Three Dimensional Integration Technology Using Copper Wafer Bonding

by

Andy Fan

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

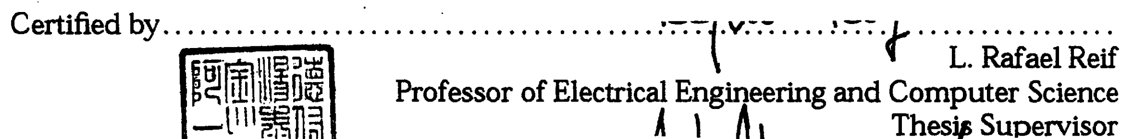
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2006
[June 2006]

© Massachusetts Institute of Technology 2006. All rights reserved.

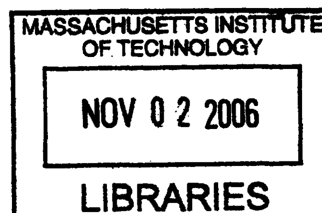
Author ....
Department of Electrical Engineering and Computer Science
May 31, 2006



Certified by ....
L. Rafael Reif
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by ....
Akintunde Ibitayo Akinwande
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students



ARCHIVES

Three Dimensional Integration Technology Using Copper Wafer Bonding

by

Andy Fan

Submitted to the Department of Electrical Engineering and Computer Science
on May 31, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

With 3-D integration, the added vertical component could theoretically increase the device density per footprint ratio of a given chip by n-fold, provide a means of heterogeneous integration of devices fabricated from different technologies, and reduce the global RC delay to a non-factor in circuits by using smarter 3-D CAD tools for optimizing device placement. This thesis work will focus primarily on the development and realization of a viable 3-D flow fabricated within MTL. Specifically, the presentation will attempt on answering these questions in regards to 3-D:

1. What enabling technologies were needed for 3-D to work ?
2. Does it really work ?
3. Will the "3-D heat dissipation problem" prevent it from working ?
4. What applications is it good for ?

Referring to the first item, a viable 3-D integration flow has been developed on both the wafer-and-die-level, and the enabling technologies were the following: Low temperature Cu-Cu thermocompression bonding, an aluminum-Cu based temporary laminate structure used stabilizing the handle wafer - SOI wafer bond, and tooling optimization of the die-die bonder setup in TRL. Next, nominal feasibility of the 3-D flow was demonstrated by fabricating a 21-stage and 43-stage CMOS ring oscillators, where each single CMOS inverter / buffer stage was constructed by connecting NMOS-only devices from one substrate with PMOS-only devices from a separate substrate. Proof-of-concept was accomplished when all 92 Cu-Cu bonds, 204 thru-SOI Cu damascene vias, and 56 pairs of MOSFETs communicated simultaneously to produce a 2.75 MHz (43-stage) and 5.5 MHz (21-stage) oscillators, ringing rail-to-rail at 5 V Vdd under proper Vt adjustments on the SOI-PMOS using integrated backgates.

Furthermore, to combat the perceived heat dissipation problem in 3-D, this work focused on using the Cu-Cu interlayer bond as heat dissipators, with Cu planes working as flux spreaders and Cu vias as direct heat conduits. Finally, 3-D RF passive integration onto existing chips can be made feasible, under certain device performance trade-offs, by using cobalt magnetic shielding, which offers at least a -10 dB throughout 0-20 GHz, with a max isolation of -24 dB at 13 GHz, at +4 dBm reference input power.

Thesis Supervisor: L. Rafael Reif
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Akintunde Ibitayo Akinwande
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Above all, I would like to thank Rafael Reif for his patience, guidance and support throughout not only my graduate career but also on my quality of life here at MIT. One of these days, I'll have to buy him a new living room carpet for that red wine spill I took ! Also, I would like to thank Tayo Akinwande for spending countless hours with me - sifting through my hopeless data, supporting my work, innoculating my thesis against colloquial infestation, and of course, talking about how the Dallas Cowboys are superior to the 49ers in every era, *especially* during the 1980s ! =) Special thanks also goes to Marty Schmidt for being my thesis reader and my graduate academic advisor for the past 9 years, and also putting up with me when I say, "I *promise* I won't take any more classes next semester!" during every fall and spring. And also, a big thanks to Anantha Chandrakasan for supporting me financialy in the time of need and to help bridge the transition between Rafael's ascension to "Mahogany Row" and me being in limbo !

Having spent more or less a third of my Earthly existence at MIT, I guess it won't be a cliche in saying that I've met more people during my stay here than any other times in my life. Debb, aka "Ma," was the person who got me here, who got me standing up, who for every step of the was essentially my second mom. Love ya, Ma ! Also, I've been keeping it real with Fletch from way back when... and when I mean *way* back, I meant when he and Wei got me out and I had my first (legally) pint of "real" Guinness at the Phoenix Landing on my 21st b-day. Steady as a rock, Fletch is; you can always lean on the guy. Thx man !

And speaking of Mr. Wei, Andy Wei has been the resident philosopher at MTL for quite some time. Always the consummate leader in our Friday night excursions, and was always ready to tackle the mysteries of the universe while we enjoy a few rounds of the elixirs of life (i.e. Guinness or Wild Turkey). I totally miss those timeless moments. Next, to Jim Fiorenza - how I miss those times when we're drinking cold Labatt's and watching playoff hockey on ESPN at home. And to Samuel Mertens - I won't forget those rubby parties held at our house anytime soon (and I'll see you in Alaska for your wedding)! Also, to Niamh Waldron, Isaac Lauer, and Dennis Ward has always been there with me, whether it's on ski trips, hockey, softball, soccer, enjoying a good pint of (you should know what beer it is by now) at The Burren, or just going for a coffee run at Building 4's Green Mountain Brew =). To Mr. Andy Ritenour, who I've known since Day 1 in Boston - along with Fletch, he is also an universal constant during my MIT career. Special thanks also goes to the rest of del Alamo group (Joyce, Anita, big/little Joerg), the Hoyt group (Cait, Maggie, Ingvar, Tonya), the other members of old Antoniadis group (Tony, Mark Armstrong, Keith, Ihsan, Hasan), Tayo's group (Johnny K. (go kick some butt at Columbia !), Leonard (already kicking butt at Wall Street), Ching-Yin, Liang-Yu), and of course, the old Reif group (Arif, Laura, Simon (wish he was still with us), Ritwik, Weize, Ming-hao, Peter, Wendy, Rajan,) and the newbies (Nisha, Ajay, Kuang-Neng, Chuan-Seng, Tan), along with all of Rafael's official eyes and ears throughout the years - Susan Kaufman, Diane Hagopian, and Ishara Smith. And finally, to all those I haven't mentioned due to lack of space (could take another 3 pages since I've been here so long), thank you all ! =)

Contents

1 Introduction to 3-D Technology 20

1.1 Some Perspectives on Interconnections . . . 21

1.1.1 2-D City Development . . . 21

1.1.2 2-D Microprocessor Development . . . 23

1.2 Identifying the Thesis Topics . . . 26

1.3 Executive Summary of Thesis Results . . . 27

2 3-D Integration Challenges 29

2.1 The MIT 3-D Approach . . . 29

2.1.1 An Overview of the MIT 3-D Process Flow . . . 30

2.2 Wafer-Level Integration: Bonding Uniformity Challenges . . . 33

2.2.1 Bond Microstructure . . . 33

2.2.2 Mechanical Contact . . . 33

2.3 Wafer-wafer Alignment . . . 42

2.4 Handle Wafer Release . . . 47

3 Die-level 3-D Integration 52

3.1 The Wafer/Die-Bonder . . . 53

3.1.1 The Bond Chuck . . . 53

3.1.2 The Bond Glass , Bottom Support, and the Wide-angle Objective . . . 53

3.1.3 The Mesa Pressure Plate and the Graphite Insert . . . 56

3.2 Multi-level 3-D Die Integration . . . 60

3.2.1 Epilogue . . . 61

4 Electrical Characterization: 3-D Ring Oscillators 64

4.1 Overview of 3-D Ring Oscillator Design . . . 64

4.1.1 Motivation . . . 64

4.1.2 Overall Structure of the 3-D Ring Oscillator . . . 65

4.2 Preliminary Cadence Simulation of the Ring Oscillators from Real MOSFET Data 67
4.2.1 Initial Design Considerations 67
4.2.2 Measurement of Fabricated single NMOS / PMOS devices 70
4.2.3 Simulated Oscillator Results: With Measured Vt's and First-order Parastics 74
4.3 3-D Ring Oscillator Measurements 75
4.3.1 Face-to-Face Ring Oscillators: 21-Stage CMOS, L = 3 μm 75
4.3.2 Face-to-Face Ring Oscillators: 43-Stage CMOS, L = 3 μm 88
4.3.3 Face-to-Back Ring Oscillators: CMOS Failures, NMOS success 90
4.4 Cu Parasitic Extraction with Bonded MOSFETS and 2-D Ring Oscillators 94
4.4.1 2-D Ring Oscillator Output From Cu-Bonded Substrates 94
4.4.2 Single NMOS / PMOSOutput From Cu-Bonded Substrates 101
4.5 Summary of Ring Oscillator Results 104

5 Thermal Characterization: Heat Dissipation in 3-D **105**
5.1 Overview of the Heat Dissipation Problem in 3-D 105
5.2 3-D Heat Transfer Simulations 108
5.2.1 Reference Simulation: 2-D SOI Heater 109
5.2.2 Reference Simulation: 2-layer SOI Heater 110
5.3 Simulation Comparisons: Adding Cu Planes vs. Cu Thermal Vias to the Referenc Structure . 115
5.4 Measured Results from Real Heat Structures 122
5.4.1 Reference Measurement and Thermal Calibration: Cu-bonded SOI Heaters, Excluding Thermal Vias 122
5.4.2 Comparison: Cu-bonded SOI Heaters, with and without Thermal Vias 124
5.5 Summary of Thermal Results 125

6 RF Sprial Inductor Integration and Magnetic Shielding Studies **131**
6.1 Introduction 131
6.2 Cobalt Magnetic Shielding Measurements 133
6.2.1 Reference Stucture: Al Spiral, No Shielding 133
6.2.2 Comparison of Refence Structure to Solid Co Shields 137
6.2.3 Comparison of Solid Co Shielding to Solid Al Shielding 139
6.2.4 Comparison between Different Co Shield Configurations 141
6.3 Summary of Inductor Results 145

7 Concluding Remarks **146**
7.1 Summary of Accomplishments and General Conclusions 146
7.2 Future Work 148

7.2.1 Handle Wafer Release Optimization . . . 148
7.2.2 Die-Die Bonding and Alignment: Equipment Optimization . . . 149

A The 3-D Process Flow: Detailed Explanations 150

B Handle Wafer Release Mechanism 155

C Die-Level Jig Improvements: A Detailed Description 158
C.1 Fix # 1: The Homemade Die-Bonding Chuck . . . 158
C.2 Fix # 2: Bond Glass Bow Reduction by Overpressing . . . 159
C.3 Fix # 3: Plateau Dies and Post-alignment Check . . . 160
C.4 Fix # 4: Graphite Insert Monitoring and Pyrex Wafer Substitution . . . 164

D Supplemental Results and Graphs from the Ring Oscillators 168
D.1 Unbonded Single NMOS Devices . . . 168
D.2 Unbonded Single PMOS Devices . . . 172
D.3 Face-Face, 21-Stage CMOS Ring Oscillators, Floating-body . . . 176
D.4 Face-Face, 21-Stage CMOS Ring Oscillator, with Backbiasing . . . 182
D.5 Face-Face, 43-Stage CMOS Ring Oscillators, Floating-body . . . 186
D.6 Face-Face, 43-Stage CMOS Ring Oscillator, with Backbiasing . . . 192
D.7 2-D NMOS-only Ring Oscillators . . . 196
D.7.1 Table of Results . . . 196
D.7.2 2-D NMOS-only Ring Oscillators, Vdd = 3 V . . . 197
D.7.3 2-D NMOS-only Ring Oscillators, Vdd = 4 V . . . 201
D.7.4 2-D NMOS-only Ring Oscillators, Vdd = 5 V . . . 205

E A Heat Transfer Primer for EE's 209
E.1 Thermal-to-Electrical Duality: The Heat Equation vs. the Poisson Equation . . . 209
E.2 Ohm's Law and its Thermal Duality: Fourier's Law . . . 210
E.3 Poisson equation and its duality: The Heat equation . . . 211
E.4 Spreading Resistance . . . 212
E.4.1 Mathematical Introduction and Physical Interpretation . . . 212
E.4.2 Further Mathematical Treatment . . . 213

List of Figures

1-1 Maps of ancient Athens (*left*) and ancient Rome (*right*) Taken from [1] and [2]. 21
1-2 Maps of modern-day Beijing(*left*) and Boston (*right*) Taken from http://www.beijingmap.us and [3]. 22
1-3 Die photos of Intel chips, areas approximately to scale (33 mm^2, 94mm^2, 81mm^2 respectively) 23
1-4 Die photos of Intel chips, areas approximately to scale (147mm^2, 203mm^2, 206mm^2 respectively) 24
1-5 Homogeneous 3-D implementation of a dual-core processor ? Maybe ! 26

2-1 The MIT 3-D model 29
2-2 MIT 3-D process flow: A bird's eye view 31
2-3 MIT 3-D process flow, part 1 31
2-4 MIT 3-D process flow, part 2 32
2-5 Electronics-Vision EV501 wafer bonding setup 34
2-6 Piston backstop position during resting and bonding phases. The graphite insert and wafer thickness were 2.55mm and 650μm each, respectively 35
2-7 Random delamination between the SOI-handle interface after TMAH etchback. 36
2-8 Bond diffuser geometry: Solid circular vs. donut 37
2-9 Bond diffuser geometry: Solid circular vs. donut, wafer after TMAH etchback 37
2-10 Bond diffuser geometry: Solid donut vs. solid graphite, wafer just after grindback 38
2-11 Graphite-bonded wafers after etchback, showing both global uniformity in Cu bonding and high tensile stress of annealed Cu films 39
2-12 Details of the laminate stack in reference to bow measurements presented in Table 2.2. . . . 41
2-13 Photo of PMOS wafer, after successuful handle-wafer bonding and substrate etchback after a piston-based bow compensation press. The Al release layer thickness in this sample was 10 μm. 41
2-14 Increase of via density by factor of 400 from 100 μm via to 5 μm via. 43
2-15 Wafer-wafer aligning protocol 44

2-16 Photo of bonded Cu-Cu chain resistor. For reference, the sample was prepared by a face-to-face waferbond of 2 NMOS wafers and a subsequent SF_6 / TMAH bulk Si etchback from one side. . . . 46
2-17 Photo of bonded 3-D ring oscillators and dummy solenoid structures, where the mislalignment was about 9 μm in the vertical direction. For reference, the sample was prepared by face-to-back, die-die bonding of a PMOS/NMOS pair and a subsequent 6 hr HCl acid encroachment release of the handle die. . . . 46
2-18 Bird's-eye view of the 3-D flow, emphasizing the formation, duration, and the destruction of the sacrifical handle-SOI bond 48
2-19 Constructing the laminte structure: The Al release layers can be protected from mechanical grindback and TMAH etchback by an exclusion-ring deposition followed by a non-exclusion ring deposition of Cu/Ta. . . . 50
2-20 Results of 4" wafer-level acid encroachment release. In each sample, the indicated Al thickness was that of the release layer, the handle material refers to either a 5000 A thermal oxide base or an 1000 A LPCVD nitride base, and the "Edge" value refers to the maximum bond encroachment distance. . . . 51

3-1 The MIT 3-D fabrication flow, with wafer or die-level processing options. Die-level integration is not recommended before Step 4 because it is very difficult to maintain structural integrity of the handle-SOI complex during die-level substrate etchback. . . . 53
3-2 Die-level bond chuck schematics. The chuck itself was retrofitted from a regular 6" waferbond chuck, and a sample photo of it was shown on the right. . . . 54
3-3 Die-die alignment procedure. . . . 54
3-4 Die aligner's new bottom support and the new quartz bond glass. . . . 55
3-5 Top and side view of the aligned dies, just after alignment and before bonding 56
3-6 The bonder setup during thermocompression bonding. . . . 57
3-7 A gallery of successful die-level integration. All photos were taken from a face-back, die-bonded sample with a 20 μm Al release layer. (a) and (b) shows that the misalignments between the two wafers were on the order of almost 10 μm. Photo (c) shows is a "looking-glass" structure where in regions devoid of Cu bonding pads, the lower-tier NMOS devices can be simultaneously visualized with the top-tier PMOS devices. Photo (d) shows a section of a completed 21-stage CMOS ring oscillator (when tested later, the output was flatlined but responded to changes in Vdd). Photo (e) shows a two-layer SOI Schottky heaters with satellite temperature- detecting diodes. . . . 59
3-8 The MIT 3-D die-level integration scheme. . . . 60

3-9 Construction of a 3-layer stack. In (a), two separate die-level Cu bonding / acid release loops were created on top of a base SOI substrate, thus forming a 3-layer stack. A proof-of-concept 3-layer blanket stack was made in (b) with the bonding configuration of (c), which was similar to what we had in Figure C-1. 61
3-10 SEM photo of the 3-layer stack from sample (b) of Figure 3-9. 62
3-11 Construction of a 3-layer stack. In (a), three separate die-level Cu bonding / acid release loops were created on top of a base SOI substrate, thus forming a 3-layer stack. A proof-of-concept 3-layer blanket stack was made in (b) with the same bonding configuration from (c), of Figure 3-9. 63
3-12 SEM photo of the 4-layer stack from sample (b) of Figure 3-11. 63

