......................................................... 166
Figure 55. Location of inserted buffers in a 1000um line (45nm technology node) (a) standard BIL at room temperature (b) temperature-aware BIL varies with only $R_w(T)$ variable (c) temperature-aware BIL varies with both $R_w(T)$ (device layer average temperature) and $C_w$ (non-uniform wire density) ........ 172
Figure 56. (a) Correlation of temperature (peak and average) using VM [1] model and TGV (power factor) with Hotspot. (b) Runtime comparison of VM vs TGV ................................................. 177
Figure 57. Improved correlation of VM using VMTG with Hotspot tool .................. 179
Figure 58. Temperature correlation comparison between Hotspot and (a) standalone VM model and (b) proposed fusion methodology VMTG ...................... 180
Figure 59. Impact of using TGV as an optimization parameter in fitness evaluation. Thermal hotspots are either reduced or better spread out ......................... 184
Figure 60. Wire delay distribution in n100_exp GSRC benchmark on individual device layers ................................................................. 188
Figure 61. Total interconnect delay distribution in n100_exp benchmark at room temperature (D), at device layer temperature(DT) and at device layer temperature along with variation in wire capacitance .......................... 189
Figure 62. 3D-IC configuration with (a) TIM and (b) TIM and IL .......................... 192
Figure 63. Weak inter-plane van der Waals interactions in the cross plane direction of Graphene (along z-plane) ................................................. 194
Figure 64. Peak temperatures in n200 benchmark for different configurations. Graphite based TIM achieving the lowest peak temperature can be observed in TC5 .. 198
Figure 65. Peak temperature reduction with increasing IL thickness of three different materials in n200 benchmark ......................................................... 199
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DIC</td>
<td>Three dimensional integrated circuits</td>
</tr>
<tr>
<td>BEOL</td>
<td>Back-end of line</td>
</tr>
<tr>
<td>BIL</td>
<td>Buffer Insertion Length</td>
</tr>
<tr>
<td>CNFET</td>
<td>Carbon Nanotube Field Effect Transistor</td>
</tr>
<tr>
<td>CNT</td>
<td>Carbon Nano-tube</td>
</tr>
<tr>
<td>CPD</td>
<td>Contributed Power Density</td>
</tr>
<tr>
<td>CVD</td>
<td>Chemical Vapor Deposition</td>
</tr>
<tr>
<td>DIBL</td>
<td>Drain-induced barrier lowering</td>
</tr>
<tr>
<td>FEOL</td>
<td>Front-end of line</td>
</tr>
<tr>
<td>GUT</td>
<td>Grid cell under test</td>
</tr>
<tr>
<td>ID</td>
<td>Interconnect density</td>
</tr>
<tr>
<td>IL</td>
<td>Intermediate layer</td>
</tr>
<tr>
<td>ITRS</td>
<td>International technology Roadmap for Semiconductors</td>
</tr>
<tr>
<td>m-CNT</td>
<td>Metallic CNT</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>NR</td>
<td>No Redundancy</td>
</tr>
<tr>
<td>NWFET</td>
<td>Nano-wire FET</td>
</tr>
<tr>
<td>OAR</td>
<td>Overlap area ratio</td>
</tr>
<tr>
<td>PD</td>
<td>Power density</td>
</tr>
<tr>
<td>PDP</td>
<td>power-delay product</td>
</tr>
<tr>
<td>RAF</td>
<td>Redundancy in all FETs</td>
</tr>
<tr>
<td>ROCP</td>
<td>Redundancy only in all critical paths</td>
</tr>
<tr>
<td>SCE</td>
<td>Selective Chemical Etching</td>
</tr>
<tr>
<td>SWCNT</td>
<td>Single-walled carbon nanotube</td>
</tr>
<tr>
<td>s-CNT</td>
<td>Semiconducting CNT</td>
</tr>
<tr>
<td>TGV</td>
<td>Thermal goodness value</td>
</tr>
<tr>
<td>TIM</td>
<td>Thermal interface materials</td>
</tr>
<tr>
<td>TSV</td>
<td>Through Silicon Via</td>
</tr>
<tr>
<td>TTSV</td>
<td>Thermal through silicon via</td>
</tr>
<tr>
<td>VLSI</td>
<td>Very Large Scale Integration</td>
</tr>
<tr>
<td>VM</td>
<td>Vertical thermal model</td>
</tr>
<tr>
<td>VMR</td>
<td>VLSI-compatible metallic carbon nanotube removal</td>
</tr>
</tbody>
</table>
### List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu)</td>
<td>Mean value</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>(P_p)</td>
<td>Presence of metallic tubes initially present and later removed</td>
</tr>
<tr>
<td>(V_{th})</td>
<td>Threshold voltage</td>
</tr>
<tr>
<td>(I_{on})</td>
<td>ON Current</td>
</tr>
<tr>
<td>(P_{pc})</td>
<td>Paths possibly critical</td>
</tr>
<tr>
<td>(I_{ON,CNT})</td>
<td>Total ON current of all s-CNTs</td>
</tr>
<tr>
<td>(g_{CNT})</td>
<td>Trans-conductance of a CNFET transistor.</td>
</tr>
<tr>
<td>(D_{CNT})</td>
<td>CNT diameter</td>
</tr>
<tr>
<td>(T_{pd})</td>
<td>Propagation delay</td>
</tr>
<tr>
<td>(C_L)</td>
<td>Load capacitance</td>
</tr>
<tr>
<td>(V_{dd})</td>
<td>Supply voltage</td>
</tr>
<tr>
<td>(Y_f)</td>
<td>Functional Yield</td>
</tr>
<tr>
<td>(W_g)</td>
<td>Width of the transistor</td>
</tr>
<tr>
<td>(N_{cnt})</td>
<td>Number of CNTs</td>
</tr>
<tr>
<td>(S_{CNT-CNT})</td>
<td>CNT-CNT spacing</td>
</tr>
<tr>
<td>(L_s)</td>
<td>Critical channel length</td>
</tr>
<tr>
<td>(N_{min})</td>
<td>Minimum redundant CNTs</td>
</tr>
<tr>
<td>(Y_d)</td>
<td>Delay-limited yield</td>
</tr>
<tr>
<td>(D_s)</td>
<td>Allowed delay degradation tolerance</td>
</tr>
<tr>
<td>(I_{on,FET})</td>
<td>ON current through a single CNFET</td>
</tr>
<tr>
<td>(T_{pd,FET})</td>
<td>CNFET delay</td>
</tr>
<tr>
<td>(T_{pd,ideal})</td>
<td>Ideal critical path delay</td>
</tr>
<tr>
<td>(P_f)</td>
<td>Failure probability of CNFET</td>
</tr>
<tr>
<td>(P_{row})</td>
<td>Row failure probability</td>
</tr>
<tr>
<td>(K_r)</td>
<td>No. of rows</td>
</tr>
<tr>
<td>(F_r)</td>
<td>Maximum no. of FETs in a row</td>
</tr>
<tr>
<td>(F)</td>
<td>Minimum feature size</td>
</tr>
<tr>
<td>(R)</td>
<td>Buffer output resistance (8X)</td>
</tr>
<tr>
<td>(C)</td>
<td>Buffer input capacitance (8X)</td>
</tr>
<tr>
<td>(R_c)</td>
<td>Unit wire resistance M4-M6</td>
</tr>
<tr>
<td>(C_c)</td>
<td>Unit wire capacitance M4-M6</td>
</tr>
<tr>
<td>(p_{par})</td>
<td>Parasitic capacitance factor</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>Switching activity</td>
</tr>
<tr>
<td>(f)</td>
<td>Frequency (GHz)</td>
</tr>
<tr>
<td>(E_\infty)</td>
<td>Permittivity of free space</td>
</tr>
<tr>
<td>(\epsilon_r)</td>
<td>Relative permittivity</td>
</tr>
<tr>
<td>(n)</td>
<td>Sub-threshold swing factor</td>
</tr>
<tr>
<td>(t_{ox})</td>
<td>Gate-oxide thickness</td>
</tr>
</tbody>
</table>
$t_{pd\_wire\_segment}$ Delay of a buffered 3D wire segment
$t_{pd\_net}$ Total 3D net delay
$t_{pd\_total}$ Total delay on a device layer
$\beta$ Temperature coefficient of metal
$T_0$ Room temperature
$P_{wire}$ Wire dynamic power per unit length
$P_{buffer}$ Buffer dynamic power
$\mu_0$ Zero-bias mobility
$V_t$ Thermal voltage
$C_{ox}$ Gate oxide capacitance
$I_{sub}$ Sub-threshold leakage
$r$ Correlation coefficient
$S_x$ Standard deviation
$(L_{x,1}, L_{y,1})$ Lower left corner of grid cell
$(R_{x,1}, R_{y,1})$ Upper right corner of grid cell
$(L_{x,2}, L_{y,2})$ Lower left corner of module
$(R_{x,2}, R_{y,2})$ Upper right corner of module
$g$ Grid cell
$m$ Module
$O_w$ Overlap width
$O_h$ Overlap height
$A_{o,w}$ Overlapping area of module and grid cell
$P_g$ Grid cell power
$PD_g$ Grid cell power density
$PD'_g$ Modified Power density of grid cell $g$
$t_x$ Number of intra layer neighbors for GUT in x-direction
$t_y$ Number of intra layer neighbors for GUT in y-direction
$dist_z$ Vertical distance between layers
$SG_g$ Sub-grid sum
$R_{w}(T)$ Unit length wire resistance at device layer temperature
$C_{w}\_ID$ Unit length wire capacitance with varying interconnect density (ID)
$S_B$ Buffer size
$t_s$ Silicon thickness
$t_{TIM}$ TIM thickness
$t_{dielectric}$ Dielectric thickness
$k$ Thermal conductivity
$M_w$ Metal width
$M_t$ Metal thickness
$M_s$ Metal spacing
$TSV_d$ TSV diameter
$TSV_{aspect}$ TSV aspect ratio
$TSV_p$ TSV pitch
$R_c$ TSV contact resistance
$V_{MTG}$ Fusion of VM and TGV
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{peak}}$</td>
<td>Peak temperature of 3D IC</td>
</tr>
<tr>
<td>$T_{\text{avg, L}}$</td>
<td>Average device layer temperature</td>
</tr>
<tr>
<td>$R_t$</td>
<td>Thermal resistance</td>
</tr>
<tr>
<td>$D/D_{\text{room}}$</td>
<td>Total wire delay at room temperature</td>
</tr>
<tr>
<td>$DT/D_{\text{avg}}$</td>
<td>Total wire delay at device layer average temperature</td>
</tr>
<tr>
<td>$DTC$</td>
<td>Total wire delay at device layer average temperatures with variation in wire capacitance</td>
</tr>
<tr>
<td>$q$</td>
<td>Heat flux per unit area</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Need for Planar CMOS Alternatives

The performance and cost benefit of technology scaling is diminishing with the “Moore’s” law slowing down. Planar CMOS transistor scaling is approaching practical limits. The increasing complexity and interconnect density are driving the development of more advanced VLSI packaging and interconnection approaches. These include three dimensional integrated circuits (3D-IC) with through silicon vias (TSVs) and emerging nano-scale devices fabricated with Carbon-nanotubes (CNTs). 3D ICs can potentially overcome the difficulty of integrating components with different functionality and processes as shown in Figure 1. Researchers have also started exploring new devices and channel materials in the sub-10nm technology nodes that have the potential to become the successor of Si-CMOS.
Introduction

Chapter 1

According to ITRS [2] some of the emerging logic devices which have the potential to replace Si in the post Si era are: (a) Nanowire FETs (NWFETs) [3] (b) III-V compound semiconductor FETs [4] (c) Graphene FETs [5] and (d) Carbon Nanotube FETs (CNFETs) shown in Figure 2. This thesis focuses on challenges faced in commercial realization of CNFET based circuits.

Fig. 1: Evolution of advanced packaging/integration styles

Fig. 2: Innovations in CMOS technology scaling until 22nm showcase the need for beyond CMOS Nano-scale devices
1.1.1 Emerging Logic Devices: CNFET Technology

The CNFET has the potential to become the channel material of future nanoscale transistors because of the excellent electronic properties of carbon nanotubes, such as near ballistic transport [6], high carrier mobility (103~104cm²/Vs), in semiconducting CNTs [7], and easy integration of high-k dielectric material [8] resulting in better gate electrostatics. CNFET uses a single-walled carbon nanotube (SWCNT) as channel material. The control electrode (gate) is placed above the conduction channel and separated from it by a thin layer of dielectric (gate oxide). The side view and top view of CNFET with an array of 4 CNTs is shown in Figure 3.

![CNFET side view and top view with an array of 3 s-CNTs and one m- CNT](image)

CNFETs allow the ballistic transport of carriers in the channel without any scattering. As a result, performance of these devices is superior to 2-D and 3-D devices. The absence of dangling bonds at the CNT surface allows an easy integration of High-K dielectric resulting
in better gate electrostatics, which in turn results in lower subthreshold slopes and lower OFF current in CNT based devices.

1.1.2 Vertical Integration: 3D ICs with TSVs

Three dimensional integrated circuits (3D-ICs) have emerged as a promising solution for gaining performance in a smaller form factor and low power. The through-silicon via (TSV)-based 3DIC technology has emerged as the most promising among the various 3D integration styles, and is the focus of this work. As mentioned in the previous section, the benefits of TSV based 3D ICs include smaller footprint, reduced global interconnects and heterogeneous integration as shown in Figure 4.

Fig. 4 (a) Homogeneous 3DIC with through silicon vias(TSVs) (b) Heterogeneous 3DIC
With aggressive CMOS scaling, while the performance of gates has improved, interconnects have become a major performance bottleneck because global interconnects do not scale accordingly with technologies. In 3D ICs, TSVs enable vertical interconnections between multiple stacked dies, replacing long global interconnects with shorter vertical interconnects as shown in Figure 4. Shortened global wires result in reduced wire delay, therefore improving the chip performance and power. Depending on the fabrication stage of TSVs with respect to devices and metallization, TSVs can be (1) Via-first (before FEOL and BEOL) (2) Via-middle (after FEOL and before BEOL) or (3) Via-last (after FEOL and BEOL). Please note, FEOL refers to the device fabrication stage and BEOL is the metallization. Throughout this work, we assume via-middle TSVs with a Face to Back chip stacking.

1.2 Technology Challenges

The above discussed emerging technologies are still under research and come with their own challenges. The specific challenges addressed in this thesis to close the gap are discussed in this section.
1.2.1 CNT-specific Fabrication Imperfections

Despite the noteworthy potential of the technology, CNTs come with their unique limitations that pose a foremost obstacle in realization of the CNFETs. We analyzed the major sources of CNT specific variations and plotted them in Figure 5:

(1) CNT Diameter Variation

(2) m-CNT induced CNT count variations

(3) CNT Length variation

![Diagram of CNT count variations](image)

Fig. 5 Contribution of individual CNT specific variations to ON current variation in CNFETs.

A minimum-sized CNFET using the Stanford CNFET model [9] at 32nm technology is used to simulate the $\sigma(I_{ON})/\mu(I_{ON})$ variation. CNT diameter variation [10]; $\mu=1.5\text{nm}$ and $3\sigma=0.5\text{nm}$. Presence of m-CNT: $P_m\% \leq 10\%$ with no CNTs removed [11][12]. CNT length variation is based on CNT diameter variation [13] and the variation parameters are given below.
Currently, there are no CNT growth techniques that can guarantee 100% semiconducting tubes. Single wall-CNTS are hollow cylinders with diameters in the range of 1nm to 2nm [13]. Depending on the chirality, a single-walled CNT can be either metallic(m-CNT) or semiconducting(s-CNT). In case of metallic tubes, the gate terminal has no control over the channel due to an ohmic- short between source and drain. Therefore, the presence of metallic tubes drastically affects the CNT-based circuits in complementary CNFET circuits. Current CNT growing techniques with a preferential growth, unfortunately entail 5% metallic CNTs with methane CVD and up to 20% metallic CNTs with plasma enhanced CVD. There are techniques that allow for removal of metallic tubes, and therefore removal of ohmic-contacts, from already fabricated gates. Techniques such as current-induced electrical burning, selective chemical etching or VLSI-compatible Metallic Carbon nanotube removal (VMR) [14] can be used to remove these metallic tubes from devices. The selective chemical etching and VLSI-compatible metallic carbon nanotube removal techniques are the most suitable. Since the presence of metallic tubes in gates is random, the removal of them causes variation in the number of tubes present in transistors, variation in reduced drive current and increased delay. We assume a $P_m\%$, presence of metallic tubes initially present and later removed in the range of 5-10%. Considering the latest advancements in m-CNT removal [14], we assume that no semiconducting tubes are removed during the metallic removal process.
Fabrication of CNFET-based circuits still faces major challenges in which the CNT diameter variation is another unique problem. In CNT growth, not all CNTs can be grown with the same diameter. Typical CNT diameters range between 1nm-2nm. CNFET diameter holds an inverse proportionality relation with the threshold voltage ($V_{th}$) and the ON current, $I_{ON}$ through a CNFET, depends directly on the threshold voltage. Variation in diameter and the number of semi-conducting tubes left under the gate when metallic tubes are removed, results in delay variation in CNFET-based gates which is addressed in this thesis. A Gaussian distribution with $\mu = 1.5$nm and $3\sigma = 0.5$nm is assumed to model the CNT diameter variation as shown in Fig. 6. From the generated CNT population, a given percentage of metallic tubes are randomly chosen and removed. This models the m-CNT induced CNT count variations.

![CNFET Diameter distribution](image)

*Fig. 6 Gaussian distribution for CNT diameter variation with $\mu = 1.5$nm and $3\sigma = 0.5$nm*
The third major source of variation is CNT length. The length of CNTs is often assumed constant and CNFETs aligned within the given CNT length are considered 100% correlated. However, in reality, equally long CNTs cannot be grown due to early catalyst precipitation and varying CNT diameter. This results in variable CNT length and few CNFETs being partially correlated in the direction of CNT growth. Therefore, for the first time, we consider variation in CNT length, in contrast to an optimistic assumption of 100% correlated CNFETs with a constant CNT length. This is especially important when CNFETs are correlated. The impact of CNT length variation on ON-current ($I_{ON}$) in CNFETs dominates the other two variations as shown in Fig. 5. CNT variations degrade the CNFET-based circuit performance and result in low functional yields.

Fig. 7. Ideal case and actual case in CNFET correlation under CNT length variation and fractured CNTs. Fractured CNTs are CNTs damaged during the transfer process of CNTs from growth substrate to target substrate.
CNTs grown should be equally long for CNFETs aligned in the direction of CNT growth (y-direction) to have identical current driving capability. Usually, CNFETs aligned in y-direction are assumed to be 100% correlated and perfectly uncorrelated otherwise (x-direction). This is termed as “Asymmetric Correlation". However, in the y-direction, partially correlated FETs exist along with perfect correlation as shown in Fig 7. Short CNT lengths can also be present due to fractured CNTs. Fractured CNTs are the CNTs damaged during the CNT transfer process from growth substrate to target substrate. In this work, we model the CNT length variation in the CNT growth process and fractured CNTs will be modelled in our future work. Larger diameters produce shorter lengths and vice-versa. Therefore, for a diameter range of 1nm - 2nm, we consider an average CNT length range of 0.6µm-0.4µm. The detailed CNT length variation modelling is described in Chapter 6.

All these variations result in CNFET circuit performance variation and functional yield loss. In addition, due to the presence of CNT-specific variations, conventional CMOS evaluation approaches for yield and delay cannot be used for CNFETs. This is a challenge for making design decisions. In CMOS, the worst case delay is computed assuming worst case gate delays for all transistors in the critical path. However, in CNFETs, the worst case is no s-CNT remaining in the channel after m-CNT removal and gate delay being infinite. This results in inaccurate estimation of functional yield and overestimation of critical path delay. Alternate approaches are thus needed for accurate CNFET based circuit evaluation.
1.2.2 Thermal management in 3D ICs

Chips are getting smaller, denser, and operate at high frequencies. 3D ICs with increased interconnect and transistor density result in high power dissipation. This ultimately results in increased temperature. This is because up to millions of components produce a great quantity of heat in such a tiny room, the temperature may sharply increase highly deteriorating the performance. This calls for solutions in thermal aware physical design at early stages like floorplanning.

3D Floorplanning - Floorplanning is the first major step in physical design. The floorplanning stage defines the location of the major blocks and components, affecting the optimization results of the subsequent stages including placement and routing. Floorplanning also provides early feedback that evaluates architectural decisions, estimates chip area, delay and congestion caused by wiring. As technology advances, the design complexity is increasing and the circuit size is getting larger. To cope with the increasing design complexity, hierarchical design and intellectual property (IP) modules are widely used. This trend makes floorplanning much more critical to the quality of a very large-scale integration (VLSI) design than ever. With the added dimension, the 3D solution space imposes significant runtime penalty to find an optimized solution compared to optimizing their 2D counterparts.

Thermal-aware Floorplanning - The need for designing chips to reduce temperature increase to reduce hot spots is more prominent in 3D ICs, where multiple
heterogeneous dies or same technology are stacked using through-silicon vias. This gives rise to the question, how to evaluate temperature distribution in 3DICs? And how to design a 3D chip layout to reduce peak temperature? An alternate parameter, other than temperature, to measure the thermal goodness of a floorplan is thus needed. It must be computed quickly compared to solving heat equations in time intensive RC networks, for fast thermal aware floorplanning. Additional package level solutions are required to address this heat dissipation problem shown in Fig. 8 in 3D ICs.

Fig. 8 Heat dissipation problem in 3D ICs due to increased power density compared to 2D

**Run time** – Most state of the art layout design tools and 3D floorplanning algorithms are probabilistic. Be it Simulated Annealing or Genetic or evolutionary algorithms, they require thousands or hundreds of thousands of iterations to find an optimized solution. To use temperature reduction as our objective, we need to evaluate
temperature distribution at each iteration and this is in the traditional way, will spike the run time up to 6hrs or even more depending on circuit size! Ex: For example, consider two test circuits shown in Fig. 9(a) [15].

![Fig. 9(a) Test circuits – planar and 3D (b) Thermal evaluation using 3D-ICE [15] (c) Thermal evaluation using Hotspot [1]](image)

When these two simple test circuits are evaluated using the state of the art thermal evaluation tools widely used in academia [15][1], even with their proposed speedup, one solution takes additional 49s. If the evaluation is done for every 100 runs, 100 times, one simulation for a thermally optimized solution will be extended for ~6hrs! We therefore
Introduction

need faster computing techniques that can be incorporated into the iterative probabilistic algorithms.

Above all, to the best of our knowledge, almost all works in literature [16][17][18][19, pp. 3-] evaluate 3D chip performance with fixed parameters at room temperature. However, due to increased power density, each layer will have a different temperature profile higher than the room temperature. Presence of a vertical thermal profile in 3D ICs significantly impacts the interconnect delay, power and buffer count on individual layers as metal resistance is a strong function of temperature. Since, interconnect delay and power are heavily used to characterize the 3D floorplan solution quality, ignoring the thermal impact on interconnect performance can lead to suboptimal solution selection during the probabilistic optimization of 3D floorplanning. Thermal-aware evaluation can possibly change the final solution selection in a non-deterministic floorplanning flow as a well packed 3D IC does not necessarily have the best thermal distribution impacting the interconnect performance.

1.3 Motivation and Significance of Work

This work is motivated by the need for planar CMOS alternatives both, at the device level and system level to overcome the channel length scaling challenges and the interconnect performance bottleneck. The proposed work on CNFET circuit design under CNT fabrication imperfections offers significant improvement in functional yield. Tube
redundancy in only statistical critical paths is proposed to achieve the desired functional yield without degrading circuit performance, with less than 2% increase in circuit power at no area overhead. We limit the redundancy to a minimum required, to avoid unnecessary increase in the channel area. With proposed minimum redundancy we are able to reduce the allowed delay degradation(tolerance) by 15-25% and still have a good functional yield (>99%). An efficient and fast algorithm compared to time intensive Monte Carlo simulation, to identify all paths that can become critical under CNT variations is also presented. On the other hand, for a rapid thermal screening of 3D IC solutions, a power-based approach is proposed. We achieve a correlation coefficient of 0.9 when the results are compared to the HotSpot tool. Compared to other state-of-the-art 3D IC thermal evaluation approaches [20][21], the proposed power model is much faster and can be incorporated into the 3D floorplanning flow without drastic increase in runtime. The need for thermal-aware evaluation for accurate selection of optimized solutions is shown on final floorplans of GSRC benchmarks. With considering wire density on different layers of 3DIC, we show that the impact of temperature and wire density on interconnect performance are very crucial for a more realistic evaluation and avoid sub-optimal solution generation. This work enables the design of reliable and variation tolerant CNFET circuits by achieving delays almost equal to delay with no variations thus facilitating the realization of CNFETs even in the presence of variations. The power based 3DIC thermal goodness evaluation technique can provide thermally optimal solutions without a drastic runtime overhead, especially in the thousands of runs in probabilistic optimization models. These
approaches can equip a designer to make reliability-oriented decisions early in the design process, leading to better chip yield.

1.4 Thesis Contributions

The contributions for work on CNFETs and 3D ICs are listed below.

1.4.1 Circuit-level design methodology for reliable and variation-tolerant CNFET circuits under CNT variations (Part A)

1) A statistical evaluation approach to improve prediction accuracy of functional yield and critical path delay under CNT fabrication imperfections.

2) A calculated minimum CNT redundancy at the transistor level to improve functional yield for a given failure rate.

3) We focus our evaluation on a critical path that in reality, due to metallic tube presence, could be a different path in each circuit. We limit the redundancy to a minimum required, to avoid unnecessary increase in the channel area.

4) Redundant tubes are only added to transistors on paths that are statistically evaluated as critical (P_{PC} - paths possibly critical) due to the presence of metallic tubes and CNT diameter variations. This strategy results in functional yield enhancement, and allows for reduction in allowed increase in delay.

5) Development of an efficient statistical algorithm for fast identification of all paths
that can become critical, $P_{PC}$, under statistical critical path delay variation. The proposed algorithm is much faster than the time intensive Monte Carlo simulations.

6) Power minimization with optimized CNT redundancy at path level without degrading the functional yield.

7) Evaluation of two standard cell layout styles for increase in circuit area due to redundancy and optimized redundancy. Identifying a tube redundancy approach to provide delay minimization with no area overhead.

8) CNT length variation is addressed for the first time in CNFET circuits, which is especially important if correlated CNTs are used in CNFETs within a logic gate.

9) Evaluation of proposed tube redundancy approach on correlated and un-correlated CNFET based circuits to estimate the functional yield and circuit performance benefits possible with CNT correlation.

1.4.2 CAD approaches for fast thermal goodness evaluation and thermal management in 3D ICs (Part B)

1) Development of a fast approach for thermal goodness evaluation of 3D floorplans with only available block power information. The unique strength of this technique compared to state of the art approaches is that, we do not rely on solving any heat equations in this method, yet determine the thermal goodness of two given 3D designs using the power densities of the blocks, and impact of neighboring blocks.
and using a correlation measure. This approach enables run time optimization with peak temperature reduction in 3D ICs and generation of thermally optimal 3D floorplans.

2) Fast fusion model for vertical thermal profile generation to significantly increase correlation with the more accurate simulation based Hotspot tool [1]. The proposed model overcomes the shortcomings in a vertical only model [22] by considering the impact of heat dissipation from intra-layer modules as well and quickly evaluating the thermal goodness of two given 3D floorplans that is in good acceptance with the Hotspot tool.

3) Design strategy to integrate true thermal-aware interconnect performance evaluation in the 3D floorplanning optimization while considering the impact of vertical thermal profile and non-uniform wire spacing on TSV-aware buffer insertion length. This strategy enables a more realistic evaluation of the wire delay and hence, avoiding sub-optimal solution generation. Most publications evaluate their 3D designs at room temperature and neglect the presence of a vertical thermal profile. This can result in seriously sub-optimal solutions being generated through the floorplanning stages.

4) In-depth analysis of 3D interconnects delay and early buffer count estimates under unified RC variations resulting from rising vertical temperature and varying wire capacitance to quantify the interconnect performance degradation.

5) Evaluation of various Thermal Interface materials and additional layers between Si, to identify the critical/minimum thickness and suitable material needed for a good
vertical heat transfer in 3DICs. We propose the use of a graphite based Thermal interface material sandwiched between the IC stacks for better heat dissipation between the ICs towards the heatsink. With this architecture, we were able to reduce peak temperature by 12% in GSRC benchmarks, without the need of any micro channel cooling.

6) Efficient interface of white space generator to bridge between our 3D floorplanner and HotSpot tool to evaluate 3D floorplan peak temperature on HotSpot and validate the effectiveness of proposed thermal goodness evaluation approach.

7) Scalable approach for layout to nxn grid conversion of “m” layers block power densities with no limit on “n” and “m”. This technique uses a block to grid overlap model to distribute the block area and power to the underlying grid sections. The block power contribution to the grid relies on the block to grid overlap ratio.

1.5 Thesis Organization

This thesis is organized two-fold to explore the possibilities for “More than Moore” at device level using CNFETs and as well as package level using 3DICs.

In the first part of the thesis, we develop methods focused on CNFETs and CNT induced variations. Methodologies are proposed for statistical delay evaluation and to reduce power increase due to redundancy in CNFET based circuits in the presence of CNT variations. For the first time, CNT length variation is investigated and its impact on functional yield is also studied. The second part of the thesis acknowledges that the thermal
performance of 3D chips is directly controlled by the thermal quality of the generated floorplan and shows that ignoring the impact of temperature on interconnect delay, power and buffer estimation in the evaluation process can result in severely sub-optimal solutions being generated. Models for fast thermal goodness evaluation of 3D layouts are proposed and integrated in the 3D floorplanning flow. The proposed thermal evaluation models are also validated against the more accurate simulation based tool, Hotspot using a correlation factor.

This thesis is organized as follows. Chapter 2 describes the related works in the field of proposed contributions pertaining to 3DICs and CNFET circuits. Part A (Chapter 3 – Chapter 6) encompasses work done on CNFETs. Part B (Chapter 7 – Chapter 10) includes the design strategies and analysis performed on 3D ICs. Chapter 3 describes the statistical evaluation of critical path delay in CNFETs. The strategy for minimum tube redundancy is discussed in Chapter 4. In Chapter 5, redundant CNT optimization for power minimization and a fast approach to identify all critical paths under CNT variations is discussed. CNT length variation is presented in Chapter 6. A first-hand analysis on importance of thermal-aware delay in 3D interconnect performance is done in Chapter 7. Chapter 8 presents the fast thermal goodness evaluation model for 3D ICs as a temperature alternative. This proposed model is integrated with 3D delay-aware floorplanning for true thermal-aware 3D interconnect performance optimization along with considering interconnect density in Chapter 9. Chapter 10 explores package level thermal management options using graphene and graphite based TIMs. Finally, thesis conclusions and future scope of work are presented in Chapter 11.
Previous Work

Chapter 2

2.1 Carbon Nanotube FETs

Zhang et al., [23] reported that the major sources of CNT specific variations in CNFETs are due to semiconducting-CNT (s-CNT) count variations caused due to removal of initially present metallic-CNTs (m-CNTs) [24], CNT diameter variations, variation due to misalignment of CNT arrays, and variation due to CNT doping. However, considering the recent advances in m-CNT process [14] and growth of high density s-CNT arrays, it is required to revisit and re-evaluate the influence of these variations on the ON current of a CNFET and take CNT length variation into account. Redundancy techniques are often used to improve the functional yield under CNT variations.

Redundancy in CNFET circuits can be either transistor level or tube level. In transistor level redundancy [25], each transistor in the original design is replaced by N2 series-parallel/parallel-series transistor structures, where N is the number of replicated transistors in each dimension. Open and short-immune structures can be obtained through this technique at the expense of large area and power overheads. In Tube level redundancy
additional CNTs are added in the channel region to increase drive current and avoid transistor failure and is a more practical approach for yield enhancement.

Few publications have proposed circuit level techniques to enhance the functional yield in the presence of CNT variations. With a trade-off in performance, Ashraf et al. [28] proposed transistor and tube-stacking configurations to reduce the statistical probability of a short between source and drain in CNFET gates. Tube level redundancy was proposed to increase functional yield at gate level [26] and in memory arrays [27]. Cheng et al. [27] proposed to add redundant CNTs to transistors based on transistor sizing to optimize critical path delay. However, it is important to note that, under CNT-specific variations, the critical path can possibly be different for every instance of the same circuit. In addition, it is highly possible to have multiple critical paths in a circuit and thus adding tube redundancy to transistors on a single-path will not sufficiently improve the circuit performance.

2.2 Vertical Integration: 3D ICs with Through Silicon Vias

HotSpot [1] is a thermal evaluation tool, extensively used in academia and can handle both 2D and 3D solution evaluation. The tool relies on RC networks and intensive heat equation solving. Though the tool is over 94% accurate, it is very time consuming to be integrated into physical design optimization cycles. Similar is the case with 3D-ICE [15] which was proposed by Sridhar et al, for thermal evaluation of 3D ICs with integrated micro-channel cooling.
Xu et al., [20] proposed a fast thermal analysis model for fixed-outline 3D floorplanning. In their approach, they simulate the thermal profile of each block placed in all possible locations on a floorplan. Thermal profiles of all blocks placed in all locations is generated using the Hotspot tool. During floorplanning, for thermal analysis, the block’s position determines its temperature. For t layer 3D IC with tk^2 thermal profiles are generated for a k x k grid. While the thermal profile varies with position it will also be impacted by the neighboring blocks which is not considered in this work. In addition, though the look up of block thermal profile is fast, the approach involves time consuming process of individual block profile generation prior to floorplanning which can take longer than incorporating HotSpot directly with increasing circuit complexity and size. A similar approach was proposed by Wu et al., [29] but the blocks are simulated by placing them only in the center of the grid and not in k^2 locations.

Xiao et al., [21] proposed another fast approach for thermal goodness evaluation as an alternative to using Hotspot. The temperatures of blocks are computed first on the coarsest level similar to HotSpot and then interpolated to finer grids based on effective ratio of resistances in the vertical and lateral directions. However, multiple interpolations are needed to capture the lateral heat flow. They show a correlation coefficient of 0.95 with Hotspot for GSRC benchmarks. The above mentioned two approaches are faster than Hotspot but not considerably fast enough for incorporating into floorplanning. Their runtime penalty reported for their approach is 4X because multiple iterations are required for refining the grids and interpolating the temperatures. Therefore, much faster thermal
evaluation approaches are required in 3D IC floorplanning that can generate thermally optimal solutions without degrading area, wire length and runtime.
PART A: Circuit-level design methodology for reliable CNFET circuits under CNT variations

Part A. Circuit-level design methodology for reliable and variation-tolerant CNFET circuits under CNT variations
Chapter 3

Statistical Evaluation of Critical Path Delay for Functional Yield Estimation

This chapter is published in IEEE-NANO, 2017.


3.1 Introduction

With their superior electrical, thermal and mechanical properties, carbon-nanotube FET transistors have been emerging as a promising candidate for the future integrated circuits. For the same length of channel, CNFETs are reported to be 13 times faster than a PMOS and 6 times faster than an NMOS transistor [24]. The drive current of the CNFETs is proportional to the number of semi-conducting tubes under the gate (channel of the transistor). An array of parallel carbon nanotubes (CNT) placed under the gate results in a stronger ON current through the parallel-tube CNFET, as compared to a single CNT, and can tolerate more of the fabrication imperfections because of statistical averaging [30]. The parallel-tube inverter is shown in Figure 10.

Fig. 10. CNFET layout of inverter with parallel tube configuration
Despite of the noteworthy potential of the technology, CNTs come with their unique limitations that pose a foremost obstacle in realization of the CNFETs. Single-wall CNTs are hollow cylinders with diameters in the range of 1nm to 2nm [13]. Depending on the chirality, a single-walled CNT can be either metallic or semiconducting. Currently, there are no CNT growth techniques that can guarantee 100% semiconducting tubes. In case of a metallic tube in the tube array, the gate terminal has no control over the channel due to an ohmic-short between source and drain.

Therefore, the presence of metallic tubes drastically affects performance of CNT-based circuits in a complementary CNFET arrangement. Current CNT growing techniques with a preferential growth, entail 5% metallic-CNTs, with methane CVD, and up to 20% metallic CNTs with plasma enhanced CVD [31]. There are techniques that allow for removal of metallic tubes, and therefore enable removal of ohmic-contacts, from already fabricated gates. Techniques such as current-induced electrical burning [32], selective chemical etching, SCE, [33] or VLSI-compatible metallic carbon nanotube removal, VMR [14] can be used for tube removal after gate fabrication. The selective chemical etching and VLSI compatible metallic carbon nanotube removal techniques are the most suitable. Since the presence of metallic tubes in gates is random, the removal of metallic tubes causes variation in the number of tubes present in transistors, reduced drive current and increased delay.

Another challenge faced during the fabrication of CNFET-based circuits is the variation in the diameter of CNT’s. In CNT growth, not all CNTs can be grown with the same diameter. Typical CNT diameters range between 1nm-2nm. The threshold voltage \(V_{th}\) of a CNFET is inversely proportional to the CNT diameter, and thus the ON current,
I\textsubscript{ON}, through a CNFET depends directly on the threshold voltage, and is inversely proportional to the CNT diameter. The impact of variation [34] in the diameter and in the number of semi-conducting tubes left under the gate after metallic tubes are removed is addressed in this chapter. The focus of this chapter is on the statistical in contrast to worst-case path delay evaluation. The effect of the presence of variations on critical path delay and functional yield are also considered.

**Worst-Case vs Statistical Evaluation**

It is increasingly important to develop better and more accurate critical path evaluation for CNFET-based circuits in the presence of fabrication imperfections. The worst-case delay approach is widely used to compute the critical path delay in CMOS-based integrated circuit design. This technique proved good for technology nodes above 45nm where the variation in CMOS gate delay was a small percentage of the total gate delay. In CMOS technology nodes below 16nm and in circuits designed with emerging materials such as carbon nanotubes, the delay variation is a significant percentage of the total gate delay.

In the worst-case approach, the maximum allowed critical path delay is defined assuming the worst-case delays for all gates in the critical path. In CMOS circuit this is a very straightforward approach. The worst possible delay of a gate is, PMOS and NMOS transistors have to be the slowest possible for a given technology node. So all fabricated circuits have the critical path delay smaller or equal to the worst-case delay value of the desired delay, which is large.

For CNFET-based circuits the definition of the gate worst-case delay has to be modified. Due to the necessity for metallic tubes to be removed, and a finite probability of
all tubes in a gate being metallic, the probability of a gate with no tubes is not zero. So, the worst possible case is no tube in a channel and the gate delay equal to infinity. Such a gate however is non-functional and therefore, a circuit with non-functional gate is also nonfunctional. Therefore, for the worst-case gate delay we will use the delay of the gate with the largest delay but still functional. With this definition, a large number of fabricated circuits are accepted to be functional because their critical path delays are equal to or lower than the worst-case path delay. However, in reality, there is a non-negligible probability that few gates have all tubes metallic that will be removed, leaving these gates nonfunctional. Such cases will directly affect the functional yield.

Fig. 11: ISCAS 85 C17 Benchmark circuit

This emphasizes the need for enhanced evaluation of the critical path delay in CNFET based circuits considering the variations [10]. Our focus is to analyze the impact of the CNT diameter variations and variation in the number of tubes in transistors, when initially present metallic tubes are removed, on critical path delay of CNFET circuits. In general, each logic gate is assigned a given delay value for performance evaluation. However, in fabricated CNFET circuits, the delay of each gate differs because of CNT density variation under the gate. For example, consider the C17 benchmark circuit shown in Figure 11. This
circuit is built with only NAND gates. Observe that though all the gates are identical, the delay of each gate is different because of the presence of variations in the CNFET-NAND gate.

Hence, we propose to use a statistical approach to improve the evaluation of performance and functional yield of CNFET-based circuits. We take into account the CNT diameter variations and variation in the number of semi-conducting tubes in the transistor after the metallic tubes are removed. Our results show that with statistical evaluation, the critical path is evaluated with a higher accuracy considering the CNT delay variations.

**CNT Fabrication Imperfections**

The variation in diameter of the carbon nanotubes in typically fabricated CNTs show a Gaussian distribution [13] with diameters of CNTs shown in Figure 12. For CNFET transistors to have acceptable drive current, we used an array of parallel CNTs [13]. In this paper, each gate is assumed to have 10 CNTs in the channel [35][28].

![CNT Diameter Distribution](image)

Fig. 12: Diameter distribution of CNTs diameter with respect to $\mu$ and $3\sigma$ diameter of 1.5nm and 0.5nm respectively
The variation is defined as the standard deviation with respect to the mean value and is given as,

\[ \text{Variation} = \frac{\sigma}{\mu} \]  

(3.1)

where, \( \sigma \) is the standard deviation and \( \mu \) is the mean. Shahi et al., [10] discuss the simulated results for variance, \( \sigma(I_{\text{on}})/\mu(I_{\text{on}}) \) of a minimum width CNFET (with 10 tubes in the channel of a gate) at 32nm technology node for various sources of CNT-specific variations. \( I_{\text{on}} \) is the ON current through a single CNT under the gate. They show a large fraction of variation from presence of metallic tubes and CNT diameter variations.

Fig. 13: CNT ON current distribution for diameter distribution given in Figure 12 of 10,000 CNT population

The variation of ON current through a CNT for individual CNTs is shown in Figure 13. The variation in \( I_{\text{on}} \) of a 1X (10 tubes in each transistor) inverter, that has equal ON current through the p-type and n-type CNFETs is shown in Figure 14. The ON current of 31
a CNFET varies by nearly 3.5% when all the CNTs under the gate are semiconducting with diverse diameters. The variation in ON current increases up to 10.9% - 14.1% when initially present percentage of metallic tubes (P_m) of 10% and 20% are removed, as shown in Figure 14.

![Fig. 14: ION variation of 1X Inverter at 32nm technology node of 1XCNFET Inverter showing increase in the ON current variation with increase in initial presence of metallic tubes](image)

To simulate the behavior of CNFET-based circuits, we used HSpice compatible model developed at Stanford [9]. The delay of the logic gates is computed for 10 CNTs in the channel region with diameters in the range of 1nm-2nm. In our experiment, 10,000 samples of CNTs are considered for the analysis. 10,000 samples are used for the ease of gate delay calculation i.e., to remove 10% metallic tubes, 1000 tubes are randomly removed from the 10,000 CNT population. The percentage of metallic tubes (P_m) initially present and subsequently removed, using VMR technique, are assumed to be 10% and 20%. We also assume that all metallic tubes are completely removed and no semi-conducting tubes are removed.
### ON Current and Gate Delay Calculation for CNFETs

For critical path evaluation in the CNFET integrated circuits, the gate delay distribution considering the variations have to be calculated primarily. To generate the gate delay distribution for each gate in the circuit, we start with a randomly assigned CNT population of 10,000 tubes with diameters in range of 1nm-2nm. From equation (3.2) [35], the $I_{ON}$ of a gate is directly proportional to the threshold voltage ($V_{th}$) and the threshold voltage of the gate is inversely proportional to the diameter of the CNT as shown in equation (3.3).

\[
I_{ON\text{-}CNT} = g_{CNT} (V_{DS} - V_{th} - \text{DIBL} \times V_{DS})
\]  
\[(3.2)\]

In Eq.3.2 $g_{CNT}$ is the trans-conductance of a CNFET transistor. $V_{DS}$ is 0.9V for 32nm technology. DIBL is the drain induced barrier lowering and is taken as 100mV/V and $V_{th}$ is the first order approximation of the threshold voltage. according to Ashraf et al. [35],

\[
V_{th} \alpha \frac{1}{D_{CNT}}
\]  
\[(3.3)\]

where $D_{CNT}$ is the diameter of a CNT. The propagation delay of the CNFET gate is given by,

\[
T_{pd} = \frac{C_L \cdot V_{DD}/2}{I_{ON\text{-}N}}
\]  
\[(3.4)\]

\[
I_{ON\text{-}N} = I_{ON\text{-}1} \times N
\]  
\[(3.5)\]
Where, $T_{pd}$ is the propagation delay, $C_L$ is the load capacitance, $V_{DD}$ is the supply voltage, $I_{ON,N}$ and $I_{ON,1}$ are the ON current through N number of CNTs and 1 CNT respectively.

The ON current for each semiconducting CNT is calculated considering only the diameter variation in range of 1nm-2nm. The drive current of a gate with 10 semiconducting tubes is a summation of 10 CNT’s ON currents. Therefore, using the 10,000 population of CNTs we generated a distribution of 1000 values of ON current for 1000 logic transistors. These values are used to calculate the delay distribution for 1000 transistors based on equation (3.4) and (3.5).

### 3.2 Methodology- Statistical Critical Path Delay Evaluation

We first define all the delays computed in this chapter as follows.

- **Functional gate**: Under the impact of fabrication imperfections, a gate is functional if it has at least one semi-conducting tube in the channel.

- **Ideal gate delay**: All tubes in the channel region are semi-conducting and all have a mean diameter of 1.5nm

- **Ideal critical path delay**: All gate delays in the critical path are replaced with ideal gate delay and the longest delay path is calculated.

- **Worst-case gate delay**: Only one semiconducting tube in the channel with the minimum diameter of 1nm.

- **Worst-Case critical path delay**: All gate delays in the critical path are replaced with
worst-case gate delay and the longest delay path is calculated.

- **Functional Yield** \( Y_f \): Functional yield is calculated as a function of delay [28]. It is given as the ratio of the number of circuits with fabrication imperfections with delay less than or equal to the 1.3 times the ideal case delay to the total number of circuits, and is given as,

\[
Y_f = \frac{\# \text{ circuits with } \leq 1.3 \times \text{ideal case delay}}{\text{total number of circuits}}
\]

In the current design evaluation, the critical path delay of a circuit based on CNFETs is estimated using the worst-case approach. Any circuit with delay less than the delay calculated in worst-case approach is accepted to be functional. However, this approach does not take into account statistical variation in the fabrication imperfections of CNTs. This will result in overestimation of the critical path delay. In this paper, we propose to use the statistical approach to evaluate the critical path delay rather than the worst-case approach. To evaluate the performance of the CNFET-based circuits, we accept the circuit to be functional only if the delay of the critical path is not larger more than 30% of the ideal case delay. This approach evaluates the critical path taking the variation in the gate delay into consideration. This results in more realistic critical path delay evaluation, accepts circuits with a very small increase in the critical path delay and increases the yield.

### 3.2.1. Worst-case Approach

In the worst-case approach, the individual gates in the circuit are assigned the worst-
We calculated the critical path delay by the summation of all the gate delays on every path from the input to output nodes. This is achieved using \textit{distances} \cite{36} (previously \textit{graphallshortestpaths} \cite{37}) function in MATLAB. Since the \textit{graphallshortestpaths} and \textit{distances} functions compute the shortest paths rather than the longest path, we negate the gate delays. Therefore, the computed shortest path by the MATLAB function is actually the longest paths. The path delays are negated again before proceeding with further analysis. The path having the longest delay among all paths is our worst-case critical path delay.

### 3.2.2. Statistical Approach

In statistical approach, each gate in a circuit is randomly assigned a delay value from the delay distribution generated by including diameter and tube number variations. The longest path delay of the circuit is computed with the statistical gate delays. This will result in a more realistic and shorter critical path delay than the critical path delay obtained through the worst case-approach. And as it will be shown later, due to statistical averaging of gate delays in fabricated circuits many circuits will satisfy the condition of the critical path delay being equal or not larger than 1.3 times the critical path delay of the ideal case. The critical path delay is evaluated for three different scenarios. One being the diameter variation case assuming all the CNTs are semi-conducting and any variation in ION is caused only by the CNT diameter variations. The other two cases, take into account the removal of 10\% and 20\% of metallic tubes (P\textsubscript{m}) that are initially present.
Diameter Variation Case ($P_m=0\%$)

For analysis of the impact of CNT diameter variations on the critical path delay, all tubes are assumed to be semi-conducting with diameter in range of 1nm-2nm but no metallic tubes present ($P_m=0\%$).

Non-Ideal Cases ($P_m =10\%$ and $P_m = 20\%$)

Considering the initial presence of metallic tubes is a more realistic scenario. The initial presence of metallic tubes reduces the drive current because of the reduced number of semiconducting tubes under the gate after the application of metallic tube removal process. In the non-ideal case, we assume the initial percentage of the metallic tubes being present ($P_m$) to be 10% and 20%. We then assume that all metallic tubes are completely removed, and no semi-conducting tubes are removed. In this case, the randomly generated 10,000 population of tubes with diameters in range of 1nm-2nm are considered. From these, 10% tubes are randomly identified as metallic and are assumed to be completely removed from the CNT population using the VMR technique. The same process is repeated for 20% metallic tubes. The ON current for each gate and delay distribution are calculated. The results for all the three cases are generated and analyzed together with the effect of CNT diameter variations.

3.3 Effect of CNT Fabrication Imperfections on Functional Yield

For achieving a high functional yield, we want larger number of circuits to be functional. To demonstrate these results, the distribution of the critical path delay of the
CNFET-circuit generated using Monte Carlo simulation is plotted in Figure 15. This plot shows the variation in the critical path delay of the ISCAS 85 C432 benchmark circuit simulated using the gate delay distribution, and closely resembles a Gaussian distribution. From Figure 15 we can observed and calculate that 99.9% functional yield (3σ point) can be achieved even in the presence of CNT diameter variation and CNT density variation after the removal of metallic tubes. The functional yield is calculated using (3.6). As discussed in [28], it can be observed that for C432 benchmark, the critical path delay value obtained from statistical distribution of gate delays should be significantly less than the critical path delay based on the worst-case. If the critical path delay is evaluated with a statistical approach, a larger number of circuits is classified as functional assuming a small degradation of delay (not exceeding 30% of ideal case) as compared to the sideal case with no variations. This is shown in Table 1 presenting our experimental results.

Fig. 15: Critical path delay distribution of ISCAS 85 C432 benchmark circuit showing the delay spread for $P_{\text{in}}=0\%$. 
3.4 Experiment Results

Our approach was tested on the logic benchmark circuits of ISCAS’85, C17, C432, C880, C1355 and C2670. For simulation purpose, the 32-nm technology node is considered, and supply voltage of 0.9V as given in ITRS guidelines [2]. A load capacitance of 200fF is assumed [10]. Our results show that using the statistical evaluation, the critical path delay is accurately evaluated compared to the worst-case approach thereby increasing the functional yield. The values for critical path delay and projected functional yield for the set of ISCA’85 logic benchmark circuits are given in Table 1. Results are shown for ideal case ($P_m =0\%$) and two non-ideal cases $P_m=10\%$ and $20\%$ including CNT diameter variations. $P_m$ is the percentage of initially present metallic tubes. The time complexity of the critical delay path calculations $O (V \log V +VE)$ (where $E$ is the number of gates and $V$ is the number of nodes in the circuit). A critical path delay of 13.2ns is calculated using the worst-case approach in ISCAS’85 C432 benchmark circuit. Statistical evaluation results show an improved evaluation of critical path delay and therefore better more realistic estimation of functional yield. In the diameter-variation only case, the critical path delay for C432 circuit, with six gates on the critical path, shows the delay of 930ps with a projected yield of 99.99%. It shows that this approach allows for more accurate delay evaluation for larger circuits and demonstrates that a high yield can be achieved.
Table 1. Improvement in critical path delay evaluation and functional yield for ISCAS benchmark circuits with statistical delay calculation approach with 10 tubes in the channel region and diameter in range of 1-2nm. 1.3X Ideal case is the nominal delay allowed for the circuit to be functional.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Critical path delay</th>
<th>Functional Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Worst-case delay</td>
<td>Ideal Case</td>
</tr>
<tr>
<td></td>
<td>(ns)</td>
<td>(ns)</td>
</tr>
<tr>
<td>C432</td>
<td>13.2</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C17</td>
<td>3.6</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C880</td>
<td>30.2</td>
<td>2.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1355</td>
<td>19.4</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2670</td>
<td>35.5</td>
<td>2.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.5 Chapter Conclusions

We use a statistical approach to estimate the critical path delays of CNFET-combinational circuits. The proposed approach predicts more realistically the critical path delay compared to the worst-case technique. It is observed that the impact on critical path
delay depends more strongly on CNT diameter variations when the initial presence of metallic tubes in CNFET circuits is increased.

Generated delay distributions for 10000 iterations were presented. The presence of metallic tubes alone increases the worst-case critical path delay by nearly 60%, and as a result of which, the worst-case approach would predict incorrectly a very low functional yield. Our approach indicates that up to 99% yield can be achieved with less than 10% of the initially present metallic tubes and for only slightly increased acceptable critical path delay (1.3 x ideal case value). For 20% of the initially present metallic tubes, additional design techniques have to be adopted to increase the yield. The statistical approach allows for more realistic prediction of the expected functional yield. The improved prediction is based on more realistic distribution of critical path delay values in fabricated chips.
Chapter 4

Tube Redundancy in Statistical Evaluation of Critical Path Delay in the presence of CNT variations

This chapter is published in IEEE-NANO, 2019.


4.1 Introduction

This chapter addresses the variation in the “ON” current, I\textsubscript{ON}, of a CNFET caused by CNT diameter and CNT count variation due to m-CNT removal. In the previous chapter, we were able to show a functional yield between 92.04% and 99.95% with initial presence of 10% and 20% m-CNTs (P\textsubscript{m} %), for a set of ISCAS’85 benchmark circuits. However, we assume a high delay degradation acceptance (delay tolerance) of 1.3X for functional yield evaluation. Moreover, for increased percentage of initially present metallic tubes, a low functional yield was observed. The delay degradation tolerance is the acceptable delay, 30% higher than the ideal critical path delay, beyond which the circuit is not accounted for the functional yield estimation. This emphasizes the need for design methodologies for improving the functional yield with a low delay degradation tolerance. This will help fabricate circuits that are closer to the ideal performance. It is also important to have credible evaluation methods. Statistical approaches for evaluation of critical path delays and functional yield allows for more realistic predictions of these parameters in fabricated chips under fabrication imperfections.
In this chapter, we propose to use a calculated minimum CNT redundancy at the transistor level to improve functional yield for a given failure rate. We focus our evaluation on a critical path that in reality could be a different path in each circuit due to the tube variations. We limit the redundancy to a required minimum in order to avoid unnecessary increase in the channel area. Redundancy is frequently used to improve a failure tolerance, but it is also important to add just enough to meet a given goal. We thus investigate the impact of a given redundancy (25% or 50% tube redundancy) compared to our calculated minimum on total gate area. We present results of improved delay degradation tolerance limit and higher functional yield.

4.2 Methodology - Tube Redundancy to improve Functional Yield

A. Minimum Tube Redundancy

Circuits are often designed to achieve a desired functional yield at a given failure rate. In this work, we add the proposed pre-calculated tube redundancy to all the transistors in a circuit to avoid a failure rate of 0.001%. The minimum redundant CNTs required, $N_{R_{min}}$, in a transistor prior to tube removal process for less than 0.001% probability of failure rate, is calculated using (4.1) [26].

$$N_{R_{min}} = \log \left( \frac{10^{-5}}{P_m} \right)$$  \hspace{1cm} (4.1)
where, $P_m$ is the probability of $m$-CNTs initially present and later removed. The $P_m\%$ is assumed between 0\%-10\% \cite{11} considering the recent advances in CNT growth technology. The diameter variation is considered in the range of 1nm-2nm, represented by a Gaussian distribution of $\mu=1.5\text{nm}$ and $3\sigma=0.5\text{nm}$ \cite{10}. We also compare our proposed minimum redundancy with $1/4^{th}$ and $1/8^{th}$ tube redundancy added to the CNFETS to analyze the performance-area trade-off. Our approach will enable improved performance in CNFET-based circuits without aggressive increase in channel area.

To better understand the minimum tube redundancy scheme, preliminary results of average reduction of gate delay in 10,000 instances of 1X inverter is shown in Fig.16. The overlapped delay distributions of 20 runs shown are generated with diameter variation and presence of $m$-CNTs. The delay distributions in red are generated with 32CNTs per CNFET with a 10\% of $m$-CNTs initially present and later removed, ($P_m\%$). This is considered as our reference case for comparing with other given tube redundancies. The inverter instances delay distribution with pre-calculated redundancy (35CNTs/CNFET) is highlighted in blue and the result of $1/4^{th}$ redundancy (40CNTs/CNFET) is shown on the extreme left. The increasing peak and decreasing mean gate delay should be noted among all the three distributions. The mean gate delay in a 1X inverter is improved by nearly 8.9\% and 19\% when minimum redundancy and $1/4^{th}$ redundancy is added, respectively. Although $1/4^{th}$ redundancy shows the highest improvement in delay, it comes at the cost of aggressive channel area increase. The pre-calculated redundancy, on the other hand, results in a reasonable delay reduction. It achieves a similar functional yield as $1/4^{th}$ redundancy with minimum increase in area as discussed in the later sections.
B. Gate area estimation

The CNT diameter, $D_{CNT}$, the channel width of the CNFET transistor, $W_g$, the number of CNTs, $N_{CNT}$, in the channel of a CNFET, and CNT-CNT spacing, $S_{CNT-CNT}$, are related by (4.2) and varies between each gate instance due to the diameter variation.

$$W_g = N_{CNT} \times D_{CNT} + (N_{CNT} - 1)S_{CNT-CNT}$$  \hspace{1cm} (4.2)

Fig. 16. Gate delay distributions of 10,000 instances of 1X inverters for 20 runs, using 40, 35 and 32 CNTs/CNFET (left to right) showing the improvement in variation tolerance (narrow distributions) and average mean gate delay reduction with redundancy.

The critical channel length, $L_g$ used for area evaluation is 32nm. The number of CNTs in a transistor before added redundancy is assumed to be 32 based on the CNT density requirement of 250CNTs/µm [38] with an optimal CNT-CNT spacing of 4nm [38]. The gate area is the product of the effective gate width and gate length of the channel. The total area occupied by the channel in a circuit is obtained by the summation of area occupied by each gate in the circuit.
4.2.1 Statistical Critical Path Delay Evaluation

A. Critical Path Delay

A statistical estimation of the critical path delay in CNFETs, proposed in [39], is used as an alternative to the traditional worst-case critical path delay approach in this work. While our statistical critical path delay evaluation improves the functional yield by only accepting circuits having lesser deviated critical path delay due to variation, it also allows for a realistic evaluation. In this approach, we define two terms –

1. Ideal Critical Path Delay - the longest path in a circuit with all gates having 32 CNTs that are semi-conducting only, with a diameter of 1.5nm.
2. Worst-case critical path delay - the longest path in a circuit when all the gates in the circuit are replaced by the worst-case gate delays. A worst-case gate delay is calculated when only one s-CNT with a minimum diameter of 1nm is present in the channel. If the worst case delay is used as a reference limit to estimate functional yield, the performance of the accepted circuits may be degraded. We thus use a lower allowed delay degradation tolerance for functional yield estimation to obtain a reasonable functional yield with good circuit performance.

B. Functional Yield \( Y_f \) Estimation

\( Y_f \) is calculated [26] as a function of delay and allowed delay degradation from ideal case and is given as,
Tube Redundancy in Statistical Evaluation of Critical Path Delay

\[ Y_f = \frac{\text{circuits with delay} \leq \text{tolerance} \times \text{ideal case delay}}{\text{Total no. of circuits}} \] (4.3)

where, tolerance is the allowed delay degradation between 5%-30%. We use a parallel tube configuration CNFET proposed by Rehman et al., [28] for a densely packed CNT array.

### 4.3 Experiment Results

Using minimum redundancy approach for a given failure rate of 0.001%, we statistically evaluate the impact of added redundant tubes on the critical path delay and the functional yield. For a 32 tubes per transistor configuration when \( P_m = 10\% \), we also compare the increase in channel area in three cases: (1) \( 1/8^{\text{th}} \) redundancy – 4 additional tubes, (2) \( 1/4^{\text{th}} \) redundancy – 8 additional tubes, (3) pre-calculated, minimum number of tubes. From Fig. 17, it can be observed how a similar functional yield of 99\% is achieved with minimum increase in gate area using our proposed minimum tube redundancy approach. However, 13\% and 25\% area overhead should be noted with \( 1/8^{\text{th}} \) and \( 1/4^{\text{th}} \) tube redundancy in c432 benchmark circuit. This gives us an insight to achieving a good functional yield with minimized increase in the channel area.
Fig. 17: $Y_f$ achieved with minimum increase in circuit area when minimum tube redundancy is added vs a given redundancy in c432 benchmark circuit.

Considering the recent improvements in the m-CNT removal techniques and the possibility for obtaining 93% semi-conducting CNT arrays [11], we present results for 0%, 5% and 10% of metallic tubes initially present and later removed. The CNT sample size in our approach is generated based on circuit complexity. With more gates in a circuit, more instances are needed for a good statistical evaluation. We use enough instances to have a meaningful Monte Carlo distribution. 10,000 Monte-Carlo simulations are performed using the Stanford CNFET model [9] and distances algorithm (previously graphallshortestpaths algorithm) from MATLAB toolbox for our statistical critical path evaluation. With the minimum tube redundancy, our results show a reduction of 7-10% in mean critical path delay as shown in Fig. 18 and an improvement of 99.9% (Fig. 19) in $Y_f$ for c432 benchmark. The statistical critical path delay calculated relatively to the ideal case critical delay is shown in Fig. 18. We can observe that the reduced critical path delay
closeness to the ideal case when minimum redundancy is employed. In addition, also notice 15-25% improvement in the delay degradation tolerance from 1.3X to 1.05X, as shown in Fig. 19. Improvement in functional yield, reduction in statistical critical path delay for other selected ISCAS’85 logic benchmarks are given in Table 2. The worst-case delays, ideal case critical path delay and the allowed delay degradation tolerance considered for each benchmark are listed in Table1. We also record the mean critical path delay out of the 10,000 instances of a circuit and show the average reduction percentage achieved with minimum tube redundancy in Table2. Please note that when $P_m\%$ is zero, no redundant tubes are added to the CNFETs and hence no comparison of mean delays is made in this case.