4-1 Cadence layout of a 21-stage CMOS ring oscillator, W_{NMOS} = 60 μm, L = 1 μm. 65
4-2 Skeleton of a CMOS oscillator schematic, showing all 4 probe pad connections. 66
4-3 A more detailed circuit diagram of a 3-D ring oscillators, showing the head-to-tail connections between the inverter I/O ports and weaving between different wafers. 66
4-4 A stripped-down layout of the same CMOS ring oscillator, this time showing the Cu bonding layer and the Cu damascene via positions only. Since this is a dark-field mask, the large, black field regions represent Cu metal coverage, whereas the white boxes denote regions of air moat isolations. Also, a group of Cu damascene vias on the lower right shorts the Vdd pad and the Cu backbias region together. 67
4-5 The basis simulation from which all other ring oscillators were designed around. Each colored curves represent the output response of a 21-stage CMOS ring oscillator (W/L_{nmos} = 60/3, W/L_{pmos} = 120/3) for a given value of Vdd. The NMOS and PMOS Vt's were assumed to be 1.23 V and -0.9 V, respectively. 69
4-6 Extraction of *unbonded NMOS* series resistance. The on-resistance of each 60 μm-wide NMOS, at Vds = 0.5 V, was plotted as a function of both the gate length L and the gate bias Vg. The total source-drain series resistance $2R_x$ was extrapolated at the approximated to be about 5 Ω, or at intersection of all curves except for the outlier line Vg = 2. 73
4-7 Simulated 21-stage, 3 μm-channel CMOS ring, with real NMOS Vt = 0.900 V and a properly-biased PMOS Vt = -0.665 V (taken from value at +10 V backbias for safe measures). 74
4-8 The makeshift face-face 3-D process: In (a), the PMOS and the NMOS substrates were bonded with the standard Cu recipe. Then, a SF_6 plasma and a TMAH etch was able to clear the backside of the PMOS substrate in (b). After the Ti backgate depositon (c) and topside via etch (d), the 3-D ring oscillators were ready for probing at the red circles in (e). 78
4-9 The makeshift Ti backgate position relative to all the other probe pads. Notice that the buried Cu-Cu bonding backplane was useless here because the top Ti backgate, sitting right on top of the PMOS BOX, has a more direct access to the PMOS gate than the Cu plane. 79

4-10 21-Stage CMOS, L = 3 μm: Vdd = 1 thru 2.5 V. Left-column plots are the signals from the 9x output buffer and the tiny traces from probing the floating Ti backgate and the "useless" Cu-Al plane's pad; right-column plots are zoomed-in traces of the aforementioned floating pads. 80
4-11 21-Stage CMOS, L = 3 μm: The plot at Vdd = 7.0 is in (a). In (b), the frequency and the peak-to-peak voltage Vpp of the output was plotted as a fuction of Vdd. Note the saturation of the oscillation frequency at high Vdd's. 81
4-12 21-Stage CMOS, L = 3 μm: The DC offset and the RMS voltage of the fast oscillations coming off of the floating Ti bckgate and the useless Cu floating pads 82
4-13 Positive DC offset on the face-to-face CMOS floating backgates was caused by the slow body-charging process 85
4-14 Saturation of the V_{rms} occurs because of fast filling / discharging of interficial charge states at the BOX-SOI interface 85
4-15 21-Stage CMOS, L = 3 μm, Vdd = +5V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) +7.5 V backbias, (e) -15 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails, and at Vdd = +5 V, V-/V+ almost resided on the rails themselves, albeit with some voltage drop from internal series resistance. 87
4-16 20 μm Al released, Face-back bonded 21-stage CMOS at L = 3 μm with +5 V backbias on the PMOS backgates. Since all output responses were flatlines, the output DC voltage Vout was plotted against the power supply voltage Vdd. This particular circuit had no Vdd-Vout shorts and was in a metastable state that inhibited oscillation. 91
4-17 20 μm Al released, Face-back bonded 43-stage CMOS at L = 3 μm with 5+ V backbias on the PMOS backgates. Since all output responses were flatlines, the output DC voltage Vout was plotted against the power supply voltage Vdd. This particular circuit exhibits an almost-shorted Vdd-Vout path near the PMOS transistors, and like the other oscillator, this sample was also in a metastable state that inhibited oscillation. 92
4-18 20 μm Al released, Face-back bonded 43-stage CMOS at L = 1 μm with floating PMOS backgates. Since all output responses were flatlines, the output DC voltage Vout was plotted against the power supply voltage Vdd. This particular circuit also had no Vdd-Vout shorts and was in a metastable state that inhibited oscillation. 93
4-19 The buried 2-D NMOS oscillators can be probed by pulling the signals up from the bottom layer with the Cu-Cu bond pads and the Cu damascene vias. 94
4-20 Simulated 2-D NMOS rings, all biased at Vdd = +4 V. In each plot, the caption "NMOS W1/L1 - W2/L2" refers to the width/length ratio of the NMOS switch (W1/L1) and the ratio of the active NMOS load (W2/L2). 95

4-21 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c), and a face-face bonded sample in (d) 97
4-22 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b) that died during processing, a face-back bonded, 20 μm Al released sample in (c), and a face-face bonded sample in (d) 98
4-23 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 99
4-24 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 100
4-25 Extraction of *face-face bonded NMOS* series resistance. The on-resistance of each 60 μm-wide NMOS, at Vds = 0.5 V, was plotted as a function of both the gate length L and the gate bias Vg. The total source-drain series resistance $2R_x$ was extrapolated at the approximated to be about 15 Ω. 102
4-26 Extraction of *face-back, 10 μm Al- bonded NMOS* series resistance. The on-resistance of each 60 μm-wide NMOS, at Vds = 0.5 V, was plotted as a function of both the gate length L and the gate bias Vg. The total source-drain series resistance $2R_x$ was extrapolated at the approximated to be about 255 Ω. 103

5-1 Poisson Eq. - Heat Eq. inequality: A quick reasoning as to why thermal vias may not be that useful in decreasing the overall temperature of a chip. See text for details. 108
5-2 Reference FEM simulation # 1: One-level, unbonded SOI heaters. Colorbar temperatures are in Kelvins. 112
5-3 Reference FEM simulation # 2: Two-level, oxide-bonded SOI heaters. Colorbar temperatures are in Kelvins. 112
5-4 Reference FEM simulation # 2: Two-level, oxide-bonded SOI heaters, isotherm plots. Heat flux lines $-\nabla(kT)$ are in blue, and temperature gradient lines $-\nabla T$ are in green. To accentuate the results, the z-axis for each plot was cut off at z = -4 μm, or at a depth 0.6 μm below the bottom interface of the BOX from the lower device tier. All colorbar temperatures are in Celsius, and "SDT" stands for "Satellite Diode Temperature" detectors. 113

5-5 Reference FEM simulation # 2: Two-level, oxide-bonded SOI heaters, heat flux plot on a vertical plane. Heat flux lines -∇(kT) are in blue, and temperature gradient lines -∇T are in green. The values on the colorbar corresponds to the thermal conductivity of the subdomains, in W/m-K. . . . 114
5-6 Structural between the reference (a), Cu-bonded but no thermal vias (b), and Cu-bonded with thermal vias (c). The colobar coding denotes tge subdomain's thermal conductivity: Pink regions were made of Cu, and brown regions were made Si. Also, SDT stands for "Satellite Diode Temperature" detectors. Also, the heat flux lines were plotted in blue, and the temperature gradient lines were plotted in green. . . . 116
5-7 Bird's eye view plot: Simulated isotherm comparisons for the reference, the 6000 A Cu-bonded with no vias, and the 6000 A Cu-bonded with thermal vias structures. . . . 119
5-8 Simulated isotherms of 6000 A Cu-bonded structure, with no vias. *These four graphs are the magnified versions of those from COLUMB (B) of Figure 5-7.* The simulated results suggests that the Cu thermal plane does spreads out the heat flux and reduces the z-axis thermal gradient when compared to the reference, non-Cu structure. DISCLAIMER: The outlines of the satellite diodes do not readily show up on the isothermal plots because their neighboring inter-subdomain temperature gradients were small. . . . 120
5-9 Simulated isotherms of 6000 A Cu-bonded structure, with thermal vias. *These four graphs are the magnified versions of those from COLUMB (C) of Figure 5-7.* The simulated results suggests that the Cu thermal vias do siphon the heat flux from its local surroundings, but they do not decrease the maximum temperature of the structure. Also, the thermal Faraday cage effect (elongation of the isotherm) can be seen from both the x-z and y-z plots. DISCLAIMER: The outlines of the satellite diodes do not readily show up on the isothermal plots because their neighboring inter-subdomain temperature gradients were small. . . . 121
5-10 (a) Schematic of Face-face bonded Triple SOI heaters. There are 3 *pairs* of current-driven Schottky heaters (3 diodes on top tier, 3 unseen diodes stacked underneath) wired in parallel, and each individual satellite diodes in the schematic are also stacked top/bottom tier pairs. (b) A layout of the reference structure to be tested. (c) A cross-sectional view of the face-face bonded reference structure . . . 127
5-11 Measured satellite diode current ("Near" or "Far") at 600 mV plotted as function of the Schottky heater power dissipation at different chuck temperatures . . . 128
5-12 Same plot as Figure 5-11, but now in semilog y-axis. . . . 128
5-13 Extracted temperatures from data collected in Figure 5-11. . . . 129

5-14 Reference Structure Measurement: (a) Face-face bonded Triple SOI heaters layout with thermal vias sandwiched between the heating elements. (b) The tempearture of the "Near" was measured and the results were compared (c) with the non-via analogue measured in Section 5.4.1. The addition of a Cu thermal via decreased the regional temperature by 12 °C. 130

6-1 Sample RF chip photo, showing five real estate-consuming RF spiral inductors. Photo taken from http://www.techonline.com 132

6-2 S21 measurement of reference Si crosstalk structure. Notice the substrate crosstalk increases with frequency because of capacitive coupling. 135

6-3 S11, S21, and S22 measurement of reference structure in Smith Chart form. Notice that the inductor ports were slightly asymmetric because Port 2's entrance geometry was very different than Port 1. 136

6-4 S21 measurement of reference Si and of a solid cobalt shield. The blue curve represents the reference crosstalk structure with no shields, and the green curve represnts the attenuation of crosstalk with a solid Co magnetic shield. 138

6-5 S21 measurement with new reference (blue trace) and a 2 um Al shield (green trace). 140

6-6 A schematic list of all the Co shield configurations tested. The rectangular and square "patched" Co shields were electrically floating, while in all other shield configurations the Co metal was shorted to the ground test pads. This list is in the same order as the legend in Figure 6-7. . . . 143

6-7 S21 measurement of reference Si and various configurations of cobalt shields. The solid cobalt shield again offered superior magnetic isolation when compared to all other Co shield configurations when considering the entire span of frequency range. 144

A-1 MIT 3-D process flow, part 1 150

A-2 MIT 3-D process flow, part 2 151

B-1 Extremely simplified overview of the handle wafer release process: Dissolution, agitation, and separation 156

B-2 An extremely simplified overview of the liquid penetrance problem: A traffic jam of trapped H_2 bubbles henders mass transport of acid and the bubbles to and from the capillary 157

C-1 The effecs of chuck flatness on the die-bonding quality. (a) Left photo shows the mesa chuck result, (b) middle photo shows the home-made chuck result, and (c) right phot shows the combination of home-made chuck and a 3.9 mm quartz fragment used as a makeshift bond glass. 159

C-2 Bond glass bow: The setup and its corresponding force diagram 160

C-3 The left photo (a) displays the bonding quality before backstop collar and downforce optimization, and the right photo (b) displays the optimized force parameters. For both cases, the home-made chuck elevation was h_{chuck} = 0.46 mm, the 2-die combination thickness was h_{die} = 1.24 mm, the quartz bond glass was 2.303 mm, and a *new* graphite insert thickness was 1.0 mm. Each ruler tick mark in the photos correspond to 1 mm. 161

C-4 Essence of die-die θ-misalignments. In (a), given an angular misalignment of θ between the two substrates, the size of the projected misalignment arc lengths S_1 and S_2 ar related to the moment arm dimensions r_1 and r_2 by the equation $S = r\theta$. In (b), the larger r-value on a 6" wafer registration set creates a larger S. In (c), a small r-value on an 1" die results in an imperceptiable value of S. 162

C-5 Mechanical travel limits of the microscope objective and their small field of views both createa dark zone that limits the accuracy of die-to-die θ-alignment. The dimension of the outer "wafer-like" outline was not drawn to scale. 163

C-6 Sacrificing some die area, the plateau region created by area reduction of the second die increases the accuracy of die-die alignment and faciliates a easy method of post-alignment verification. 164

C-7 A set of 6 die-bonding pairs made in mid-April, 2006. The released dies in (a)-(f) were bonded in succession with identical mechanical bonding parameters indicated within the main text. The thickness of the die stack prior to bonding matches those mentioned in the captions for C-3, and each ruler tick in the photos correspond to 1 mm in length. 165

C-8 A second set of 7 die-bonding pairs made two weeks after the first set. Upon breaking the quartz bonding glass after pair (f), a 2.00 mm pyrex glass subsitute was used in pair (g), in which the pyrex glass also broke. 166

D-1 Unbonded NMOS Id-Vd plots. The width/length ratio in microns for each NMOS were: (a) 80/1, (b) 60/1, (c) 20/1, (d) 10/1, (e) 5/1, (f) 60/3. 169

D-2 Unbonded NMOS Id-Vg plots, with subthreshold slope extraction. The width/length ratio in microns for each NMOS were: (a) 80/1, (b) 60/1, (c) 20/1, (d) 10/1, (e) 5/1, (f) 60/3. As seen from the subthreshold slope, these devices do not turn off as well as they should. We probably have some serious edge leakage problems as well as a double-Vt hump due to the much-smaller Vt caused by our layout's **poly gate extension**. 170

D-3 Unbonded NMOS Vt-extraction plots. The width/length ratio in microns for each NMOS were: (a) 80/1, (b) 60/1, (c) 20/1, (d) 10/1, (e) 5/1, (f) 60/3. 171

D-4 Unbonded NMOS Id-Vd plots. The width/length ratio in microns for each PMOS were all 60/3, and the difference between the plots were the backbias voltages: (a) floating, (b) grounded, (c) -5 V, (d) -10 V, (e) +10 V 173

D-5 Unbonded NMOS Id-Vg plots. The width/length ratio in microns for each PMOS were all 60/3, and the difference between the plots were the backbias voltages: (a) floating, (b) grounded, (c) -5 V, (d) -10 V, (e) +10 V 174
D-6 Unbonded NMOS Vt plots. The width/length ratio in microns for each PMOS were all 60/3, and the difference between the plots were the backbias voltages: (a) floating, (b) grounded, (c) -5 V, (d) -10 V, (e) +10 V. 175
D-7 21-Stage CMOS, L = 3 μm: Vdd = 1 thru 2.5 V. Left-column plots are the signals from the 9x output buffer and the tiny traces from probing the floating Ti backgate and the "useless" Cu-Al plane's pad; right-column plots are zoomed-in traces of the aforementioned floating pads. 177
D-8 21-Stage CMOS, L = 3 μm: Vdd = 3 thru 4.5 V. 178
D-9 21-Stage CMOS, L = 3 μm: Vdd = 5 thru 6.5 V. 179
D-10 21-Stage CMOS, L = 3 μm: The plot at Vdd = 7.0 is in (a). In (b), the frequency and the peak-to-peak voltage Vpp of the output was plotted as a fuction of Vdd. Note the saturation of the oscillation frequency at high Vdd's. 180
D-11 21-Stage CMOS, L = 3 μm: The DC offset and the RMS voltage of the fast oscillations coming off of the floating Ti bckgate and the useless Cu floating pads 181
D-12 21-Stage CMOS, L = 3 μm, Vdd = +5V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) +7.5 V backbias, (e) -15 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails, and at Vdd = +5 V, V-/V+ almost resided on the rails themselves, albeit with some voltage drop from internal series resistance. 183
D-13 21-Stage CMOS, L = 3 μm, Vdd = +4V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) +7.5 V backbias, (e) -15 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails. 184
D-14 21-Stage CMOS, L = 3 μm, Vdd = +3V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) -15 V backbias. Note that at Vdd = +3 V, the PMOS devices can no longer tolerate a backbias of more than 6.5 V without severe degradation to the output signal. 185
D-15 43-Stage CMOS, L = 3 μm: Vdd = 1 thru 2.5 V. Left-column plots are the signals from the 9x output buffer and the tiny traces from probing the floating Ti backgate and the "useless" Cu-Al plane's pad; right-column plots are zoomed-in traces of the aforementioned floating pads. Notice now that the Ti and Al pads were now electrically separated, thus supporting the previously mentioned notion that the 21-stage oscillator's shorted pads were probably from leftover stringers. 187
D-16 43-Stage CMOS, L = 3 μm: Vdd = 3 thru 4.5 V. 188
D-17 43-Stage CMOS, L = 3 μm: Vdd = 3 thru 4.5 V. 189

D-18 43-Stage CMOS, L = 3 μm: The plot at Vdd = 7.0 is in (a). In (b), the frequency and the peak-to-peak voltage Vpp of the output was plotted as a fuction of Vdd. Note the saturation of the oscillation frequency at high Vdd's. Also note this oscillator rang approximately half the speed of the 21-stage oscillator data shown on 81. 190
D-19 43-Stage CMOS, L = 3 μm: The DC offset and the RMS voltage of the fast oscillations coming off of the floating Ti bckgate and the useless Cu floating pads 191
D-20 43-Stage CMOS, L = 3 μm, Vdd = +5V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +10 V backbias, (d) +13 V backbias, (e) +14 V backbias, and (f) -10 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails. However, unlike the 21-stage variety, it tkea +13 V of backbias to regain control of the V+/V- extrema. 193
D-21 43-Stage CMOS, L = 3 μm, Vdd = +4V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +10 V backbias, (d) +12 V backbias, and (e) -10 V backbias. 194
D-22 43-Stage CMOS, L = 3 μm, Vdd = +3V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +10 V backbias, and (d) -10 V backbias. 195
D-23 A complete table of results for teh 2-D NMOS-only oscillator biased at Vdd = +3, +4, and +5 V. 196
D-24 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 197
D-25 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 198
D-26 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 199
D-27 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 200
D-28 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 201

D-29 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 202
D-30 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 203
D-31 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 204
D-32 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 205
D-33 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 206
D-34 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 207
D-35 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d) 208

E-1 Duality between electricity and heat. The left cartoon shows a charge q moving through a slab of material with electrical conduct under a potential difference of V. The right cartoon shows an unit of heat Q moving through a slab with a thermal conductivity of k under a temperature difference of 100 °C. 209
E-2 Equivalence between Poisson and Heat equations. If the dimensions a, b, and c were all equal, then the electric field E and the heat flux lines $k\nabla T$ would be geometrically and numerically be identical if $k = \sigma = 1$. Moreover, the equipotential and isothermal contours would also be equal. 212

E-3 Setup of the spreading resistance problem. Material of conduction has conductivity k, thickness t, and cross-sectional area of A. Material of heating has cross-sectional area A_s, effective thickness of zero, power density generation of $\dot{q}$(x,y), and power dissipation generation of $\dot{Q}$ = $\dot{q}$(x,y) · A_s. 214

List of Tables

2.1 NMOS waferbow measurements, at different process stage prior to bonding. A " ˆ " sign denotes compressive bow, and a "U" sign denotes tensile bow. 40