![Graph showing statistical critical path delay reduction](image)

Fig. 18: Statistical critical path delay reduction achieved in c432 benchmark with and without redundancy when $P_m = 5\%$ and $10\%$. 

49
Fig. 19: Comparison of $Y_f$ in c432 benchmark for various delay degradation acceptance when tube redundancy is introduced.

Table 2: ISCAS’85 logic benchmark circuits’ worst-case delay, ideal case delay and delay degradation tolerance limits in the range of 5%-30%

<table>
<thead>
<tr>
<th>BM</th>
<th>Worst case delay(ns)</th>
<th>Ideal case delay(ns)</th>
<th>1.3X (ns)</th>
<th>1.2X (ns)</th>
<th>1.05X (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C17</td>
<td>4.32</td>
<td>0.089</td>
<td>0.117</td>
<td>0.108</td>
<td>0.094</td>
</tr>
<tr>
<td>C432</td>
<td>12.2</td>
<td>0.254</td>
<td>0.330</td>
<td>0.305</td>
<td>0.267</td>
</tr>
<tr>
<td>C880</td>
<td>26.6</td>
<td>0.553</td>
<td>0.719</td>
<td>0.664</td>
<td>0.581</td>
</tr>
<tr>
<td>C1355</td>
<td>21.6</td>
<td>0.449</td>
<td>0.584</td>
<td>0.539</td>
<td>0.471</td>
</tr>
<tr>
<td>C2670</td>
<td>33.8</td>
<td>0.703</td>
<td>0.914</td>
<td>0.844</td>
<td>0.738</td>
</tr>
</tbody>
</table>
Table 3: Improvement in Mean Critical Path Delay and Functional Yield of selected ISCAS’85 benchmark circuits with delay degradation tolerance in the range of 5%-30% when proposed minimum tube redundancy is added

<table>
<thead>
<tr>
<th>BM</th>
<th>Pm %</th>
<th>With 32 CNTs/CNFET and no redundancy</th>
<th>With proposed minimum redundancy</th>
<th>Reduction in mean critical path delay</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean delay (ns)</td>
<td>Functional Yield (%)</td>
<td>Mean delay (ns)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.3X</td>
<td>1.2X</td>
<td>1.05X</td>
</tr>
<tr>
<td>C17</td>
<td>0%</td>
<td>0.089</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.094</td>
<td>100%</td>
<td>99.2%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.101</td>
<td>100%</td>
<td>85.1%</td>
</tr>
<tr>
<td>C432</td>
<td>0%</td>
<td>0.256</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.275</td>
<td>100%</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.291</td>
<td>100%</td>
<td>86.7%</td>
</tr>
<tr>
<td>C880</td>
<td>0%</td>
<td>0.555</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.591</td>
<td>100%</td>
<td>98.6%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.623</td>
<td>100%</td>
<td>80.3%</td>
</tr>
<tr>
<td>CJ355</td>
<td>0%</td>
<td>0.453</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.488</td>
<td>100%</td>
<td>97.9%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.515</td>
<td>100%</td>
<td>73.2%</td>
</tr>
<tr>
<td>C2670</td>
<td>0%</td>
<td>0.705</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.779</td>
<td>100%</td>
<td>99.1%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.871</td>
<td>100%</td>
<td>85.3%</td>
</tr>
</tbody>
</table>

4.4 Chapter Conclusions

Variations in CNT growth process have to be considered for a more realistic delay-modelling of the CNFET-based circuits. With the proposed minimized tube redundancy, we are able to reduce the delay increase tolerance above the ideal case critical path delay.
and still have a good yield. The minimum tube redundancy facilitates the design of variation-tolerant CNFET-based circuits through more realistic evaluation of delay and $Y_f$ in the presence of variations. The statistical critical path delay approach improves accuracy of functional yield and delay evaluation. When $P_m < 1\%$, the variation in critical path delay is totally attributed by CNT diameter variation and no redundant tubes are needed. Yet, a high functional yield is achieved when only diameter variation is present. From the results presented, we can observe, how, minimum tube redundancy helps to reduce the gap between delay variation due to CNT count and the ideal critical path delay.

It is interesting to note that, when 10% metallic tubes are initially present and removed, the impact of CNT diameter and CNT count variation subsides with proposed CNT redundancy. The functional yield is improved from 0% to >95% when delay degradation tolerance is reduced to 1.05X in this case. 100% functional yield is achieved for lower percentage of metallic tubes and higher delay degradation tolerance, due to a good packing density of CNTs in the channel. This work is a step towards the process of designing reliable and variation-tolerant CNT-based logic circuits.
Chapter 5

Critical Path Only Tube redundancy for Power minimization in CNFETs under CNT diameter and number variations

This chapter is published in IEEE-TNANO, 2021.


5.1 Introduction

Despite their highly desirable performance characteristics, similar for p-type and n-type transistors [13], Carbon nanotube (CNT) based devices (Fig. 20(a)) and CNFET technology are still in the research stages as they face many fabrication challenges with CNT related variations being the dominant one [14]. Prior works proposed CNT correlation [40] and CNT redundancy (also called tube redundancy) [26][27][41] for yield enhancement under CNT variations. When two or more FETs share the same set of CNTs and therefore exhibit identical behavior, they are said to be correlated.
This CNFET correlation was shown to improve functional yield of individual gates [40] in the presence of CNT variations. However, correlation impact on circuit-level performance remains unexplored. Previously published evaluations of correlated transistors [40][23][42] were based on theoretical analysis of experimental results for large numbers of unconnected CNFETs.

In CNT redundancy technique, additional CNTs are added in the channel region of all CNFETs to maintain the required drive current when metallic tubes are removed and to avoid transistor failure. Many works [26][27][41] reported in the literature aim to improve the CNFET circuit delay and failure tolerance using CNT redundancy. Adding redundant tubes, however, increases power dissipation. Therefore, it is also important to add just enough [26] to meet a given goal while balancing the power-delay trade-off. In the previous chapter [41], we proposed to add a pre-calculated number of redundant tubes to all transistors to improve yield, referred as RAF (Redundancy in all FETs).
Table 4. Nomenclature/Values used in ROCP Approach

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value/Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>s-CNT</td>
<td>Semi-conducting CNTs</td>
<td>---</td>
</tr>
<tr>
<td>m-CNT</td>
<td>Metallic CNTs</td>
<td>---</td>
</tr>
<tr>
<td>(P_m)</td>
<td>Percentage of metallic tubes initially present</td>
<td>0-10%</td>
</tr>
<tr>
<td>(Y_D)</td>
<td>Delay-limited Yield</td>
<td>in %</td>
</tr>
<tr>
<td>(D_T)</td>
<td>Allowed delay degradation tolerance</td>
<td>5-30%</td>
</tr>
<tr>
<td>(D_{CNT})</td>
<td>CNT diameter</td>
<td>1-2 nm [10]</td>
</tr>
<tr>
<td>(N_{R_{\text{min}}})</td>
<td>Minimum number of redundant tubes required in a FET for a given failure rate</td>
<td>[26]</td>
</tr>
<tr>
<td>(P_{PC})</td>
<td>All paths possibly critical</td>
<td>----</td>
</tr>
<tr>
<td>(I_{\text{ON,FET}})</td>
<td>ON current through a single CNFET</td>
<td>varies</td>
</tr>
<tr>
<td>(T_{pd,FET})</td>
<td>CNFET delay</td>
<td>---</td>
</tr>
<tr>
<td>(T_{pd,ideal})</td>
<td>Ideal critical path delay</td>
<td>varies</td>
</tr>
<tr>
<td>(L_g, W_g)</td>
<td>Physical gate length and gate width</td>
<td>(L_g = 32 \text{ nm})</td>
</tr>
<tr>
<td>(V_{DD})</td>
<td>Supply Voltage at 32 nm technology</td>
<td>0.9 V [2]</td>
</tr>
<tr>
<td>(C_L)</td>
<td>Load capacitance</td>
<td>200 fF [10]</td>
</tr>
<tr>
<td>ROCP</td>
<td>Redundancy only in all critical paths</td>
<td>----</td>
</tr>
<tr>
<td>RAF</td>
<td>Redundancy in all FETs</td>
<td>---</td>
</tr>
<tr>
<td>NR</td>
<td>No Redundancy</td>
<td>---</td>
</tr>
</tbody>
</table>

However, if redundant tubes are added to all transistors to maintain performance after metallic tube removal, they contribute significant additional power dissipation. This emphasizes the need to limit the number of redundant tubes. As redundancy is added
mostly to limit delay degradation, in this paper we propose adding redundant tubes only to transistors on critical paths. This can reduce power dissipation, as a much smaller number of redundant tubes are added, while not degrading the delay that is mostly defined by the delay of critical paths. It is highly possible to have multiple critical paths in a circuit and thus redundant-tubes have to be added to all transistors on these paths. It is critically important however to note that, under CNT- variations, the critical path may be different for every fabricated instance of the same circuit and for each instance a new path(s) can become critical.

In this chapter, we thus propose to add the optimized number of redundant tubes to transistors only in all potential critical paths (ROCP – Redundancy only in all critical paths), identified for a given failure rate in the presence of specified variation. An efficient method is proposed to identify all paths that can become critical under CNT variations. Our approach, ROCP, minimizes the increase in power due to tube redundancy and very importantly, retains the yield enhancement achieved with all-transistor CNT redundancy [41] (RAF). Addition of redundant tubes can increase circuit area and that potential increase has to be included in the evaluation. Most of the published papers only evaluate the increase in channel area due to tube redundancy [43] and not the increase in the total circuit area. When redundant tubes are added, the overall standard cell area may also be impacted, but not necessarily.

We evaluate the proposed approach for delay, static power as well as area and yield on a set of ISCAS’85 benchmarks. We design the benchmark circuits using (i) Uncorrelated-
Critical Path Only Tube Redundancy

Chapter 5

CNFETs and (ii) Correlated-CNFTs, and for two layout styles: (i) Conventional CMOS-style layout and (ii) Staggered-style layout [44].

5.2 Chapter Contributions

- Design strategy to add redundant tubes only to transistors in paths that are statistically evaluated as critical ($P_{PC}$ - *paths possibly critical*). Due to reduction in the number of redundant tubes being added, static power dissipation is reduced. This strategy results in significantly reduced static power dissipation but the same yield and performance as compared to adding redundancy to all FETs ($RAF$). The usually allowed delay increase (tolerance) under variations, will also be reduced.

- An efficient method for identifying all potential paths that can become critical $P_{PC}$, under statistical critical path delay variation using limited runs of Monte Carlo ($MC$).

- Evaluation of two correlated-CNFT based standard cell layout styles, conventional CMOS-style and staggered, for increase in circuit area due to tube redundancy.

- In depth analysis and comparison of performance and delay limited-yield, under CNT variations, for CNFT-based circuit in two standard cell layout styles, Conventional CMOS and Staggered styles with different degrees of CNFT correlation

To the best of our knowledge, this is the first work to propose optimized CNT redundancy added only to CNFETs in all potential critical paths, $P_{PC}$, for static power
minimization under CNT variations. Our approach minimizes power without degrading yield and increase in circuit delay. Please note that all benchmark results are compared to RAF (redundancy in all FETs) approach at the path level, as other works present only results for a large number of individual CNEFTs and not actual benchmarks.

The rest of the chapter is organized as follows: Section 5.3 gives a background on CNFET delay and static power evaluation, and CNT correlation. A statistical approach for critical path analysis using Monte Carlo simulation is described in Section 5.4. Our proposed approach, ROCP (Redundancy only in all critical paths) is described in Section 5.5. A fast method to identify all possible statistical critical paths is presented in Section 5.6. Section 5.7 discusses the area overhead evaluation of two correlated-standard cell layouts. The results and major conclusions are presented in Section 5.8 and 5.9, respectively.

5.3 Background

We use Monte Carlo (MC) simulations to evaluate the power, performance and delay-limited yield of the CNFET- circuits. The MC simulates the fabrication of thousands of chips. Doing MC before fabrication of a “real” chip may be computationally intensive on more complex circuits, but still far better than making costly design mistakes.
5.3.1 Delay and Static Power evaluation

The delay and static power dissipation in a CNFET depend on its tubes’ ON currents \((I_{ON\_CNT})\). \(I_{ON\_CNT}\) depends on the threshold voltage, \(V_{th}\) and \(V_{th}\) depends on the tube diameter [45][46]. Therefore, both delay and power depend on the number of tubes and their diameters. The delay of a CNFET is calculated using (5.1) [10],

\[
T_{PD\_FET} = \frac{C_L \times 0.5 \times V_{DD}}{I_{ON\_FET}}
\]

where, \(C_L\) is the load capacitance, \(V_{DD}\) is supply voltage and \(I_{ON\_FET}\) is the summation of ON currents of all s-CNTs \((I_{ON\_CNT})\), remaining in the channel after m-CNT removal. The drive current is calculated using the circuit-compatible CNFET model presented in [45] and validated in [10] using HSPICE. It includes non-idealities associated with CNFETs.

The delay of CNFET-based circuit is defined by the critical path delay. Critical path is the longest sensitizable path under statistical delay model, in which each path has a delay distribution due to the presence of CNT variations.

The static power consumed by each FET is directly proportional to the number of s-CNTs \((n)\) in the channel \((n \times I_{ON\_CNT} \times V_{DD})\). The total static power consumption of the circuit is the sum of power dissipation in all individual logic gates in the circuit.

5.3.2 CNFET Channel Area Evaluation

The channel width \((W_g)\) of a CNFET with a CNT diameter, \(D_{CNT}\), \(n\) number of CNTs in the channel, and a CNT-pitch, \(S\), is calculated using (5.2).
In this work, we model 32 nm technology and to overcome inter-CNT screening effects, $S$ of 4 nm is used.

### 5.3.3 Metallic Tubes

To model the CNFET circuit fabrication process, we assume VLSI-compatible m-CNT removal technique [11] in which the undesired m-CNTs are etched by electrical breakdown due to self-heating when sufficient source-drain bias is applied. This process removes $\geq 99.99\%$ of m-CNTs vs. $\leq 1\%$ of s-CNTs. With its good accuracy we assume all m-CNTs and no s-CNTs are removed during the m-CNT removal process. We assume the percentage of metallic tubes ($P_m$) initially present and later removed in the range of 0-10% [11][47], considering the recent advances in CNT growth techniques.

### 5.3.4 CNFET Correlation

Transistors sharing the same set of tubes are called correlated and exhibit identical current driving capability. The correlation is achieved when the CNTs are long enough to cover more than one transistor [40]. In [40] CNTs are grown between Fe catalyst strips on a quartz substrate and later transferred to Si substrate. Transistors with individual sets of tubes are called uncorrelated, as shown for example in Fig. 21. To avoid correlation, CNTs
Critical Path Only Tube Redundancy

must be shorter to cover only one transistor. Tube length is controlled by the growth time. To model the transistor correlation, we assign the same set of CNTs to all transistors that are to be aligned in the layout of a standard cell.

![Diagram of correlated and uncorrelated CNFETs]

Fig. 21. Correlated CNFETs sharing the same set of tubes and uncorrelated CNFETs with individual sets of tubes.

All these correlated transistors are represented by the same delay in our MC simulations. For the uncorrelated transistors case each transistor will be assigned randomly a set of tubes and will have a different delay.

5.3.5 **Definitions**

A few important terms used in the paper are defined below.

i. **Ideal Critical Path Delay, \( T_{pd,ideal} \):** The critical path delay assuming all CNTs in all transistors are semiconducting and with the same diameter. Ideal critical path delay is calculated assuming CNT density of 250 CNTs/\( \mu \)m [48], with a mean diameter of 1.5 nm.
ii. *Functional Yield*: The ratio of the number of CNFETs with at least one s-CNT in the channel (functioning) to the total number of transistors.

iii. *Delay Degradation Tolerance, D_T*: \( D_T \) is the allowed increase in a circuit delay beyond the ideal critical path delay \( (T_{pd, ideal}) \). We established a limit for delay degradation which in this paper is initially set to be 1.3X. A circuit does not satisfy the delay criteria if its delay is above the tolerance limit \( D_T \). Such a circuit will be considered as non-functional in the calculation of the delay-limited yield [28].

iv. *Delay-limited Yield, \( Y_D \)*: The number of functional circuits, subject to fabrication variations and the initial presence of m-CNTs, is equal to the number of circuits with its critical path delay smaller than a given delay degradation tolerance, \( D_T \). \( Y_D \) is calculated using (5.3),

\[
Y_D = \frac{\text{No. of circuits with delay } \leq (D_T + T_{pd, ideal})}{\text{Total no. of circuits}}
\]

where, \( D_T \) is in the range of 1.05 X - 1.3 X.

### 5.4 Monte Carlo Simulations

To perform MC simulations, we start with creating a large population of CNTs within a diameter range of 1-2 nm, represented by a Gaussian distribution (\( \mu=1.5 \) nm and \( 3\sigma=0.5 \) nm). The CNT population includes semiconducting tubes and the assumed \( P_m \) between 0-10% [11] metallic tubes. Metallic tubes are randomly distributed in the generated CNT
population. For each benchmark circuit we build and test its 10,000 instances. Based on multiple evaluations we performed, 10,000 samples were sufficient to generate stable results in the investigated benchmarks using our approach. Larger circuits may require larger MC sample size. The p- and n-channel CNFETs are built by randomly choosing CNTs from the previously generated CNT population. To build logic gates using uncorrelated-CN FETs, each transistor in the gate is assigned a randomly chosen set of CNTs. For building gates using correlated-CN FETs, we randomly choose a set of CNTs and the same set of CNTs is assigned to all correlated transistors as shown in Fig. 22 for an XOR2 gate. Each instance of CNFET-based circuit is built by randomly assigning the gates, from the gate distribution, to the circuit. Next, a critical path or paths are extracted and the statistical circuit delay is evaluated [39].

5.4.1 Statistical Critical Path Delay Evaluation

In this approach, each gate in a circuit is randomly assigned a delay value from the gate delay distribution generated using diameter and the number of tubes variations. All multi-input logic gates in the circuit are decomposed to sequentially cascaded 2-input gates to enable faster evaluation. We assume that all gates are minimum size as gate sizing is not the focus of this paper. Next, the longest path (or paths) are identified based on delays assigned to gates. So, it is possible a different path is critical. For each circuit instance a distribution of circuit delays is generated and only the circuits with delay smaller than the
given delay limit are accepted as delay-limited functioning circuits. In our statistical evaluation we set the maximum acceptable delay to be equal to the ideal delay multiplied by a factor that we call delay tolerance factor. In our research we start with the tolerance factor equal to 1.3 and later show that it can be reduced with still maintaining an acceptable yield.

5.5 Methodology - Redundancy only in Critical Paths

Redundancy is often used to improve failure tolerance, but it is important to add just enough redundancy to meet a given goal while balancing the increase in power dissipation. The minimum number of required redundant CNTs \( N_{R_{\text{min}}} \) in a transistor to achieve a failure rate below 0.001\% is calculated using (5.4) [26].

\[
N_{R_{\text{min}}} = \log \left( \frac{10^{-5}}{P_m} \right)
\]  

(5.4)

The calculated \( N_{R_{\text{min}}} \) is referred to as the minimum redundancy throughout the chapter.

5.5.1 Need to identify all critical paths under variations

To reduce delay, we previously added pre-calculated minimum redundant tubes to all transistors in the circuit [41] (RAF). This reduced critical path delay by up to 10\%, but unfortunately increased the average power dissipation by nearly 8\%, in a set of ISCAS’85 benchmarks [41]. To limit the increase in power, and maintain the reduction in circuit
delay, we propose to add redundant tubes only to transistors in gates on the critical path. Redundancy in a limited set of transistors leads to lower power dissipation as compared to redundancy added to all transistors.

However, like variations in CMOS technology, CNT-specific variations cause different critical (longest delay) paths in fabrications of the same circuit. For example, let us consider two instances of a sample circuit, shown in Fig. 22 with $P_m = 10\%$. The delay of every gate instance varies due to variation in the number of s-CNTs and therefore, the value of ON currents. We represent the total number of s-CNTs present and m-CNTs removed by a (s, m) pair, respectively. There is a critical path (in red) in sample instance-1 with two gates (in green) of the same (s, m) pair, but different delays. This is due to variation in CNT diameter. In

![Sample circuit](image-url)
instance-2 of the sample circuit, an additional critical path shows up due to the presence of CNT variations. This example demonstrates that the number of critical paths and the critical path delay can vary between circuits fabricated based on the same design.

Therefore, adding redundancy during the design phase only to transistors on critical paths, identified based on design simulation using typical transistors parameters, would fail to speed up critical paths that can emerge in fabricated circuits due to variations. It is thus required to add redundancy to all potential critical paths to speed up all instances of the fabricated circuit. Only if all potential critical paths in a circuit are identified and redundancy is added to them during the design stage, can we expect with a given probability that critical paths in all fabricated circuits have delays within assumed delay tolerance. To the best of our knowledge, this is not addressed in any of the published research on statistical evaluation of CNFET-based circuits.

5.5.2 Design Methodology

To accumulate the list of all possible paths that can become critical under specified variations, $N$ number of MC simulations are first run. The flow of the proposed design methodology to generate the required number of ($N$) instances of a circuit with $ROCP$ is shown in Fig. 23. The required number of simulations, $N$, depends on the size and the
complexity of the circuit under design. The gate-level netlist is used as an input together with the previously generated distribution of gate delays under CNT variations.

We perform statistical critical path evaluation [39] and use the MAP container function in MATLAB to track every unique critical path (CP) detected. After all potential critical paths are identified over \( N \) iterations, the minimum number of redundant tubes per transistor, \( N_{\text{R,min}} \), calculated for failure rate of 0.001\%, need to be added to all critical path transistors.

![ROCP design methodology with adaptive optimization](image)

In a straight-forward approach, another \( N \) number of MC simulations are run on the circuit with this added redundancy. For all tested benchmarks, it was observed that the set of
critical paths identified with added redundancy, is the same, as the set identified with no redundancy when $P_m \leq 10\%$.

For example, in ISCAS’85 C432 benchmark circuit, there are 36 inputs and 7 outputs, for a total of 252 in-out paths (Number of ways a signal can travel from any input node to any output node). However, as shown in Table 5, only 19 of these in-out pairs are repeatedly critical in 10,000 ($N$) instances of the circuit when $P_m = 10\%$. MC simulations are performed to identify all these statistical paths and the number of their occurrences. Any path with a non-zero probability of being critical in any MC iteration is identified as one of the statistical critical paths. The probability of occurrence of a critical path, $P(C_P)$, is the ratio of the number of occurrences of a given $C_P$, to the total number of occurrences of all possible critical paths. The sum of probabilities of path occurrence for all statistical critical paths in a circuit is 1. In our approach, critical paths with less than 0.001 probability of occurrence are ignored to reduce runtime. The $P(C_P)$ of all potential critical paths with no redundancy and with minimum redundancy added are compared in Table 5. The path depth and CNT variations both contribute to a path being critical. Please observe that no new path becomes critical after the addition of the minimum redundancy. This implies that in fabricated circuits under the assumed variations and with $P_m \leq 10\%$, the set of potential critical paths without added redundancy remains the same after the redundancy is added. However, the probability of occurrence of a critical path, $P(C_P)$, can vary.

This is because the reduction in critical path delay due to added redundancy needs to be at least equal to the minimum possible gate delay in the circuit to allow paths with lower path depths to become critical. Based on our evaluation, the reduction in critical path delay
(ns) with pre-calculated redundancy ($N_{R_{min}}$), a failure rate of 0.001%, and $P_m=10\%$ does not exceed twice the minimum gate delay without redundancy. Hence, the paths with lower path depth (less than 9 for C432 in Table 5) remain non-critical. With improving CNT selective growth process, we expect $P_m$ to reduce further. Hence, it can be expected that for future technology non-critical paths remain non-critical after adding redundancy to previously identified critical paths.

It is important to note that, with the changing $P(C_P)$ of the critical in/out pairs, the total number of path occurrences with redundancy is reduced as indicated by the red arrow in Table 5. Therefore, paths become critical less often when there is minimum redundancy. It was also observed that the critical paths with high $P(C_P)$ under no redundancy remain
Critical Path Only Tube Redundancy

highly probable under added redundancy as highlighted by the red dotted lines. (A reason specific to this benchmark causes lower \( P(C_P) \) for paths of depth 10.)

5.5.3 *Need for fast critical path accumulation*

Using the straight-forward approach to generate the list of all possible critical paths and a distribution of delays, we would need to perform \( N \) number of MC simulations twice for each circuit under design. MC simulations, however, are very time intensive. From our multiple MC experiments on a set of ISCAS’85 benchmarks, we observed that while generating the distribution of the statistical critical path delay requires a large number of iterations for proper convergence, all possible statistical critical paths can be identified in much fewer iterations, as shown in Fig. 24 for the C432 benchmark. For verification, the 19 possible critical paths (Table 5) were first identified using a large number of MC simulations (red). However, if the critical path enumeration is closely monitored (black), the 19 paths can be identified in a much shorter time (less than 40 iterations for this benchmark). Therefore, we use convergence monitoring as a stopping criteria in MC for
generating a MAP of all possible critical in/out pairs. This approach will help identify all the potential critical paths in far fewer iterations.

![Graph showing critical path count](image)

**Fig. 24** Statistical critical path count in C432 benchmark obtained using MC simulations

### 5.6 Adaptive Monte Carlo for Fast Critical Path Identification

We start by identifying all potential critical paths under specified variations with no redundancy. To speed up this process, we check the cumulative critical path count after each iteration. When the path count does not increase over a dynamically-determined number of iterations, we assume all likely critical paths have been found.

Specifically, if the previous iteration added new critical paths but the current iteration did not, we store the current iteration number in the Discard_itr ($D_{itr}$) parameter. We set parameter $M_T$ (total number of monitoring iterations) to $10 \times D_{itr}$. If the critical path count remains unchanged after a further $M_T$ number of iterations, we consider the path counts to be converged and halt the monitoring. Otherwise, $D_{itr}$ is reset and $M_T$ is recalculated. The number of iterations required for the critical path count to converge is called the accumulation iterations, $A_T$. 

71
We then simulate the required $N$ instances of the circuit with added redundancy (ROCP) for all critical paths identified during the first $A_T$ iterations (indicated with the dark arrows in Fig. 23). The total number of MC iterations in our approach can thus be represented as: $A_T + M_T + N$. For example, using a straight-forward approach on the C432 benchmark, 20,000 MC iterations (1452.3s) would be required: 10K (707.5s) to identify all potential critical paths and another 10K (744.8s) to generate circuit instances with added redundancy for delay evaluation. However, using our approach, we will only be performing 10,407 (794.1s) iterations (37 accumulation iterations ($A_T$) to collect the 19 statistical critical paths, 370 monitoring iterations ($M_T$) and 10,000 iterations ($N$) to generate the distribution. These additional runs ($A_T + M_T$) for identifying potential critical paths are dynamically added hence termed adaptive optimization.

The critical paths accumulated using this approach are in good agreement with the critical path data identified using a larger number of MC runs.

5.7 CNFET Correlation

In this section we consider the impact of correlation between CNFETs on area at the gate-level, and yield. The directional growth of CNTs was shown to enable CNFET correlation and improvement in functional yield of 26.5 X [40] over uncorrelated CNFETs when evaluated on a large number of transistors. In this research, for the first time, performance and yield of benchmark circuits, not just a large number of individual transistors, are compared for uncorrelated and correlated CNFETs. For comparison we use our proposed fast ROCP approach on two layout styles. In CNFET correlation, we assume
that all CNTs are of equal length and long enough to completely cover all the transistors within a correlated group [49]. We analyze conventional CMOS-type layout shown in Fig. 25(a) and a staggered layout [44] shown in Fig. 25(b). Both layouts represent an XOR2 gate drawn using the Cadence Virtuso tool. The layouts are considered only for area evaluation. As concluded in [44] and confirmed in our research, the staggered layout style seems better suited for CNFET specific imperfection-immune standard cell design.

Fig. 25. (a) XOR2 gate in conventional CMOS-type layout with PU and PD networks separately correlated (b) XOR2 gate in staggered layout style with PU and PD transistors correlated as a group.
(i) **CMOS-type Correlated-layout:** In this layout style, Pull-up (PU) and Pull-down (PD) CNT stacks are aligned vertically as shown in Fig. 25(a). All the p-CNFETs share the same set of CNTs (set-1). Similarly, all n-CNFETs share a set of tubes (set-2) but not the same set as the p-type group.

(ii) **Correlated-Staggered layout:** In the staggered layout topology, PU and PD are aligned horizontally as shown in Fig. 25(b). The intra-cell routing is modified as required to incur a minimum increase in area. All p-CNFETs and n-CNFETs share the same set of CNTs.

### 5.7.1 Validating Yield Enhancement in Correlated-gates

According to Zhang et al.[40], theoretical expression of CNFET failure probability ($P_F$), the functional yield of CNFETs is improved with increasing transistor width (increased number of CNTs) and with an increased number of correlated CNFETs. In this work, we compare the failure probability of logic gates using uncorrelated and two correlated styles. As an example, we present here results for $P_F$ of all logic gates occurring in 10,000 instances of C432 benchmark with all transistors having a channel width equal to 40 nm and with $P_m = 10\%$. Based on multiple evaluations we performed, $P_F$ is zero for wider CNFETs (>50 nm) when $P_m=10\%$. Therefore, 40 nm transistor width is chosen to present the impact of correlation on a non-zero failure probability. We assume uniform CNT density and a CNT diameter of 2 nm, and the results are presented in Table 6.
Table 6: Reduction in Gate Failure Probability in C432 Circuit due to Increasing Transistor Correlation when $P_m=10\%$

<table>
<thead>
<tr>
<th>$W_g$</th>
<th>$P_F$ (gates) Uncorrelated</th>
<th>$P_F$ (gates) Correlated-CMOS type</th>
<th>$P_F$ (gates) Correlated-Staggered</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 nm</td>
<td>$2.6 \times 10^{-6}$</td>
<td>$1.5 \times 10^{-6}$</td>
<td>$1.09 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

Please remember that the number of transistors in a correlated set in the staggered layout is twice the number of transistors in a correlated set in CMOS-style. Having that in mind we can observe that the failure probability is decreasing with the increasing number of transistors in the correlated set as shown in Table 6. These results, however, represent only the functional yield, not delay-limited yield, for a large number of individual, unconnected gates with transistors organized in correlated groups to model two different layout styles.

### 5.7.2 Gate area with added redundancy

The two layout styles are evaluated to show that adding redundancy can be implemented at no additional area cost using either of them. When redundant tubes are added, the increase in channel area does not necessarily result in an increase of the standard cell area. The number of tracks (T) is the conventional metric of a standard cell height. The
track height is decided based on height of the largest used standard cell. XOR2 is the largest gate occurring in all benchmarks investigated in this paper. We use 32 nm technology node [2] and the two layout topologies are compared in terms of number of tracks (T). For simplicity and quick evaluation, we are using the MOSIS scalable CMOS submicron rules [50].

Typically, in CMOS, two sets of standard cells are designed, one with 12 tracks for high performance and another with 9 tracks (~9T) for high density. Since, we are not focusing on gate sizing, we use the minimum and equal sized p and n-CNFETs for a high-density design. The height of the standard cell is set based on the number of tracks, T. A track is the sum of the minimum width and spacing of a metal line M1 or M2. We use a M1 pitch of 112 nm (7λ) from Intel’s 32 nm technology design rules [51], in terms of λ (1T = 7λ). For this area analysis we assume all CNTs are 2 nm in diameter.

For an XOR2 standard cell, 5.3T height (37.75λ = 11.75λ channel + height of other layout objects + spacing between routing layers) is needed using the staggered layout topology, which is rounded off to 6T for ease of fabrication. The CMOS-type topology requires 6.7T height (47.5λ) rounded off to 7T. When the pre-calculated minimum redundancy, $N_{R_{min}}$, is added, the channel height increases by +1.12λ. This results in 5.5T for staggered topology and 6.9T for CMOS-type layout. It is important to note that 5.5T and 6.9T is within the 6T and 7T limit, respectively. Hence, the added minimum redundant tubes will not impact the overall standard cell area for both considered layout types. However, if a larger number of tubes are needed for gate sizing, the staggered layout topology is a better choice because, when an additional track is needed, two tracks will be
required in CMOS layout style. One for PU expansion and another for PD expansion. This will result in additional cell height of $14\lambda$. The staggered layout on the other hand, will use only one added track ($7\lambda$) that can be shared between the PU and PD stacks. Since, the overall increase in total area is the same for both the layout topologies with added redundancy, the staggered style is therefore better for area in CNFET circuit design with minimized area from modified intra-cell routing [44].

5.8 Experiment Results

Our design methodology is programmed in MATLAB using distances (previously graphallshortestpaths) function, available in the MATLAB toolbox to find the critical paths in a circuit. 10,000 MC simulations are used to test a set of ISCAS’85 logic benchmark circuits. They include all primitive gate types: BUF, INV, AND, NAND, OR, NOR and XOR. All experiments were performed on a single core Intel (Broadwell, IBRS) CPU at 2.2 GHz with 24 GB RAM.

5.8.1 Redundancy only in critical paths vs. in all FETs

Fig. 26 shows the results of MC simulations of the critical path delay distributions and the delay-limited yield. Fig 26(a) shows the critical path delay distribution in C432
benchmark circuit for uncorrelated CNEFTs with no tube redundancy. Fig 26(b) for RAF and Fig 26(c) for ROCP. With redundancy added only to transistors in the critical paths (ROCP) we want to verify that the circuit performance enhancement already achieved with the redundancy added to all transistors (RAF) is not compromised. The delay distribution with no redundancy (Fig. 26(a)) is completely contained between 1.1X and 1.3X delay degradation tolerance limits. If we add the pre-calculated optimized number of redundant tubes to all circuit transistors, the critical path delay decreases as shown in Fig 26(b) for RAF. The shift to the left shows reduction in the mean critical path delay and, please notice, 99.6% delay-limited yield with only 5% (1.05X) delay tolerance. When the same optimized number of redundant tubes is added but only to CNFETs in all critical paths, as shown in Fig. 26(c), the delay distribution stays within the same delay range as RAF from Fig. 26(b). Note the 99.5% delay-limited yield achieved with 5% (1.05X) delay degradation tolerance using ROCP, which is almost the same as for RAF. It is also interesting to note the decreasing standard deviation, \( \sigma \) (STD), of the delay distribution in ROCP as compared to RAF.
Fig. 26. Statistical critical path delay distribution for C432 when $P_m=10\%$.  
(a) No redundancy  
(b) RAF  
(c) ROCP  

Therefore, a greater number of circuits have critical path delay closer to the mean critical path delay as compared to all CNFETs with redundancy, RAF. Table IV gives the results for the circuit performance enhancement and delay-limited yield for other ISCAS ’85 benchmark circuits comparing our ROCP approach against RAF. Please observe that ROCP retains the critical path delay reduction and delay-limited yield improvement.
achieved with RAF, showing improved variation tolerance. Our limited redundancy approach can thus be used to reduce power with improved performance.

**5.8.2 Power Savings with Critical Path Redundancy**

Fig. 27 shows the total circuit static power vs. critical path delay of C432 benchmark when $P_m=10\%$ for ROCP and RAF. In RAF (yellow), all the FETs have the same initial number of tubes and therefore circuits have more similar power dissipation and hence smaller spread. Please observe that the lower and upper delay limits of the distribution (x-axis) are similar for the ROCP and RAF cases. However, ROCP results in an average power reduction of 7.3% when compared to RAF. Therefore, ROCP can reduce the power dissipation in circuits without degrading the circuit performance. The average increase in circuit static power when CNT redundancy is added to transistors (for both cases of RAF and ROCP) for other ISCAS’85 benchmarks is shown in Fig. 28 with $P_m =$ (a) 5% and (b) 10%. It is compared to the no-redundancy case (represented by the red line). The average power savings (% difference between the two peaks) obtained with ROCP compared to RAF is shown on the secondary Y-axis in Fig 28. In C2670 benchmark, the average total power for ROCP is almost equal to the reference case power without redundancy. This is because for $P_m \leq 10\%$ that circuit has only one critical path. Therefore, the increase in power due to redundancy is minimal since gates on only one path have redundant tubes. Please observe an average reduction of 7.3% in the total power for ROCP as compared to RAF with $P_m=10\%$. 
Fig. 27. Total circuit static power vs. delay in C432 benchmark for RAF and ROCP.
Fig. 28. Power savings obtained with ROCP compared to RAF. Power minimization of (a) up to 6.7% when P_m = 5% and (b) up to 8.3% when P_m = 10% P_m = 10%.

This is an important observation: we improved the critical path delay at no or minimum expense in power.

To capture the combined impact of our ROCP strategy on circuit delay and power, Table 7 shows the Power Delay Product (PDP) in a set of ISCAS’85 benchmarks. The PDP with ROCP for P_m=10%, is reduced by an average of 6.52%, as compared to the RAF approach.
Table 7: Power-Delay Product of ROCP vs. RAF

<table>
<thead>
<tr>
<th>ISCAS'85 Benchmark</th>
<th>$P_m$%</th>
<th>PDP (nJ)</th>
<th>% Reduction in PDP using ROCP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RAF</td>
<td>Proposed ROCP</td>
</tr>
<tr>
<td>C17</td>
<td>5%</td>
<td>0.0043</td>
<td>0.0042</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.0044</td>
<td>0.0042</td>
</tr>
<tr>
<td>C432</td>
<td>5%</td>
<td>0.4719</td>
<td>0.4510</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.4706</td>
<td>0.4376</td>
</tr>
<tr>
<td>C880</td>
<td>5%</td>
<td>2.2085</td>
<td>2.0836</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>2.1968</td>
<td>2.0300</td>
</tr>
<tr>
<td>C1355</td>
<td>5%</td>
<td>2.6355</td>
<td>2.5230</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>2.6206</td>
<td>2.4539</td>
</tr>
<tr>
<td>C2670</td>
<td>5%</td>
<td>9.2617</td>
<td>8.6588</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>9.0805</td>
<td>8.3830</td>
</tr>
</tbody>
</table>

Average savings in PDP

\[
P_m = 5\% \quad 4.83\%
\]

\[
P_m = 10\% \quad 6.52\%
\]
### Table 8: Critical Path Delay and Delay-Limited Yield in ISCAS '85 Benchmarks with Declining Tolerance Limit, $D_T (1.3X - 1.05X)$

<table>
<thead>
<tr>
<th>BM</th>
<th>$P_n$ (%)</th>
<th>Y_D (%)</th>
<th>Y_D (%)</th>
<th>Y_D (%)</th>
<th>Mean-CP Delay (ns)</th>
<th>Mean-CP Delay (ns)</th>
<th>Mean-CP Delay (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>RAF</td>
<td>ROCP</td>
<td>1.15 1.05</td>
<td>1.3X 1.05X</td>
<td>1.15 1.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.3X X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delay (%)</td>
<td>Delay (%)</td>
<td>Delay (%)</td>
<td>Delay (%)</td>
<td>Delay (%)</td>
<td>Delay (%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(ns)</td>
<td>(ns)</td>
<td>(ns)</td>
<td>(ns)</td>
<td>(ns)</td>
<td>(ns)</td>
</tr>
<tr>
<td>C17</td>
<td>5%</td>
<td>0.094</td>
<td>100</td>
<td>99.2</td>
<td>78.</td>
<td>0.09 100</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.101</td>
<td>100</td>
<td>85.1</td>
<td>0</td>
<td>0.09 100</td>
<td>100 100</td>
</tr>
<tr>
<td>C432</td>
<td>5%</td>
<td>0.275</td>
<td>100</td>
<td>99</td>
<td>43</td>
<td>0.25 100</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.291</td>
<td>100</td>
<td>86.7</td>
<td>0</td>
<td>0.26 100</td>
<td>100 100</td>
</tr>
<tr>
<td>C880</td>
<td>5%</td>
<td>0.591</td>
<td>100</td>
<td>98.6</td>
<td>60</td>
<td>0.55 100</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.623</td>
<td>100</td>
<td>80.3</td>
<td>0</td>
<td>0.56 100</td>
<td>100 100</td>
</tr>
<tr>
<td>C1355</td>
<td>5%</td>
<td>0.488</td>
<td>100</td>
<td>97.9</td>
<td>57</td>
<td>0.45 100</td>
<td>100 100</td>
</tr>
<tr>
<td>C2670</td>
<td>5%</td>
<td>0.779</td>
<td>100</td>
<td>99.1</td>
<td>49</td>
<td>0.70 100</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.871</td>
<td>100</td>
<td>85.3</td>
<td>0</td>
<td>0.71 100</td>
<td>100 100</td>
</tr>
</tbody>
</table>

### 5.8.3 Delay-limited yield: Uncorrelated v. Correlated CNFETs

The critical path delay improves with added redundancy as previously shown in Fig. 27. Increasing correlation among unconnected CNFETs also improves the functional yield [40]. Our next simulations investigate delay-limited yield without redundancy using...
uncorrelated and correlated CNFETs at circuit level. Fig. 29 shows the statistical critical path delay distributions of C432 circuit with $P_m=10\%$ and no redundancy using (a) uncorrelated (b) CMOS-type correlated, and (c) Staggered correlated style topologies. Uncorrelated CNFETs show the highest delay-limited yield of 97.1\% under a delay tolerance of 1.15X (Fig. 29(a)). It must be observed that with increasing correlation among CNFETs, the delay-limited yield and critical path delay variation are degraded. This is evident from the increased standard deviation in correlated CMOS-type layout (Fig.29(b)-yellow) and even higher increase in Staggered-correlated style (Fig. 29(c)-green). In cases when there are many metallic tubes (that had to be removed) in the set of tubes used to correlate a group of CNFETs, the performance of all CNFETs in the group will be degraded. However, in case of uncorrelated transistors, if a set of tubes in one transistor had many removed m-CNTs, only this specific CNFET has degraded performance, while other CNFETs will not necessarily exhibit degradation. The probability of all uncorrelated transistors having initially a large and similar number of m-CNTs is very low. Therefore, circuits with uncorrelated CNFETs have a narrower delay distribution and a higher number of instances with mean delay value. Also, please note that the number of all statistically potential critical paths increases with increasing correlation. Our fast adaptive optimization technique successfully identifies all new paths dynamically (#MC trials).

We also compare the average delay-limited yield among two layout styles with two different levels of transistor correlation and with added redundancy for a delay tolerance of 1.05 X - 1.1 X. The results presented in Table 9 are an average of all the investigated ISCAS’85 benchmarks. In Fig. 29, the delay-limited yield ($Y_D$) degrades with increasing
degree of correlation (p-type and n-type groups in CMOS-style and all p-type and n-type transistors in one group for Staggered) among transistors at the circuit level. In Table 9, we show that our proposed ROCP approach can improve the $Y_{D\%}$ to $>98\%$ with only $10\%$ increase in allowed delay tolerance ($D_T = 1.1 \times$) as compared to circuits having no redundant tubes, even under increased correlation among CNFETs. In Fig. 30 we show the combined impact of redundancy and correlation in CNFETs on average critical path delay variation ($\sigma/\mu$). Please note that the average critical path delay variation in investigated benchmarks increases by an average of $1.8 \times$ with increased correlation in Fig 30.

Table 9: Average Delay-limited Yield compared by Correlation Style In Investigated ISCAS ’85 circuits ($P_M=10\%$)

<table>
<thead>
<tr>
<th></th>
<th>Delay-limited Yield ($Y_D%$) by Correlation style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Uncorrelated</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_T$</td>
<td>1.1X</td>
</tr>
<tr>
<td>NR</td>
<td>12%</td>
</tr>
<tr>
<td>ROCP</td>
<td>100%</td>
</tr>
</tbody>
</table>

86
However, with added redundancy the variation in critical path delay reduces as compared to no redundancy, and the difference between RAF (red) and ROCP (green) is almost negligible in both layout styles. The improvement in delay variation achieved with RAF is retained using our ROCP approach (similar $\sigma/\mu$) using uncorrelated and both correlated layout styles.

Fig. 29. Statistical critical path delay distribution using (a) Uncorrelated (b) CMOS-type Correlated, and (c) Correlated-staggered style CNFETs in C432.
Critical Path Only Tube Redundancy

Chapter 5

Fig. 30. Average critical path delay variation in ISCAS’85 benchmarks with and without redundancy compared by increasing correlation with $P_m=10\%$

5.9 Chapter Conclusions

This chapter proposes to add redundant tubes only to CNFETs in all potential critical paths ($ROCP$). It reduces the increase in the static power dissipation without degrading the delay-limited yield as compared to all transistor redundancy ($RAF$). The $ROCP$ approach improves timing and reduces the increase in power due to redundancy to at most 2%. This is achieved with less than 0.2% (average) change in delay-limited yield as compared to $RAF$ approach and no increase in circuit area in all tested ISCAS’85 benchmarks.

With our fast and efficient methodology of identifying critical paths emerging under CNT variations, we are able to capture up to 99.99% of all possible critical paths that can
exist in all circuits fabricated based on the same design. On the other hand, our results show that correlation in CNFETs increases variation in critical path delay because of shared performance degradation among CNFETs in correlated groups. Therefore, though the functional yield is improved, the delay-limited yield and circuit performance with CNFET correlation is degraded. Hence, we suggest using uncorrelated CNFETs in the design of logic gates to enhance delay-limited yield. The savings achieved in PDP with ROCP, are promising for the design of CNFET based circuits when high yield, high performance, and low power consumption are expected.
Chapter 6

CNT Length Variation

This chapter is published in IEEE-NANO, 2020.


6.1 Introduction

The influence of the two major variations, tube diameter variation and variation of the number of tubes in a gate channel, has been studied extensively considering uncorrelated and correlated transistors [25][52][23][53][28][35]. When two or more FETs share the same CNTs and exhibit identical behavior, they are said to be correlated in the direction of CNT growth. The CNFET correlation was shown to improve functional yield [53][40] in the presence of diameter and tube number variations.

Previously published evaluations of correlated transistors [40][42] were based on theoretical analysis and experimental results for large numbers of individual CNFETs. They assumed that all tubes in each correlated group of CNFETs had the same length, long enough to cover all correlated transistors. Wang et. al [42], proposed a miss-alignment immune layout design for CNFETs using tube correlation. Zhang et. al [40] and Lin et al., [52] analyzed unconnected transistors, perfectly correlated within the given fixed length of tubes.
Fig. 31. (a) *Ideal (Case(i))*: All CNTs are equally long and cover channels of all the CNFETs in the direction of CNT growth (b) *Actual (Case ii)*: CNFETs with varying CNT lengths in the direction of CNT growth

To perform chip level analysis, Lin et al., [52] analyzed 1M transistors. In this chapter we present Monte Carlo Simulations (MC) run on a large number of unconnected transistors and on various ISCAS benchmarks. Our results agree with the theoretical expressions in [52] and [53] showing improvements in functional yield with the increased number of perfectly correlated CNFETs. Assuming equally long CNTs means that there is no variation in the length of the tubes. However, not all CNTs can be grown equally long and therefore CNFETs might not be perfectly correlated due to variation in tube length as shown in Fig.31 (b).

In this work, we also use MC Simulations to investigate the influence of variation in CNT length on performance of a group of 1M CNT based transistors to compare with
previously published results [40]. To evaluate gate delays we consider the variation in the “ON” current, $I_{ON}$, of a CNFET. In previous chapters [39][41] $I_{ON}$ variations are considered to be caused by CNT diameter variation and CNT count variation. In this chapter, however, we also include the variation in tube length. When the presence of CNT length variation is included its impact depends on the number of correlated transistors with respect to the distribution of the tube length. If the number of correlated transistors is equal to or smaller than the number of CNFETs that can fit within the shortest length of tubes, there will be no length variation and the functional yield will be improved. If the mean value of the tube length is used to fit correlated CNFETs the tube length variation may reduce yield and influence performance. The CNT length variation influences, in a notable percentage, the total gate delay and reduces the improved functional yield in CNFETs with correlated CNTs.

The rest of the chapter is organized as follows. Section 6.2 discusses the CNT correlation and variation in CNT diameter and CNT counts. The influence of CNT length variation on correlated CNFETs is described in Section 6.3 along with four choices of CNT lengths analyzed. The results are presented and discussed in Section 6.4 with concluding remarks in Section 6.5.

6.2 Background

6.2.1. CNT Diameter and Count Variations

The variation in “ON” current, $I_{ON}$, of a CNFET caused by CNT diameter and CNT counts due to the presence of m-CNTs, $P_{m\%}$ is discussed first. Even a small percentage
of metallic tubes present has to be removed for proper CNFET operation. The diameter variation is considered in the range of 1nm-2nm, represented by a Gaussian distribution of \( \mu = 1.5\)nm and \( 3\sigma = 0.5\)nm [10]. An initial presence of 10\% m-CNTs (\( P_m \)) that are removed using the VMR process is considered [11]. We assume that all m-CNTs are completely removed during the m-CNT removal process and for simplicity we assume that no semiconducting CNTs are inadvertently removed. In reality a very small percentage of semiconducting tubes might be removed but the percentage is so small that can be neglected. These CNT variations establish a base case for functional yield analysis. Transistor correlation is considered next for tubes that are long enough to cover more than one CNFET.

### 6.2.2. Carbon Nanotube Correlation

Patil et al., [54] proposed the wafer scale growth of aligned single-walled carbon nanotubes. In this growth process, highly aligned CNTs are grown horizontally on a quartz substrate between Fe catalyst strips. The length of the CNTs essentially depends on the spacing between the catalyst strips. Zhang et al., [40] took advantage of this directional CNT growth
Fig. 32 CNFET 1 and CNFET 2 are correlated in direction of CNT growth(x). CNFET 1 and CNFET 3 are uncorrelated in the direction perpendicular to CNT growth(y).

They and proposed aligned active layout styles and showed a functional yield improvement of 26.5X compared to uncorrelated CNT growth. The CNFETs aligned along the direction of CNT growth are considered correlated and CNFETs perpendicular to the length of the tubes are considered uncorrelated as shown in Fig. 32. The key idea in CNT correlation is to take advantage of the correlation in both CNT count [53] and CNT type (m-CNT or s-CNT) [52].

Under the assumed CNT variations, a CNFET will fail if all the CNTs under the gate are metallic and need to be removed. In circuits fabricated with uncorrelated CNFETs, the probability of failure of each CNFET is independent since all the transistors are independent of each other. However, in correlated CNFETs with aligned CNT growth, if
one CNFET fails, the entire row fails, as the same CNTs are shared among all the FETs aligned in the direction of CNT growth. Therefore, the probability of a failing row is the same as the probability of a single transistor failure in that row [40]. The functional yield calculated [40] using uncorrelated CNFETs is given as,

\[ \text{Yield} = 1 - \sum_{i=1}^{N} P_F \]  

(6.1)

where, \( N \) is the total number of independent CNFETs and \( P_F \) is the failure probability of CNFFET. The circuit level yield calculated using correlated CNFETs [40] is given as,

\[ \text{Yield} = 1 - K_R P_{FR} \]  

(6.2)

where, \( P_{FR} \) is the row failure probability of \( N \) transistors distributed in \( K_R \) rows. Correlation in CNFETs can thus help reduce the probability of failure in CNFET-based circuits by \( N/K_R \) times. Table 10 shows gate failure probability using MC simulations for uncorrelated and correlated gates for C432 ISCAS’85 benchmark.

<table>
<thead>
<tr>
<th></th>
<th>Uncorrelated CNFETs</th>
<th>Correlated CNFETs</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_F )</td>
<td>( 1.93 \times 10^{-6} )</td>
<td>( 3.87 \times 10^{-7} )</td>
</tr>
</tbody>
</table>
6.3 Methodology

6.3.1 CNT Length Variation

Aligned-CNTs grown under controlled conditions can reach long lengths \([55][56]\) up to a micrometer \([57][58]\) on average. Long CNTs can be used to fabricate correlated CNFETs to improve functional yield as discussed in the previous section. The length of all CNTs must be the same for all correlated transistors to have the same current driving capability. This is an ideal case shown in Fig.31(a). Zhang et al., \([40]\), assume a fixed CNT length of 200\(\mu\)m.

Though CNTs reach long lengths, equally long CNTs cannot be grown on the substrate due to early catalyst precipitation and varying CNT diameter \([55]\). Smaller diameter CNTs are typically longer than the larger diameter CNTs. When the presence of CNT length variation is considered, correlated CNFETs may not necessarily improve the circuit performance and functional yield. Correlation coefficient which is a value between -1 and 1 is used by Zhang et al., \([53]\) to show the relation between length of (x : direction of CNT growth) CNTs and correlation in CNFETs. They presented a correlation coefficient of 0.9 in CNT count for a CNT length up to 6 microns based on experimentally extracted data. Therefore, we performed our experiments assuming an average CNT length of 6\(\mu\)m. The CNT length distribution is represented by \(\mu= 6\mu\)m and \(\sigma = 1.3\mu\)m. We then analyze the
Impact of CNT length variation on functional yield and total delay using four cases mentioned later in this section.

We assume the 32nm technology node. Based on the assumed CNT length a given number of CNFETs, \( N \), are distributed in \( R \) number of rows. The maximum number of FETs, \( F_r \), that can be accommodated in a single row, \( r \), in a standard cell design depends on the length of the CNT, \( L_{\text{CNT}} \), the size of the transistor (3F) and the spacing between FETs, \( S_{\text{min,FET}} \) and is given as,

\[
F_r = \frac{L_{\text{CNT}}}{(3F+S_{\text{min,FET}})}
\]  

(6.3)

where, \( F \) represents the minimum feature size and is equal to 32nm in our experiments.

### 6.3.2 Cases Analyzed

**Case (i) Ideal:** All CNTs are equally long and their length is equal to an average CNT length of 6µm as shown in Fig 34(a). A CNFET in this case will fail only if all channel CNTs are metallic and will need to be removed.

**Case (ii) CNT length variation:** CNT length varies within the row according to the tube length distribution as shown in Fig 34(b). The width of the row is equal to the mean CNT length of 6µm. Failure of CNFETs in this case may be due to (a) all channel tubes being metallic, (b) CNFETs, at the end of the row, with no CNTs left due to short CNT lengths, (c) the combination of both.
To reduce the number of CNFETs failing due to CNT length variation, the width of the row (horizontal) can be chosen in many ways and two possibilities are analyzed here.

*Case (iii-a) Min. CNT Length:* The row width is equal to the shortest CNT length from the CNT length distribution, as shown in Fig 33(a) and Fig 34(c). The number of rows $R$, in this case where there is no tube length variation, required to place $N$ transistors is larger than $R$ used in the first two cases.

Fig. 33. (a) Minimum CNT length from the CNT length distribution is chosen to decide the maximum number of FETs in a row and improve reliability. (b) The CNT length chosen is equal to $\mu - 3\sigma$ of the CNT length distribution. FETs with degraded performance may be formed in this case.
Case(iii-b) (µ-3σ) Length: The row width is chosen equal to the CNT length reached by 90% of the grown CNTs (µ-3σ). This increases the number of CNFETs correlated in a row compared to case(iii-a). The number of FETs that can fit this row width is greater than in case(iii-a), but few FETs will have compromised performance due to possible missing of tubes in the far right transistors as shown in Fig 33(b) and Fig 34(d).

The width of source and drain and the length of a channel, (S/G/D) together is given as 3F, where F represents the minimum feature size. We also assume $S_{min,FET}$ equal to 3F. Therefore, the total width assumed to be occupied by a cell with PU and PD pair along with spacing between them is 0.12µm. In Case(i), the $F_r$ is 50 FETs in a row with the CNT length of 6µm as shown in Fig 34(a). In Case(ii) also, $F_r$ is 50, but allowing variation in CNT length within a row with the mean CNT length of 6µm as shown in Fig 34(b).

In Case(iii-a), $F_r$ is decided based on the minimum CNT length from the CNT length distribution. The number of rows required for distributing $N$ transistors in this case is greater than any other case, as all tubes are assumed to be of the minimum length. In Case(iii-b) shown in Fig 34(d), the number of FETs in a row is decided based on a µ-3σ value of the tube length distribution.
Experiment Results

In MC simulations we conducted, 32-nm technology node is considered with supply voltage of 0.9V as projected by ITRS guidelines [2]. The analysis is done on a sample size of 1M transistors, \( N = 1M \). A CNFET channel width of 40 nm is considered with uniformly
spaced CNTs and the optimum CNT-CNT spacing of 4nm [38]. 1000 MC simulations are run to generate the total delay distribution. Results are presented as delay variation($\sigma/\mu$) of 1M transistors for length variation only (Case(ii)) and together with CNT count and diameter variations (Total) as shown in Fig. 35. Please observe that CNT length variation alone contributes about 30.2% to the total delay variation in the cases we analyzed. Therefore, for correlated transistors it is crucial to consider the variation in the tube length or limit correlation to the number of transistors that can all be covered with tubes of the minimum length.

Variation in tube length also impacts the functional yield improvement achieved with perfectly correlated CNFETs. The functional yield is calculated using the number of failing rows [40]. The probabilities of row failures are given as $P_{FR}$ in Table 11. Variation in the CNT count is probably the most important factor in the decrease of the functional yield of

![Fig. 35. Variation ($\sigma/\mu$) in gate delay due to individual CNT length variation(Case(ii)), and together with $P_m\% =10\%$ and CNT diameter variation(Case(i)). CNT length variation contributes nearly 30\% to ON current, $I_{ON}$ variation in CNFETs.](image)

101
CNT Length Variation

Chapter 6

CNFET-based circuits. The variation in tube length further degrades the functional yield as compared to constant average CNT length.

In our MC simulation we use the same number of transistors \( N = 1 \, \text{M} \) for all four cases. Therefore, the number of transistors per row (or correlated transistors sharing the same set of tubes.) is different in considered cases. Case (i) and Case (ii) have the same number of transistors in each row (50), and in both Cases (iii-a) and (iii-b) the number of transistors depends on the chosen tube length. Cases (i) and (iii-a) have constant tube length but the probabilities of row failures are different. This is because the number of transistors per row is lower in Case (iii-a) and therefore failure probability is lower due to increased number of rows. Comparing Cases (ii) and (iii-b) with tube length variations, it can be seen that row failure probability is higher when the length variation is larger. As shown in Table 11, the smallest probability of row failure, \( P_{FR} \), can be achieved when there is no tube length variation.

The larger the tube variation, higher the \( P_{FR} \), which is Case (ii) with highest \( P_{FR} \) and then case (iii-b) for \((\mu-3\sigma)\) CNT length. The \( P_{FR} \) for Case(iii-a) is smaller than for Case(i) because the number of transistors in a row is smaller. In Case(iii-b), the \( P_{FR} \) is increased by CNFETs failing due to no CNTs left in the channel after m-CNT removal as well as possible CNFETs with no CNTs due to short tubes. (shorter than the \((\mu-3\sigma)\) length). It must be noted that though \((\mu-3\sigma)\) length accommodates larger number of correlated FETs in a row as compared to (case(iii-a)), \( P_{FR} \) for \((\mu-3\sigma)\) length Case (iii-b) is higher than case (iii-a) because of tube length variation.
Table 11. Probability of Failure ($P_{FR}$) of Correlated CNFETs under various Cases of CNT Length Variation

<table>
<thead>
<tr>
<th>Case #</th>
<th>Row Failure Probability (%) $P_{FR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Ideal</td>
<td>$5.0 \times 10^{-5}$</td>
</tr>
<tr>
<td>(ii) CNT length variation</td>
<td>$8.4 \times 10^{-1}$</td>
</tr>
<tr>
<td>(iii-a) Min. CNT length</td>
<td>$1.1 \times 10^{-6}$</td>
</tr>
<tr>
<td>(iii-b) ($\mu-3\sigma$) Length</td>
<td>$4.8 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Table 12 shows results for delay variation for Case(ii), Case(iii-a), and Case(iii-b). As could be expected case (iii-a) with the shortest tube length (the smallest number transistors in a correlation set) shows the least variation ($\sigma/\mu$) in total delay and a low failure rate due to no tube length variation. Though a length of ($\mu-3\sigma$) is attained by 90% of CNTs, in Case (iii-b), few of CNFETs with a reduced number of tubes due to the length variation degrade performance.

Table 12. Impact of CNT Length Selection on Variation of Total Delay in 1M Transistors for 1000 MC Simulations with $P_M\% = 10\%$

<table>
<thead>
<tr>
<th>CNT Length Selected</th>
<th>$\mu_{\text{delay}}$ (s)</th>
<th>$\sigma_{\text{delay}}$ (s)</th>
<th>$\sigma/\mu$ (delay)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Ideal Case</td>
<td>8.8E-17</td>
<td>8.7E-20</td>
<td>9.88E-4</td>
</tr>
<tr>
<td>(ii) CNT length variation</td>
<td>1.09E-16</td>
<td>1.64E-19</td>
<td>1.49E-3</td>
</tr>
<tr>
<td>(iii-a) Min. CNT length</td>
<td>1.76E-18</td>
<td>1.96E-22</td>
<td>1.13E-4</td>
</tr>
<tr>
<td>(iii-b) ($\mu-3\sigma$) Length</td>
<td>9.58E-17</td>
<td>2.17E-20</td>
<td>2.26E-4</td>
</tr>
</tbody>
</table>
Hence, from Table 12, the total delay variation \((\sigma/\mu)\) when CNT length is considered (Case(ii)) is degraded by 1.51X when compared to the ideal case (Case(i)) with all CNTs of same length and equal to the mean CNT length. Case(iii-b) is degraded by 2X when compared to Case(iii-a) due to CNFETs with compromised performance towards the end of the row. The total delay variation \((\sigma/\mu)\) when compared to Case(ii) is improved by 13.4X in Case(iii-a) and by 6.5X in Case(iii-b). Therefore, the choices of smallest CNT length Case (iii-a) and \((\mu-3\sigma)\) length Case (iii-b) are both better for performance than Case (ii) with longer tubes but more length variation.