2.2 PMOS waferbow measurements for different Al thickness within the laminate structure that holds the PMOS and handle wafers together. A " ˆ " sign denotes compressive bow, and a "U" sign denotes tensile bow. Each indicated Al thickness correspondsto the thickness of a *single* release layer shown in Figure 2-12. 42

4.1 A collection of NMOS-only and CMOS ring oscillator configuratons. Each W/L ratio correspond to the width / length of MOSFETs in microns. In the NMOS-only cells, the first set of W/L corresponds to the switching transistor, while the the second W/L corresponds to the enhancement-mode NMOS active loads. In the CMOS cells, the PMOS transistors' widths doubled their NMOS counterparts. Forced-backbias means that Vdd was shorted to the Cu backbias pad, whereas in independent backbias cells, those two ports are separate. 68

4.2 Summary of basic MOSFET parameters for unbonded NMOS and PMOS wafers. Vt = Threshold voltage, and ST-slope = Subthreshold slope 71

4.3 Summary of CMOS ring simulations, with real NMOS / PMOS Vt's included in the models. For the all 3 μm long device, Vt_{nmos} = 0.9 V and Vt_{pmos} = -0.665 V (assuming that our Vt can be properly biased to a well-behaved negative number). For the 1 μm long devices, Vt_{nmos} = 0.662V and the PMOS Vt remained the same. 75

4.4 Output from 21-Stage CMOS ring, L = 3 μm, with floating backbias on the PMOS body . . . 77

4.5 Output from 21-Stage CMOS ring, L = 3 μm, with varied backbias points on the PMOS body 86

4.6 Output from 43-Stage CMOS ring, L = 3 μm, with floating backbias on the PMOS body . . . 89

4.7 Output from 43-Stage CMOS ring, L = 3 μm, with varied backbias points on the PMOS body 89

4.8 Summary of simulated and measured rsults from the 2-D NMOS ring oscillators biased at Vdd = +4 V only. The red "X" means that those devices were unavailable for testing. 96

Chapter 1

Introduction to 3-D Technology

The advent of copper chemical-mechanical polishing (CMP) by IBM in 1997 created a new avenue of attack in the never-ending struggle between semiconductor device engineers and the designed maximum speed of a given microprocessor. By replacing aluminum interconnects with copper, the overal RC delay of a chip decreased by a factor of 2. Moreover, chip reliabilty also increased dramatically because copper's bamboo-like grain structures offer a much higher resistance to electromigration compared to its aluminum cousin [4]. For a few years, this was a revolution in the industry as engineers took advantage of 50% RC reduction and out came with blazing fast chips such as the Intel's Pentium and AMD's K2 processors. However, as Moore's law continued past year 2000, it was evident that good times don't last very long in this industry. Even as device scaling continues to push the gate delay down and the device density up, benefits provided by miniaturization has been offset by a substantially larger increase in back-end delays. Hence, the term "the interconnect bottleneck" has gotten much attention lately [5, 6]. According to the 2005 ITRS roadmap [7], the total active-wiring interconnect length per cm^2 consisting of metal layers 1-5 is about 1 km long, and it will increase in an exponential fashion from year 2005 onwards! Furthermore, by year 2011, when the DRAM (Dynamic Random Access Memory) half-pitch and a microprocessor's metal-1 half-pitch coincides at 40 nm, that same total interconnect length is projected to be around 2.5 km with 12 metal layers, and there are no known solutions for achieving such wiring complexity, let alone trying to decrease the RC delay caused by these long wires. To top off the bad news, that's projected to occur 5 years from now !

This should not be a shocking news to anyone; rather, it is just a consequence of the quintessential planar, two-dimensional (2-D) design that engineers have relied on since the inception of the integrated circuit industry. Let's digress for a moment and see why this is so.

1.1 Some Perspectives on Interconnections

1.1.1 2-D City Development

If you currently work or live in an urban environment, take a quick peek outside your office window for a moment. Now, try to search for a one-story building within your line of sight. I am willing to bet that you'll see less than three, if you're lucky. Why is that? Simple: *Real estate = money.* There are many reasons for this phenomenon, but none are more important than the value of *faster communcation.* When economic, agricultural, political, and educational institutions (and whatever else there may be inside a city) are place together in a nexus, efficiency often increases - but up to a certain point. And what's the limiting factor ? Without exception, it's always related to land area. The never-ending open loop of increasing population density, leading to increase in city efficiency, which then leads to a need for city expansion has plagued every civilization since the time when Egyptian pharaohs ruled. As we will see, the parallels between city expansion and back-end integrated circuit (IC) architecture is actually quite amazing.

Consider the following: Before the invention of steel-reinforced concrete and the birth of skyscrapers, cities around the world were always build and designed with a 2-dimesional roadmap. Nucleation of towns usually start with a cluster of properties that were built around the town square, or its counterpart in microprocessors, the Arithmetic / Logic Unit (ALU). Next, the absolute essential elements in a city, whether it's a city hall (the registers), a military central post (decoders) or what not, were always placed at a distance closest to the town square for immediate access by citizens (electrons) who commuted to and from work (act of memory access); hence, the city (processor core) growth pattern takes shape of a circle. As cities grow, its dimensions increases steadily in a radial fashion [1].

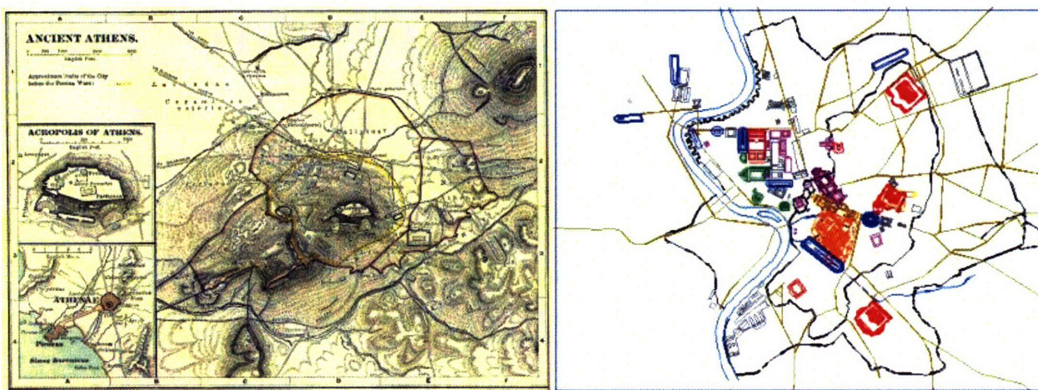


Figure 1-1: Maps of ancient Athens (*left*) and ancient Rome (*right*) Taken from [1] and [2].

[1] Ironically, although circuit designers know that geometrically one should pack neighboring transistors radially to minimize the nearest-neighbor RC delay, all modern microprocessor designs are instead generally laid out as Manhattan-style grids. Why ? This was a result of both lithography limitations in the early years of IC and the fact that those engineers were more concerned with ***packing density*** (rectangular tesselations can be close-packed to 100% efficiency while circles can't) rather than worrying about the total interconnect length

Figure 1-2: Maps of modern-day Beijing(*left*) and Boston (*right*) Taken from http://www.beijingmap.us and [3].

Take a look at Figure 1-1, which shows some sketches of ancient Athens (circa 500 BC.) and ancient Rome (circa 300 AD). Compare these skethes with those of modern-day Beijing and Boston in Figure 1-2, one can see that the trend of radial density growth has changed little over the course of almost 2 millenia. It is also interesting to note that while maintaining their radial topography, both building and population densities of modern cities exponentially surpass those from the pre-Industrial Revolution era. Why is that ? It was because the invention of steel by Bessemer in 1855 and its offspring invention, the reinforced concrete in 1857 by Monier, revolutionized the way engineers build structures. *While keeping the same 2-D street blueprints, they can now build* up *instead of across,* thus saving real estate and at the same time increased the population, productivity, and communication of the city [8] [2]. Just think - without the vertical integration capability offered by steel, New York City would never have become the economic hub of the world as it is today.

Although our analogy between city design to microprocessor design may seem a bit far-fetched, their evolutionary steps are surprisingly similar. Both were built and designed around a 2-D "substrate," and both are destined to suffer the following fates:

1. As the city grows, the town square's activity increases, but so does its real estate value.
2. Higher real estate values near the city's core forces the relocation of non-essential elements towards the city's outer limits (ie. The poor citizens !). Commuting times thus increase.
3. Re-incorporation of relocated subjects back into the city limits requires an overall area expansion
4. Repeat cycles 1-3 for every city growth event

[2] It can be argued that the Romans were quite capable in the art of vertical integration, constructing huge structures like the Coloseum and the aqueducts that lasted even to this day. However, they could not achieve anywhere near the vertical densities provided by skyscrper designs.

5. Breaking point:

   - Growth cycle will **cease** when city limit expansion hits natural / political boundaries or it hits a critical commute time that hurts commerce
   - Growth cycle will **recommence** when new building technologies emerge (ie. 3-D adobes such as skyscrapers) that will exponenetially increase the density of citizens, property, and city activities while maintaining a constant real estate.

Let's see if the growth and evolution of a typical microprocessor follow this trend.

1.1.2 2-D Microprocessor Development

First, let's look at some die photos of Intel's earlier processors as depicted in Figure 1-3:

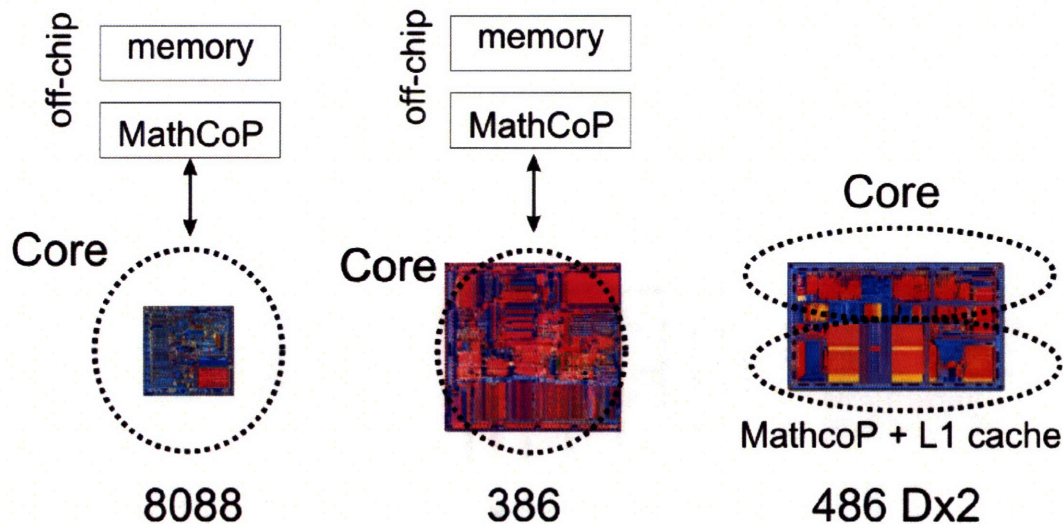


Figure 1-3: Die photos of Intel chips, areas approximately to scale (33 mm^2, 94mm^2, 81mm^2 respectively)

In the first generation Intel processors (8088), the entire silicon footprint was used to fit just the core; any peripheral connections such as math-coprocessors (one may call this an off-chip "core") or memory have to made off-chip, which -drastically increased the transit time between read / write cycles. The first and easiest solution was to expand the silicon area and increase the *functionality per footprint ratio.* Two generations later in the 386x series, the core area tripled, the chip can indeed do more than its predecessors with its new 32-bit bus, but its peripherals continue to linger outside the main I/O pins. The off-chip RC delay problems have yet to be solved.

Rather than pushing to an ever-larger silicon footprint, device engineers had another weapon up their sleeves: Scaling. By revamping core architecture and using 1 μm CMOS instead of the previous 1.5 μm

devices, the 468x series contained the following innovations: An overall decreased gate delay, both on-chip L1 cache and the math-coprocessor were integrated on-chip, and if that wasn't enough, the total chip area has decreased a bit. The Intel 486x family really started the current scaling war among chip-making companies. Before anyone knew it, the Pentium and its AMD / Cyrix clones have pushed the envelope even further. By the end of this fifth-generation chip lineage, CMOS devices have been scaled down to 0.35 μm, the L1 cache capacity has been maximized, the dual pipeline design ("superscalar" architecture) was a mainstay, and to fit all those features under one roof while maintaining good yield, the Si footprint has been minimized to around 141 mm^2, just a tad bigger than the long-obsolete 386x. However, applications that uses these newer / faster processors need even more memory, and a solution to fix this problem was the utilization of the L2 cache, of which was eventually implemented off-chip because the engineers ran out of their allotted Si real estate. Therefore, by scaling and increasing the overal density to make a chip faster, we have taken a step back with the introduction of a new source of back-end RC delay. Can this problem be resolved?

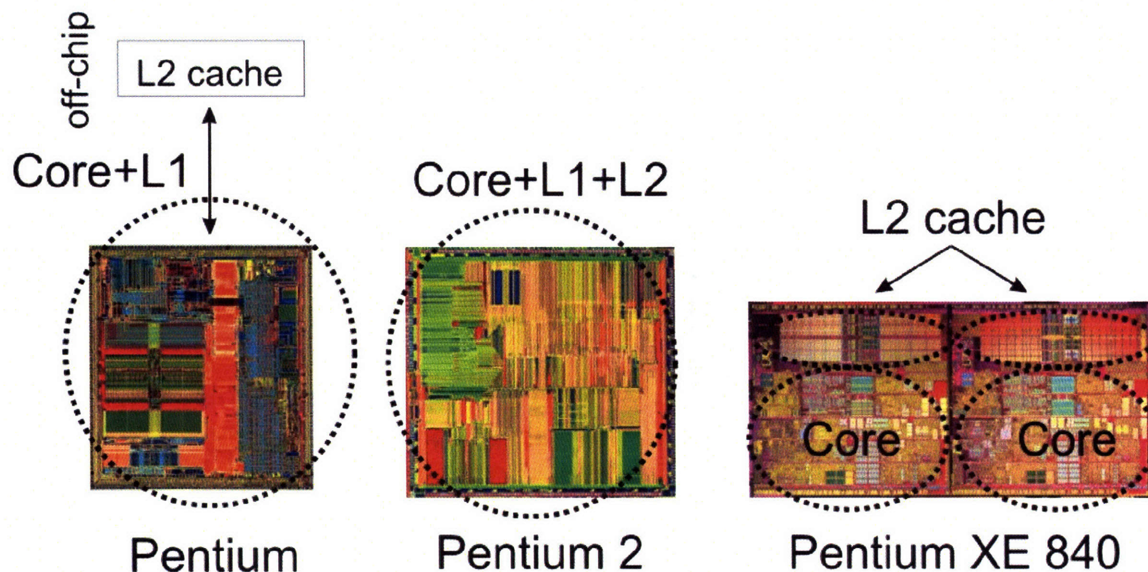


Figure 1-4: Die photos of Intel chips, areas approximately to scale ($147mm^2$, $203mm^2$, $206mm^2$ respectively)

Alas, when in doubt, integrate ! And the fruits of the labor are the latest consumer microprocessors on the market, aptly dubbed the"Sixth Generation" chips as seen in Figure 1-4. The family includes the Pentium II and its newer cousins, the Pentium dual-core Extreme Edition 840 and AMD's dual-core Athelons, both of which were introduced in 2005. All three 64-bit processors come with a fully integrated L2 cache and are both superpipelined and hyper-threaded for faster and more efficient multitasking. Furthermore, the multicore chips have two independent execution units that takes multi-tasking to a new level, albeit with one major concession made in the chip's physical layout: You have to double-up the components for each core, which requires a bit more Si area than normal. While multi-core designs seem to be the way of

the future (Intel predicts they'll introduce their quad-core chips sometime in 2007), the good news come with an asterisk, though - ***The emergence of a multi-core structure is a signal that benefits of scaling are going to end and some new technology is going to carry the torch for further improvement in IC design.*** In a nutshell, today's chips are so over-intergrated to the point that its activity is generating heat levels that are degrading device performance; in essence, modern chips are creating a self-made fever that apparently no medicine can cure. To clarify the above statements, consider the following facts.

1. The higher the switching frequency, the more power P each transistor will dissipate since $P = CV^2 f$, where C = gate capacitance, V = voltage, and f = frequency
2. Higher density of devices means the total power dissipation density (W/m^3) has increased, which in turn increases the maximim on-chip temperature
3. Self-heating of SOI devices produces positive feedback in Item 2
4. Higher device density also leads to a higher number of (and longer) interconnects, which in turn increases the current need to drive these wires, thus exacerbating Items 1 and 2

The heat dissipation problem is the main reason why designers are relying on multicore processors. Using sub-optimial cores running at lower clock frequencies, one can still achieve equal or better performance with innovations in computer architecture while keeping device performance constant. Again, the price to pay is the required redundancy in all circuit components - in other words, *we need to double the real estate.* Upon re-examining Figure 1-4, one can immediately see where 3-D integration might help. Can we just fold the dual cores on their axis of symmetry and gain twice the functionality, and at the same time, reducing the Si footprint by 50% as show in Figure 1-5? Or how about constructing 3-D structures from substrates that are otherwise incompatible with each other during CMOS / III-V processing ? One can dream of the following scenario:

Homogeneous Integration Bonding of like parts, such as the 3-D multicore example, or DRAM on top of CMOS logic

Heterogeneous Integraton Bonding substrates of dissimilar technology, such as CMOS to III-V, MEMS, or any other combinations

Thus, as steel was a revolutionary tool for increasing density, effieciency, and elegance in macro 2-D structures, 3-D integration could be the same holy grail for microelectronic 2-D structures. The method of choice in the MIT 3-D integration scheme is to use copper (Cu) wafer bonding, and the rest of this thesis will explore its feasiblity, its practicality, and some selective structures that will benefit the most from it. In fact, this is the perfect place to ask the follwoing question: *What is the purpose of this thesis ?*

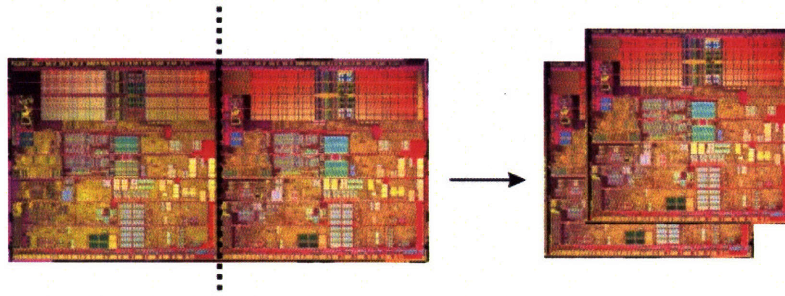


Figure 1-5: Homogeneous 3-D implementation of a dual-core processor ? Maybe !

1.2 Identifying the Thesis Topics

Theoretically, one can make numerous statements about how 3-D can increase the device density, decrease the global interconnect RC delay, increase the circuit functionality-per-footprint, and etc, but can they prove it? This thesis will examine a real-life examples that cover all sides of story - The good, the bad, and the ugly. Specifically, the four rather blunt questions this thesis will attempt to answer are the following, where the pronoun "it" refers to "3-D integration":

1. **What will enable it to work ?**

   Two chapters will be devoted to identifying the technological challenges associated with MIT's 3-D process flow. Specifically, the major technological roadblocks associated with wafer-wafer alignment, overall bond quality, handle wafer realease, and the die-level counterparts of all above issues will be discussed.

2. **Does it really work ?**

   The easiest method to implement a proof of concept is to build a simple circuit that demonstrates the feasiblity of Cu-Cu bonding. The vehicle of choice is the 3-D CMOS ring oscillator, where for demonstration purposes only, one wafer would contain all PMOS devices and on the other substrate purely NMOS devices. If the oscillators work, then it proves that the 50+ PMOS - NMOS interconnections are indeed reliable and the substrate-substrate alignment was accurate enough.

3. **Does it really help ?**

   Since the ring oscillator design will not be state-of-the-art, it will be very difficult to make any 2-D to 3-D performance comparisons such as decrease in overal RC delay vs. circuit topology. Instead, this thesis will try to investigate the "thermal dissipation problem" associated with 3-D and how, or if it's even possible, to decrease the overall maximum temperature of a chip by using Cu-Cu bonding in a creative way. To be more specific, we will try to see which topography in the Cu bond layer can act as a better heat conduit: Cu thermal vias or thermal planes ?

4. **What else is it good for ?**

   One of the ways in which 3-D can really help is to reduce the Si footprint by building area-consuming circuit blocks or elements on top of its neighbors. A good candidate for such scheme is the area minimization of the impedance matching spiral inductors in RF power amplifiers. Normally, a cascoded RF amplifier uses inductors in both the base and collector terminals to match the input / output impedance and to tune the respective port's Q in order to maximize the transducer gain of the circuit. By placing the large spiral inductors on top of the already-large RF SOI LDMOS (or BJT), one can use the area saved by 3-D stacking for other devices. For this to be successful, however, one needs an effective magnetic shield to eliminate any eddy currents induced caused by the inductor's RF magnetic field on the Si device surface. Hence, the focus of the thesis here will be on testing the effectiveness of cobalt magnetic shields.

1.3 Executive Summary of Thesis Results

Nominal feasibility of 3-D integration using Cu-Cu bonding was demonstrated with the successful ringing of both a 21-stage CMOS and a 43-stage CMOS ring oscillator implemented using face-face wafer bonding. The claim of feasibility stems from the fact that in the 43-stage oscillator, 56 pairs of PMOS / NMOS- SOI devices (fabricated on two separate wafers) successfully communicated with each other through a combination of 200+ Cu damascene vias topped by 90+ inter-level Cu-Cu bonds. Integrated backgates provided the user an option for manual V_t adjustments of all top-tier PMOS-SOI gates, and body-backbiasing proved to be an integral element in mitigating the unwanted V_t-shifts caused by SOI-body and SOI-BOX interface charging, which in turn, were due to impact ionization events occuring within the channel.

The face-face bonding configuration, when considering a 2-layered stack, provided a convenient method in decoupling the body-to- interconnect parasitic capacitance that plagues the face-back bonded ring oscillators, of which none of them oscillated but their flatline outputs did respond linearly to changes in Vdd. Also, the wafer-level, face-to-face bonded Cu exhibited a much lower Cu-Cu series resistance when compared to the face-back, die-die integrated analogue. This suggest that the overall Cu bonding quality can and will influence the thermal and electrical characteristics in 3-D circuits, and bonding equipment refinement - including better die-die and wafer-wafer alignment systems, better die-die mechanical contact schemes, in-situ surface cleaning options, and etc, are the real keys to success Cu-Cu 3D integration.

Although the face-face configuration [3] seemed to be superior to the face-back counterpart in all electrical aspects of 3-D, it theoretically suffers immensely from heat dissipation problems, for the top-tier devices in the face-face bonded circuits will be located further away from the base heat sink. Regardless of either bonding configurations, the first-order solution in tackling the 3-D heat dissipation problem should revolve

[3] Again, when considering a bilayer stack

around using the pre-existing Cu bonding layer as heat flux diffusers. By re-distributing the heavily-biased z-dependence in heat flow onto the x and y-axes, the effective vertical thermal gradient can be decreased as much as 15 % according to FEM (finite-element model) simulations. As for Cu thermal vias, while theroetically and experimentally they were very effective in removing local hotspots and creating thermal Faraday cages to isothermally sequester extreme hotspots, these vias were unable to affect the global temperature profiles unless one increases their cross-sectional area. Therefore, unless there is a real temperature crisis somewhere on a chip, the usage of thermal vias should be considered only as an emergency measure because of their cost in real estate is high.

Lastly, the application of 3-D we chose to focus on was how to integrate an area-expensive passive element (RF spiral inductor) on top of an existing circuit by virtues of 3-D. The path to success revolved around finding the proper magnetic shielding that can protect the underlying circuits from **B**-field-induced substrate currents generated by the top spiral inductor at RF frequencies. A solid cobalt (high permeability) shield of 400 nm exhibited a -24 dB improvement in substrate isolation at 13 GHz when compared to a reference, dual-inductor substrate crosstalk detector that lacked any magnetic shielding. Moreover, the same cobalt shield proved to be superior to an Al shield of 2 μm thick across mid-range RF frequencies of 6 - 20 GHz. Although the solid cobalt shield proved to be a formidable magnetic shield, keen judgement must be used before applying Co films as EMI shields because of its inherent high RF power dissipation, of which when placed near high-Q devices, will definitely be a detrimental factor rather than an enhancement.

Chapter 2

3-D Integration Challenges

The focus for this chapter will be to explore the challenges associated with wafer-level 3-D integration using Cu-Cu thermocompression bonding. To begin our discussions, we will first take a quick glance at the MTI 3-D process flow.

2.1 The MIT 3-D Approach

To begin, one has to be aware that there are many variations of 3-D integration in the literature, and each offer its own advantages and disadvantages over each other. Intead of doing a compare-and-contrast exercise on each 3-D method, the reader is invited to read some wonderful reviews given on selective silicon epitaxial growth [9, 10], multi-chip modules (MCM) [11], MEMS mciro-spring contacts [12], and polymer and oxide-based wafer bonding methods [13, 14, 15, 16]. Right now, let's focus on the rationale behind our choice in Cu wafer bonding [17, 18, 19, 20]. Figure 2-1 shows such a structure:

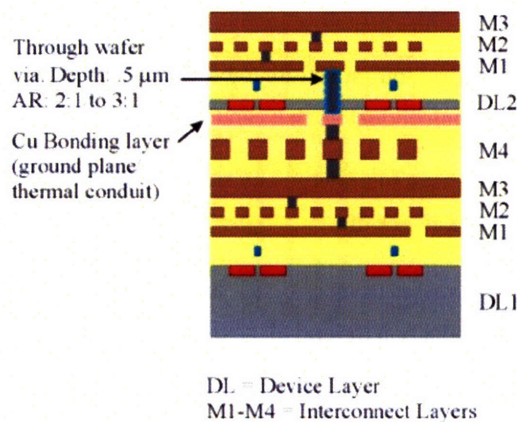


Figure 2-1: The MIT 3-D model

The MIT model 3-D structure shows that two device layers, denoted by DL1 and DL2, complete with its own multi-level interconnect layers M1-M4, were fused together by thermocompression bonding of two pre-existing Cu films deposited on their respective substrates. The model also shows that the combination of an inter-level via and part of the Cu-Cu bonding interface provide the electrical connnection between DL1 and DL2. Copper films adjacent to the electricall-active Cu-Cu bond acts as dummy bonding pads that enlarge the effective bonding area and naturally the cross-wafer bonding strength. Now that we've described the final product, let's take a brief look at the fabrication process flow.

2.1.1 An Overview of the MIT 3-D Process Flow

The MIT 3-D process was based on a modular flow where each successive device layer bonds on top of an existing vertical base stack. The process was also designed such that exposure of the continually-growing base stack to mechanical or chemical attack were minimized.

Before one can construct a multi-layer device stack, we first have to start with two wafers. One of these two members will be an SOI wafer (designated as the "top" wafer in Figure 2-2) that will undergo backside substrate thinning and a subsequent backside via etch and fill. The filled backside vias were then capped with a thin copper / tantalum (Cu/Ta) bilayer to facilitate bonding. Afterwards, the resultant thin-film device layer will then be bonded to a base wafer (designated as the "bottom" wafer in Figure 2-2) in a face-to-back fashion. Finally, the dummy support wafer used in the wafer-thinning step has to be released from the 3-D bond stack. In the end, since the "initial" seed stack is topologically identical to the green "bottom" wafer, the seed stack can re-enter the process loop and multiple Cu-bonds can be made with relative ease. The salient features of this process are:

- The process is inherently modular, where one can either construct a multilayer stack by serial bonding of single-layer stacks, or if one wishes to, an m-layered seed stack can be bonded to an n-layered "top" stack
- The process allows devices from the "top" substrate to be made from *any* given technology, whether it's III-V, strained Si, or organic semiconductors that can tolerate the required bonding temperatures
- The growing seed stack does not have to undergo any substrate etchback steps, thus preserving the structural integrity of the base Si wafer or die

Figures 2-3 and 2-4 depict these processes in more detail, and if the reader wishes to dive deeper into each process, then please refer to Appendix A at the end of the thesis for the nuts-and-bolts of the fabrication flow.

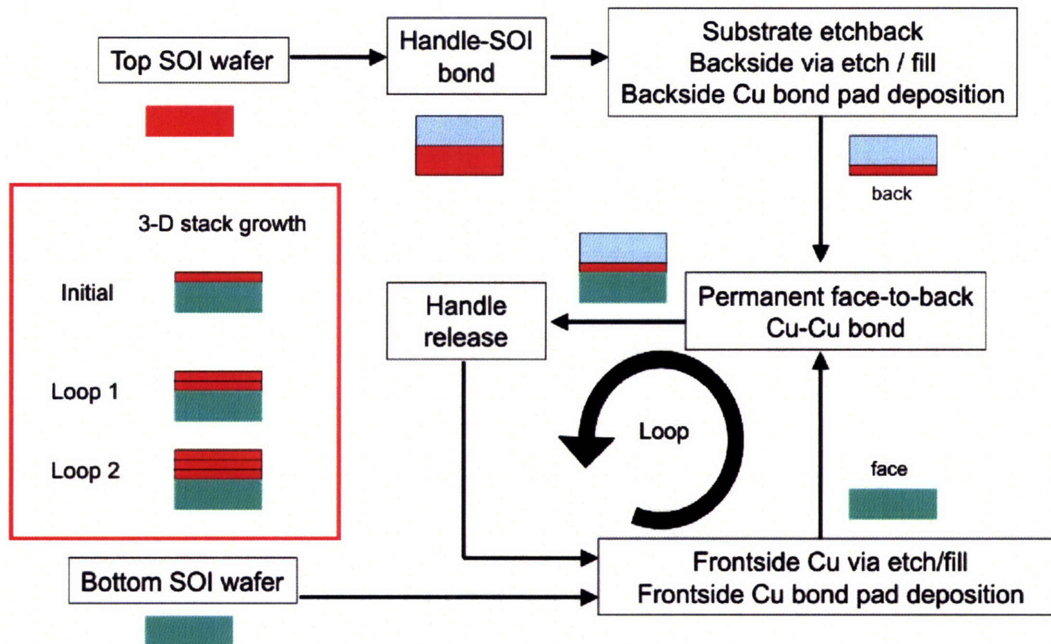


Figure 2-2: MIT 3-D process flow: A bird's eye view

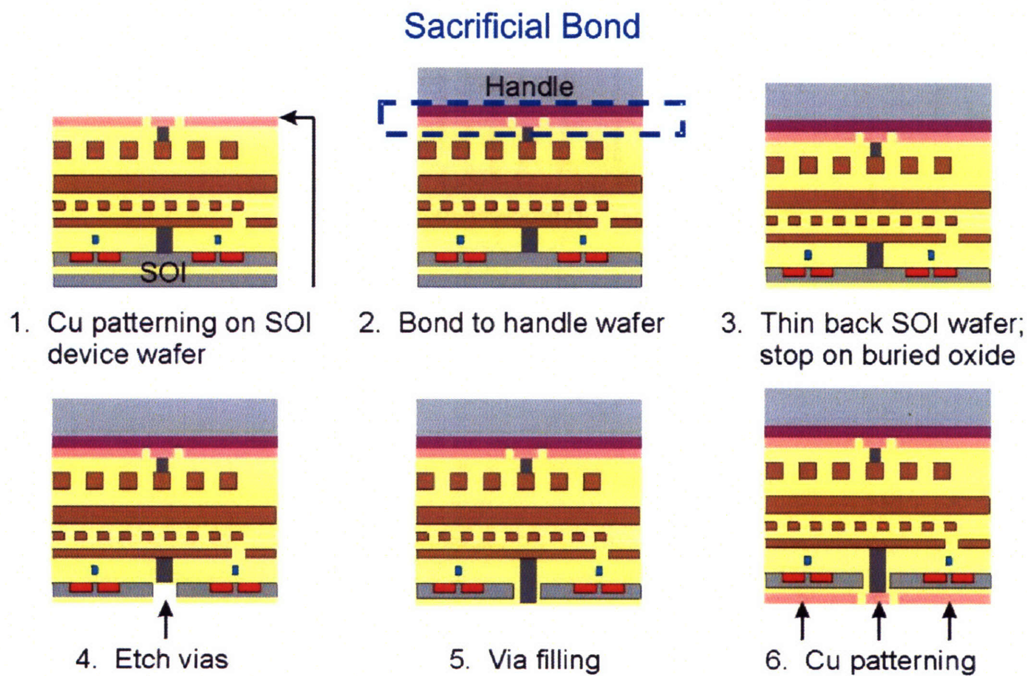


Figure 2-3: MIT 3-D process flow, part 1

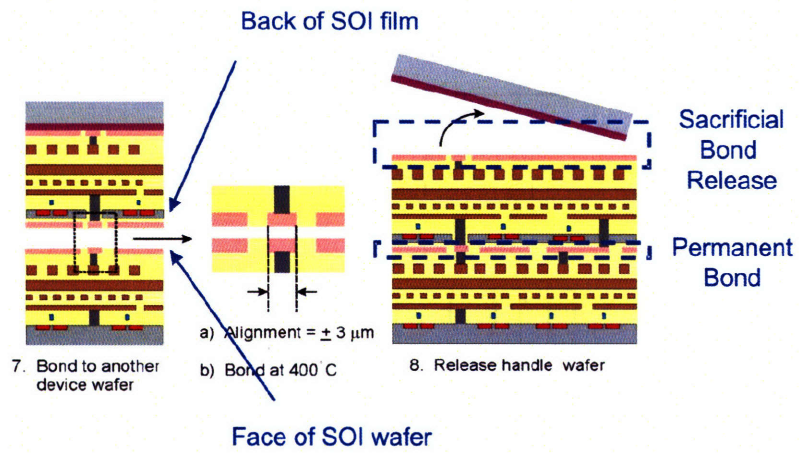


Figure 2-4: MIT 3-D process flow, part 2

2.2 Wafer-Level Integration: Bonding Uniformity Challenges

2.2.1 Bond Microstructure

One may say that Cu-Cu wafer bonding should be easy: Just slap two wafers together, press them and heat them, and presto - you're done, right? Sadly, nothing could be farther than the truth. While two clean Cu or Au surfaces do indeed bond to each other readily, metal thermocompression bonding is notorious for having global bond uniformity problems. The metallic system has neither the "chemical zipper" like the van der Waals attraction force in Si-Si direct bonding [21, 22] nor the solvent-matrix coagulation between two spin-on polymer surfaces [1]. Instead, Cu-Cu or Au-Au bonding proceeds only with atomic diffusion and grain growth at elevated temperatures, and whether or not the reaction proceeds depends largely on the micro-roughness of the metallic interface. When the micro-roughness variation becomes a long-range order defect, these microscopic voids and other defects can easily grow into macroscopic ones upon two 400°C bonding anneals.

In fact, Chen, et. al. has shown that while a freshly e-beamed Cu pair exhibit the highest bonding strength, a Cu pair that has been exposed in cleanroom air for more than two weeks bonds just as well, even with no pre-bonding surface cleans [23]. However, if one tries to remove the native copper (II) oxide (CuO) with dilute HCl prior to bonding, the RMS surface roughness increases from sub-nanometer to about 1.5 nm, and just that slight increase in pre-bonding roughness can result a post-bonding Cu-Cu interface undulation of more than 75 nm [23]. Worse, if at some random local spots the surface roughness far exceeds the RMS value, no amount of heat and mechanical pressure can supply enough kinetic energy to facilitate plastic deformation and inter-atomic diffusion. This results in micro-voids, and when voids aggregrate in the wrong spot in a Cu-Cu 3-D structure, one would have an open circuit or a film delamination spot.

Lesson Learned: Fresh deposition of Cu ensures the highest possible bond quality, while surfaces with slight oxidation also results in a smooth bonding interface if no acid cleans were performed. Pre-bonding acid cleans do strip the Cu surface of impurites but roughens up the bonding interface dramatically, and thus is actually detrimental to the final Cu-Cu bond quality.

2.2.2 Mechanical Contact

Even when Cu surface conditions were perfect, the quality of the bond was often dictated by the a second variable: How well can one press the two wafers together while maintaining an even contact force across the 6" substrates. Again, this problem may appear on the surface as a trivial matter, but there's much more than meets the eyes. The critical factors of concern are the compliance of the chuck material, the material compliance of the inserts in-between the wafers and piston, the piston's travelling distance, and

[1] Metal-metal surfaces can indeed have a zipper-like bonding wavefront upon contact, but this can only occur under high vacuum and must be preceded by an in-situ, pre-bonding Ar sputter clean, in which the unstable dangling bonds formed at one metal-air interface are hungry for interactions with the other interface

the amount of piston downforce needed for a successful bond. To see the interplay of these factors, let's look at the bonder setup after a wafer pair has been aligned and it's ready for bonding. This is depicted in Figure 2-5.