6.5 Chapter Conclusions

While correlation in CNFETs is extensively explored in the literature for improving functional yield, variation in CNT length is mostly ignored. When no tube length variation is considered, long CNTs or smaller S/D/G are both beneficial for functional yield enhancement. With scaling of VLSI technology, \(F_r\) increases. This is due to scaling S/D regions of the transistor and CNT length remaining unaffected with technology. However, in the presence of CNT length variations, using the rows of longer CNTs can degrade the yield and delay. This is due to early terminating CNTs that do not reach the length required for the number of correlated transistors. Assuming all CNFETs within a given CNT length to be perfectly correlated is highly optimistic. The non-negligible percentage contribution
by CNT length variation to the variation of ON current of CNFETs emphasize the need to include the tube length variations when modelling CNFET-based circuits behavior. It will allow for a more realistic estimation of the CNFET based circuit’s functional yield and performance. The reduction in row failure probability, $P_{FR}$, of CNFET based circuits under CNT length variation truly depends on the CNT length chosen. Despite the increased number of correlated FETs in the direction of CNT growth $P_{FR}$ increases due to the presence of shorter CNTs that do not cover the channels of all the FETs in a row. Our results provide insights to the possible uncertainty in circuit performance improvement with correlated CNFETs and the need for considering CNT length variation as an important source of $ON$ current variation when correlated CNFETs are explored.
PART A.2: Significance of work and Major Conclusions

CNT-specific variations and their effects on a CNFET based design have become a major problem at advanced nodes. Therefore, CNT Variations are important for realistic delay-modelling. The typical VLSI design flow is efficient, but should not always be adopted as is. Traditional CMOS techniques, may not work out with a different technology, our methodology with all the key contributions listed below is tailored specifically for CNFET based circuit design.

The major contributions in the first part of this thesis include:

1. Incorporating CNT-specific variation effects into design stage
2. Improved prediction accuracy of functional yield and critical path delay
3. Speedup all potential critical paths
4. All critical paths have delay within assumed delay tolerance
5. Power minimization without degrading the functional yield at path level
6. CNFET circuits with high yield, high performance and at low static power
7. Evaluate benefits of CNFET correlation on performance yield
8. Fast design optimization

Variation aware design decisions are necessary at every step of the design process. Therefore, our contributions in this field are likely to have a lasting effect on the design of robust CNFET based integrated circuits under CNT specific variations.
Part B. CAD approaches for fast thermal goodness evaluation and thermal management in 3D ICs
B.1 Introduction

In addition to the classical downscaling of devices and CNFET having the potential to become the channel material of future nanoscale transistors, a recent trend acknowledged in current and future electronic devices is the need to tightly integrate functionally and technologically diverse modules in a single chip. In this context, TSV-based 3D ICs possess the unique advantage over 2D ICs through the potential of “More than Moore” [59]. Shorter interconnects enabled by through-silicon vias (TSV) in three dimensional integrated circuits (3D ICs) has led to increased performance and reduced power dissipation [60], along with ease of heterogeneous integration [61]. In 3D ICs, stacking of multiple dies replaces long global interconnects with shorter vertical interconnects. The shortened global wires result in lower wire delay, therefore improving the chip performance and potentially lower the power [62] compared to their 2D counterparts. In conventional 2D circuits, a large number of wire routed over-the-block are required for inter-block connections. Thus, more high metal layers (or global metal layers) are necessary to complete inter-block routing. On the other hand, wires in the 3D ICs are connected to TSVs, which significantly cuts down the cost of over-the-block wiring [63].

From the perspective of IC design, the quality of a 3D floorplan is defined by performance metrics, i.e., area, wirelength, interconnect delay and dynamic power. It is extremely important that these performance metrics are estimated as early and as accurately
as possible, as the quality of the generated floorplan sets the foundation for better chip performance at all subsequent stages of design.

The close stacking of power-dissipating dies can result in higher power densities causing the 3D designs to dissipate more heat than their planar counterparts, thereby increasing the total chip temperature. Yet, with no practical, effective methods available for cooling the chip, efficient heat removal from 3D ICs remains a problem today. Consequently, optimizing the area and wire length of the chip alone may not result in the best performance. Also, buffers contribute non-negligible RC delay, which inevitably impacts the net delay. Most of the optimization techniques consider the device-level thermal profile but neglect the interconnect dependence on temperature in performance evaluation. In designing for thermal-aware floorplans, evaluating the interconnect performance as a function of temperature is equally important to considering temperature distribution due to power dissipation in circuits blocks.

Otherwise, 3D optimization will result in erroneous design choices, failure to achieve convergence and increased cost due to design iterations. Hence, there is an immense need for thermal-delay aware floorplanning tools for a realistic 3D interconnect performance evaluation that can be used to build high performance digital systems as well as shorten design time. This circumstance defines the overall motivation for the remainder of thesis work.
B.2 3D floorplanning & Problem Formulation

**3D Floorplanning Problem Formulation** - Let $B = \{b_1, b_2, \ldots, b_m\}$ be a set of ‘m’ rectangular blocks whose width, height and area are denoted by $w_i, h_i$ and $a_i$, $1 \leq i \leq m$. Each module is free to rotate. Let $N$ represent the netlist, connecting pins located at the center of each block in $B$. Let $(x_i, y_i)$ denote the coordinate of the bottom-left corner of each block $b_i$ on the chip. A floorplan $F$ is an assignment of $(x_i, y_i)$ for each $b_i$, $1 \leq i \leq m$ to ‘L’ device layers such that there is no module overlap. The goal of 3D floorplanning is to optimize a predefined cost function such as a linear weighted combination of chip area (minimum bounding rectangle of $F$ considering all layers), wire length which is the sum of interconnect lengths, interconnect delay and in this work, a measure of thermal goodness of the floorplan.

**3D Floorplanning** - The proposed methodologies in this thesis are integrated on top of the non-deterministic 3D floorplanning tool [64] encompassing TSVs and modules co-placement. This tool was built upon the 3D floorplanner in [65], which is an extension of the 2D floorplanning software developed by Wang et. al, [66]. In this 3D floorplanner, nets are assigned to TSV islands within the optimization stage. Careful buffer injection is completed at each iteration step before interconnect delay is estimated. TSVs are included in the wire delay estimation for buffer insertion. The generated final four-tier floorplans are used for thermal analysis using Hotspot tool [1]. A custom-built whitespace generator is used to bridge between or 3D floorplanner and Hotspot tool. The input to the whitespace
Part B: CAD Approaches for fast thermal goodness evaluation in 3D ICs

generator is floorplan descriptions (without whitespace information in layout) from 3D floorplanner and output is the floorplan description with the co-ordinates of all whitespaces appended to the floorplan description. This whitespace description is necessary to evaluate a floorplan in Hotspot tool because Hotspot adds up all bottom most modules x-coordinates and all left most modules y-coordinates to generate the height and width of the chip. Hotspot does not have a provision to explicitly input the floorplan width and height. Therefore, the whitespace description is necessary.

B.3 General Assumptions

We assumed 45 nm technology node and module power density in the range of $10^5$ - $10^7$ W/m$^2$ [22]. As the floorplanning benchmarks do not include module power density values, random power density values from the assumed range were assigned to all modules in all benchmarks. Once assigned, power density values of all modules remained the same for all experiments. Since module activity information is not available either, one can assume that average power activity is included in the power density values.

Based on the technology requirements and recommendations of the International Roadmap of Semiconductors (ITRS) [2], the current version of the 3D floorplanning tool used in this work has the following assumptions:
Part B: CAD Approaches for fast thermal goodness evaluation in 3D ICs

1. The device layers are stacked using face-to-back (F2B) die-stacking strategy, because it is the most commonly used configuration, and also doesn’t limit the number of device layers that can be stacked.

2. Only signal TSVs are assumed for our experiments. The impact of power-ground TSVs and thermal TSVs is beyond the scope of the thesis.

3. Cu-based via-middle TSVs are used because of their established process and superior performance.

4. The maximum number of stacked device layers assumed in our experiments is four.

5. The stacking of more than four device layers may counter the benefit of wire reduction due to the increased silicon surface area of TSVs in the design [67].

6. During assignment of nets to TSVs, it is assumed that a 3D net spanning between two consecutive layers needs allocation of a single TSV on the upper layer only.

7. A common requirement in modern ASIC designs is fixed-outline of chip. Keeping this in view, a fixed-outline constraint is applied during our floorplanning, i.e., all modules (circuit blocks and TSV clusters) are packed within the fixed-outline region with an aspect ratio of close to one. The circuit blocks are hard modules with fixed area and aspect ratio. A maximum allowed whitespace of 5% is assumed.

Some important terminologies specific to our 3D floorplanning tool are –
Part B: CAD Approaches for fast thermal goodness evaluation in 3D ICs

- Optimum chip area \((chip\_optiarea)\) – summation of area of blocks and TSV clusters.
- Fixed area \((fix\_area)\) – summation of area of blocks, TSV clusters and the white space area.
- Packing area \((chip\_packarea)\) – Represents the evaluated floorplan area. It is calculated by the \(chip\_w*chip\_h*nlayer\), where \(chip\_w\) and \(chip\_h\) represents the width and height of the minimum rectangle that encloses the blocks and TSVs. The \(chip\_w\) and \(chip\_h\) is the maximum value of all the device layers represented as \(nlayer\). The packing efficiency of a floorplan is represented as the ratio of \(chip\_optiarea\) and \(chip\_pack\) area.

B.4 Benchmark Specification

The original and modified GSRC benchmark circuits used in our experiments are shown in Table B1. GSRC benchmarks are commonly used circuits in existing works and are ideal for experimentation, as compared to other older benchmarks (ami and MCNC). The block area expansion and modification in number of nets for each benchmark is reported. The modification of benchmarks was necessary for comparison with existing works and perform a meaningful buffer insertion. Keeping in mind that the current growth of VLSI circuit sizes and complexity, the benchmarks have been modified by expanding each block.
Part B: CAD Approaches for fast thermal goodness evaluation in 3D ICs

area by 100x and adding more multi-pin nets to the original netlists. The general flow of the 3D floorplanner and overview of the proposed methodologies in shown in Fig. B2.

Table B1: Specifications of GSRC benchmarks used in our experiments

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Total Block Area</th>
<th>Total #of nets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original (mm²)</td>
<td>Expanded (mm²)</td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>Expanded</td>
</tr>
<tr>
<td>N100</td>
<td>0.176</td>
<td>17.64</td>
</tr>
<tr>
<td>N200</td>
<td>0.187</td>
<td>18.67</td>
</tr>
<tr>
<td>N300</td>
<td>0.273</td>
<td>27.32</td>
</tr>
</tbody>
</table>

Fig. B2. General flow diagram of EA-based 3D floorplanning as used in this work and proposed evaluation tuning as highlighted.
Chapter 7

Buffered Interconnect Power and Performance with Temperature profile

This chapter is published in IEEE-ISCAS, 2018.


7.1 Introduction

This chapter is a first-hand analysis on the impact of ignoring temperature in interconnect performance evaluation.

In the recent years, an extensive work has been done in the thermal aware floor-planning and thermal via insertion [68][22][69] to address thermal management and heat removal in 3D ICs. Most research works concentrate on power generated within the blocks and the device leakage power [70][16] and do not consider the impact of temperature on interconnects.
Yuchun Ma et al., [16] analyzed the dependence of leakage power with primary emphasis only on the device level leakage power of blocks. A methodology to inject buffers in 3D interconnects and carefully estimate the delay and power of buffered interconnects was proposed by [71]. In [17][18], the authors discuss the importance of accurate buffer count estimation in evaluating total interconnect power consumption. However, their predictions of interconnect delay in 3D ICs no longer holds accurate with a vertical temperature profile, as usually the temperature is different on each die. Though unit wire resistance changes slightly with temperature, the buffer performance degrades drastically with temperature. Hence this chapter concentrates on the back-end of line performance and aims at analyzing the impact of temperature on the buffer count and interconnect delay dynamic and leakage.
In this chapter all the analysis is done on four-tier final floorplans of GSRC benchmarks. All the floorplans are generated with the non-deterministic floorplanning tool described in Section B.2 TSVs are included in the wirelength estimation for buffer insertion but ignored for the performance evaluation in this chapter. In this work, we focus on evaluating interconnect delay and power on each device layer independently. Therefore, in this chapter, especially in the buffer insertion algorithm, we used, TSVs contribution could be omitted. Interconnects on each device layer are analyzed separately at its respective temperature. We do not distribute the wires to different metal layers and hence reduced to be analyzed as a 2D problem as shown in Figure 37. However, the temperature profile will definitely be in a range when actual routing is done and also depends on the benchmark. We assume the capacitance and thermal conductivity of the metal are constant with temperature.
Table 13: 45nm Technology parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value / Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device Technology</td>
<td>45nm</td>
</tr>
<tr>
<td>$R_0$</td>
<td>Unit length wire resistance M4-M6</td>
</tr>
<tr>
<td>$C_0$</td>
<td>Unit length wire capacitance M4-M6</td>
</tr>
<tr>
<td>$R$</td>
<td>Buffer output resistance (8x)</td>
</tr>
<tr>
<td>$C$</td>
<td>Buffer input capacitance (8x)</td>
</tr>
<tr>
<td>$p_{inv}$</td>
<td>Parasitic Capacitance Factor</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Switching activity</td>
</tr>
<tr>
<td>$f$</td>
<td>Frequency (GHz)</td>
</tr>
<tr>
<td>$V_{DD}$</td>
<td>Supply Voltage</td>
</tr>
<tr>
<td>$E_0$</td>
<td>Permittivity of free space (F/m)</td>
</tr>
<tr>
<td>$E_r$</td>
<td>Relative Permittivity of HfO2</td>
</tr>
<tr>
<td>$E_{r}$</td>
<td>Relative Permittivity of SiO2</td>
</tr>
<tr>
<td>$n$</td>
<td>NFACTOR or sub-threshold swing factor</td>
</tr>
<tr>
<td>$t_{ox}$</td>
<td>Gate oxide thickness (nm)</td>
</tr>
<tr>
<td>$T$</td>
<td>Temperature of the device layer</td>
</tr>
</tbody>
</table>

To keep the analysis simple an 8x size buffer [72, p. 8] is considered for buffer insertion and no buffer sizing is done. The GSRC benchmarks do not provide information of the power dissipation in blocks which is necessary to generate a temperature profile. Hence, a typical vertical temperature profiles of a 4 tier 3D ICs, shown in Figure 36 and taken from [70][73] is considered (profile1) for this initial analysis. A second profile in the same temperature range but with slightly higher and lower temperatures on device layers 3 and 2 respectively is also considered (profile2). Heat sink is assumed to be placed at the bottom of the 3D IC. These profiles are used to analyze, if considering the vertical temperature
profile is important in early design stages. We assume the temperature is uniform across a given tier. We analyzed 45nm technology that is the least technology for which all the models and parameters were available to us. A similar behavior is expected at the lower technology nodes due to scaling. All the technology parameters used and listed in Table 13 are taken from the NCSU technology model file [74][75].

7.3 Background

7.3.1 Interconnect Delay

The 3D interconnect is a span in inter-blocks and between different tiers. The layer-wise wirelength distribution considered includes the 2-D wire segments of 3D nets realized with TSVs and 2D wires. The number within the layer of buffers on a device layer also depends on the TSV position [64].

Fig. 37: TSV enable 3D interconnect spanning multiple dies. The wirelength will be reduced to 2D on each die ignoring the TSVs.

The delay of a wire segment is given as,
Buffered Interconnect Power and Performance with Temperature profile

\[ \text{t}_{\text{pd, wire segment}} = \frac{l^2}{2} \cdot R_0 \cdot C_0 \]  

(7.1)

where, \( t_{\text{pd, wire segment}} \) represents the delay of a buffered 3D interconnect segment. ‘\( l \)’ represents the length of the wire segment i.e., the buffer insertion length at room temperature. Buffer insertion length (BIL), is given by,

\[ BIL = \sqrt{\frac{2RC(1+p_{\text{inv}})}{R_0C_0}} \]  

(7.2)

and is calculated to be 250 µm at room temperature [18]. The total delay of 3D interconnect, \( t_{\text{pd, net}} \) on a given device layer can be estimated from the summation of total wire delay and total buffer delay of a net post buffer insertion. The total delay on a device layer \( t_{\text{pd, total}} \) is the summation of the delay of all the nets. With a vertical temperature profile, the buffer insertion length will be different on each device layer.

7.3.2 Interconnect Power

The total interconnect power consumption is given by three main components; wire power, buffer power and TSV power. The TSV power consumption is ignored as it primarily contributes to the vertical power profile. The total interconnect power dependence on temperature will be dictated by the buffer count and the buffer leakage power, which is discussed in the later sections.
7.4 Interconnect Delay Dependence on Temperature

We use Eq. 7.3 to calculate wire resistance at various temperatures.

\[ R_t(T) = R_0(1+\beta(T-T_0)) \]  \hspace{1cm} (7.3)

\( R_0 \) is the wire resistance at room temperature, \( T_0 \), (27 \( ^\circ \)C), \( \beta \) is the temperature coefficient of the metal. The change in wire resistance will influence the BIL which must be calculated for temperatures specific to each device layer, as given in the temperature profile. The unit wire resistance, buffer propagation delays and buffer insertion lengths for given temperatures are shown in Table 14. Buffer propagation delays at all temperatures in early design are simulated using Cadence Virtuoso at 45nm technology. These BIL values are used to evaluate the number of buffers on a given device layer at a given temperature. Please note that at higher temperatures, the BIL decreases due the increase in resistivity of wire only.

Table 14: Evaluated values at given temperatures

<table>
<thead>
<tr>
<th></th>
<th>Temperature (C)</th>
<th>( R_t ) ( \Omega/\mu\text{m} )</th>
<th>8x Buffer delay (ps)</th>
<th>BIL(T) (( \mu\text{m} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL1</td>
<td>42</td>
<td>0.466</td>
<td>79</td>
<td>272.54</td>
</tr>
<tr>
<td>DL2</td>
<td>65</td>
<td>0.505</td>
<td>82.1</td>
<td>261.97</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>0.522</td>
<td>83.5</td>
<td>257.34</td>
</tr>
<tr>
<td>DL3</td>
<td>108</td>
<td>0.579</td>
<td>87.3</td>
<td>244.43</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.598</td>
<td>88.5</td>
<td>240.47</td>
</tr>
<tr>
<td>DL4</td>
<td>160</td>
<td>0.668</td>
<td>92.0</td>
<td>227.53</td>
</tr>
<tr>
<td>Room</td>
<td>27</td>
<td>0.44</td>
<td>74</td>
<td>280.4</td>
</tr>
</tbody>
</table>
Only 8x buffer is used throughout our early design exploration analysis and a final routing will surely include buffer sizing.

### 7.5 Interconnect Power Dependence on Temperature

#### Interconnect Dynamic Power

The dynamic power dissipated in a wire, per unit length is given in Eqn. 7.4 and, in buffer by Eqn. 7.5,

\[
P_{\text{wire}} \text{ (per unit length)} = \alpha * C_0 * V_{dd}^2 * f \tag{7.4}
\]

\[
P_{\text{buffer}} \text{ (per buffer)} = \alpha * C_{\text{buffer\_total}} * V_{dd}^2 * f \tag{7.5}
\]

Where, \( C_0 \) is wire capacitance per unit length and the total buffer capacitance for a unit buffer, \( C_{\text{buffer\_total}} \), is composed of the buffer input capacitance, unit-length wire capacitance and the load capacitance. The maximum load capacitance, \( C_{\text{load}} \), is equal 484fF for 8X buffer [72]. We choose a nominal value for the load capacitance assuming the buffer drives an average fan-out and is taken as 240fF. Since, the wiring power is calculated separately, to avoid including it twice, the wire capacitance is not considered in the buffer power calculation. The dynamic power consumption depends on the switching activity and the number of buffers. The buffer insertion and the leakage depend on device layer temperature.
Interconnect Leakage power

The leakage power in buffers constitutes of multiple components of which the gate leakage and the sub-threshold leakage are increasingly predominant with scaling. However, the rate of gate leakage increase is controlled with the High-k dielectrics and metal gates in place for nano-CMOS. On the other hand, for technology nodes 45nm and below, the sub-threshold leakage is a strong function of temperature. Hence, only the sub-threshold leakage is considered in this research. The total leakage power contributed by all buffers on a single device layer is evaluated. Leakage power is a product of leakage current and supply voltage. The sub-threshold leakage is given as [76],

\[ I_{\text{sub}} = \mu_0 \cdot C_{\text{ox}} \cdot \left( \frac{W}{L} \right) V_t^2 \cdot e^{\frac{V_{gs}-V_{t,h}}{n \cdot V_t}} \cdot e^{1.8 \cdot \frac{e}{V_t}} \]  

(7.6)

Where, \( \mu_0 \) is the zero-bias mobility, \( n \) is a process dependent term called NFACTOR/sub-threshold swing factor, \( V_t \) is the thermal voltage, \( C_{\text{ox}} \) is the gate oxide capacitance, \( (W/L) \) is the width to length ratio of the MOS device. The leakage power computed for a single 8X buffer at different device layer temperatures is given in Table 15. It must be observed that for every ~1.5X increase in the temperature there is ~ 1.2X - 1.6X increase in leakage power.
Table 15: Leakage power of an 8X buffer with increasing temperature

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>Leakage power of an 8X buffer (T) μW</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>78.1</td>
</tr>
<tr>
<td>42</td>
<td>96.1</td>
</tr>
<tr>
<td>65</td>
<td>128</td>
</tr>
<tr>
<td>75</td>
<td>145</td>
</tr>
<tr>
<td>108</td>
<td>206</td>
</tr>
<tr>
<td>120</td>
<td>231</td>
</tr>
<tr>
<td>160</td>
<td>331</td>
</tr>
</tbody>
</table>

7.6 Experiment Results

The floorplanner tool and evaluation algorithm were performed on a 4xDual Core Sun SPARC IV at 1.35 GHz and total 32 GB RAM. All the graphical plots represent results for n200 GSRC benchmark. In our analysis, it is important to remember that the two buffers of the same size on two different device layers will not have the same propagation delay. This is due to varying temperature through device layers as shown in Fig. 36. It was also observed that in shorter interconnects with wirelength close to the BIL, segmenting the wire and injecting a buffer at a higher, than room, temperatures will increase the interconnect delay rather than decreasing it. For n200 benchmark, in wire length range 250-400μm that increase in delay was by 8X-10X for the temperature range shown in Fig. 36. It is evident from these observations that buffer delay increasingly dominates the wire delay for shorter interconnects at higher temperatures. From our observations, specifically for the
n200 GSRC benchmark a wire should not be segmented if,

\[
\frac{\text{Wirelength}}{BIL} < 1.5
\]

This limitation might be specific to benchmark, technology node and temperature profile and must be evaluated accordingly. From Table 15, please note the underestimation of delay and the number of buffers at room temperature compared to the values evaluated with a vertical temperature profile. For n200 GSRC benchmark on device layer 4 (highlighted in red in Table 16), an underestimation of 55% in interconnect delay and 31% in number of buffers is observed. Figure 38 compares interconnect dynamic power consumption in wires and buffers (total) in n200 benchmark at room temperature and at temperatures specific to a given device layer, f(T). Please notice that wire dynamic power changes only by ~1% because the wire capacitance is assumed constant with temperature. However, the buffer dynamic power consumption increases from 3.4% to 22% when compared to the buffer dynamic power consumption at room temperature. It is due to the increase in the number of buffers. Figure 39 shows comparison between total (individual total bars) interconnect power (dynamic + leakage) and buffer leakage (lower bars) contribution for n200 GSRC benchmarks at room temperature and temperatures specific to a given device layer. For every 10C increase/decrease in temperature (profile2), total interconnect delay and dynamic power change by 2% - 2.5%. This is due to the difference in buffer count that varies between 1.48%-2% for different temperature profiles within the same temperature range. Observe that the leakage power increases by 1.2X to 1.5X from device layer 1 to 4.
Fig. 38: Wire dynamic power compared with total dynamic power consumption for n200 GSRC benchmark. Minimal increase in wire dynamic power and major contribution by buffer dynamic power to be observed.

Fig. 39: Comparison of buffer leakage and total (dynamic + leakage) interconnect power consumption in n200 GSRC benchmark at room temperature and with vertical temperature profiles 1 and 2.
The results for n100 and n300 GSRC benchmarks given in Table 17 show a similar trend of huge underestimation of buffer leakage power, up to 5.4X in n300. That underestimation of 1.5X–5.4X is not acceptable as it undermines the quality of generated optimized layout solutions.

Table 16: Total delay of buffered interconnect for n100, n200 and n300 benchmark circuits
Table 17: Interconnect dynamic power ($P_{\text{wire\_dyn}}$, $P_{\text{buf\_dyn}}$), leakage power ($P_{\text{buf\_leakage}}$) and the total interconnect power ($P_{\text{total}}$) for n100 and n300

<table>
<thead>
<tr>
<th>Device layers</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{wire_dyn}}$</td>
<td>9.15</td>
<td>18.31</td>
<td>13.14</td>
<td>5.61</td>
</tr>
<tr>
<td></td>
<td>9.18</td>
<td>18.34</td>
<td>13.09</td>
<td>5.60</td>
</tr>
<tr>
<td>$P_{\text{buf_dyn}}$</td>
<td>33.61</td>
<td>74.22</td>
<td>51.59</td>
<td>22.27</td>
</tr>
<tr>
<td></td>
<td>35.11</td>
<td>82.71</td>
<td>62.03</td>
<td>28.82</td>
</tr>
<tr>
<td>$P_{\text{buf_leakage}}$</td>
<td>52.56</td>
<td>116.05</td>
<td>80.68</td>
<td>34.83</td>
</tr>
<tr>
<td></td>
<td>67.56</td>
<td>240.12</td>
<td>255.85</td>
<td>190.99</td>
</tr>
<tr>
<td>$P_{\text{total}}$</td>
<td>95.32</td>
<td>208.58</td>
<td>145.41</td>
<td>62.72</td>
</tr>
<tr>
<td></td>
<td>111.84</td>
<td>341.16</td>
<td>330.97</td>
<td>225.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Device layers</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{wire_dyn}}$</td>
<td>53.98</td>
<td>103.12</td>
<td>69.84</td>
<td>29.42</td>
</tr>
<tr>
<td></td>
<td>53.97</td>
<td>102.88</td>
<td>69.69</td>
<td>29.19</td>
</tr>
<tr>
<td>$P_{\text{buf_dyn}}$</td>
<td>216.11</td>
<td>446.10</td>
<td>290.72</td>
<td>290.72</td>
</tr>
<tr>
<td></td>
<td>224.00</td>
<td>491.74</td>
<td>342.11</td>
<td>123.31</td>
</tr>
<tr>
<td>$P_{\text{buf_leakage}}$</td>
<td>337.94</td>
<td>697.59</td>
<td>454.62</td>
<td>1411.1</td>
</tr>
<tr>
<td></td>
<td>431.01</td>
<td>1427.6</td>
<td>192.83</td>
<td>1038</td>
</tr>
<tr>
<td>$P_{\text{total}}$</td>
<td>608.02</td>
<td>1246.8</td>
<td>815.18</td>
<td>1223.8</td>
</tr>
</tbody>
</table>

7.7 Chapter Conclusions

In this chapter, we studied the variation of interconnect performance and power consumption in 3D ICs with temperature. We focused on accuracy of predictions when temperature of device layers is not properly included in parameters evaluations. Deriving from the analysis, the dependence of buffer insertion length and buffer count on temperature must be considered carefully when evaluating the interconnect performance.
The buffer insertion length will vary with temperature on each device layer thereby influencing the number of buffers. The increase in interconnect delay with temperature is primarily due to the increase in the number of buffers due to wire resistance change with temperature. The percentage error in delay estimation and in buffer count in 3D interconnects due to using room temperature parameters can negatively influence quality of generated 3D floorplans. The dynamic power of wire doesn’t change severely as the capacitance is constant with temperature. We also observed buffer leakage domination which constitutes more than 50% (comparing gray and orange bars in Fig. 39) of the total interconnect power consumption when temperature dependent parameters are used in evaluation. This contribution arises from increase in both buffer leakage as well as the buffer count. Our research shows that significant underestimation of interconnect delay and power consumption which happens when actual device layer temperature is not used can lead to incorrect values used in a cost function that govern the process of 3D IC floorplan optimization. It can result in severely sub-optimal 3D layout solutions being generated. To prevent this from happening, when evaluating interconnect performance and power to assess quality of 3D layout solutions, temperature must be considered.
Chapter 8

**Fast thermal goodness evaluation of a 3D-IC Floorplan**

This chapter is published in IEEE-ISCAS, 2018.


8.1 Introduction

Three-dimensional integrated circuits (3D ICs) have enabled vertical integration for better packing density and reduced footprint. Their enhanced chip performance, due to the short vertical connections known as through-silicon-vias (TSVs), make them a promising technology for overcoming the challenges of planar technology [77]. Figure 40 shows a cross-section of a four layer TSV-based 3D IC with high vertical integration density. Despite their capability for overcoming the interconnect bottleneck in 2D ICs, 3D ICs still face a major challenge of effective heat removal from the densely packed device layers as discussed in the previous chapter. There are several works in the literature addressing thermal optimization in 3D ICs by adding thermal TSVs (TTSVs) [78], using liquid cooling [79], redistributing white space around high power dissipating blocks [5], using a carbon-based thermal interface material (TIM) [80], and early thermal-aware design optimization to lower chip temperature [22].
Physical design techniques like floorplanning use non-deterministic algorithms that require thousands of iterations to reach a desired stage of optimization. The increased solution space due to vertical adjacency of device layers already adds to the computational time. To co-optimize the chip temperature together with other objectives, there is an immense need for fast thermal evaluation of the generated solutions. Most state-of-the-art works use finite element analysis [81] or compact resistive networks [82] for obtaining peak temperature. Finite element analysis aims for accuracy while the compact resistive network strikes a balance between speed and accuracy. Run time of a thermal-driven floorplanning algorithm is still much longer than the non-thermal-driven version due to the millions of thermal evaluations performed during the floorplanning process.

Fig. 40: 4-layer TSV-based 3D IC showing layer thicknesses and vertical temperature distribution.
8.2 Chapter Contributions

As discussed in Chapter 2, several works proposed fast thermal analysis of 3D ICs using various techniques. However, all these techniques impose a huge run time penalty or encompass shortcomings that could limit the solution’s thermal quality due to their low correlation with the more accurate simulation based tools like Hotspot [1].

In this chapter, we propose using a simple power-based measure called thermal goodness value (TGV) to evaluate 3D floorplans for their thermal goodness with much shorter runtime than previous approaches. The unique strength of the proposed algorithm is that it can quickly evaluate the generated solutions for their thermal goodness without computing the chip temperatures. The technique involves a layout to \( nxnxl \ (l = \text{number of device layers}) \) grid translation to incorporate the impact of all adjacent blocks.