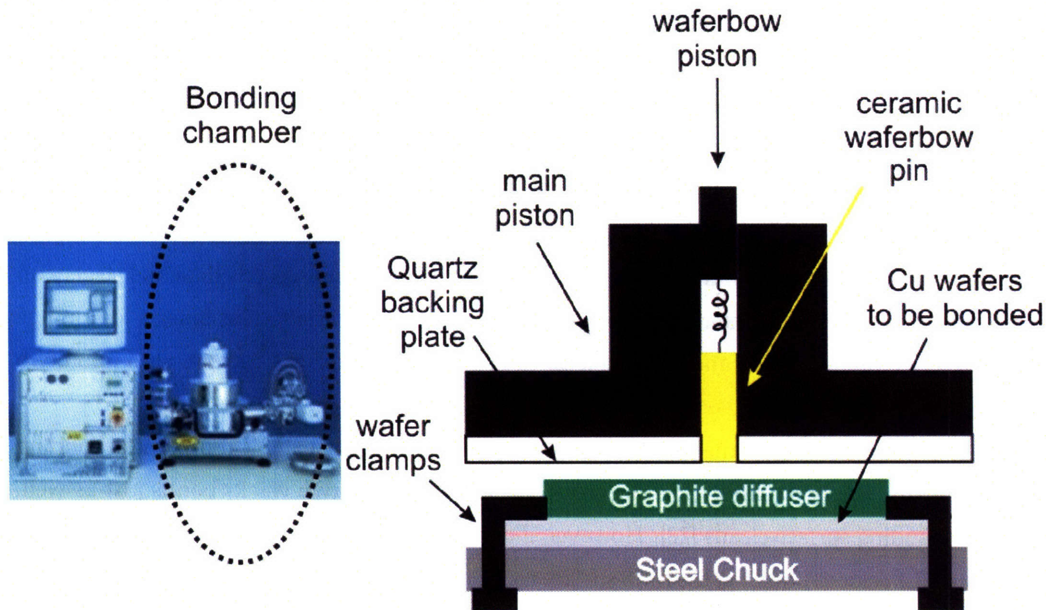


Figure 2-5: Electronics-Vision EV501 wafer bonding setup

The Piston Backstop

In wafer bonding, the correct piston travelling distance needs to be set to ensure maximum contact between the two wafers. This was accomplished by dialing in the total thickness of the to-be-bonded substrates and of the graphite diffuser by rotating the piston backsop collar. From Figure 2-6, it should be clear that if the limit distance was dialed in too large, the the piston will never make contact with the graphite diffuser; on the other hand, if one dials in the thickness a little too short, then upon a piston downforce more than 300 N, either the quartz backing plate will fracture or there will be a mechanical displacement of the backstop relative to its initial zero position. While breaking the quartz backing plate is an irrecoverable fault, the vertical displacement of the backstop can be monitored and re-zeroed after every bonding session. With constant vigilance in monitoring the equipment state, the bonding quality for 4" and 6" Cu-Cu thermo-compression has become more consistent, although the overall uniformity is still subpar. Some examples of these ever-present random delamination between the SOI-handle bonding interface after Si etchbacks are in Figure 2-7 [2].

---

[2]For reference, the wafers in Figure 2-7 has undergone the following processes:

1. Starting handle wafer: 5000 A thermal oxide on Si

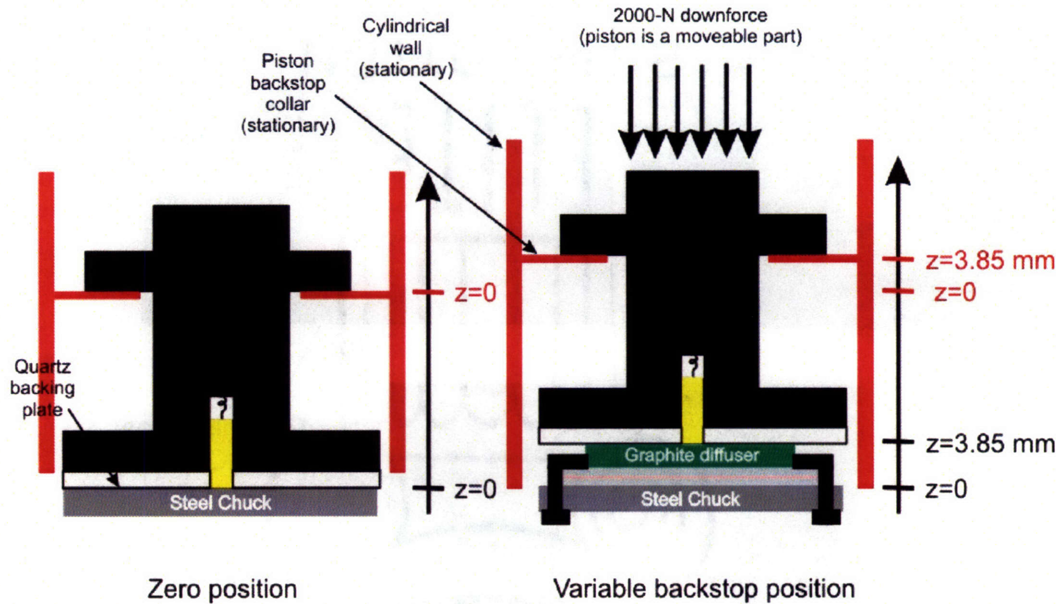


Figure 2-6: Piston backstop position during resting and bonding phases. The graphite insert and wafer thickness were 2.55mm and 650μm each, respectively

Since the bonding uniformity has not yet been fully optimized, there must be other mechanical factors in the system in which one can improve on. This actually leads to our next topic of discussion - the bond diffuser design.

The Bond Diffuser

From Figure 2-8, there were three basic kinds of diffusers available in MTL: Solid steel circular, steel donut with an 1 cm deflectable steel center tab, and solid graphite circular. The following discussions will hopefully convince the reader that both the flatness and the diffuser material compliance were also key factors in maximizing the Cu-Cu bond quality.

To start with, let's briefly review the Cu-Cu bonding results from the stainless steel diffusers and prove that diffuser flatness is an essential element to be contended with. The first experiment involved the comparison of a solid steel vs a donut-shaped steel diffuser with deflectable center. The idea behind the de-

2. Starting fake SOI wafer: Si base with the trilayer
   - 5000 A thermal oxide as BOX
   - 2500 A LPCVD poly as the fake SOI
   - 1 μm PECVD oxide (from concept-1) as the fake ILD
3. Sputter 20 μm of Al on the fake SOI as part of the laminate structure
4. E-beam 500 / 3000 A of Ta/Cu on both wafers
5. Cu-Cu bonding, followed by mechanical grindback and TMAH etchback, thus exposing the fake BOX's back surface

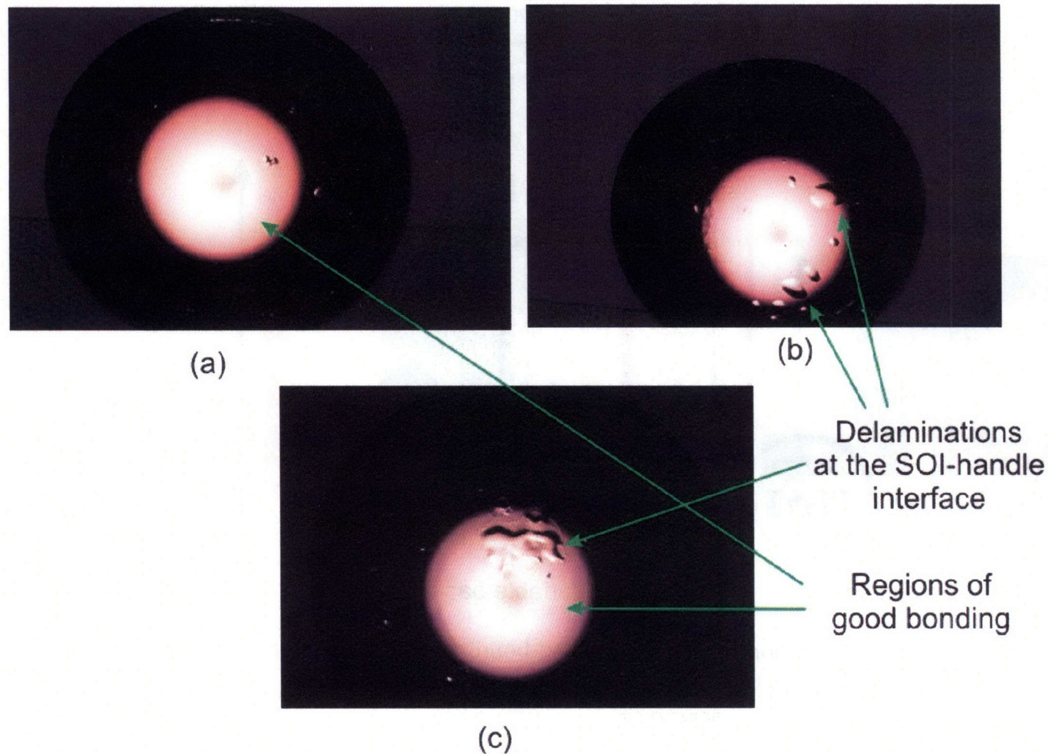


Figure 2-7: Random delamination between the SOI-handle interface after TMAH etchback.

flectable center tab is that for Si-Si direct bonding, the initial push by the waferbow pin at the wafer center creates a "bonding wavefront" that travels radially outward, and along that wavefront, the cleaned Si-Si surface will automatically adhere to each other via van der Waals forces. In Cu-Cu thermocompression bonding, however, the deflectable center tab can actually be quite a nuisance because the waferbow pin's maximum downforce was only 7 N. The performace of the three types of diffusers can be compared from Figure 2-9 [3].

The rough patterns observed on the wafers were the results from micro-fractures at places where the Cu-Cu bond was either incomplete or completely failed. Furthermore, while the facture patterns made by the steel solid chuck were random by nature and changed from sample to sample, a target-like microfracture pattern persist on all wafers bonded with the donut steel diffuser. These results suggest that the while

[3] The wafers in Figure 2-9 consist of the following layers and underwent these steps:

1. Starting top and bottom wafers: 5000 A thermal oxide on Si
2. Sputter blanket 500 / 3000 A of Ta/Cu, with no exclusion rings
3. Bond top and bottom wafers
4. Mechanical grindback 500 μm of Si from one side, followed by TMAH aqueous etchback until the buried thermal oxide is present

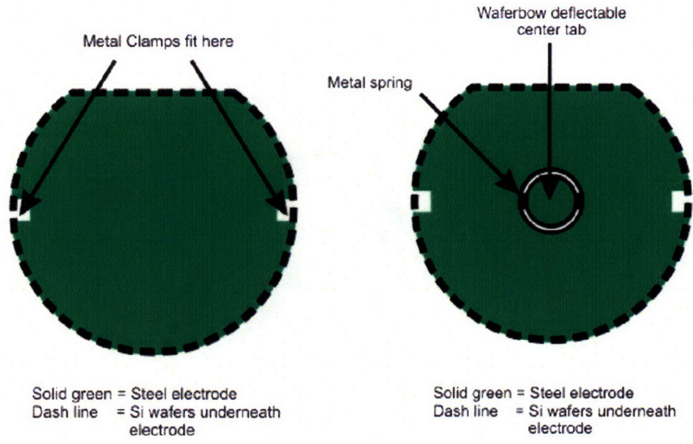


Figure 2-8: Bond diffuser geometry: Solid circular vs. donut

Solid steel electrode

- Random non-bonding spots
- Electrode was not as flat as we thought)

Donut steel electrode

- Definitive non-bonding circle at center
- Electrode is flatter

Figure 2-9: Bond diffuser geometry: Solid circular vs. donut, wafer after TMAH etchback

the center-deflectable tab in the donut-diffuser is indeed detrimental to the bonding process due to a weak waferbow pin contact pressure, the global surface smoothness of the donut diffuser was much better than the solid steel diffuser because of its repeatable fault patterns. Thus, this experiment showed that the diffuser flatness is an absolute must for a good bond.

Next, the importantace of diffuser compliance was tested by comparing the results between the flat steel donut diffuser vs a solid graphite diffuser, of which was both thinner and softer than its steel cousin. Figure 2-10 the results. These wafers have undergone the same treatment as Figure 2-9 except for the TMAH aqueous etchback. Just after mechanical grindback, one can already see that the graphite diffuser bonds better than the steel donut because its compliance can compensate for some degrees of surface roughness (The circular imprint did not show up too well in the photo because it very difficult to show such a minute difference in polarization in normal photographic film). Moreover, to prove that the graphite diffuser gave

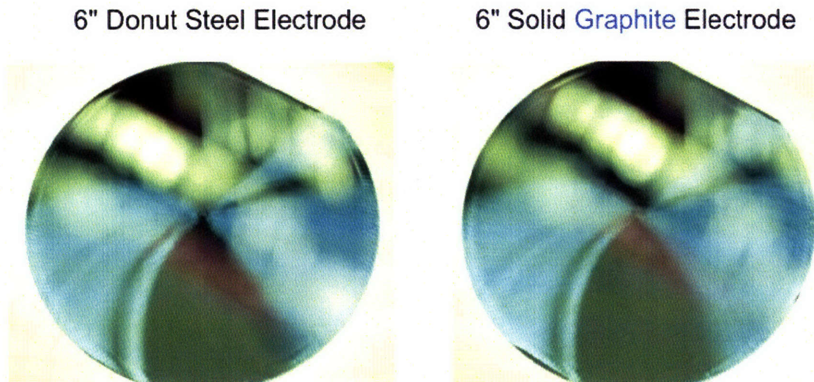


Figure 2-10: Bond diffuser geometry: Solid donut vs. solid graphite, wafer just after grindback

an uniform bond for blanket Cu films, we have constructed the following structure that will give some extra information about the bonded Cu films. In short, we have replaced the mechanical grindback step in the preparation for Figures 2-9 and 2-10 with a CHF_3 dry etch to remove the 5000 A thermal oxide from one side of the bonded pair only (designated the "top" side). This was followed by a TMAH etchback, and the result can be seen from Figure 2-11.

The top view shows the exposed smooth thermal oxide surface with no apparent delaminations in any regions; this is in contrast of the bottom view, which shows that the backside of the "bottom" wafer was attacked viciously by TMAH. The fact that the Cu-Cu bond beneath the smooth thermal oxide surface did not delaminate after an inordinate amount chemical attack gave proof that the graphite diffuser indeed made the Cu-Cu bond smoother and stronger.

On a side note, one other interesting thing came out of the the above experiment. If one focuses on the edge view of the etched wafer in Figure 2-11, the photo shows that the entire stack exhibits a concave curl (the "bottom" surface, which still contains about 100 μm of Si substrate, was pointed up towards the computer keyboard). This is proof that the sputtered Cu film was deposited in tensile and remains in tensile stess even after two 400 °C thermal cycles worth of grain growth. Unforseen by yours truly at the time, the tensile stress combination provided by the Cu-Cu bond and the laminate structure proved to be a disaster waiting to happen when one was ready to bond a PMOS wafer to a handle. The culprit here: Massive waferbow. Its cure: Maximum bond downforce, bow-compensation metal deposition, and maximum downforce during thru-wafer Cu via damascene.

Waferbow Compensation

Last, but not least, the final mechanical variable in bonding is the amount of piston downforce needed to ensure good wafer contact and to exert enough external energy into the bonding interface to facilitate atomic diffusion. To first order, if one dials in the correct piston backstop collar height, empirical results

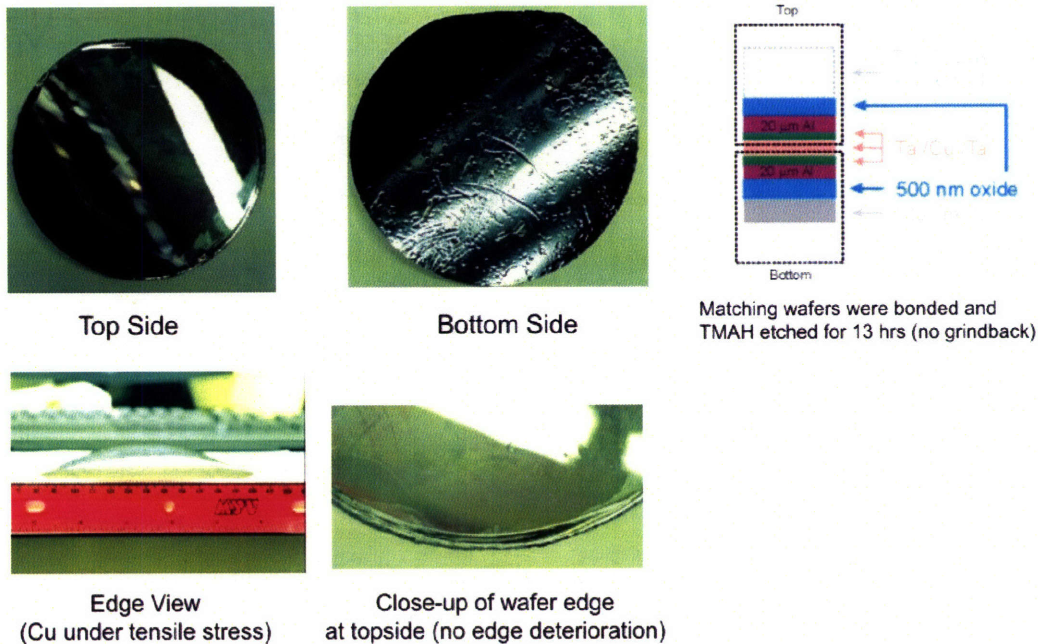


Figure 2-11: Graphite-bonded wafers after etchback, showing both global uniformity in Cu bonding and high tensile stress of annealed Cu films

show that a normal 6" diameter, 650 μm-thick Si wafer with a 5-10 μm waferbow in either flexture concavity can bond to its Siamese twin substrate with contact forces as low as 2000 N. But as the thin film stacks add up during front-end and back-end processing, the total film stress and the overall waferbow will both increase. As proof, recorded in Table 2.1 was a sequence of NMOS waferbow measurements after some selected process steps taken prior to wafer bonding (a " ˆ " sign because a compressive film deposited on Si will create a convex curvature, whereas a "U" sign denotes a tensile bow because tensile film deposited on Si will exhibit a concave curvature). [4]. Since wafers with excess bow (usually also very stiff as a result) will resist mechanical forces applied by the piston and tends to create air pockets within the Cu-Cu interface, bow management prior-to and during bonding can also make or break the quality of the overall bond.

During the course of fabricating the NMOS and PMOS devices for the ring oscillators, we actually encountered a huge parity in terms of waferbow management techniques between the two wafer groups. In both cases, the waferbows were large enough such that bow-compensation techniques were needed prior to bonding. Focusing our attention to the NMOS process first, as the MOSFET fabrication process moves from the front-end to the back-end, the waferbow of an SOI substrate tends to become mildly compressive up to the point where the topside Cu damascene vias are ready to be etched and filled. In terms of waferbow,

[4]These measurements were peformed on the Tencor FLEX laser scanner available in TRL, and a double-side polished Si wafer with an 1 μm PECVD top oxide was chosen to be our reference wafer because it represents the lowest common denominator case where a single wafer contains some standard amount of ILD on top. The measured waferbow value of +16.44 μm was a sanity check that our oxide bow measurements was indeed compressive and was correct.

Substrate	Waferbow	Radius of curvature
X2-side polished Si + 1 μm oxide (reference)	16.44 μm ^	120.03 m
X2-side polished NMOS (substrate thickness=525 μm)	49.05 μm ^	36.72 m
+ 2 μm sputtered Cu frontside	52.34 μm U	33.65 m
+ 1 μm sputtered Cu backside	4.06 μm ^	648.05 m
+ Anneal at 300 °C, 1 hr	7.13 μm ^	270.34 m
+ Cu via damascene, frontside	14.66 μm ^	132.63 m

Table 2.1: NMOS waferbow measurements, at different process stage prior to bonding. A " ^ " sign denotes compressive bow, and a "U" sign denotes tensile bow.

the damascene via fill was a turning point in the process because to properly fill-in an 1.1 μm-deep via prior to Cu damascene, a bilayer of 500 A Ti and an excessive2μm Cu film had to be deposited as an overfill prior to Cu CMP. At this point, the waferbow dramatically turned into a tensile one (see Table 2.1 for numerical details), and its magnitude was large enough that it caused a serious uniformity issue when one tries to perform Cu damascene CMP. This is not a show-stopper, however, and it was remedied somewhat by sputtering an 1μm bow-compensating Cu film on the backside. Upon polishing off the Cu via overfill and adding an extra 500 / 3000 A Ta/Cu lift-off process to define the Cu bond pads, the ready-to-be bonded NMOS SOI wafer has a nominal bow of 14.66 μm compressive, which in the wafer bonding world, this was considered to be relatively flat. In summary, if the NMOS wafers did not receive a bow-compensation treatment, the extreme tensile bow exhibited by the Cu damascene vias would have made the wafer bonding step much more difficult.

The seemingly harmless need for NMOS waferbow compensation during the Cu damascene step was a prelude for bigger problems when it came to the PMOS wafers. This was because the laminate layers that hold the PMOS and the handle wafers contain two, 10 to 20μm-thick Al release layers, both of which add a tremendous amount of tensile stress to the overall stack. Thus, the PMOS-handle complex exhibited an enormous tensile waferbow that hindered the handle wafer bonding, the backside Cu via damascene, *and* the final Cu-Cu bond processes. Table 2.2 lists some bow measurements taken from various PMOS wafers after handle bonding and backside Si grindback / etchback. Moreover, Figures 2-12 and 2-13 are the pictorial representation of the PMOS-laminate-handle stack and a photo of a bonded-etchbacked PMOS wafer used in the bow measurements in Table 2.2, respectively.

Because the PMOS tensile bow was too large to be completely corrected by backside film compensation, the only course of action in achieving an uniform PMOS-NMOS bond across the entire 6" wafer was to use a piston downforce of 10,000 N (maximum machine limit) and a piston backstop collar overpress distance

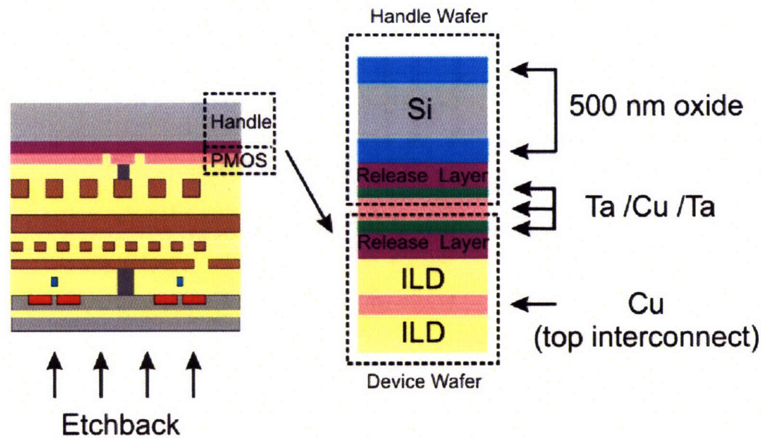


Figure 2-12: Details of the laminate stack in reference to bow measurements presented in Table 2.2.

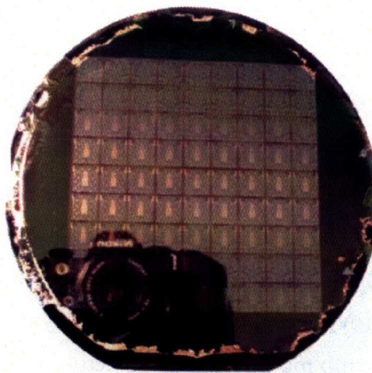


Figure 2-13: Photo of PMOS wafer, after successuful handle-wafer bonding and substrate etchback after a piston-based bow compensation press. The Al release layer thickness in this sample was 10 μm.

of -100 μm (in other words, dialing in the total wafer + diffuser thickness and then subtract 100 μm to compensate for the inherent waferbow of the individual PMOS and handle) were required.

Lesson Learned: In the EV aligner-bonder system, the waferbow pin height (ie. The backstop collar) needs to be calibrated frequently to ensure a good wafer contact. Moreover, contact energy transfered from the piston to the wafer pair can be maximized by using a solid, compliant diffuser made out of graphite. Finally, a thick Al release layer in the laminate structure or a thick Cu damascne via overfill can create tensile waferbows in excess of 100-200 μm, thereby requiring waferbow compensation techniques such as backside film deposition or stress-induced cracks on wafer surfaces to enable further fabrication processes.

Substrate	Waferbow		Radius of curvature
X2-side polished Si + 1 μm oxide (reference)	16.44 μm	^	120.03 m
20 μm Al laminate– blank dummy	300.55 μm	U	5.98 m
20 μm Al laminate – real PMOS	108.95 μm	U	15.94 m
10 μm Al laminate – real PMOS	177.25 μm	U	10.27 m
10 μm Al laminate – real PMOS with thinner handle substrate	286.46 μm	U	6.98 m

Table 2.2: PMOS waferbow measurements for different Al thickness within the laminate structure that holds the PMOS and handle wafers together. A " ˆ " sign denotes compressive bow, and a "U" sign denotes tensile bow. Each indicated Al thickness correspondsto the thickness of a *single* release layer shown in Figure 2-12.

2.3 Wafer-wafer Alignment

The ability to accurately align one wafer to another can also be the determining factor in whether or not 3-D would be a viable technology in the future. "Why is this important?" the reader might ask. It's simple: A n-fold increase in vertical via pitch results in a n^2 increase in via density. This might not seem much, but it can be significant as the via pitch approach the 1 μm barrier. To see this, let's examine the effect of vertical integration as we scale down the via pitch. For instance, using wafer bonding techniques, it is conceivable for monolithic 3-D integration to improve the vertical via density by a factor of 400 using monolithic 3-D integration over a MCM-based (multi-chip module) 3-D architecture. This is assuming that the MCM was made from typical ball grid solder joints with a via width and pitch of P_{MCM} = 100 μm within a 0.8x0.8 cm^2 area, as shown on the left diagram in Figure 2-14. Now, 3-D integration with Cu-Cu bonding, it is very easy to create and bond Cu via pads with a pitch of P_{Cu3d} = 5 μm, provided that our wafer-to-wafer alignment tolerance is much smaller than 5 μm.. and bingo ! Just by increasing the n by 20, the vertical via density by a factor 400, or in absolute numbers, our 16 BGA solder balls have just grew to 6400 Cu pads. From a standard packaging perspective, a 400x increase in the number of solder joint connections is considered to be a vast improvement in terms of increased connection density.

Furthermore, if the wafer-to-wafer alignment technology improves to about +/- 1 μm, then it's possible to shrink the via pitch to about 2μm and we would gain another 6.25x in via density. As one should expect, the critical factor in Cu-Cu 3-D integration is how well you can align the two wafers prior to bonding. As mentioned before, for every factor n in alignment improvement, there is a factor n reward in via pitch and n^2 increase in via density. Unfortunately, the same factor of n in alignment improvment means that the wafer alignment system has just increased exponenetially in complexity. Unlike regular lithography where the sub-micron alignment accuracy by a stepper is the norm, there is a ceiling to how well one can align

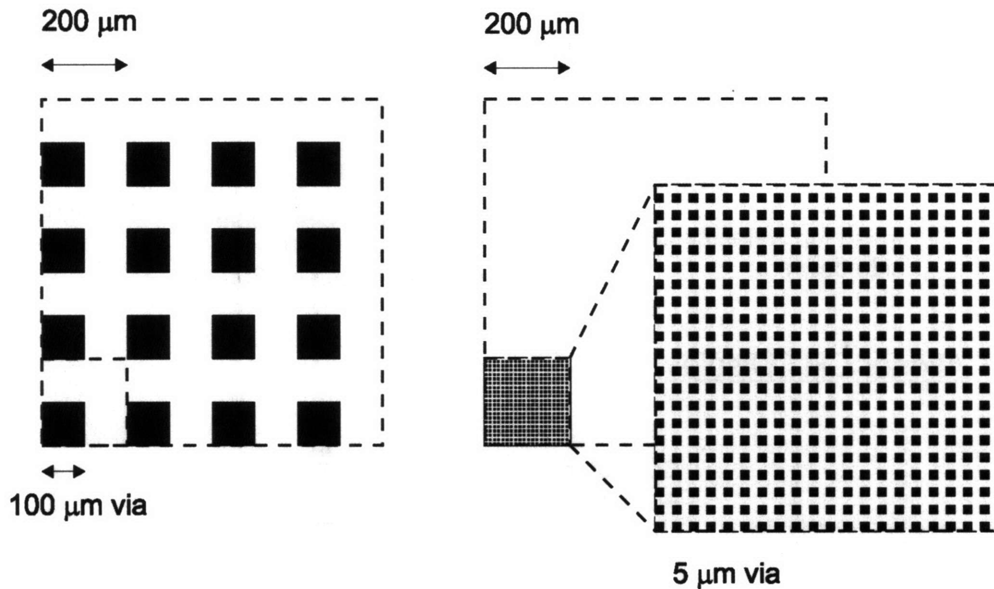


Figure 2-14: Increase of via density by factor of 400 from 100 μm via to 5 μm via.

one 6" wafer to another. The fundamental enemy here is not of mechanical dexterity in wafer handling, but rather that in MTL's current alignmen scheme, we are trying to scan and align opposing surface registration marks on two thick *and* opaque substrates to an accuracy on the order of 1 μm. The emphases on wafer thickness and opacity are multi-faceted:

- If all substrates considered were opaque and were made of Si, then one can design an infrared aligning system that allows a thru-wafer view with a one-shot alignment process. The IR system must have enough depth of focus and light transmission to see through at least two 650 μm worth of Si substrate to be effective.
- If IR is not available, then one must design a system with some sophisticated double-side alignment mark scheme. This will ultimately decrease the total alignment accuracy because two or more registration procedures are now needed. The EV-620/501 aligner at MIT is of this classification.
- The choice of material for the laminate layer and the handle wafer will also determine which optical system one should use. For example, the MIT 3-D process's laminate structure contains blanket films of Cu and Al, both of which are impervious to infrared transmission.

The following discussion may seem mundane to the reader, but understanding in detail how we perform wafer-wafer alignment in MTL will expose the fact that while our alignment system is excellent for MEMS applications, it has glaring limitations when used for integrated circuits where bond pads are smaller and denser. For starters, the EV-620/501 aligner/bonder system in MTL is a non-IR, double-sided registration

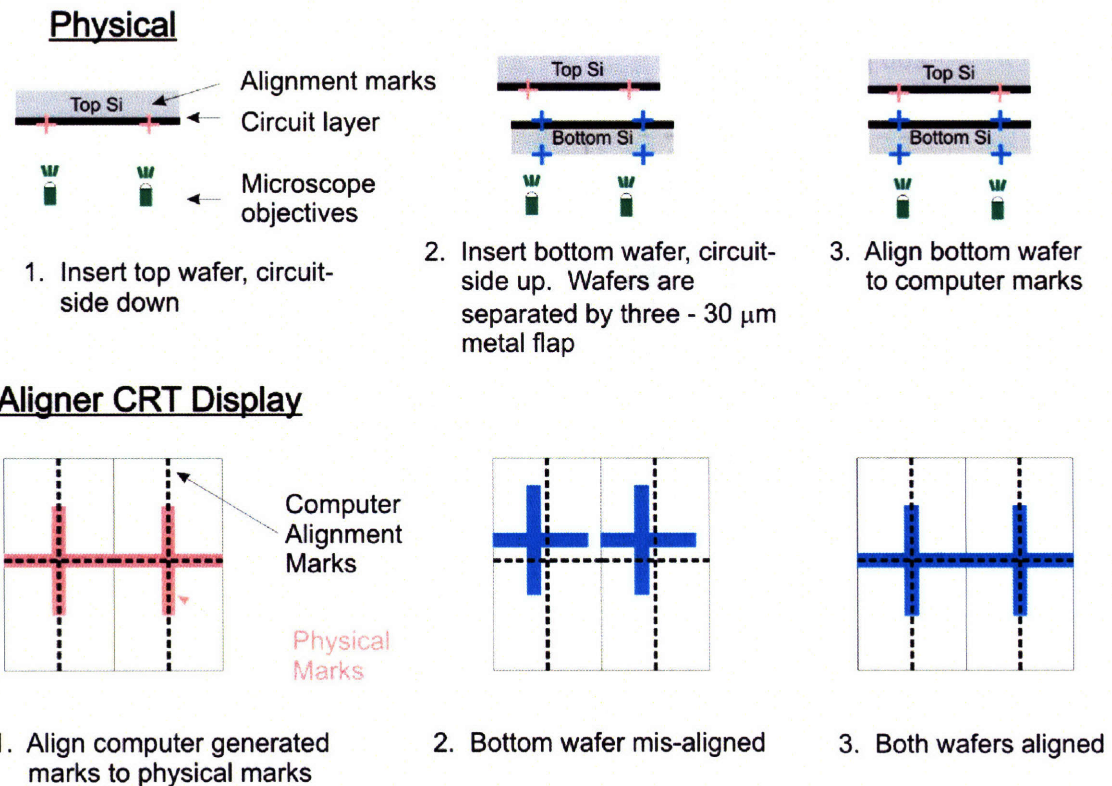


Figure 2-15: Wafer-wafer aligning protocol

alignment system that exhibits a +/- 1 μm alignment tolerance specification for 4" and 6" substrates. The actual alignment process is shown in Figure 2-15.

Although the registration mark positions of the top (or the first) wafer can be made arbitrary within the microscope's field of view, the position and alignment accuracy of the second wafer's pre-made, double-side registration marks are the key to success [5]. Assuming that the contact aligner has an inherent +/- 1 μm misalignment tolerance, an accumlation of both orthgonal +/- 2 μm shift and a slight θ shift can occur between the top and bottom registration marks. Now, combine this shift with a +/- 1 μm orthgonal and some

[5]Although not shown in Figure 2-15, the key instrument used in the front-to-back registration lithography step in MTL is the Electronic Visions EV1 contact aligner. Unlike a normal contact aligner, this particular model has 4 microscope objectives - 2 topside, 2 bottomside. The description of the process is actually very similar to that of Figure 2-15:

1. Align, expose, develop, and etch frontside Cu registration mark using the top objectives
2. Insert same mask in TRL's EV1 contact aligner
3. Optically align the ***bottom*** microscope objectives to the mask
4. Align computer-generated superficial crosses to the actual marks on the mask
5. Insert opaque wafer, with pre-etched frontside registration marks facing down and resist-side up
6. Using the same bottom objectives, manaully align the pre-etched frontiside marks to the computer-generated crosses
7. Expose, develop resist on "backside" of wafer; etch backside registration marks

more θ shift during wafer-to-wafer alignment, and sudden you're up to a +/- 3 μm orthogonal mislalignment plus maybe up to an 1 degree shift in θ, which further degrades the orthogonal components. If one takes the +/- 3 μm wafer-wafer misalignment as a base parameter, then the minimum Cu pad dimensions can be no smaller than about 5 μm x 5 μm. Moreover, the air moats dimensions described on Item #6 in Figure 2-4 also cannot be smaller than 5 μm, and this can reduces the effective Cu-Cu bond area across the wafer by a significant amount. [6].

In all, due to the machine-dependent aligmnment constraints, the 3-D ring oscillator and the heat diode mask set were made to have an overly-relaxed set of layout. Pre-empting the discussion in later chapters, the Cu bonding layer's design rules were:

1. Assume a maximum wafer-wafer misalignment tolerance of +/- 10 μm
2. Thru-wafer Cu damascene via dimensions were 10 μm x 10 μm
3. Cu bonding pad has to overlap thru-wafer Cu vias by at least 5 μm
4. Minimum length of any electrically-active Cu areas is 10 μm
5. Mimimum air moat length in any direction is 10 μm

The +/- 10 μm maximum wafer-wafer misalignment tolerance assumption can easily be satififiedon the 6" wafer level but not on the die-level (more about die-die alignment in a later section). Furthermore, as mentioned previously, any Cu-Cu misalignments will change the air gap isolation distances between the electrically- active and inactive Cu pads. One consequence of this could be an unwanted increase in parasitic capacitance skwed towards one particular layout direction, and its exact effect on circuit performance can be very difficult to predict or extract. This could be an implicating factor in why our face-back, die-level CMOS oscillators failed to function (also to be discussed later), but it's difficult pinpoint if that is true or not. To continue with our misalignment discussion, since the air gap moat widths across the die is no longer a constant 10μm, the local bonding strength among different regions of a given die could become unbalanced. A quick example of this can be seen in the bonded Cu-Cu chain resistor, whihc was fabricated on the same wafer as the face-face 3-D ring oscillators to be discussed in Chapter 4, was displayed in Figure 2-16.