TGV is a power-based metric and takes into account the power distribution of the entire 3D chip as compared to one peak temperature used in typical thermal-driven optimization. Therefore, our technique can efficiently eliminate thermally unfavorable solutions. Our results also show a good correlation of our TGV measure to peak temperatures generated with the Hotspot tool [1] as many published works use peak temperature to guide the floorplan for thermal optimization [22][83]. With a high runtime speedup, our algorithm is thus a promising alternative to guide thermal-aware 3D floorplanners without incurring long computational times.
8.3 Background

In this chapter, all the evaluations were done on four-layer final floorplans of GSRC benchmarks. We evaluated at least 100 floorplans of each benchmark. These floorplans were generated by a non-deterministic floorplanning tool using evolutionary computation described in Section B.2. Therefore, it is safe to assume that through thousands of iterations many scenarios that can lead to worst-case thermal hotspots due to switching block positions in X-Y as well as Z direction have been considered.

Random power density values in the range of $10^5 - 10^7 \text{ W/m}^2$ are assigned to the modules at 45nm technology node. Once assigned, power density values of all modules remained the same for all experiments. The generated final floorplans are then used to evaluate peak temperature using the well-known Hotspot [1] tool and the proposed thermal goodness value. For all our analysis in this chapter, we assume that the heat sink is on the top of the 4-layer 3D IC stack. It is important to note that the proposed floorplan thermal evaluation technique aims to efficiently prune thermally unfavorable 3D solutions rather than generating accurate temperature distributions.

Correlation coefficient

We used the correlation coefficient as a measure to compare our 3D floorplan thermal goodness ($\text{W/m}^2$) evaluation with the Hotspot tool’s peak temperature (K). Correlation
Coefficient \((r)\) is a unit less metric with a value between -1 and 1. It measures the strength and direction of relationship between two variables “\(X\)” and “\(Y\)” and is given by,

\[
r = \frac{1}{k-1} \left[ \frac{\sum_{i=1}^{k} (X_i - \bar{X})(Y_i - \bar{Y})}{s_x s_y} \right] \tag{8.1}
\]

\(-1 \leq r \leq 1\)

In Eq. 8.1, \(k\) represents the number of samples and \(\bar{X}, \bar{Y}\) are the mean and \(S_x, S_y\) are the standard deviation of the sample data of variable \(X\) and \(Y\), respectively. The stronger the two variables are correlated, the closer the value \((r)\) is to 1. If the two variables are inversely correlated, \(r\) is closer to -1. An \(r\) value of 0 implies no correlation exists between the two variables. The relation between two variables with any unit of measurement can be compared using the correlation coefficient \((r)\).

### 8.4 Fast thermal goodness evaluation

The proposed thermal goodness evaluation technique for 3D circuits aims to distinguish which of two given floorplans has a better thermal-aware module placement using a simple power-based measure that can be calculated in a very short time. The flow of the algorithm is visualized in Figure 41. For clarity, the algorithm is described in two parts. In section 3.1, the translation of all \(l\) device layers of a 3D layout to an \(n \times n \times l\) power grid is described and the process is depicted in Figure 41.
The 3D power grid evaluation for TGV is based on including the impact of neighbor grid cells as described later in section 3.2. The neighbor cells include intra-layer cells that are on the top, bottom, left, and right of the grid cell under test (GUT) and the inter-layer grid cell beneath the layer under consideration. The cells diagonal to GUT and cells in the layer above GUT are ignored as their heat exchange with GUT is negligible.
8.4.1 Floorplan to grid translation

The 4-layer 3D IC layout is translated into four \( n \times n \) grids where ‘\( n \)’ represents the number of rows and columns of the grid as shown in Figure 43(a). For ease of computation, \( n \) is a power of 2. All evaluations are therefore performed on 64x64, 128x128, and 256x256 grids. Since 64x64 grid yields a good correlation coefficient compared to Hotspot [1], we do not use a grid beyond 256 due to memory constraints. The 3D layer under consideration is superimposed on the \( n \times n \) grid for calculating the overlap area. Each step in the layout to grid translation is described below.

**doOverlap():** Each module is first checked for overlaps with the underlying grid cell for computational speedup. Let \((L_g.x, L_g.y), (R_g.x, R_g.y)\) and \((L_m.x, L_m.y), (R_m.x, R_m.y)\) be the lower left and upper right coordinates of grid cell \(g\) and module \(m\), respectively, as shown in Figure 42.

![Fig. 42: Grid cell and module coordinate notations.](image)

The module’s position is verified with respect to the grid cell coordinates underneath it using Eq. 8.2 and 8.3.

Module \(m\) is to the left/right of grid cell \(g\) if,

\[
R_m.x < L_g.x \quad \text{||} \quad R_g.x < L_m.x
\]  

(8.2)
Similarly, module $m$ is on the top/bottom of the grid cell $g$ if,

$$R_g.y < L_{m}.y \quad \text{or} \quad R_{m}.y < L_{g}.y$$ (8.3)

If either of the conditions are true, the module does not overlap the grid cell and the overlap area evaluation is not performed.

**overlappingArea()**:

Every module overlapping the grid cell contributes different power density to the grid cell. Therefore, the module’s contributed power is calculated individually using the proportion of area occupied by the module in the underlying grid cell. The proportion of the grid cell $g$ occupied by module $m$ is termed the occupied area ratio (OAR) and is given by,
Occupied Area ratio (OAR) = \frac{A_{O,m}}{\text{grid cell area}} \quad (8.4)

where $A_{O,m}$ is the overlapping area of the module and grid cell. $A_{O,m}$ is calculated as shown in Figure 43(c) using Eq. 8.5, 8.6, and 8.7.

\[
O_W = \min (R_{m,x}, R_{g,x}) - \max (L_{m,x}, L_{g,x}) \quad (8.5)
\]

\[
O_H = \min (R_{m,y}, R_{g,y}) - \max (L_{m,y}, L_{g,y}) \quad (8.6)
\]

\[A_{O,m} = O_W \times O_H \quad (8.7)\]

The grid cell area is calculated as the product of grid cell width, and grid cell height. The grid cell dimensions are dependent on the 3D layout dimensions.

\textit{Contributed power density (CPD)}: The calculated OAR is used to evaluate the power density contributed by the module to the underlying grid cell. It is termed the module’s contributed power density (CPD) and is given by,

\[CPD_m = OAR \times PD_m \quad (8.8)\]

where $PD_m$ is the power density in W/m$^2$ of module $m$. If ‘$k$’ number of modules overlap a given grid cell as shown in Figure 43(b), the total grid cell power $P_g$ is the product of the sums of individual CPD and the grid cell area given by Eq. 8.9.

\[\text{Grid cell power } (P_g) = (\sum_{i=0}^{k} CPD_i) \times \text{grid cell area} \quad (8.9)\]

By the end of this step, power values of all grid cells of the 3D layout are available for TGV calculation.
8.4.2 Thermal Goodness Value (TGV)

The generated power numbers at grid level are adjusted for the impact of intra-layer and inter-layer neighbor grid cells. Let $PD_g$ be the power density of a grid cell $g$ in a $nxn$ grid on device layer $l$. The heat transferred from the intra-layer neighbor cells is added to or subtracted from the GUT’s power depending on the magnitude compared to GUT. Based on the location of the GUT, it can have 1 or 2 neighbors in both $x$ and $y$ direction as shown in Figure 44.

![3D thermal goodness evaluation model](image)

Fig. 44: 3D thermal goodness evaluation model.

The power of the neighbor cell in the vertical direction ($z$) is added to the GUT’s power. This is the grid cell directly beneath the GUT on layer ($l-1$). Therefore, the final modified power density, $PD'_g$ of the grid cell $g$ is given by,

$$PD'_g = PD_g + \frac{\sum_{i=1}^{t_x} (\pm p_{gi})_x}{g_w*g_w} + \frac{\sum_{j=1}^{t_y} (\pm p_{gj})_y}{g_h*g_h} + \frac{(p_{gl})_z}{\text{dist}_z*\text{dist}_z} \tag{8.10}$$

The $\text{dist}_z$ is the vertical distance for the heat to transfer between layers. It is the combined thickness of the active layer (Si), back-end-of-line (SiO$_2$), and the thermal interface material (TIM), highlighted in Figure 40. Let $t_x$ and $t_y$ be the number of intra-layer...
neighbors of GUT in the x- and y-direction. The constraints for determining the sign of the power value of the neighbor grid cell, and the number of neighbors, are given by,

\[ P_{g_i} > 0 \quad \text{if} \quad P_{g_i} > P_g \] (8.11)

\[ P_{g_i} < 0 \quad \text{if} \quad P_{g_i} < P_g \] (8.12)

Let \( R \) be the row and \( C \) be the column of cell \( g \). The number of neighbors (\( t \)) in x- and y-direction are,

if \((C = 0 \ || \ n) \quad \&\& \quad (R = 0 \ || \ n)\), then \( t_x = 1 \ ; \ t_y = 1 \) (8.13)

if \((0 < C < n) \quad \&\& \quad (0 < R < n)\), then \( t_x = 2 \ ; \ t_y = 2 \) (8.14)

if \((R = 0 \ || \ n) \quad \&\& \quad (0 < C < n)\), then \( t_y = 1 \ ; \ t_x = 2 \) (8.15)

if \((C = 0 \ || \ n) \quad \&\& \quad (0 < R < n)\), then \( t_y = 2 \ ; \ t_x = 1 \) (8.16)

After power values of all the grid cells on all the device layers are updated, we check for local hotspots on each layer using a 2x2 sub-grid window. The grid cell power densities \( g_{ij} \) of all possible 2x2 sub-grids in the horizontal and vertical direction within a layer are added to generate the sub grid sum, \( SG_s \), given by Eq. 8.17.

\[ \text{Sub grid sum (} SG_s \text{) } = \sum_{i=0}^{i+1} \sum_{j=0}^{j+1} g_{ij} \] (8.17)

\( i \) and \( j \) represent the row and column of the lower left grid cell of a 2x2 sub-grid. The total sum of all 2x2 sub-grids in the 3D chip is the required thermal goodness value given by Eq. 8.18.

\[ \text{TGV} = \sum_{l=0}^{l+2} (\sum_{s=0}^{(n-1)^2} SG_s)_l \] (8.18)
This TGV is used to evaluate the thermal goodness of the generated 3D floorplan as an alternative to time-intensive peak temperature calculation. When TGV (sum of all 2x2 windows) is used as an optimization parameter in 3D floorplanning, it tries to optimize each submatrix/sub-grid such that no submatrix/sub-grid of a 2x2 size has a large sum (or a hotspot). The sub-grid with the largest sum corresponds to the hottest region on the chip. From Figure 43(a), we can observe that the influence of the center grid cells is more pronounced than the ones on the boundary on a nxn grid. Therefore, whether 2x2 or 3x3 sub-grid size, this impact is captured in our evaluation. A 3x3 or higher sub-grid size may enhance TGV evaluation and its correlation with Hotspot because it will multiply the center grid cell impact by a higher factor (because it is added multiple times). The sub-grid sizes vs TGV enhancement should be further analyzed and will be addressed in our future work.

8.5 Experiment Results

The 3D floorplanner tool and the proposed thermal goodness evaluation algorithm were developed in C++ and executed on a 4xDual Core Sun SPARC IV at 1.35 GHz and total 32 GB RAM. Since TSVs were placed in fixed-size islands, they were considered as modules with zero power dissipation for TGV evaluation.

The main aim of this work was to evaluate the 3D floorplan’s thermal goodness with close accuracy to Hotspot tool [1] peak temperature in significantly-reduced time. We thus first compare the simulation time for steady-state evaluation using Hotspot with the thermal goodness algorithm. The runtime reported includes the layout to grid translation for both
Hotspot and TGV. The Hotspot tool’s improved runtime reported in [84] does not have information of the underlying benchmarks. Therefore, for a fair comparison, we generated and compared the runtimes on GSRC benchmarks for three different grid sizes given in Table 18 using Hotspot’s default parameters. Please observe a speed up of 22X to 412X for the n100 benchmark with increasing grid size.

<table>
<thead>
<tr>
<th>Grid Size</th>
<th>Benchmark</th>
<th>Average runtime (25 runs)</th>
<th>Speedup (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Hotspot [1] (s)</td>
<td>Proposed TGV (s)</td>
</tr>
<tr>
<td>64x64</td>
<td>n100</td>
<td>1.89</td>
<td>0.0851</td>
</tr>
<tr>
<td></td>
<td>n200</td>
<td>2.581</td>
<td>0.0901</td>
</tr>
<tr>
<td></td>
<td>n300</td>
<td>4.526</td>
<td>0.128</td>
</tr>
<tr>
<td>128x128</td>
<td>n100</td>
<td>20.4</td>
<td>0.158</td>
</tr>
<tr>
<td></td>
<td>n200</td>
<td>33.77</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>n300</td>
<td>54.46</td>
<td>0.328</td>
</tr>
<tr>
<td>256x256</td>
<td>n100</td>
<td>259.27</td>
<td>0.629</td>
</tr>
<tr>
<td></td>
<td>n200</td>
<td>459.45</td>
<td>0.759</td>
</tr>
<tr>
<td></td>
<td>n300</td>
<td>623.36</td>
<td>0.877</td>
</tr>
</tbody>
</table>

The speedup achieved is even better for larger benchmarks (n200 and n300). Though the grid sizes evaluated are the same for different benchmarks, the difference in runtime is from the layout to grid translation. With the speedup achieved, it is evident that our
proposed TGV algorithm can be incorporated into any thermal-aware floorplanner for thermal goodness evaluation at each iteration without large overall run-time increase.

To compare the power measure from TGV and the peak temperature from Hotspot we report the correlation coefficient between the two variables. Figure 45 shows the correlation coefficient of 0.89 when compared with Hotspot for n100 benchmark on 30 floorplan samples. It can be observed that the TGV is in a good positive correlation with the Hotspot’s peak temperatures. The correlation coefficient depends on how close the floorplan’s peak temperatures from the two tools are to each other.

![Fig. 45: Positive correlation between TGV and Hotspot for n100 benchmark.](image)

In other words, our proposed TGV model is sensitive to a minimum temperature difference of 5 K between the floorplans under comparison. When the temperature difference is higher than 5 K, the probability that TGV can match the Hotspot’s value is also high. Therefore, the larger the temperature difference, the higher the correlation coefficient. This can be observed in Figure 46 which reports a correlation coefficient of 0.96 for all GSRC benchmarks combined. With the increase in benchmark size, the
temperature difference increases (between the benchmarks). Therefore, a larger correlation is observed.

Fig. 46: Positive correlation indicated by the dotted trend line between TGV and Hotspot for GSRC benchmarks using 60 floorplan samples.

Fig. 47: Sample 2D floorplan powers and associated thermal map comparing TGV and peak temperature
Fig. 48: Thermal map of 4-layer n100 benchmark comparing peak temperature and TGV

For two given 3D floorplans if peak temperature is higher for the first floorplan than the latter, we want our TGV model to predict the same. For a better understanding, this is first shown on an example of two power grids for 2D floorplan. One floorplan with a concentrated large hotspot shown in Figure 47(a) and second with distributed smaller hotspots shown in Figure 47(b). We can visually determine that the second floorplan is a thermally better solution. The same is determined by our TGV model.

Next, we compare two sample 3D floorplans (FP1 and FP2) of n100 benchmark, with four device layers each, shown in Figure 48. Layer 0 is the farthest from the heat sink and thus has the highest temperature in both the floorplans. In this example, FP1 has a lower
peak temperature than FP2. Therefore, FP1 is a thermally better solution and the same is determined by the computed TGV. When the temperature difference between floorplans under comparison is less than 5 K, in rare cases our TGV model can predict incorrectly. However, a small 5K difference in temperatures implies that the two floorplans have modules almost equally well-distributed thermally.

Figure 49 compares power-based TGV (grey) and peak temperatures (red) as a function of the 3D floorplan area. Let us consider two sample floorplans (FP1 and FP2) highlighted in the dotted windows. FP1 has an area of 0.319 mm$^2$ and a peak temperature of 564K. FP2 has a slightly larger area of 0.326 mm$^2$ and a peak temperature of 567K.

Among these two 3D floorplans, FP1 is a better solution according to Hotspot’s temperature value as FP1 has a lower peak temperature (in red) than FP2. However, FP2 is better according to TGV (in grey). Since area is usually also included in the cost function, this small inaccuracy of TGV for less than 5K difference can be balanced by the area factor.
The goal of temperature evaluation is to eliminate a really bad temperature distribution solution, which is clearly accomplished with TGV. In all the tested samples for GSRC benchmarks, for temperature difference less than 5 K, the largest area difference observed was 0.015mm$^2$. This again reinforces the conclusion that both the solutions are equally well-packed thermally. Since even Hotspot is prone to an error of 3.5% [84], the sensitivity to such small temperature differences in our TGV model can be safely ignored. To determine the prediction rate of the proposed TGV, we also report the success factor, shown in Figure 50.

Fig. 50: Success rate analysis on 10 samples of GSRC benchmarks.

We compared pairs of floorplans using TGV measure and Hotspot. Success rate is defined as the number of times the result of TGV comparison on all tested pairs of floorplans (which floorplan is thermally better – lower TGV value) matched the result of the Hotspot tool’s peak temperature comparison on the same pair. For this analysis, 15 floorplan samples were generated and 10 were randomly chosen for exhaustive testing.
Each floorplan from the set of 10 chosen was compared with every one of the remaining 14 floorplans. We compared TGV values and peak temperatures. With this exhaustive testing, 140 floorplan pair comparisons were done. If Hotspot and TGV both determine the same floorplan as a better solution, the success rate improves. It can be observed in Figure 50 that we achieved a success rate of 97% when the temperature difference between the compared floorplans was greater than 5 K and success rate of 87% when the temperature difference is less than 5 K.

8.6 Chapter Conclusions

To facilitate rapid thermal screening of 3D floorplans in early stages of design, we propose a power-based measure as an alternative to temperature measures to be used in thermal-aware floorplanning. Our proposed model is capable of selecting a better thermal-aware floorplan without a need for time-intensive simulations to generate the temperature distributions of the floorplan. The proposed TGV (thermal goodness value) exhibits good positive correlation with the more accurate Hotspot tool. The correlation coefficient is larger than 0.88 when the Hotspot’s peak temperature difference between the floorplans is larger than 5 K. The correlation factor improves with increasing temperature difference between the floorplans being evaluated. The evaluation accuracy increases with increasing grid size at the cost of increased runtime. However, the TGV evaluator runtime is very short even for a grid size of 256x256. An average speed up of 29X achieved with our TGV on a problem size of 64x64 when compared to the Hotspot tool which reinforces the
advantage of the proposed model’s simple yet fast evaluation that allows quick selection of thermally-better solutions during the time-intensive optimization cycles. The calculated power-based TGV can be used as a good objective factor in guiding the 3D floorplanner for generating thermally-optimized 3D layouts faster. With reduced runtime, TGV thus provides a framework to be used for both design space explorations and early thermal-aware software development for thermal-aware 3D IC solutions.
Chapter 9

Delay Optimization during 3D-IC Floorplanning using Thermal-aware Evaluation

9.1 Introduction

TSVs are electrical connections between the vertically stacked dies. The recent advances in this technology offer manifold opportunities including high memory bandwidth, providing a viable alternative to the interconnect bottleneck in their 2D counterparts [77], improved performance, and significant savings in power consumption [60]. However, heat is trapped in 3DICs due to the vertical stacking of the power dissipating dies and thus, elevating the thermal profile. Figure 51 illustrates the rising vertical thermal dissipation problem due to increased power density in 3D ICs in comparison with planar technology. Therefore, the necessity for thermal aware solutions, in early physical design stages like floorplanning, to reduce the increase in temperature and hot spots is even more prominent in 3D ICs.

Floorplanning is the first major step in the VLSI physical design and affects the optimization results of the subsequent stages including placement and routing. The goal of 3D floorplanning is to find locations for circuit blocks and TSVs to optimize circuit performance and power. In this chapter, to evaluate the goodness of a floorplan, we use a
pre-defined cost function such as a linear weighted combination of chip area, interconnect delay and a measure of thermal goodness of the floorplan. The goodness of the evaluated optimization parameters plays a key role in driving the 3D floorplanning problem to generate better optimized 3D layouts.

Fig. 51. Trapped heat between stacked layers in 3D ICs due to increased power density compared to 2D heat dissipation path.

Most works in the literature [85][71][64][18][86] evaluate 3D chip interconnect performance calculated at room-temperature parameter and fixed wire density. However, due to increased power density, each device layer will be at a different temperature and higher than the room temperature. Also, packing density on each device layer is different and therefore a non-uniform wire density exists across the vertically stacked dies [87]. The presence of a vertical thermal profile and non-uniform wire density in 3D ICs significantly impacts the interconnect delay and buffer count as metal resistance and wire capacitance are a strong function of temperature and wire spacing respectively [87][88]. Since,
interconnect delay and power are heavily used to characterize the 3D floorplan solution quality, ignoring the varying values of RC parameters impact on interconnect performance can lead to incorrect solution selection during the probabilistic 3D floorplan optimization. It is challenging to efficiently remove generated heat from a well-packed 3D IC and it is impacting the interconnect performance. Thermal-aware performance evaluation can possibly improve the solution selection process in a non-deterministic floorplanning flow.

9.2 Previous Works

Most state-of-the-art layout design tools and 3D floorplanning algorithms are probabilistic. They require thousands of iterations to find an optimized solution. To use temperature as a floorplanning objective, we need to evaluate temperature distribution at each iteration, or possibly at a specified number of iterations. This will increase the run time proportionally to the time complexity of the temperature distribution generating software, which might be significant especially for simulation-based approach, Hotspot. In addition to improving accuracy and quality of design processes using thermal-aware [22][89][20][21] and delay-aware floorplanning [85][71][64][18][86], there has been an extensive research on effective heat removal techniques [78][90][80][79] in 3DICs to alleviate the high temperatures.

There has been a number of approaches for evaluating, and simulating temperature distribution in 3D ICs. They all vary in accuracy, time complexity and how heat transfer is considered. The most accurate simulation-based approach, with unfortunately very high
time complexity, was proposed by Huang et al [1] and correlation with the Hotspot results is used as the accuracy measure for other simplified approaches. Ni et al. [89] proposed a power-based approach for the 3D thermal-aware design. In their approach, heat transfer between two adjacent blocks is considered only if the blocks are contiguous. A hotspot however can result also from the mutual impact of adjacent hot blocks that are not contiguous but in close proximity. Xu et al. [20] proposed pre-simulating, using Hotspot [1], the thermal distribution due to each block placed at all possible locations on a given device layer of the 3D layout. Then, they quickly estimate the block temperatures during floorplanning iterations using bilinear interpolation. However, the time taken for pre-simulation of all blocks in all possible positions, which can be significant as circuit complexity increases, should be included in the overall time complexity of the approach. Since they use a post layout TSV insertion method, we cannot perform a fair comparison with their results. Xiao et al. [21] put forth an approximate thermal model for thermal optimization and show a correlation of 0.97 with the Hotspot tool. Despite its accuracy, the thermal-aware model still imposes a runtime penalty of 4X as compared to a non-thermal driven 3D floorplan optimization. Cong et al [22] proposed a simplified vertical model that is fast and has a correlation coefficient of 0.82 with the more accurate model Hotspot as reported by [21]. However, the correlation depends on the floorplans considered for evaluation. With the lateral heat impact ignored, the simple vertical model can produce inaccuracies resulting in inferior solution selection during the floorplanning iterations.

All these works are based on the traditional wire-length based floorplanning and do not optimize wire delay directly. Delay-aware floorplanning is primarily beneficial for
unifying the objectives of reducing the wire length and the number of TSVs. Even when delay-aware floorplanning [64] is considered, the assumption used is that the wire density and wire resistance are constant on all layers. However, the length of the wire segment used for buffer insertion on different device layers should be different due to the existence of a vertical thermal profile. The distribution of wires on each device layer will also differ resulting in varying wire densities on each layer. Prior works on delay-aware floorplanning [64][91] do not consider the optimization of thermal-delay, leading to inaccuracies in solution selection.

In summary, it is important to note that significant underestimation of interconnect delay can result when the interconnect resistance dependence on actual device layer temperature is ignored. Increasing delay in the wires leads to poor pre-layout estimates of delay. Wire capacitance also varies with wire density in 3D ICs [92], impacting the early buffer estimates and thus, the total delay evaluation. This can lead to incorrect values used in a cost function that governs the process of 3D IC floorplan optimization.

To prevent this, in our work, thermal-aware interconnect parameters are considered. We also noticed that the increase in wire delay with temperature reduces the buffer insertion length and increases the number of needed buffers. More buffers with a considerable increase in buffer delay also impacts interconnect performance. Deriving from the analysis, the dependence of buffer insertion length and buffer count on temperature is also analyzed carefully in this work when evaluating the interconnect performance. Motivated by the above arguments, in this paper, we propose the following.
9.3 Chapter Contributions

- Fast fusion model for vertical thermal profile generation that includes horizontal temperature dependencies and significantly increases correlation with the more accurate simulation based tool, Hotspot [1]. The proposed model overcomes the shortcomings in vertical only model [22] by considering the impact of heat dissipation from intra-layer modules as well. It quickly evaluates the thermal goodness of two given 3D floorplans, for comparison. Results of comparisons are in excellent agreement with the Hotspot tool.

- Design strategy to integrate realistic thermal-aware interconnect performance evaluation in the 3D floorplanning optimization. We consider the impact of vertical thermal profile and non-uniform wire spacing on TSV-aware buffer insertion length. This strategy enables a more realistic evaluation of the wire delay and hence, avoiding sub-optimal solution generation.

- Expanding 3D IC floorplanning to allow for in-depth analysis of 3D-interconnect delay and early buffer count. Estimates are performed under unified RC variations resulting from rising vertical temperature and varying wire capacitance to quantify correctly the interconnect performance.

To the best of our knowledge, this is the first work to propose a fast vertical thermal profile generation considering the lateral heat impact, highly correlated with Hotspot tool, that can be easily integrated in the already time intensive 3D optimization cycle. This is
also the first work to incorporate and analyze the impact of individual device layer temperatures on the 3D interconnect performance.

The rest of the chapter is organized as follows. Section 9.4 briefs about the 3D floorplanning flow and buffer insertion. The proposed fusion method of simplified vertical thermal model and thermal goodness evaluation is described in Section 9.5. The influence of temperature and interconnect density on 3D interconnect performance is discussed in Section 9.6 along with the 3D floorplanning fitness parameters used in this work. Section 9.7 presents the experimental results and chapter conclusions are given in Section 9.8.

9.4 Preliminaries

Floorplanning is an important step in physical design of integrated circuits and even more important for 3D system. Floorplanning fixes geometrical locations for all circuits blocks to optimize a predefined cost function that describes quality of the floorplan. Floorplanning in a major way influences delay and locations of hotspots in a 3D system.

All the floorplans, in this work, were generated by a non-deterministic, evolutionary-computation based floorplanning tool described below in Section 9.4.1 using evolutionary computation. All the interconnect performance evaluations were done on four-tier floorplans generated for GSRC benchmarks.

9.4.1 3D IC floorplanning

The proposed fusion model is integrated on top of the non-deterministic 3D floorplanning tool [64] encompassing TSVs and modules co-placement. This tool was built
upon the 3D floorplanner in [65], which is an extension of the 2D floorplanning software developed by Wang et al [66]. In this 3D floorplanner, nets are assigned to TSV islands within the optimization stage. Careful buffer injection is completed at each iteration step before interconnect delay is estimated. TSVs are included in the wire delay estimation for buffer insertion. The generated floorplans are used to evaluate each device layer’s average temperature using the proposed fusion model in each optimization cycle. The interconnect parameters are calculated as a function of the device layer average temperature and used in the subsequent iteration to compute thermal-aware interconnect delay. The final floorplans are then used to evaluate each device layer’s peak temperature using the well-known Hotspot [1] tool. For all our analysis, we assume that the heat sink is on the top of the 4-layer 3D IC stack (Layer 0 is the farthest from the heatsink). The floorplanning tool uses TSV locations and delay, non-uniform interconnect density across multiple stacked device layers to assess and optimize the buffer count and delay [92].

We assumed 45 nm technology node and module power density in the range of $10^5 - 10^7$ W/m$^2$ [22]. As the floorplanning benchmarks do not include module power density values, random power density values from the assumed range were assigned to all modules in all benchmarks. Once assigned, power density values of all modules remained the same for all experiments. Since module activity information is not available either, one can assume that average power activity is included in the power density values.

9.4.2 Correlation coefficient
The correlation coefficient is used as a measure to compare our 3D floorplan average temperature evaluation (K) with the Hotspot tool’s peak temperature (K). Correlation coefficient, $r$, is a unit less metric with a value between -1 and 1. It is described in Section 8.3 and is given by,

$$
r = \frac{1}{k-1} \left[ \frac{\sum (X - \bar{X}) \cdot (Y - \bar{Y})}{S_X S_Y} \right]
$$

(9.1)

$$-1 \leq r \leq 1$$

In Eq. 9.1, $k$ represents the number of samples, $\bar{X}, \bar{Y}$ are the mean and $S_X$ and $S_Y$ are the standard deviation of the sample data of variable $X$ and $Y$, respectively. The stronger the two variables are correlated, the closer the value $r$ is to 1. If the two variables are inversely correlated, $r$ is closer to -1. An $r$ value of 0 implies no correlation between the two variables.

### 9.4.3 Buffer Insertion Length (BIL)

The insertion of buffers along a long wire to reduce delay is called buffer insertion. BIL is the minimum required distance between consecutive buffers. Placing an optimally sized buffer reduces the delay of the corresponding un-buffered wire. A serious drawback of excessive use of buffers is that these large buffer gates add to power consumption and also occupy non-negligible silicon area. If not optimized, a large number of buffers may even degrade the overall interconnect performance as buffer delay dominates wire delay in
shorter wires at higher temperatures. In this work, we adopt the buffer insertion methodology proposed in [71] for our buffer count estimates.

The nomenclature and all the c parameters used in this work are listed in Table 19 and 20 [74] [75] respectively.