In this particular case, our wafer-wafer misalignment was merely around 2.5 μm in the x-direction only [7]. However, since the designed width of each resistor leg was 10μm, the effective Cu-Cu overlapping area has decreased by a whopping 50% (A misalignment of 2.5 μm means you have to subtract 2.5 μm from both edges of the Cu line, which totals to 5 μm Cu-overlap loss) ! Exacerbating the situation is the large area of air moats surrounding these resistors, which in which the purpose of such layout was to gauge how

[6] Worst of all, the aforementioned misalignment tolerances are actually what one would get ***on a good day and does not include any misalignments caused during the physical transportation of the wafer pair going from the aligner to the bonding chamber.***

[7] In other words, anyone who regularly bonds wafers in MTL would say, "The bonder aligned pretty well today !"

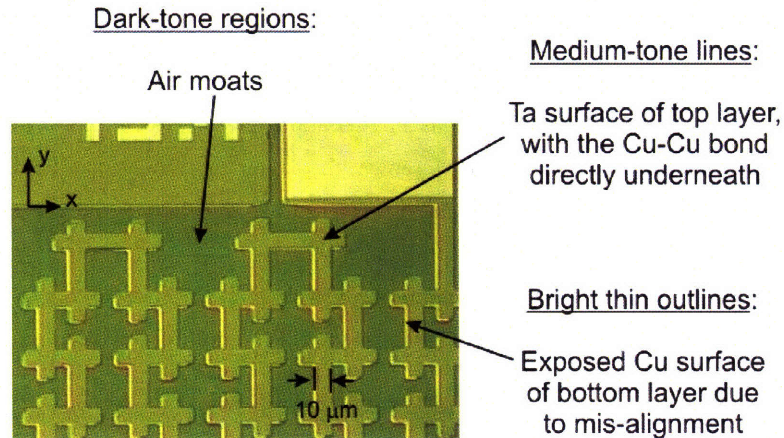


Figure 2-16: Photo of bonded Cu-Cu chain resistor. For reference, the sample was prepared by a face-to-face waferbond of 2 NMOS wafers and a subsequent SF_6 / TMAH bulk Si etchback from one side.

large can the non-bonding area of a given plot be until either stress-induced film delamination or physical puncture from sag points along the free-standing film occur. From the photo, one can observe that the overall structural integrity in regions of sparse Cu pads can be quite good, despite exposures to a 5+ hr plasma etch and a 3 hr hot TMAH attack, if both wafer-wafer alignment and bond uniformity were close to perfect. Conversely, if wafer or die-die alignment was way off, acid or alkaline encroachment due to liquid penetration into free-standing film defects can wreak havoc on the devices. Such disaster can be seen in Figure 2-17, where acid corrosion occured in rampant in regions of either sparse Cu coverage or with very wide thru-wafer Cu damascene vias (severe via dishing can create huge voids within a big Cu-Cu bond pad).

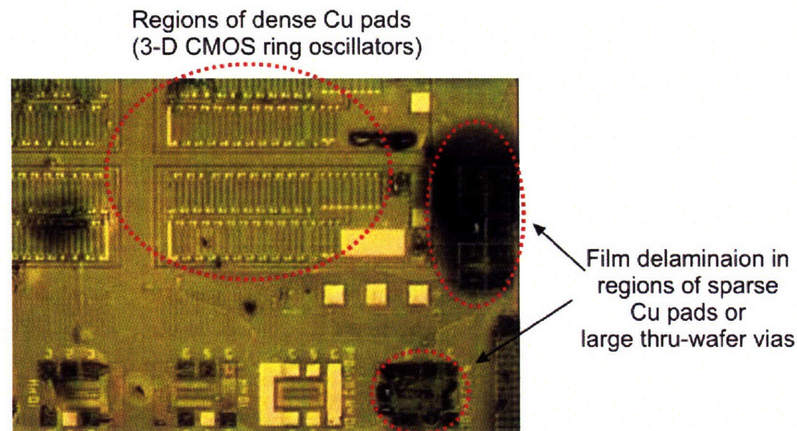


Figure 2-17: Photo of bonded 3-D ring oscillators and dummy solenoid structures, where the mislalignment was about 9 μm in the vertical direction. For reference, the sample was prepared by face-to-back, die-die bonding of a PMOS/NMOS pair and a subsequent 6 hr HCl acid encroachment release of the handle die.

Now, with all the fabrication and design problems associated with our base +/- 3 μm figure of merit, how can we make it better ? It is the humble opinion of the author that the cure for misalignments (which leads to tighter air moat tolerances, denser Cu vias, etc.) lies within a state-of-the-art, custom-made IR aligner and bonder setup like those from Mitsubishi's Advanced Mechanics and Systems group [17]. In their system, the pre-bonding, alignment, and the bonding chambers were all integrated into one contiguous complex and all operate under ultra-high vacuum (UHV) for cleaniness. To be more specific, the device wafers first undergo a pre-bonding Ar sputter clean to create unstable dangling bonds along the metal-air (or oxide-air) interface. Without leaving UHV, the wafers then travel to the alignment stage where the combination of their IR setup and the piezoelectric-controlled micromanipulators can obtain an alignment accuracy of sub - 0.5 μm without the use of a double-sided registration system. Once aligned, wafers were immediately set into contact mode within the UHV environment, where a bonding wavefront will start from the point of contact and travel outward towards the wafer edges. This is very similar to the van der Waals bond between two pre-cleaned Si wafers, but with a twist: The group claims that using their Ar pre-sputter clean approach, they can bond *any* metallic surface to its Siamese twin regardless of which element it is. Above all, the aligned and contacted wafers were transported to the piston chamber without human handling of the substrates and the bond chuck, thus further minimizing alignment errors.

Lesson Learned: Critical wafer-wafer or die-die alignment is the key to maximizing the vertical-to-planar interconnect density raio in 3-D integration. Alignment tolerances can directly affect parameters such as the size of Cu bond pads, the pad-pad air isolation distance, the total percentage of Cu coverage in a given footprint (directely related to the overall bond strength and quality of the wafer pair), the pad-pad parasitic capacitances, and much more. While MTL's double-registration mark system was more than adequate for MEMS-related devices where alignment tolerance were more forgiving, it may not be suitable for 3-D integration purposes. A more robust, in-situ pre-clean, IR alignment, and bonding system like the ones from Mitsubishi is probably the desirable option.

2.4 Handle Wafer Release

To recapitulate, the thesis discussions have so far focused on the material science (Section 2.2.1, the coarse mechanical (Section 2.2.2), and the fine mechanical engineering (Section 2.3) aspects of wafer-wafer bonding. In this section, we shall touch on the the chemical aspects of the MIT 3-D integration flow - the design and implementation of the handle-SOI sacrificial bond, also known as the laminate layer, and its destruction mechanism: The handle wafer release. To start with, let's a global look at the entire 3-D process flow in Figure 2-18.

To have a robust face-to-back 3-D process, one needs to create a sacrifical bond layer in Step 2 that's strong enough to withstand the vigorous mechancial grindback and TMAH etchback in Step 3, the Cu via damascene CMP in Step 5, the permanent Cu-Cu bond's thermal budget of 400 °C for 30 min in Step 7, and

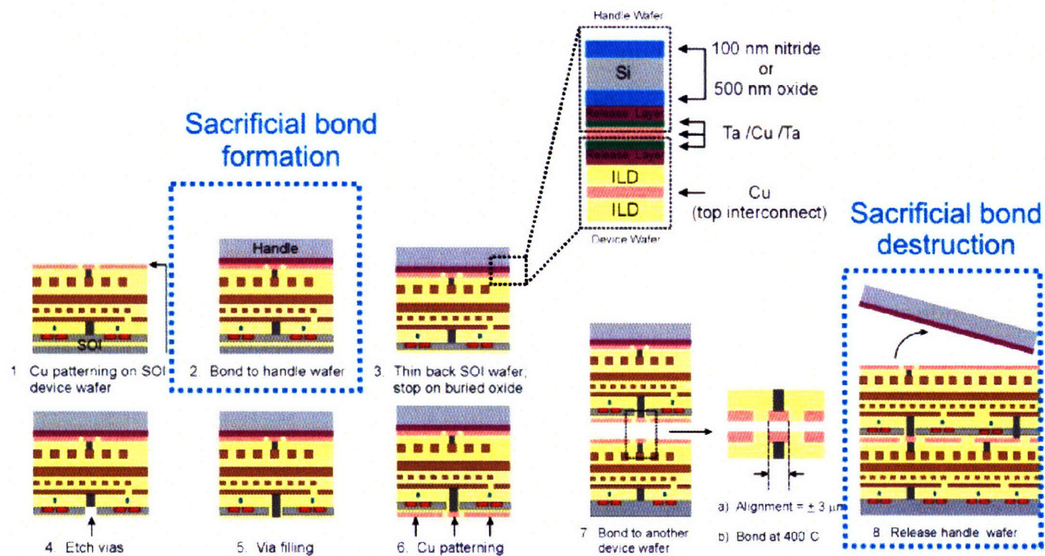


Figure 2-18: Bird's-eye view of the 3-D flow, emphasizing the formation, duration, and the destruction of the sacrifical handle-SOI bond

then be able to be destroyed somehow in Step 8 without damaging any other layers in the resulting 3-D stack. Knowing that there's no single "magic bullet" bonding material that can satisfy all these conditions, the basic design of the sacrifical bonding layer, or also referred in our literature as the *laminate structure*, will be divided into three parts: Choosing a suitable permanent bonding material that can withstand all forms of corrosion or mechanical abuse, choosing a suitable cladding material or scheme that will be destroyed by a specific chemical reagent, and choosing the release reagent itself; in essence, the permanent bond material will stay in its place forever, while the release cladding layer will act as a "permanent bond bypass" and will facilitate separation of the handle substrate and the main 3-D stack. A rough schematic of the laminate structure scheme can also be found in the magnified inset in Figure 2-18.

Laminate Structure Construction

Rather than going through the entire design matrix, it will suffice to say that the optimum choice for the bond member inside the laminate structure should be **Cu**, the optimum wafer release layer material was chosen to be **Al**, and the optimum releasing reagent was chosen to be hot HCl. Here's a checklist to justify these choices: **Copper was an excellent bonding medium choice because**

- We already knew from previous experiments that a 400 °C Cu-Cu bond was strong enough to survive both the razor test (tensile stress) and mechanical grindback of Si (shear stress)
- We know from chemistry that Cu can withstand all forms of hot alkaline attack up to 120 °C [8].

[8] Household chemistry proof: Do you have a washing machine with Cu pipes installed ? Copper pipes can last years of abuse by a

- We know that Cu can withstand multiple rounds of 400 °C anneals
- We know that Cu can withstand any common cleanroom acids that are non-oxidizing (HCl, HF, cold H_2SO_4, cold H_3PO_4, buffered-oxide etch, CH_3COOH), all developers, and all cleanroom solvents except "Microstrip"
- For extra protection, it is fortunate that the Ta diffusion layer is chemically inert to **all** known aqueous chemical attack up to 150°C with the exception of strong HF or HF + nitric acid solutions. In some respect, Ta is more "noble" than Au or Pt in the analytical chemsitry world because it is even impervious to hot solutions of aqua regia (HCl + HNO_3).
- We know that Cu is impervious to nearly all room-temperature plasma etch chemistries, due to the low vapor pressure of its etch products (chlorides, fluorides, etc.). However, Cu will oxidize quite fast in any kinds of ashers.

On the other hand, **Al was an excellent choice for the release layer** cladding material because

- Although aluminum oxide (Al_2O_3) acts a self-passivating oxide in many semiconductor processes, it succombs easily to strong, hot acid or base attack when the anodized film is thinner than 1 μm [9]
- Al is impervious to SF_6 plasma attack, which can be used to our advantage if we chose to dry etch the entire handle substrate from the top instead of using a wafer release reagent.

Last, but not least, hot HCl was the optimum choice for the release reagent because it does not attack Cu (provided the air content inside the solution wss kept in check to reduce Cu oxidation) and it can destroys layer of Al upon contact. But wait.. there is a conundrum here: *Won't the Al cladding layers undergo severe corrosion during TMAH etchback ? This would cause a premature separation of handle wafer from the PMOS-SOI, right ?* The answer to both questions was "yes," and indeed there is a scheme to circumvent this problem. Instead of depositing the layers inside the laminate structure one after another, the Al layer deposition can be first made to have an exclusion ring around the wafer edges. Then, one can protect this Al layer from any forms of corrosion by depositing the Cu/Ta overlayers without exclusion rings. A schematic of the laminate layer construction can be seen in Figure 2-19

If the reader has an insatiable thirst for more details on the science behind acid-encroachment handle wafer release, which includes a brief dialogue on the physical, interfacial, and surfactant chemistry of things, please refer to Appenix B for further information. Otherwise, the reader now must be thinking, "Well, that's interesting, but does the acid release process actually work on a bonded 6" wafer pair ?" Sadly, the answer is no, and in fact, it barely worked on a bonded 4" wafer pair. Figure 2-20 depicts the outcome

---

combination of scalding water, 3% sodium hypochlorite (aka bleach, an oxidizer *much more* potent than the feared nitric acid at equal molarity), and various forms of alkaline laundry detergents

[9] Once the protective oxide was destroyed, Al is one of the "wimpiest" metal in terms of corrosion resitance, for it is so amphoteric that it dissolves in solutions with pH <4 (Coca Cola) and ph > 10 (Drano).

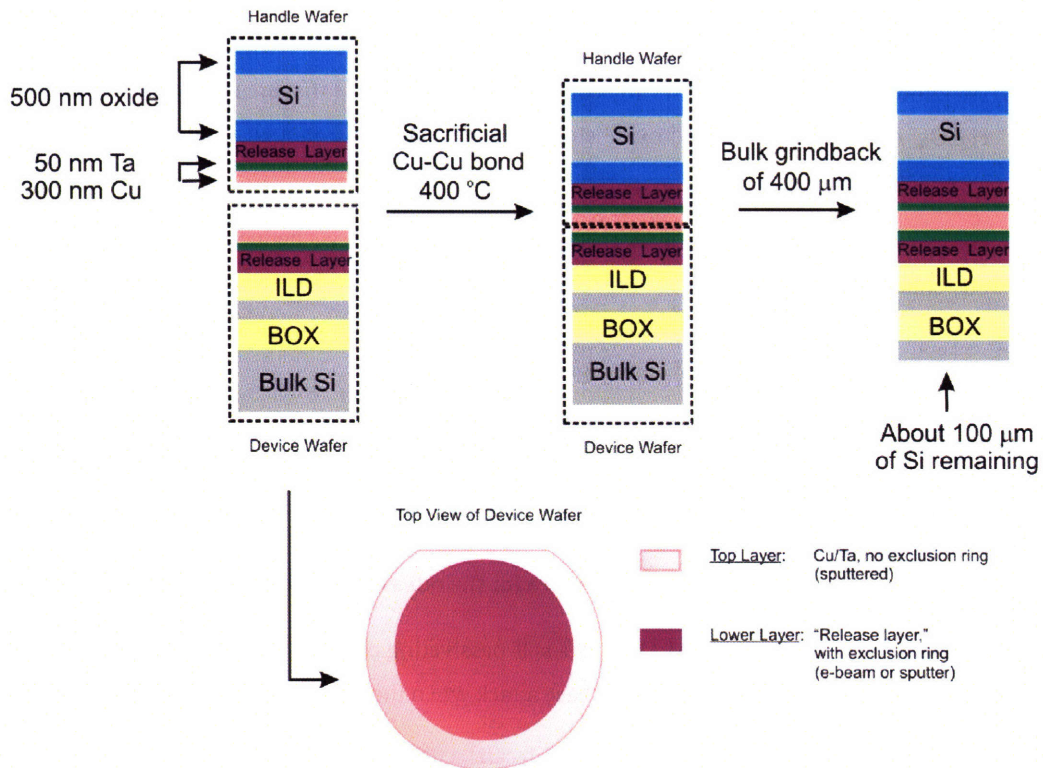


Figure 2-19: Constructing the laminte structure: The Al release layers can be protected from mechanical grindback and TMAH etchback by an exclusion-ring deposition followed by a non-exclusion ring deposition of Cu/Ta.

of acid encroachment on bonded 4" pairs as a function of the Al release layer thickness. The release times varied from sample to sample, but most of these measurements were taken after 1 day's worth of acid soaks. The release reagent in this case was a volumetric mixture of 1:1:1:0.16 of HCl, H_3PO_4, water, and the non-ionic surfactant nonylphenol ethoxylate (NPE) 12-mol moiety. The temperature of the solution was set at 100 °C.

What always happened was the following. Within the first hour of acid release, the initial transient acid encroachment rate is fast and furious. However, but as time approaches infinity, the resulting acid encroachment distance reaches an asymptotic value, but not quite. Basically, there seems to be a two-tier transient response, the first being approximately exponential in time, and the second following a square-root time dependence. Nevertheless, the conclusion is that even with a dual 20 μm Al release layers, wafer-level handle release with acid encroachment is impossible due to surface tension effects.

Lesson Learned: The optimal sacrificial handle-SOI laminate structure design comprised of two parts: Choosing Cu as a permanent bonding medium and choosing two cladding layers of Al, sputtered with exclusion rings, as the release layer. Thus, the resulting laminate structure was able to with-

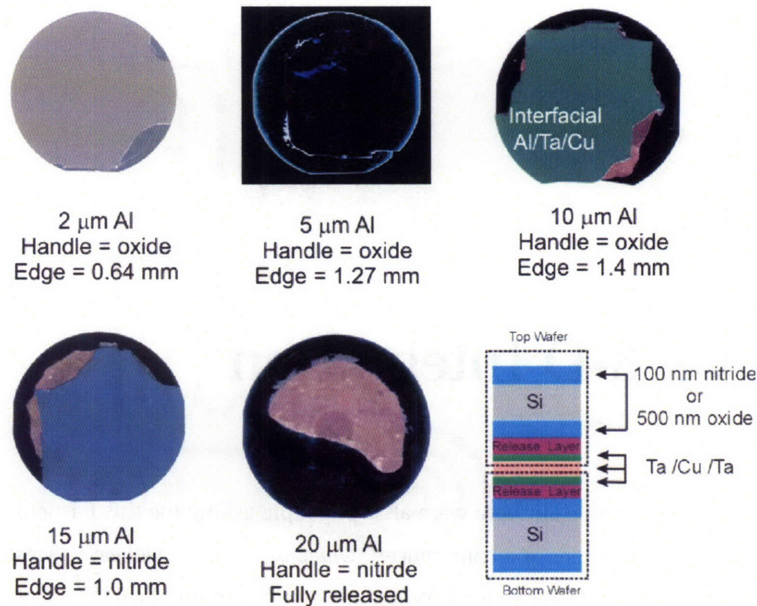


Figure 2-20: Results of 4" wafer-level acid encroachment release. In each sample, the indicated Al thickness was that of the release layer, the handle material refers to either a 5000 A thermal oxide base or an 1000 A LPCVD nitride base, and the "Edge" value refers to the maximum bond encroachment distance.

stand mechanical grindback, TMAH etchback, Cu CMP, and the thermal stress applied to the system during the final Cu-Cu interconnect bond. Although a hot HCl solution can be used to destroy the Al release layers and causing the collapse of the laminate structure, the acid encroachment distance within the release layer cavity was not deep enough to facilitate the separation of even a 4" bonded pair. Therefore, this method of wafer release can only work if:

- If additional release cavities were deep-RIE etched from the top of the handle wafer, or
- If the substrate size was reduced further, such as on a die-scale.