Table 19: Nomenclature

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$VM$</td>
<td>Vertical thermal model [1]</td>
<td>--</td>
</tr>
<tr>
<td>$TGV$</td>
<td>Thermal Goodness Value</td>
<td>W/m$^2$</td>
</tr>
<tr>
<td>$V_MT_G$</td>
<td>Proposed fusion of VM and TGV</td>
<td>--</td>
</tr>
<tr>
<td>$T_{peak}$</td>
<td>Peak temperature of 3D IC</td>
<td>K</td>
</tr>
<tr>
<td>$T_{avg,L}$</td>
<td>Average device layer temperature</td>
<td>K</td>
</tr>
<tr>
<td>$R_{th}$</td>
<td>Thermal resistance</td>
<td>K/W</td>
</tr>
<tr>
<td>$D/D_{room}$</td>
<td>Total wire delay at room temperature</td>
<td>ns</td>
</tr>
<tr>
<td>$DT/D_{avg}$</td>
<td>Total wire delay at device layer average temperatures</td>
<td>ns</td>
</tr>
<tr>
<td>$DTC$</td>
<td>average temperatures considering variation in wire capacitance</td>
<td></td>
</tr>
</tbody>
</table>
Table 20: Technology Specifications Used in our Experiments

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_w )</td>
<td>Unit length wire resistance (M4-M6)</td>
<td>0.29-0.66 ( \Omega/\mu m )</td>
</tr>
<tr>
<td>( R_{w(T)} )</td>
<td>Unit length wire resistance at device layer temperature</td>
<td>varies</td>
</tr>
<tr>
<td>( C_w )</td>
<td>Unit length wire capacitance (M4-M6)</td>
<td>0.173- 0.176 ( fF/\mu m )</td>
</tr>
<tr>
<td>( C_{w, ID} )</td>
<td>Unit length wire capacitance with varying interconnect density (ID)</td>
<td>varies</td>
</tr>
<tr>
<td>( R )</td>
<td>Buffer output resistance (8x)</td>
<td>300 ( \Omega )</td>
</tr>
<tr>
<td>( C )</td>
<td>Buffer input capacitance (8x)</td>
<td>6.65 ( fF )</td>
</tr>
<tr>
<td>( S )</td>
<td>Buffer size</td>
<td>8x</td>
</tr>
<tr>
<td>( p_{inv} )</td>
<td>Parasitic Capacitance Factor</td>
<td>0.5</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Switching activity</td>
<td>0.1</td>
</tr>
<tr>
<td>( V_{DD} )</td>
<td>Supply Voltage</td>
<td>0.9 ( V )</td>
</tr>
<tr>
<td>( t_{Si} )</td>
<td>Silicon thickness</td>
<td>150 ( \mu m )</td>
</tr>
<tr>
<td>( t_{TIM} )</td>
<td>TIM thickness</td>
<td>20 ( \mu m )</td>
</tr>
<tr>
<td>( t_{SiO2} )</td>
<td>Dielectric thickness</td>
<td>800 ( \mu m )</td>
</tr>
</tbody>
</table>
Units wire resistance at temperature $T$ (Ω/µm) varies

$k$ Thermal conductivity (Silicon) 100 W/m-K

$\#DL$ Total device layers 4

$M_W$ Metal width 140 nm

$M_T$ Metal thickness 252 nm

$M_S$ Metal spacing varies

$TSV_{dia}$ TSV diameter 3 µm

$TSV_{AR}$ TSV Aspect Ratio 10

$TSV_p$ TSV pitch 6 µm

$R_c$ TSV contact resistance 10

9.5 Fast Fusion Model for Improved Temperature Correlation

9.5.1 Simplified Vertical Thermal Model [22]

For generating a quick vertical temperature distribution of a 3D-IC, we adopt the simple vertical temperature model (VM) proposed by Cong et al [22]. In the VM, the whole chip is divided into vertical tile stacks as shown in Figure 52(a). Stack-wise temperature (highlighted in red bounding box) analysis is then performed by forming an Elmore-delay-like closed-form formula for the temperature increase at the desired node as shown in Figure 52(b).
The temperature at node \( n \) can be written as,

\[
T_n = \sum_{i=1}^{n} (R_{th,i} \sum_{j=0}^{n-1} P_j) + R_b \left( \sum_{j=0}^{n-1} P_j \right)
\]  

(9.2)

In Eq. 9.2, \( R_{th} \) represents the thermal resistance (\(^{0}\text{C/W}\)) and \( P_j \) is the power density of a tile in the stack on layer \( j \). \( P_j \) is the assigned cumulative power density of the modules overlapping a given tile based on the 3D layout. The thermal resistance, \( R_{th} \) is calculated as,

\[
R_{th} = \frac{t}{k \times A}
\]  

(9.3)

where \( t \) is the thickness of the layer, \( k \) is the thermal conductivity of silicon and \( A \) is the cross-sectional area perpendicular to the heat flow. This vertical thermal model \((VM)\) is often used to evaluate the peak temperature, \( T_{peak} \) of a 3D IC in thermal-aware 3D floorplanning. However, the peak temperature calculated using \( VM \) is reported to have a correlation of only 0.8 [21] with the more accurate simulation based tool like Hotspot [1].

Fig. 52. (a) 3D tile stack array (b) Single tile stack analysis
9.5.2 Thermal Goodness Evaluation of 3D IC

In order to more effectively and faster prune thermally unfavorable 3D solutions in the optimization flow, we proposed a power-based metric called thermal goodness value (TGV) in our prior work [93]. The TGV has a good correlation of 0.96 with the Hotspot tool. TGV aims to distinguish which of two given floorplans has a better thermal-aware module placement using a simple power-based measure that can be calculated in a very short time (29X faster than Hotspot).

In this approach, the power densities of modules, \( PD_i \), are first assigned to cells of \( nxn \) grid, on device layer \( l \) based on the generated 3D layout. We use the proportion of area occupied by the module in the underlying grid cell to calculate how much of the module PD is assigned to the grid cell, \( PD_{g} \). These grid cell power densities, \( PD_{g} \), are then updated as \( PD'_{g} \), considering the impact of intra-layer and inter-layer neighbor grid cells on each grid cell under test (GUT). Based on the location of the GUT, it can have 1 or 2 neighbors in both \( x \) and \( y \) direction and 1 neighbor cell in the vertical direction \( z \) on device layer \( l-1 \) as shown in Figure 53. Since the primary heat flow path from hot layers towards the heatsink (bottom to top) is much stronger than the secondary heat flow path towards the package (top to bottom), the secondary heat flow from layer \( l+1 \) to layer \( l \) is ignored.
The final modified power density, $PD'_g$ of the grid cell $g$ is given in Eq. 9.4. The first term in Eq. 9.4 represents the internal PD of the grid cell $g$ under test, second and third terms represent the PD contributed from the intra-layer neighbor grid cells in x and y-direction respectively and the last term represents PD from neighbor cell in the vertical direction $z$.

\[ PD'_g = PD_g + \sum_{i=1}^{t_x} (\pm P_{g_i})_x + \sum_{j=1}^{t_y} (\pm P_{g_j})_y + \frac{(P_{g_R})_z}{\text{dist}_z \times \text{dist}_z} \]  (9.4)

The $\text{dist}_z$ is the vertical distance for the heat to move between layers. It is the combined thickness of the active layer (Si), back-end-of-line (SiO$_2$), and the Thermal Interface Material (TIM). Here $t_x$ and $t_y$ are the number of intra-layer neighbors of GUT in the x- and y-direction. The sign of the power value of the neighbor grid cells, $P_{g_i}$ and $P_{g_j}$, is determined based on a simple comparison with $PD_g$, power density of GUT as shown below.

\[ P_{g_i}, P_{g_j} > 0 \quad \text{if} \quad P_{g_i}, P_{g_j} > PD_g \]  (9.5)
Delay Optimization using Thermal-aware Evaluation

\[ P_{g_i}, P_{g_j} < 0 \quad \text{if} \quad P_{g_i}, P_{g_j} < PD \]  

(9.6)

After power density values of all the grid cells, \( g_{ij} \), on all the device layers are updated, we calculate the total power intensity within a given region on a device layer using 2x2 sub-grid windows, \( SG_s \) (total power of a 2x2 window) given by Eq. 9.7.

\[ \text{Sub-grid sum}(SG_s) = \sum_{i}^{i+1} \sum_{j}^{j+1} g_{ij} \]  

(9.7)

where, \( i \) and \( j \) represent the row and column of the lower left grid cell of a 2x2 sub-grid on and \( n \times n \) grid, and subscript \( s \) represents the window number). The sub-grid with the largest sum corresponds to the hottest region on the chip. The total of power in all 2x2 sub-grids on device layers, \( l \), in the 3D chip is the required thermal goodness value given by Eq. 9.8.

\[ TGV = \sum_{s=0}^{(n-1)^2} SG_s \]  

(9.8)

TGV (sum of all 2x2 windows) as an optimization parameter aims for overall better heat spread when used in floorplanning such that no sub-grid of a given size has a large sum. Thus, TGV can be used to evaluate the thermal goodness of the generated 3D floorplan as an alternative to time-intensive peak temperature calculation.

9.5.3 Proposed Fusion of VM and TGV
To evaluate the interconnect delay as a function of temperature we need to have a device layer temperature. In the original VM model, the input for evaluating layer temperature is the actual power densities of modules assigned to a n x n grid. However, from our experiments we observed that the VM generated temperatures (\( T_{\text{avg}}, T_{\text{peak}} \)), are weakly correlated (0.85 and 0.6) with Hotspot tool. Though TGV has a high correlation (0.96) with Hotspot, it is only a power based metric. For a thermal-aware interconnect delay evaluation, we need a temperature value that is well correlated with Hotspot. Therefore, we would like to improve VM’s \( T_{\text{avg}} \) correlation using TGV. When the TGV updated power densities were used as inputs to the VM, we achieved a much better correlation with Hotspot as compare to just VM. We named this approach \( V_{MTG} \). This is done to overcome the shortcoming of ignored intra-layer thermal impact in VM. The

![Flow of proposed fusion methodology \( V_{MTG} \): 1-dimensional heat flow analysis on TGV model](image)

Fig. 54. Flow of proposed fusion methodology \( V_{MTG} \): 1-dimensional heat flow analysis on TGV model
proposed fusion methodology of $V_M T_G$ will have a high correlation with Hotspot for using $TGV$ and will calculate the required device layer temperatures by using $VM$. In the process of TGV evaluation, the intra-layer heat dissipation from neighbor cells is also considered unlike VM. The overview of the proposed methodology is shown in Figure 54.

In the proposed fusion model, after assigning module power densities to the grid cells, power density of each grid cell is updated as given by Eq. 9.9.

$$PD'_g = PD_g + \frac{\sum_{i=1}^{\text{tx}} (\pm P_{Ri})_x}{A_x} + \frac{\sum_{i=1}^{\text{ty}} (\pm P_{Ri})_y}{A_y} \quad (9.9)$$

where, $PD_g$ is the power density assigned to a grid cell, $PD'_g$ is the updated power of a grid cell, $A$ is the cross-sectional area perpendicular to the heat flow and $t_x$ and $t_y$ represent the neighbor grid cells for the GUT. Since $VM$ already considers the vertical heat flow, we ignore the power density contributed in the z-direction in Eq. 9.9. Now, these updated power densities form the input to the vertical temperature model ($VM$). The stack-wise temperature analysis is then performed on all $(l*n^2)$ stacks, where $l$ is the total number of device layers in a 3D-IC. At the end of the analysis, the thermal distribution of a $n \times n$ grid on each device layer is available. Now, the average temperature, $T_{avg}$, on a device layer $L_i$, is computed as,

$$T_{avg,L_i} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} T_{ij}}{n^2} \quad (9.10)$$

In Eq. 9.10, $i$ and $j$ are the rows and columns of a $n \times n$ grid on layer $L_i$. We use this $T_{avg,L_i}$ to update the metal resistance on each device layer for individual net delay evaluation, described in the next section. Please note that $TGV$ is a power factor measured in (W/m$^2$).
while all other metrics evaluated using Hotspot [1] and VM [22] model are temperatures (K).

9.6 Accurate 3D-interconnect delay evaluation

The 3D interconnects span across inter-blocks and between different tiers. The considered layer-wise delay distribution includes the 3D nets realized with TSVs and 2D wires. The number of buffers on a device layer also depends on the TSV position [64]. The delay, $D$, of a buffered 3D interconnect with TSVs is given by the Elmore-delay in Eq. 9.11 [71].

\[
D = N \left( \frac{R}{W} \left( Cw \frac{l}{N} + C \cdot W(1 + p_{inv}) \right) + R_w \frac{l}{N} \left( Cw \frac{l}{N} + C \cdot W \right) \right) + N_{tsv} \left( \frac{R}{W} \left( C_{tsv} + C \cdot W(1 + p_{inv}) + R_{tsv} \left( C_{tsv} \frac{l}{2} + C \cdot W \right) \right) \right)
\]  

(9.11)

where, $N$ and $N_{tsv}$ represent the number of buffered segments and number of TSVs in the wire respectively, $W$ represents the buffer size and $l$ represents the length of the wire to be buffered. Please note that 3D nets, in our analysis, are only buffered if they yield shorter delay than being unbuffered. Ratio $\frac{l}{N}$ represents the buffer insertion length ($BIL$) [71] and is given by,

\[
BIL = \frac{l}{N} = \sqrt{\frac{2(R + C(1 + p_{inv}) + C_{tsv}N_{tsv}R_{tsv})}{R_wC_w}}
\]  

(9.12)
In Eq. 9.12, $R_{tsv}$ and $C_{tsv}$ represent the TSV resistance and TSV capacitance, respectively. $R_{tsv}$ is the summation of material resistance and TSV contact resistance. TSV contact resistance highly dominates the TSV material resistance and hence the influence of temperature on TSV material resistance is ignored to save runtime. $R$ and $C$ represent the output buffer resistance and input buffer capacitance, respectively. $C_w$ and $R_w$ represent the unit-length wire capacitance and resistance at room temperature, respectively. As stated earlier, for buffer count estimation we adopt the buffer insertion scheme proposed by Abrar et al. [71] In their approach, buffer insertion is done separately on each device layer. We assume a constant buffer size of 8X throughout this work as the buffer sizing is not the focus of this paper.

### 9.6.1 Influence of temperature on wire resistance

If thermal effects are taken into account, the interconnect resistance must be expressed as a function of temperature. Usually, a linear model is accurate within the range of on-chip operating temperatures. We assume the capacitance and thermal conductivity of the metal are constant with temperature. To avoid over estimating the delay using peak temperature, the device layer’s average temperature, $T_{avg,L}$ is used throughout this work to evaluate delay as a function of temperature, $D_{T_{avg,L}}$. The wire resistance at device layer temperature, $R_w(T)$ is given by Eq. 9.13 [94]

$$R_w(T) = R_w(1+\beta(T-To))$$  \hspace{1cm} (9.13)
where, $R_w$ is the unit wire resistance at reference temperature, $\beta$ is the temperature coefficient of resistance of the metal, $T_0$ is the reference temperature (room temperature, 27°C) and $T$ is the average temperature the device layer. The calculated $R_w(T)$ for the given device layer is used in Eq. 9.11 in place of $R_w$ to calculate the delay of wire on that layer.

### 9.6.2 Influence of interconnect density on wire capacitance

The placement of blocks and TSV islands on a 3D layout influences the interconnect density and thereby the wire capacitance. We estimate the wire spacing [87] on metal M4 at the 45nm technology node to compute the wire capacitance on each device layer. The interconnect density (ID) on a device layer is first calculated as the ratio of average wire length to 3D footprint area, which defines the maximum available routing area. This varying interconnect density across multiple device layers is then used to estimate the wire spacing on each device layer separately. The wire spacing is defined by the ratio of interconnect densities with respect to maximum density value [87]. The intermediate device layers have the maximum wire density and therefore minimum spacing, resulting in maximum capacitance.

### 9.6.3 Impact of $R_w(T)$ and $C_{w, ID}$ on buffer insertion length

Since, the positions of TSV islands and blocks after each floorplanning run are different, the average device layer temperature on each 3D tier is also different. Due to varying temperatures and wire capacitances, the optimal buffer insertion length will not be
the same either. Hence, it is important to evaluate the buffer insertion length separately on each device layer after every iteration.

The needed number of buffers depends on both varying parameters; wire resistance varying with temperature \((R_w(T))\) and wire capacitance varying with interconnect density\((C_{w\_ID})\). Therefore, it is important to incorporate the influence of both effects in the design formulas. To better understand the impact of \(R_w(T)\) and \(C_{w\_ID}\) on BIL and buffer count, let us consider a 1000\(\mu\)m long wire spanning from DL1 to DL0 as shown in Figure 55. Let a TSV of 40 \(\Omega\) contact resistance connect these wire segments through silicon between device layers 0 and 1. If BIL is computed using room temperature parameters, the buffer count on this wire will be 3 as shown in Figure 55(a). However, if the average temperature of the DL1’s is 120\(^{\circ}\)C, and we consider only temperature effect on wire, BIL decreases and the buffer count increases to 4, as shown in Figure 55(b). Figure 55(c) shows the combined impact of temperature increase and capacitance decrease due to reduced interconnect density reducing on DL0. In such case the buffer count is reduced to 2.

In case 55(c), the impact of interconnect density dominates over the thermal effects on BIL. However, this is not always the case. The number of buffers on a device layer may increase or decrease subject to the layer’s temperature and wire spacing. Therefore, it can be observed that using a constant BIL at room temperature can lead to significant underestimation or overestimation of the buffer count and thereby impacting the overall interconnect delay evaluation guiding the optimal floorplan selection.
9.6.4 3D Floorplan Evaluation

The total delay is a more precise optimization objective than minimizing wire length [64] and the number of TSVs separately. Abrar et al., [64] show a delay reduction of up to 12% and improved solution quality using delay-aware floorplanning. In addition, the delay-aware approach also effectively includes the impact of TSV RC delay on the delay of a net.
Therefore, our 3D floorplan evaluation is guided by a delay-aware cost function, the estimated delay using wire capacitance-aware and temperature-aware buffer insertion.

We use two floorplanning cost functions as shown by Eq. (9.14) and Eq. (9.15).

\[
\begin{align*}
\text{CF1} &= \alpha \cdot \text{Area} + \beta \cdot \text{Delay} \\
\text{CF2} &= \alpha \cdot \text{Area} + \beta \cdot \text{Delay}_{\text{Tavg}} + \gamma \cdot \text{TGV}
\end{align*}
\]

\(\alpha\), \(\beta\) and \(\gamma\) are the tuning parameters. CF1 optimizes only area and delay. CF1 tries to reduce delay by packing the modules as tightly as possible and it can result in very nonuniform temperature distribution. Therefore, CF2 includes TGV as an optimization parameter to balance the solution quality from a temperature perspective and enable a better power spread throughout all the layers of the 3D-IC. CF1 is used with all room temperature parameters, and also using actual device layer temperature. While CF2 optimizes the delay at device layer temperature only. The weights of \(\alpha\) and \(\beta\) were fixed at 1 and 100 respectively as they have better minimized the average value of the area and delay [87]. We experimented with \(\gamma\) weight for TGV in the range of 10 to 50. \(\gamma=10\) has very minimum effect on the thermal quality of the generated 3D layouts. \(\gamma>50\) yielded 3D solutions with large increase in area and thus, decreased packing efficiency as shown in Table 21 for n100 benchmark.
Table 21: Comparison of Evaluation Parameters for n100 Benchmark with $\gamma = 40, 45$ and 50

<table>
<thead>
<tr>
<th></th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area ($x10^3$) ($\mu m^2$)</td>
<td>6.4</td>
<td>6.5</td>
<td>6.8</td>
</tr>
<tr>
<td>Packing Efficiency</td>
<td>81.4</td>
<td>80.9</td>
<td>76.2</td>
</tr>
<tr>
<td>Total Delay(ns)</td>
<td>807</td>
<td>802</td>
<td>817</td>
</tr>
<tr>
<td>Peak temperature (K)</td>
<td>563</td>
<td>555</td>
<td>546</td>
</tr>
</tbody>
</table>

Therefore, the cost function used in this thesis uses a $\gamma$ weight of 45, which has minimized the average value of TGV and peak temperature without drastically degrading the packing efficiency below 80%.

Delay calculated using room temperature parameters is denoted as $D_{room}$ and using device layer temperatures as $D_{Tavg}$. Prior works on delay-aware floorplanning do not consider the thermal impact on delay. Hence we compare our results with and without thermal optimization using the above cost functions.

Also, for a meaningful buffer insertion, we use the modified (expanded) GSRC benchmark circuits [71] as shown in Table 22. All buffer insertion results are presented using these modified GSRC benchmarks. As the chips grow in size and complexity, large-scale placement is essential to achieve multiple design objectives. Hence, we use these expanded benchmark circuits in which the size of each block is increased by 100 times and additional multi-pin nets are included.
Table 22: Modified GSRC Benchmark for Buffer Insertion

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Block area</th>
<th># Nets</th>
</tr>
</thead>
<tbody>
<tr>
<td>n100_exp</td>
<td>100x</td>
<td>885</td>
</tr>
<tr>
<td>n200_exp</td>
<td>100x</td>
<td>2136</td>
</tr>
<tr>
<td>n300_exp</td>
<td>100x</td>
<td>2914</td>
</tr>
</tbody>
</table>

9.7 Experimental Results

All the analyzed 3D floorplans were generated using the non-deterministic 3D floorplanning tool described in Section 2. The 3D floorplanner tool and the proposed fusion algorithm were developed in C++ and executed on a 4xDual Core Sun SPARC IV at 1.35 GHz and total 32 GB RAM. TSVs are placed in fixed-size islands and hence, considered as modules with zero power dissipation for TGV evaluation. All experiments are done using a white space (WS) of 5% maximum, and an average of 15 runs is reported for each benchmark.

We first compare the correlation of the standalone VM model, TGV model and proposed fusion model with the Hotspot tool. Then, we present our floorplanning results on GSRC benchmarks with the proposed fusion methodology integrated into the floorplanning flow.
9.7.1. TGV vs VM correlation and runtime comparison

TGV is fast and its power factor is in good correlation with Hotspot’s peak temperature. Since TGV’s average run time is up to 575X faster than Hotspot’s on a 256x256x4 grid [93], it is expected that it will incur much less runtime overhead as compared to using Hotspot’s peak temperature to guide the 3D floorplanner’s cost function through the thousands of optimization cycles. Since VM is also fast, we first compare our temperature alternative TGV (W/m²) metric with VM. Three metrics are compared for correlation with Hotspot’s peak temperature in Figure 6: (i) Peak temperature, $VM_{T_{\text{peak}}}$ (K), (ii) TGV[93] (W/m²) and (iii) Average temperature, $VM_{T_{\text{avg}}}$ (K). All compared values are for the layer farthest from the heatsink. Presented results are calculated based on an average of 50 floorplans of GSRC benchmarks.

(a)
In Figure 56(a) on a 64x64 grid size, please observe that the TGV’s correlation coefficient is greater than 0.9 with respect to the Hotspot tool [1] while $VM_{T_{peak}}$ has a correlation of only 0.65. The average device layer temperature using VM, $VM_{T_{avg}}$ has a better correlation of 0.84 as compared to $VM_{T_{peak}}$. Thus, TGV is calculated in our 3D floorplanner at every optimization cycle and is used in the cost function.

Figure 56(b) shows the runtime for each of the discussed cases. The runtimes presented in Fig 56(b) are the time taken for estimating the TGV’s power metric and temperatures (avg. and peak) using VM. Even on a grid size of 64, TGV is approximately 30% faster than VM with peak and with average temperatures. Since the VM’s model for temperature estimation is the same for peak temperature and average temperature, their
runtimes are similar. In TGV there is no need for power to temperature conversion and that can explain the shorter TGV runtimes. Please note that using a grid size of 32, TGV retains the correlation greater than 0.9 and yet saves 71% run time as compared to a grid size of 64. Therefore, for all our analysis of the thermal-aware floorplanning tool presented in this work, we use a grid size of 32. TGV is used as an optimization parameter as it is still highly correlated with the Hotspot tool, has a higher correlation factor than \( VM \) and is 1.4X faster than VM (both average and peak). 3D floorplanning runtime and time used for converting the 3D layout to grid are not included for a fair comparison.

### 9.7.2 Fusion model correlation with Hotspot tool

TGV is useful as a thermal goodness measure of a floorplan but since it is based on power densities it does not provide temperatures for the device layers that are needed to calculated temperature dependent parameters. The average layer temperature is required to update the interconnect resistance, \( R_w(T) \) at each floorplanning run. The proposed fusion method, \( V_M T_G \), shows improved correlation with Hotspot tool over of \( VM_{T_{avg}} \) for the average device layer temperature. The temperature correlation trend for 55 floorplans of n100 benchmark with respect to Hotspot tool using the fusion model, \( V_M T_G \), is shown in Fig 7. The reported peak temperatures (shown in black on Fig 7), on chosen floorplans, are simulated using Hotspot while the average temperatures (red) are computed using the fusion model, \( V_M T_G \) with a grid size of 64.
Fig. 57. Improved correlation of VM using $V_MT_G$ with Hotspot tool

In Figure 57, floorplans from different stages of the non-deterministic optimization algorithm are taken to show a wider temperature range. The average of the computed temperatures ($T_{avg}$) on the hottest layer is used to compare with Hotspot peak temperature.
Fig. 58. Temperature correlation comparison between Hotspot and (a) standalone VM model and (b) proposed fusion methodology $V_M T_G$

Each point on Fig 57 represents a temperature value for one individual floorplan. The points are connected with by a line to allow for a better visual observation of the very close correlation between the two methods. In Figure 58, we show the improved correlation using $V_M T_G$, we compare the temperatures of a subset of floorplans from Figure 57 highlighted in dotted line. The floorplans temperature in Figure 58(a) using original VM model are highly non-uniform with respect to the Hotspot’s peak temperature and thus have a correlation coefficient of only 0.6. The trend of the floorplans temperatures is improved in Figure 58(b) with correlation coefficient of 0.9 using our proposed fusion methodology $V_M T_G$. Our temperature range is slightly lower than Hotspot because we are comparing average device layer temperature to the peak temperature. For a fair comparison with VM [22], in our model we use a Si substrate of 50µm thickness (vs 150µm thickness used in Hotspot). Thus, it is important to look at the temperature trend rather than the absolute
values. This combined model achieves a very high correlation of 0.99 with Hotspot’s peak temperature. This implies that if we use our proposed model, $V_M T_G$, through the non-deterministic floorplanner iterations the same better temperature-aware solutions are chosen as they would be when using $T_{peak}$ evaluated by the much more time-consuming Hotspot tool. This re-iterates that our model can quickly and correctly guide the 3D floorplanner in finding a better thermally optimized 3D solution.

### 9.7.3 3D floorplanning with $V_M T_G$

**(i) Delay based on room vs device layer temperatures**

There could be a significant underestimation in evaluated interconnect delay when thermal impact on interconnect performance is not considered. Table 23 compares delays calculated with RC parameters at room temperature, $\text{Delay}_{\text{room}}$, and at evaluated actual temperatures, $\text{Delay}_{\text{Tavg}}$. It shows the approximate percentage of underestimation of interconnect delay in GSRC benchmarks due to using the interconnect parameters’ values at room temperature.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>n100</th>
<th>n200</th>
<th>n300</th>
<th>n100 _exp</th>
<th>n200 _exp</th>
<th>n300 _exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Delay}_{\text{room}}$ (ns)</td>
<td>0.51</td>
<td>0.85</td>
<td>1.37</td>
<td>71.3</td>
<td>81.6</td>
<td>100.8</td>
</tr>
<tr>
<td>$\text{Delay}_{\text{Tavg}}$ (ns)</td>
<td>0.71</td>
<td>1.16</td>
<td>1.90</td>
<td>101.7</td>
<td>117.1</td>
<td>145.6</td>
</tr>
<tr>
<td>percentage underestimation</td>
<td>37.6%</td>
<td>37.3%</td>
<td>39.2%</td>
<td>42.6%</td>
<td>43.5%</td>
<td>44.4%</td>
</tr>
</tbody>
</table>
Please observe that there is at least 38% underestimation on average in GSRC benchmarks and this underestimated delay increases to 43% in larger benchmarks (GSRC expanded). This emphasizes the importance of considering the ignored impact of vertical thermal gradient on 3D interconnect performance with continued scaling beyond Moore. Our 3D floorplanner thus, incorporates these thermal effects with fitness functions given by Eq. 9.14 and Eq.9.15 and the results are presented below.

(ii) CF1 vs CF2

The influence of TGV on quality of 3D floorplanning solutions is demonstrated using GSRC benchmarks in Table 24. In Table 24, we compare optimized interconnect delay values using CF1 and CF2. All peak temperatures in Table 24 are evaluated on final floorplans using the Hotspot tool.

Table 24: Comparison of total interconnect delay using CF1 and CF2 in GSRC

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Case1: $\text{CF1}<em>{\text{norm}} = f(A, D</em>{\text{norm}})$</th>
<th>Case2: $\text{CF1}<em>{\text{TGV}} = f(A, D</em>{\text{TGV}})$</th>
<th>Case3: $\text{CF2} = f(A, D_{\text{TGV}}, \text{TGV})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Area (mm$^2$)</td>
<td>Delay$_{\text{TGV}}$ (ns)</td>
<td>Temp. (K)</td>
</tr>
<tr>
<td>n100</td>
<td>0.061</td>
<td>0.714</td>
<td>567</td>
</tr>
<tr>
<td>n200</td>
<td>0.080</td>
<td>1.168</td>
<td>576</td>
</tr>
<tr>
<td>n300</td>
<td>0.125</td>
<td>1.908</td>
<td>548</td>
</tr>
<tr>
<td>Average</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
We analyze three cases to optimize the delay and generate final 3D floorplans. The area, delay and peak temperature of the final floorplans are reported in Table 23. The device layer temperatures are estimated through our fusion model in each floorplanning iteration and are used to calculate the interconnect parameters at their respective layer temperatures. These update wire RC are then used to calculate the delay in Table 23. The reported peak temperatures are calculated on the final floorplans using the Hotspot tool. An overview of the cases analyzed is given below:

Case 1: CF1 and delay parameters calculated at room temperature parameters, Case 2: CF1 with delay parameters at device layer temperature and Case 3: CF2 that includes TGV to the objective function and delay’s parameters evaluated using device layer temperatures.

Case 1: (CF1room) The 3D floorplans generated for this case are our baseline case (CF1room). This is because, all prior works on delay-aware floorplanning use room-temperature parameters to calculate the delay. However, as it was explained before DelayTavg, not Delayroom, is a much better estimation of the delay value in actual design. Therefore, for a fair comparison between the three cases, we report the delay computed at the device layer temperature for Case 1 as well. Please note that the fitness function is still guided by the delay at room temperature in this case.

Case 2: (CF1Tavg) The delay component, in the cost function is calculated using parameter values at device layer temperature. It is important to note that the total interconnect delay is better optimized when the thermal impact on interconnect delay is considered. A reduction of 2.6% is achieved in the thermal-aware interconnect delay when delay at device
layer temperature is used to optimize the 3D floorplan as compared to using $D_{room}$ in fitness evaluation. This reduction is attributed to better packing and better heat spread as well. This is evident from the 0.7% decrease in peak temperature and very minimum increase in area.

Case 3: (CF2) TGV is added to a cost function to better optimize heat distribution. With CF2, the peak temperature reduces by 2.5% and delay decreases by 5%. The reduction in interconnect delay in this case is attributed to the decrease in temperature due to thermal optimization using TGV. These improvements in delay are in addition to the already reported reduction in interconnect delay reported in [64] with delay-aware floorplanning as compared to the typical wire length-aware floorplanning. Even from a peak temperature perspective, our results are better optimized as compared to [20][21] without having to reserves or re-distribute white spaces. To better understand the role of TGV in solution

Fig. 59. Impact of using TGV as an optimization parameter in fitness evaluation. Thermal hotspots are either reduced or better spread out
selection of 3D floorplanning, we show the thermal maps of final floorplan of a 4-layer 3D IC of the n100 benchmark in Figure 59. It is visually evident that using CF2 in Figure 59(b), the concentrated hotspots using CF1_{Tavg} (Figure 59(a)) are either reduced or are better spread out. Therefore, it suggests that including TGV in the cost function helps to generate better thermal-aware solutions.

9.7.4 Influence of interconnect density and temperature on interconnect delay and buffer count

The combined impact of interconnect density and device layer temperature can be observed in larger benchmarks that require buffer insertion. The results of total interconnect delay and estimated buffer count in _expanded GSRC benchmarks are compared in Table 25. Three cases are considered here using CF2 cost function. These cases include delay evaluated using CF2 - (i) D, delay at room temperature, (ii) DT, delay at device layer temperature and (iii)DTC delay at device layer temperature and varying wire capacitance. Since TGV was shown to be an important objective factor in the previous discussion, cases 2 and 3 in this section include TGV in 3D floorplanning cost function. The first case $D$, is our baseline for comparison. Delay at room temperature is reported in this case. The second case, $DT$, helps to understand the impact of temperature alone on total interconnect delay and buffer count. The combined impact of interconnect density and temperature can be seen in the third case, $DTC$.

In Case 1, One can observe the significant underestimation of delay in case of room temperature when compared to Case 2 and Case 3. The substantial increase in the buffer
count of 4.5x and 3.6x in Case 2 and Case 3, respectively, is noteworthy. This is because when a vertical thermal profile is considered the buffer insertion length (BIL) is shorter and varies on each device layer as compared to the baseline’s fixed BIL. The non-uniform wire density considered in Case 3, appears to counteract this increase in delay and in the number of buffers in Case 2. This reduction is caused by the increased wire spacing on top and bottom layers due to smaller routing congestion. The buffer counts reported in Table 25 are significantly lower than values reported in [85][71]. This is because instead of segmenting all wires, we only segment a wire to insert buffers if it achieves a reduction in delay compared to the unbuffered wire. The buffer counts in Table 25 indicate the increased error in estimating buffers when room temperature parameters are used and thus need to be further explored.

Table 25: Total delay and buffer count comparison in expanded GSRC benchmarks using CF1 and CF2. TSV_{DIA}=3\,\mu\text{M}, TSV_{AR}=10, TSV_{P}=6\,\mu\text{M}.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Room temperature (D) Case 1: CF2 = f(A,D_{room})</th>
<th>Impact of T_{\text{eq}} (DT) Case 2: CF2 = f(A,T_{\text{room}},D_{\text{TSV}})</th>
<th>Impact of T_{\text{eq}} and variable C_{\text{eq}} (DTC) Case 3: CF2 = f(A,D_{\text{room}},C_{\text{mon}},D_{\text{TSV}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>Delay (room) (ns)</td>
<td># buffers (room)</td>
<td>Delay (T_{\text{eq}}) (ns)</td>
</tr>
<tr>
<td>n100_exp</td>
<td>73.66</td>
<td>495</td>
<td>94.25</td>
</tr>
<tr>
<td>n200_exp</td>
<td>83.3</td>
<td>1520</td>
<td>123.45</td>
</tr>
<tr>
<td>n300_exp</td>
<td>100.86</td>
<td>2340</td>
<td>137.17</td>
</tr>
<tr>
<td>Average</td>
<td>1.0</td>
<td>1.0</td>
<td>+37.64%</td>
</tr>
</tbody>
</table>

Though the overall interconnect delay decreases with DTC, Case 3, when compared to DT, Case 2, it is equally important to understand its impact on individual device layers.
Therefore, the wire distributions of buffered interconnect delay, in n100_exp benchmark, on each device layer in all three cases (D (blue), DT (red), DTC (green)) are shown in Figure 60 along with the total number of nets on all device layers in each case. Device layer (DL) 0 is the farthest from heat sink, and DL 3 is closest to heat sink. The x-axis represents the wire delay and the y-axis represents the number of nets in a given wire delay range. The number of nets are computed as sum of 3D nets and 2D nets. A net is still counted as one net after it is divided into smaller segments by buffer injections. So any changes in the number of nets between the three cases is due to different distribution of modules between device layers.

To minimize total delay, DTC (Case 3) reduces the number of long nets as compared to Case D and Case DT on the hottest device layer, DL0, and increasing the number of long interconnects of the coolest layer, DL3, this is because the increase in wire resistance due to temperature reduces the BIL, thus reducing the interconnect delay on DL3. DL2 in Figure 60(b) is more wire spacing limited than any other device because it’s the closest intermediate layer to heat sink (desired for high power modules) and thus has high wire-density resulting from the intra-layer nets and as well as the inter-layer nets spanning the 3D IC. This will result in long delay wires on that layer. When TGV is used and DTC is optimized through the floorplanning optimization cycle, many nets with medium-high delay in Figure 60(b)) are redistributed to DL3 (Figure 60(a)) and DL1 (Figure 60(c)). This is because DL3 has a lower temperature compared to any other layer and DL1 has better spacing compared to DL2. These results indicate the importance of using DTC as a floorplanning optimization goal to improve the interconnect distribution on each device.
layer of a 3D-IC without aggressively increasing the area. As the wire RC variations impact the number of inserted buffers and thus impact the wire delay, DTC does a reliable job of improving the accuracy of the estimated delay and buffer count in 3D-IC.

In Figure 61, the total interconnect delay is shown in all three cases of D, DT and DTC. Though, the total interconnect delay increases by 12.7% in n100 GSRC benchmark due to rising temperature profile. But then, it reduces by 14.6% due to the non-uniform wire density when compared to DT. Therefore, the true increase in wire delay due to temperature
is 3.8%. If either of the factors of temperature or interconnect density are not considered, the wire delay is significantly either over- or under-estimated leading to severely sub-optimal solutions.

![Diagram showing total interconnect delay distribution in n100_exp benchmark at room temperature (D), at device layer temperature (DT) and at device layer temperature along with variation in wire capacitance.](image)

Fig. 61. Total interconnect delay distribution in n100_exp benchmark at room temperature (D), at device layer temperature (DT) and at device layer temperature along with variation in wire capacitance.

In addition, it is noteworthy to mention that our runtime penalty is less than 1X using the proposed thermal-delay-aware floorplanning in the non-thermal 3D optimization. In a few cases, we even see reduced runtimes with CF2 than CF1 as the floorplanner tends to find better solutions with reduced peak temperatures faster due to quick early pruning of unfavorable solutions. Since the change in vertical thermal gradient is not significant with every run, one can safely use our model to evaluate the temperatures every 100 iterations to further reduce runtime despite its already acquired speed.

### 9.8 Conclusion

In this chapter we have shown how the temperature dependence of interconnect resistance and wire capacitance calculated based on actual wire density impact the 3D
floorplanning solution quality. The proposed design strategy for a more realistic performance evaluation during the floorplanning helps in achieving the more accurate optimization of the circuit performance. With scaling, the wires are much more resistive. Therefore, without including the vertical temperature profile a prediction of 3D ICs performance could be strongly over- or underestimated. It also means that quality of 3D layout solutions generated during early optimization stages of physical design can be questionable. We also showed that the optimization of interconnect performance, by using buffer insertion, may give rise to strongly different results over the operating range of vertical temperature profile of the 3D IC.

Whether the interconnect performance and buffer count are elevated or reduced, our results show that considering the impact of temperature or interconnect density individually will not give a real picture. In order to guide the 3D floorplanner for better optimal solution selection, the combined impact of these factors is needed. Therefore, it is crucial to incorporate thermal effects and wire density into analysis and design. It is important to note that the main aim of the proposed fast true thermal-aware interconnect performance evaluation technique is to efficiently prune thermally unfavorable 3D solutions using a more realistic interconnect delay evaluation. Our results provide new insights into the thermal design space of 3D ICs, and the important role of thermal impact on design parameters in generating improved quality layouts.
Chapter 10

Thermal Management in 3D Designs using Graphene based TIM to counteract thermal impact on 3D interconnect performance

This chapter is published in IEEE-NMDC, 2018.


10.1 Introduction

The previous chapter focused on thermal management of 3D ICs in early physical design stages by adopting thermal-delay aware floorplanning. This chapter investigates at the package level, an additional intermediate layer and its suitable materials to help alleviate the thermal dissipation problem in 3D ICs. Few works in literature discuss the suitable materials for the TSVs for effective vertical heat conduction. Carbon nanotubes (CNT) are one of the most widely-accepted among them. Other papers focus on the design and materials of the heat spreaders in 3D ICs. With its very high thermal conductivity of 3000-5000 W/m-K, graphene was early explored in 2D designs as a material for heat spreaders. Barua et al., [95] explored graphene based heat spreaders in 3D ICs and discussed simulation results of monolayer and few layer graphene (FLG) in substantially reducing the on-chip temperature. Though the thermal interface material’s (TIM) role in 3D ICs is little known, authors in [96] recently investigated graphite based TIM in the 2D chips.
Authors in [97], consider the 3D IC configuration and in addition to the heat spreaders and TIM, insert an intermediate layer (IL) of graphene at each device level in 3D IC, as shown in Fig 62 (b). They claim that it will help in EMI shielding and effective heat spreading. However, no practical or theoretical results have been published in support of this claim.

In this final chapter, we investigate various 3D IC configurations with three different TIM and IL materials. This enables the exploration of thermal management in 3D ICs at package level. We consider monolayer graphene, graphite and copper, and we vary thickness of these layers to see if IL of graphene really helps in heat spreading. We also discuss the need for optimization in thermal conductivity and thickness needed to see a reduction in chip peak temperature.
10.2 Graphene/Graphite/Copper Intermediate Layer Properties

With their high thermal conductivity, graphene and graphite can aid the state-of-the-art thermal optimization techniques in 3D ICs. We investigate how an additional carbon-based Intermediate layer, IL in 3D IC configuration, will help in reducing the unacceptable high peak temperature. The thermal properties of the considered IL materials are thus discussed. The thermal conductivity, $\kappa$ of a material relates the heat flux per unit area, $q(W/m^2)$ to the temperature gradient as given in Eq 10.1.

$$q = -\kappa \nabla T$$  \hspace{1cm} (10.1)

The negative sign in the relationship indicates the heat flow from high to low temperature. Graphene is known for its superior heat conducting ability with a very high in-plane (along x-y plane) thermal conductance of 3000-4000 W/m-K, as given in Table 26 and shown in Fig. 63. Due to this worthy thermal property of graphene, Du et al., [97] put forth the idea of better heat dissipation in 3D-IC with an inter-die layer of graphene. However, it is important to note that the heat flow in cross-plane (along z-axis) of graphene is weak and limited by the inter-plane van der Waals interactions. The high thermal conducting property becomes merely $\sim$6 W/m-K for cross-plane conduction and actually becomes a vertical thermal dissipation bottleneck.
On the other hand, >1000 W/m-K thermal conductivity of graphene for lateral heat spread is achievable only in its purest form, which is challenging to fabricate and also cannot be suspended freely in 3D ICs. When graphene is supported by SiO₂, the in-plane heat conducting property of graphene degrades to ~600 W/m-K [98], yet still higher than other metals like copper with κ of 389 W/m-K. With κ of 500 W/m-K, with its relatively ease of fabrication, graphite, another possible IL material, is thus also considered in our experiments to enable a good density of the carbon material for vertical heat conduction. Though metals like copper have lower in-plane κ value compared to graphene, it has to be noted that unlike graphene, copper exhibits the same κ in cross-plane as well. Hence, these three materials are used in our work to investigate the effect of additional IL in 3D IC heat removal in the following sections.
Table 26: Material thermal properties in 3D-IC

<table>
<thead>
<tr>
<th>Material</th>
<th>Thermal conductivity (W/m-K)</th>
<th>Thickness(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure Graphene</td>
<td>4000</td>
<td>3.35E-10</td>
</tr>
<tr>
<td>Supported Graphene (on SiO₂)</td>
<td>600</td>
<td>3.35E-10</td>
</tr>
<tr>
<td>Graphite</td>
<td>500</td>
<td>3.35E-10</td>
</tr>
<tr>
<td>Copper</td>
<td>389</td>
<td>3.35E-10</td>
</tr>
<tr>
<td>Si substrate</td>
<td>142.8</td>
<td>0.00078</td>
</tr>
<tr>
<td>TIM</td>
<td>4</td>
<td>2.00E-05</td>
</tr>
</tbody>
</table>

10.3 3D-IC Layer Configuration Test Cases

We use HotSpot [1] to simulate temperature distribution, in GSRC benchmarks, for various configurations of 3D IC layers with the IL introduced between the TIM and Si substrate. The test cases considered in our analysis are given in Table 27. With Fig. 62(a) being the baseline (TC 0) configuration as highlighted in Table 27, we look at cases with IL introduced (Fig 62(b)) of different materials and thickness.
Table 27: Simulated test cases using HotSpot Tool

<table>
<thead>
<tr>
<th>Test Case #</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC 0</td>
<td>Default TIM</td>
</tr>
<tr>
<td>TC 1</td>
<td>TIM + Graphene</td>
</tr>
<tr>
<td>TC 2</td>
<td>TIM + Graphite(with TIM thickness)</td>
</tr>
<tr>
<td>TC 3</td>
<td>TIM + Copper(with graphene thickness)</td>
</tr>
<tr>
<td>TC 4</td>
<td>TIM + Copper(with TIM thickness)</td>
</tr>
<tr>
<td>TC 5</td>
<td>Graphite-based TIM</td>
</tr>
<tr>
<td>TC 6</td>
<td>No TIM, only monolayer Graphene IL used</td>
</tr>
</tbody>
</table>

In TC1, an IL of monolayer graphene is used and simulated with a $\kappa$ of 600 W/m-K. While this case suffers from the weak van der Waals interactions, we consider graphite as IL material with thickness equivalent to the TIM to enable vertical heat conduction. Copper, with its property of maintaining the same $\kappa$ value in-plane and in cross-plane direction, is also studied with varying thickness of TIM and monolayer graphene. TC 5 and 6 are variations of the baseline configuration with different TIM materials and no IL. We also investigate for the critical thickness of the IL beyond which the 3D-IC peak temperature will not decrease any further. The benchmark floorplan for each material case is simulated. We start with a graphene monolayer and then increase the thickness by a single layer for each run until the peak temperature reduction is saturated. These cases provide a useful insight for 3D IC designers to understand how the temperature of a 3D IC is altered with these materials and configurations. All the thermal simulations are done with
Hotspot V6.0 tool [1] at 45nm technology node. GSRC benchmarks of n100, n200 and n300 are simulated with the above configurations. These benchmarks come with only block level connectivity data and no power density information. We thus assume power densities in range of 0.9 to 2.0 W/mm$^2$ which are randomly generated for each block. We take into account the block size for estimating the block power and a product of the generated power density and block area is used for the total power of the block. A heat sink is assumed to be on top of the 3D IC stack in this work.

### 10.4 Experiment Results

The floorplanner and the HotSpot V6.0 tools were performed on a 4xDual Core Sun SPARC IV at 1.35 GHz and total 32 GB RAM. The generated final floorplans are used for thermal evaluation in HotSpot tool.

Our initial intent was to observe peak temperature reduction in a 3D-IC when a monolayer-graphene is inserted at the interface of inter-die in between TIM and Si. However, when simulated, the peak temperature spiked up unacceptably. Due to this increase, graphene’s impact on lateral heat spreading could not be observed. It was evident from the results that this unusual behavior of graphene is because of its very low thickness (one atom thick).
Fig. 64: Peak temperatures in the n200 benchmark for different configurations. Graphite based TIM achieving the lowest peak temperature can be observed in TC5.

The corresponding increase/decrease of peak temperatures in each configuration compared to the default case for the n200 benchmark are given in Fig 64. From test cases 1, 3 and 6 it is evident that, irrespective of the material used for the intermediate layer, the peak temperature of the 3D IC will continue to escalate if the material does not have enough thickness. This is due to the thermal resistive and capacitive components used in the node temperature calculation, given in Eq. 10.2.

\[ C_{th} = \text{area} \times \text{thickness} \times \text{volumetric heat capacity} \]  

(10.2)

With the direct proportionality to thickness, a low thickness layer cannot store heat effectively and heat transfer will be only due to the thermal resistive component. Hence low thickness layers are not able to remove heat vertically, and all the power dissipated
stays within the Si layer beneath it, resulting in unacceptable temperature rise. The layer thickness was thus increased in steps, to identify the critical thickness beyond which the peak temperature reduction will saturate. Multiple simulations are run with TC1, 2 and 3 configurations starting at monolayer of graphene and increasing the layer thickness by one graphene monolayer for each run until the peak temperature saturates. Results are shown in Fig. 65. Even with the lowest thickness, IL of copper achieves the lowest temperature due to its even thermal conduction property in both, in-plane and cross-plane, directions. Interestingly, all three materials saturate at similar thickness and nearly like peak temperatures.

Fig. 65: Peak temperature reduction with increasing IL thickness of three different materials in the sn200 benchmark.
Our final runs include TC5 with graphite based TIM. No IL is inserted in this case. Compared to TC1 with TIM and graphite as IL, TC5 achieved very good peak temperature reduction of 12.84% as shown in Fig. 63. This analysis gives us an insight to the minimum thickness required and which material’s thermal properties best suit for heat removal in 3D IC at packaging level.

10.5 Chapter Conclusions

The discussion presented in this chapter aims to understand the impact of graphene or graphite based IL and TIM in 3D IC thermal management at packaging level. While this is achieved merely by the material properties, when implemented in conjunction with other state-of-the art cooling techniques including, micro-channel cooling, thermal aware floorplanning or graphite-based heat spreaders, a further reduction in 3D IC peak temperatures can be achieved. For effective vertical heat conduction, a material with decent thermal conductivity and with good thickness is enough to absorb the heat from the layer beneath it. However, low thickness of a material with a high thermal conductivity will hamper the heat removal process rather than aiding it.

Lastly, we propose to further investigate the use of a very promising graphite-based TIM as an alternative to compensate poor heat dissipation exhibited in 3D ICs. Simulation results show a peak temperature reduction of up to 56°C. It suggests that, for effective
Thermal management, this might be a potential cost-effective and an easy to fabricate method compared to graphene. The simulation results obtained are important for 3D IC designers to take early design decisions and alleviate hotspots even without resorting to other heat removal techniques.

****** END OF PART B ******
The increasing complexity and interconnect density are driving the development of more advanced VLSI packaging and interconnection approaches. These include three dimensional integrated circuits (3D-IC) with through silicon vias (TSVs) and emerging nano-scale devices fabricated with Carbon-nanotubes (CNTs). The vertical stacking of heat dissipating layers in 3dIC results in increased heat trapped between the layers. With no effective heat removal techniques, the increased power density can greatly deteriorate the 3D interconnect performance. On the other hand, having an order of magnitude better energy delay product (EDP) compared to Si CMOS logic, CNFET can be used to build highly energy-efficient integrated circuits. CNT specific variations due to limited control over their growth process hinder the benefits offered by the CNFET technology. Therefore, this thesis is organized two-fold to explore the possibilities for “More than Moore” at device level using CNFETs and as well as package level using 3DICs.

In the first part of the thesis, we develop methods focused on CNFETs and CNT induced variations. Methodologies are proposed for statistical delay evaluation and to
reduce power increase due to redundancy in CNFET based circuits in the presence of CNT variations. For the first time, CNT length variation is investigated and its impact on functional yield is also studied. The second part of the thesis acknowledges that the thermal performance of 3D chips is directly controlled by the thermal quality of the generated floorplan and shows that ignoring the impact of temperature on interconnect delay, power and buffer estimation in the evaluation process can result in severely sub-optimal solutions being generated. Models for fast thermal goodness evaluation of 3D layouts are proposed and integrated in the 3D floorplanning flow. The proposed thermal evaluation models are also validated against the more accurate simulation based tool, Hotspot using a correlation factor.

11.1 Thesis Conclusions

Part A. Circuit-level design methodology for reliable and variation-tolerant CNFET circuits under CNT variations

In Chapter 3, a statistical approach is proposed to predict more accurately the critical path delay and functional yield compared to the worst-case technique. This approach enables a realistic prediction of functional yield based on a more realistic distribution of critical path delay variations in fabricated chips.
In Chapter 4, a calculated minimum CNT redundancy is used at the path level to improve functional yield for a given failure rate. We focus our evaluation on a critical path that in reality, due to metallic tube presence, could be a different path in each circuit. We limit the redundancy to a minimum required, to avoid an unnecessary increase in the channel area. The presence of metallic tubes alone increases the worst-case critical path delay by nearly 60% as a result of which, the worst-case approach would predict incorrectly a very low functional yield. Our approach indicates that up to 99% yield can be achieved with less than 10% of the initially present metallic tubes and for only slightly increased acceptable critical path delay (1.3 x ideal case value). For 20% of the initially present metallic tubes, additional design techniques have to be adopted to increase the yield.

In Chapter 5, we show that adding redundant CNTs to all FETs in the circuit to reduce critical path delay and achieve a good functional yield, unfortunately results in an increase in power of up to 8%, as compare to no tube redundancy, in tested ISCAS ’85 logic benchmark circuits. To reduce the power increase due to redundancy, we propose adding redundant tubes only to transistors on critical paths, ROCP. It was shown that the ROCP approach reduces the increase in the power dissipation, without degrading the functional yield as compared to all transistor redundancy, RAF. The increase in power due to redundant tubes can be reduced to <2% with our approach is ISCAS’85 benchmark circuits. Our results show that the power minimization is achieved without degrading the functional yield. Our ROCP approach is driven by the statistical repetition of critical paths in the circuit and their frequency of being critical. With our fast and efficient methodology
of identifying all critical paths under variations, we are able to capture up to 99.99% of all possible critical paths in a given circuit with reduced runtime. In addition, the impact of any path omitted due to early equality point determination will be minimum, as they are not captured due to their very low frequency of being critical. The savings achieved in PDP and improved delay constraint for functional yield estimation with ROCP, are promising for the commercialization of CNFET based circuits with high-quality, high yield, high throughput and low power consumption enabling possible earlier technology adoption.

With the proposed minimized tube redundancy, we are able to reduce the delay increase tolerance above the ideal case critical path delay, and still have a good yield. The minimum tube redundancy facilitates the design of variation-tolerant CNFET-based circuits through more realistic evaluation of delay and functional yield in the presence of variations.

When \( P_m < 1\% \), the variation in critical path delay is totally attributed by CNT diameter variation and no redundant tubes are needed. Yet, a high functional yield is achieved when only diameter variation is present.

It is interesting to note that, when 10% metallic tubes are initially present and removed, the impact of CNT diameter and CNT count variation subsides with proposed CNT redundancy. The functional yield is improved from 0% to >95% when delay degradation tolerance is reduced to 1.05X in this case. 100% functional yield is achieved for lower percentage of metallic tubes and higher delay degradation tolerance, due to a good packing density of CNTs in the channel.
Chapter 6 investigates circuits with correlated CNFETs for CNT length variation. In the literature, it is assumed that all transistors in correlated sets are identical, due to sharing the same sets of tubes. Therefore, all transistors in the correlated sets have the same failure probability and the same drive current. However, though CNTs reach long lengths, not all grow to be equally long. CNTs grown on the substrate will reach various lengths due to early catalyst precipitation and varying CNT diameter. Thus, assuming 100% correlation in correlated-CNFETs is too optimistic. In this chapter we analyze the impact of variation in CNT length on functional yield and performance of CNFETs. Our results show that variation in tube length can increase failure probability of rows of correlated transistors by 85%, and increase variation in total delay by 1.51X when compared to correlated CNFETs without variation in CNT length. The impact of CNT length variation on gate delay with and without tube diameter and tube count variations is analyzed.

The following papers are published related to these topics:


Conclusions and Future Scope

10.1109/NANO46743.2019.8993879. [41]


10.1109/NANO47656.2020.9183538. [49]

Part B. CAD approaches for fast thermal goodness evaluation and thermal management in 3D ICs

Several works in the literature addressed the problem of thermal management in 3D ICs at various stages of physical design flow including floorplanning and packaging. While most of the research is focused on reducing the chip temperature, the impact of the vertical thermal gradient in evaluating the interconnect power and performance remains less explored and is addressed in this work. To avoid sub-optimal solution selection in the
3D floorplan optimization cycle, we seek to improve the accuracy of evaluation of the
goodness of a 3D floorplan by generating more realistic 3D layouts. This is achieved
through considering the thermal impact and varying wire density of 3D chips on 3D
interconnect performance.

In Chapter 7, we present a first-hand analysis of how the resistance of 3D
interconnect delay varies with device layer temperatures. We noticed that weak increase in
wire delay with temperature reduces the buffer insertion length and rises the number of
needed buffers. More buffers with a considerable increase in buffer delay and leakage
power impacts interconnect performance. It is shown that without including a vertical
temperature profile a prediction of 3D ICs performance and power dissipation could be
strongly over- or underestimated. It also means that the quality of 3D layout solutions
generated during early optimization stages of physical design can be questionable. The
percentage of overestimation and underestimation depend on location of the heatsink, the
temperature profile and the benchmark itself. In tested benchmarks, we noticed up to 55%
underestimation of the interconnect delay, 31 % in buffer count and 63% and higher in
interconnect power consumption as compared to typically used room-temperature
interconnect parameters' values.

In Chapter 8, a new 3D thermal goodness evaluation model is proposed for faster
evaluation of thermal goodness without computing chip temperature in 3D ICs. Runtime
penalty for thermal evaluation is a huge challenge in early 3D layout exploration. We
successfully show that this individualized model can outperform other thermal-evaluation algorithms to produce thermally optimal 3D floorplan designs in just a fraction of runtime consumed by other works. The unique strength of our technique compared to state of the art approaches is that, we do not rely on solving any heat equations in this method, yet determine the thermal goodness of two given 3D designs with the available power density information.

A scalable approach is proposed for layout to $n \times n$ grid conversion of “m” layers block power densities with no limit on grid size and number of device layers used. This technique uses a block to grid overlap model to distribute the block area and power to the underlying grid sections. The block power contribution to the grid relies on the block to grid overlap ratio. With our testing statistics, we are able to show a correlation coefficient of 0.99 between our prediction results and HotSpot temperature evaluation. The proposed approach is over 80% faster than HotSpot’s steady state temperature evaluation. The savings in runtime enable the integration of our technique through all iterations of the floorplanner with reasonable accuracy of evaluation. The TGV (Thermal goodness) value can be used to guide the floorplan cost function and efficiently replace the peak temperature value conventionally used by other works.

During delay-aware floorplanning, the current assumption is that the wire density is the same on all layers. However, the length of wire segments on different device layers will be different due to TSV position along the net. Consequently, the distribution of wires on each
device layer will also differ resulting in non-uniform wire density on each layer. Hence, for realistic and accurate estimation of interconnect delay and buffer estimation, the variation in wire density across the layers is incorporated during performance-aware 3D floorplanning in Chapter 9. In this chapter, we also developed a fast fusion model- a power based thermal goodness evaluation of the 3D-ICs and a vertical only temperature profile generation, to quickly evaluate the device layer temperatures through the many probabilistic floorplan optimization cycles. The proposed algorithm bridges the gap in vertical only thermal model by considering the impact of the heat from intra-layer modules as well.

The non-uniform wire density will result in varying wire capacitance across the device layers of 3D ICs. The increasing interconnect density results in increased wire capacitance. However, the increase in wire resistance due to temperature is higher than the variation in wire capacitance in the intermediate layers and vice versa on top and bottom layers. Thus, the non-uniform wire density can be used to advantage to counteract the impact of temperature on 3D interconnect performance on the device layer farthest from heat sink. We show that considering wire density in wire delay evaluation along with thermal aware interconnect parameters will result in improving the quality and yield of the 3D floorplans generated.

The proposed algorithm significantly improves the correlation (>0.9) of the vertical only temperature model with the more accurate simulation-based thermal models, like the
Conclusions and Future Scope

Hotspot tool. It is also 29X faster on a grid size of 64x64x4 for GSRC benchmarks compared to Hotspot.

We analyze the impact of using constant RC interconnect parameters at room temperature on evaluated delay and estimated buffer count when compared to parameters as a function of device layer temperatures and interconnect-density. Our results show 40% underestimation in interconnect delay on average and 3.6X-4.5X variation in buffer count using room temperature parameters. An average difference of 19% in total delay in GSRC benchmarks using the proposed true thermal-aware and wire-density-aware interconnect performance evaluation when compared to only thermal-aware delay, emphasizes the need for a more realistic evaluation the 3D interconnect performance to avoid sub-optimal solution selection.

This improved 3D floorplanning evaluation framework provides further solutions for improved estimation of buffers in 3D chips, by accurately modeling the thermal-aware delay of wires containing TSVs.

In addition, an efficient interface of white space generator is developed to bridge between our 3D floorplanner and HotSpot tool to evaluate 3D floorplan peak temperature on HotSpot and validate the effectiveness of proposed thermal goodness evaluation approach.

Chapter 10 explores the various possible materials for thermal interface and additional layers between Si for effective thermal management of 3D IC at package level. We identify
the critical/minimum thickness and suitable material needed for a good vertical heat transfer in 3DICs and propose the use of a graphite based Thermal interface material sandwiched between the IC stacks for better heat dissipation between the ICs towards the heatsink. With this architecture, we were able to reduce peak temperature by 12% in GSRC benchmarks, without the need of any micro channel cooling.

As CMOS scaling approaches its physical limits, this research on both the emerging nano-generation logic devices and fine grained integration of disparate technologies will enable early technology adoption of planar CMOS alternatives.

The following papers are published/submitted related to these topics:


The other publications indirectly related to this work:


11.2 Future work

This work addresses some of the most practically relevant design overheads and fabrication challenges introduced in 3D ICs and CNFETs. The present work can be further improved with some of the following suggested, to aid the efficiency of the proposed models.

- **Improve accuracy of statistical gate delay**: Proper load sizing/fan-out in CNFETs can be considered for more accurate delay and tube redundancy calculation.

- **CNT Length modelling**: The CNT length variation can be further investigated considering the fractured CNTs that results during the CNT transfer process to target substrate. CNT length variation can be incorporated into the redundancy optimization for better determining the required number of redundant tubes and the options for the critical length possible to avoid functional yield loss can be further explored.

- **Fractured CNTs**: Further investigate the impact of CNT fracture in transfer process to target substrate in CNT length variation
Conclusions and Future Scope

- **Suggested speed up:** In floorplan to grid translation, when we increase the grid granularity, an increased number of grid cells are occupied by a single module itself since the grid cell is very small compared to the module. In this case, if the module’s contributed area is equal to the grid cell area, we can terminate iterating through the remaining module list. As the floorplan to grid conversion is the most time consuming part, this will further speed up the evaluation process.

- **Impact of temperature on buffer performance:** While 3D wire delay is impacted by device layer temperature, the buffers are also strongly influenced by the vertical thermal gradient. Ways to dynamically incorporate the increase in buffer delay, dynamic and leakage power with temperature can considered to further improve the 3D interconnect performance evaluation.

- **Sub-threshold leakage is a strong function on temperature:** Thermal and interconnect density impact interconnect power and buffer power as well. It serves as a potential optimization goal.

- **Weightage on TGV optimization parameter:** Further analysis should be performed to improve the thermal hotspot distribution by increasing weights on TGV in 3D fitness function and analyze its impact on delay and circuit area.
Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


[72] “Available Online; BUF_X8.”