And this is where our next adventure takes us - the realm of die-level 3-D integration.

Chapter 3

Die-level 3-D Integration

During my graduate career at MIT, there were always people asking me this particular question while we were presenting our 3-D schemes at a conference: *"Ummm.. say, don't yall have a yield problem when you do 3-D integration on a wafer-level?"* Yes indeed. When one attemps to bond a 6" wafer with a 9x9 die array to its counterpart, we run into an old semiconductor yield issue called the known-good-die (KGD) problem. Simply put, no one can guarantee that a working 3-D device can be found at a predetermined spot on a bonded die without pre-testing those two dies individually beforehand. Moreover, the KGD problem can be exacerbated when an engineer decides to waferbond a 4-level 3-D stack, where a yield of 95% per wafer can suddenly drop of $(0.95)^4 = 81.4\%$ for the multi-layer stack. Therefore, it makes perfect sense on paper, at least, to convert 3-D integration into a die-level scheme where the KGD problem can be circumvented altogether. In addition, if one chooses die-scale integration, he or she can take a KGD from a III-V process, bond it to a CMOS KGD, and the result is a serious exercise in OEIC integration ! Fortunately in the MIT 3-D process, there are numerous breakpoints where one can do a wafer-to-die scale conversion, and this was depicted in Figure 3-1.

In Figure 3-1, our recommendation for the earliest breakpoint is right after Step 3, or the Si substrate etchback. As noted in the figure caption, it is very difficult to perform mechanical and Si grindback on a die level mainly because of the tooling involved in handing such small substrates can be more trouble than it is worth. Also, if the wafer-die scale conversion occcured there, the Al release layers no longer contain an Cu/Ta edge envelope, and will therefore be prone to TMAH corrosion. Furthermore, if one desires a "damascened" backside via, we would actually recommend that the wafer-die conversion be pushed back to Step 6. This was the path we took for the thesis work on the face-to-back bonded CMOS ring oscillators.

While the perceived benefits of going to die-scale 3-D integration, whether it's heterogeneous integration or what not, are plentiful, it's obvious that there's no free lunch here. The three demons of die-integration are: Throughput, the quality of die-die alignment, and the quality of the die-die bond. We will touch on the last two points in the remainder of this chapter, and both of them again require discussions on

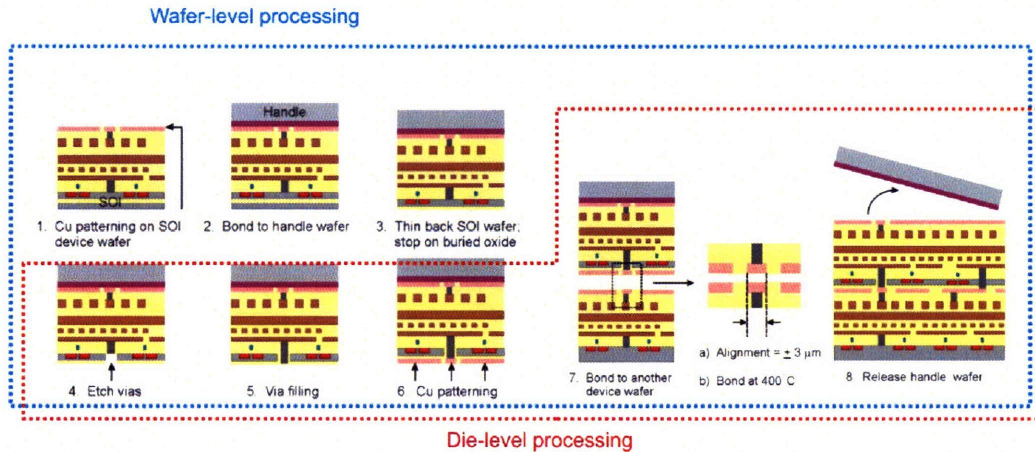


Figure 3-1: The MIT 3-D fabrication flow, with wafer or die-level processing options. Die-level integration is not recommended before Step 4 because it is very difficult to maintain structural integrity of the handle-SOI complex during die-level substrate etchback.

equipment limitations. Since we've already mentioned all the major bonding parameters at great extents in previous sections, the following discussions will be much more brief and succinct.

3.1 The Wafer/Die-Bonder

3.1.1 The Bond Chuck

The aformentioned EV 620/501 wafer aligner-bonder system in TRL can be retrofitted into a die aligner-bonder with a few more gadgets. To begin with, since the vacuum grooves that hold the 4" or 6" wafers during alignment do not have extensions at the bond chuck center, a brand new die-level chuck was designed and custom made, and a photo of this was shown in Figure 3-2. As seen in the schematic, the center vacuum hole sits on top of a 1 mm stainless steel mesa with a cross-sectional diameter of 4 mm. The idea behind the mesa design was to make sure that the dies were elevated enough to ensure total contact between the substrate, the bond glass, the diffuser, and the piston surface when the entire stack is sitting inside the bonding chamber.

3.1.2 The Bond Glass , Bottom Support, and the Wide-angle Objective

Next, in order for us to continue the equipment discussion, a description of the alignment process must be presented. To the first order, the die-die alignment procedure is exactly the same as in the wafer-wafer case in Section 2.3 but with a slight twist. As shown in Figure 3-3, one would insert the first die into the aligner upside down and the chuck's center vacuum tap will then hold the substrate upside down. After aligning the computer registration marks and the physical registration marks as shown previously in Figure 2-15 on

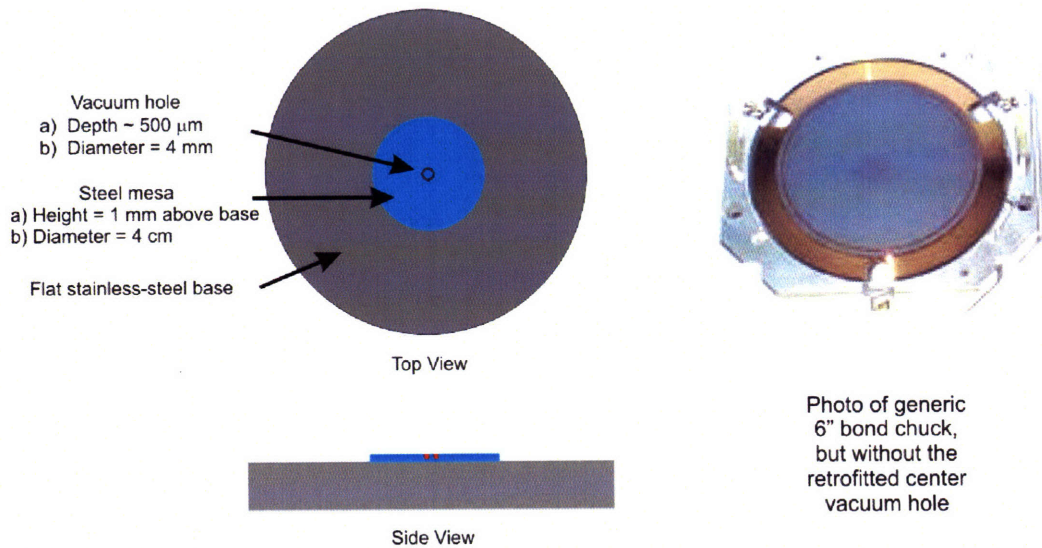


Figure 3-2: Die-level bond chuck schematics. The chuck itself was retrofitted from a regular 6" waferbond chuck, and a sample photo of it was shown on the right.

page 44, the second die with the double-side registration marks can be inserted for alignment.

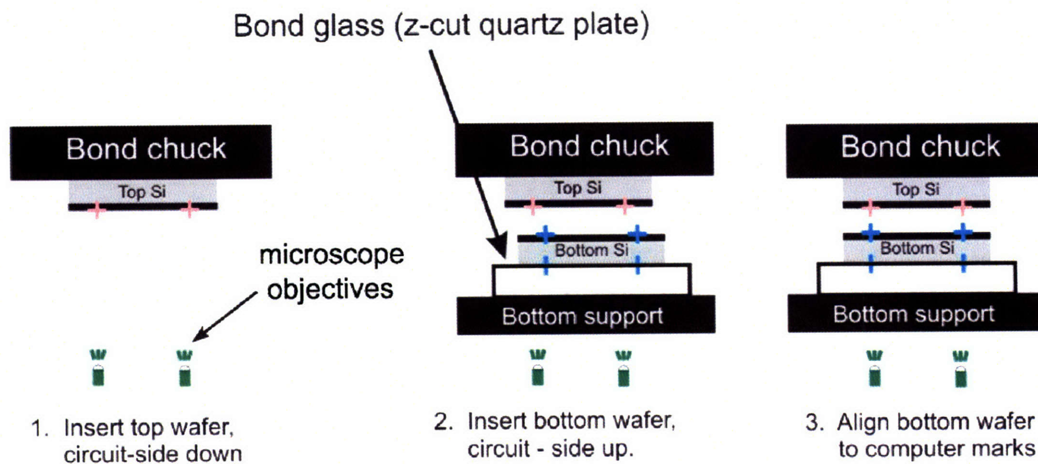


Figure 3-3: Die-die alignment procedure.

Although this sounds theoretically feasible, unfortunately the above tasks cannot be done with the normal 6" bottom aligner supporter because it is both opaque at the chuck center and it also lacks vacuum grooves to hold the bottom die in place during aligning. Going further, even if the aligner's bottom support was fully transparent, neither microscope objectives on the aligner could travel all the way out to the center of the chuck. In fact, the objectives' travelling distance was so small that if one places an 1 in x 1 in die at the center of a transparent support, the objectives were unable to see the edge of the die.

To fix all these problems, a new aligner bottom support was custom-made and a corresponding quartz bond glass was machined with the correct vacuum groove positions in place. Also, a single 3.5x wide-angle extension objective was also purchased in order to extend the aligner's field of view. Previewing things to come, two of these objectives should have been bought because without binocular vision, the theta-misalignments between the top and bottom dies will be ***extremely*** difficult to correct because of the small substrate size and its inherent difficulty in θ alignments.

Next, the schematics for both the transparent bottom support and the bond glass were presented in Figure 3-4. Again, note that both of these parts has to exhibit *naked-eye transparency* to ensure proper aligner operation. The consequence of this was severe becuause, as we shall see, a ***transparent bond glass exhibits some undesirable mechanical limits that could interfere with the overall die-die bonding quality if thermocompression was the method of choice.***

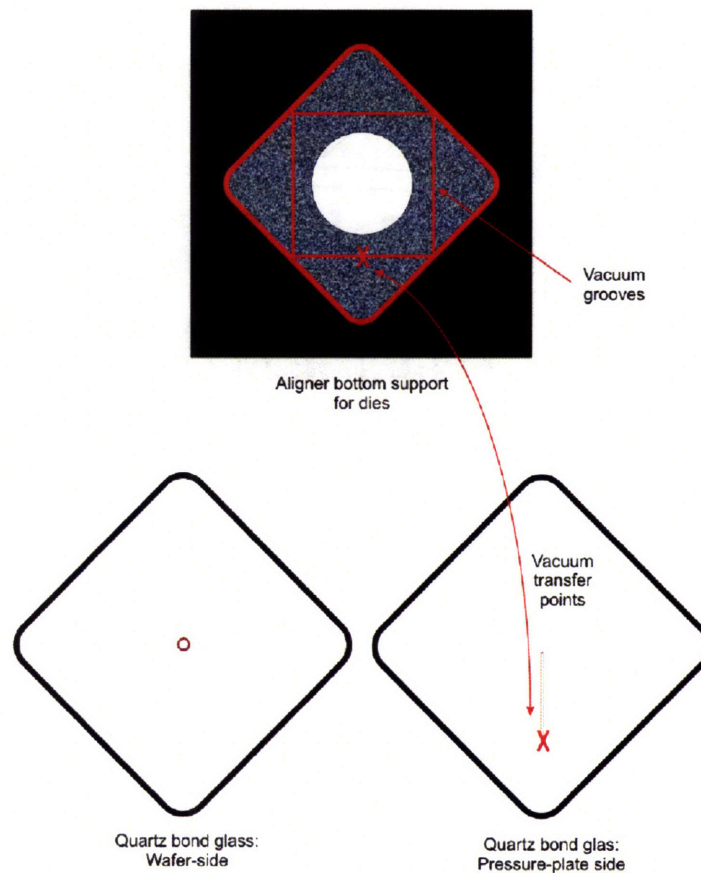
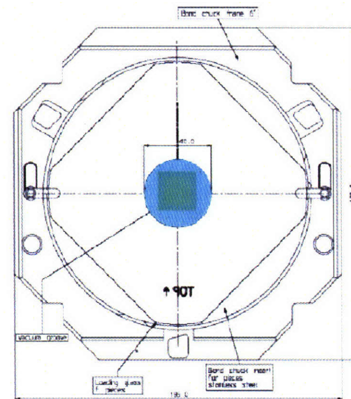


Figure 3-4: Die aligner's new bottom support and the new quartz bond glass.

Once the dies were properly aligned inside the aligner, the combination of the bond chuck and clamped dies have to be physically removed from the aligner, flipped upside down, and then inserted into the bonding chamber. Figure 3-5 shows both a topside and a cross-sectional view of the chuck-die combo right



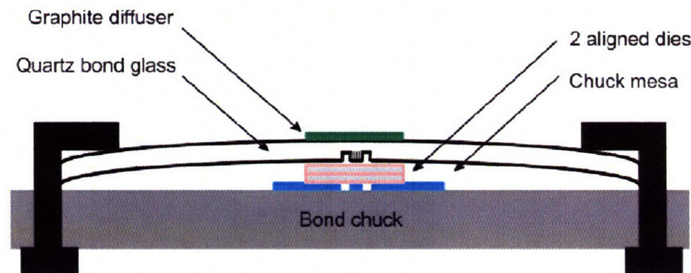


Figure 3-5: Top and side view of the aligned dies, just after alignment and before bonding

before bonding.

3.1.3 The Mesa Pressure Plate and the Graphite Insert

Finally, when the clamped dies were loaded into the bonder, the actual thermocompression routine can commence. This entire setup depicted in Figure 3-6. To ensure maximum energy transfer from the piston to the bond chuck, a stainless steel mesa pressure plate was designed such that its cross-sectional area roughly matches the mesa protrusion of the bond chuck. We still don't know the function of the narrow grooves milled onto the mesa surface, but perhaps it's related to adding friction to the mesa - bond glass interface. Last, but not least, a soft graphite diffuser was placed on top of the entire assembly in order to evenly distribute the force applied by the piston during bonding. As we will see, the condition of this graphite insert was **THE KEY ELEMENT** in obtaining a good Cu-Cu die bond. For the next couple of pages, we will discuss four modifications applied to the aformentioned die-bonding setup that improved the die-bonding yield:

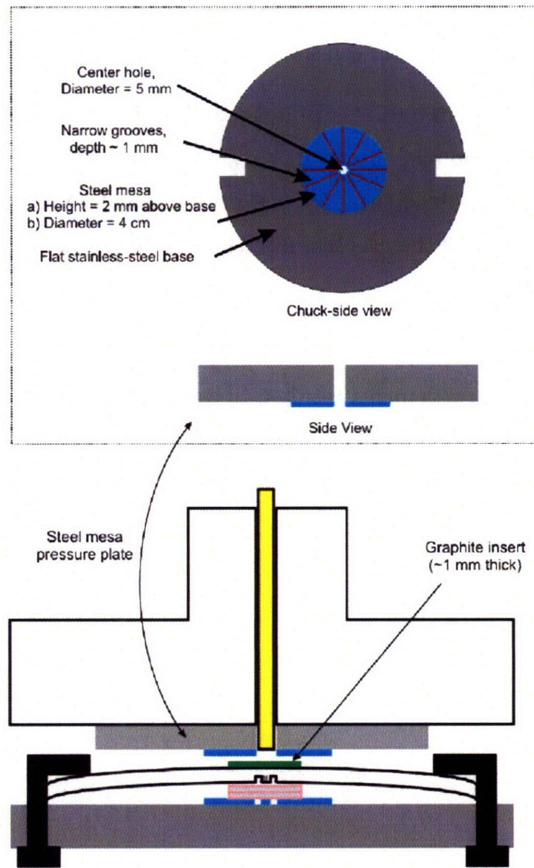


Figure 3-6: The bonder setup during thermocompression bonding.

1. **Fix 1: Home-made Bond Chuck** The mesa bond chuck's center vacuum tap was an impediment in our struggle to achieve a good bond unifomity in both 2"x2" or 1 cm x 1 cm dies. Instead, we made our own die-bonding chuck where the base was entirely flat and silver paint was used for die-attachment in place of vacuum.

2. **Fix 2: Backstop Collar Overpress to Reduce Bond Glass Bow** As seen from Figure 3-6, to ensure maximum energy transfer from the mesa pressure plate to the aligned substrates, one needs to overcome the bond glass's bow, a situation created by the anchoring clamps on either side of the chuck. Re-calibrating the piston backstop and zero-ing in the yield point of the quartz glass proved to be a time-consuming and an expensive exercise, albeit a fruitful one.

3. **Fix 3: Plateau Dies and Post-alignment Check** One cannot rely on faith and assume that the dies are perfectly aligned when the alignment prcoess was completed. Without visual confirmation of the alignment prior to bonding, the end result was always a die-die misalignment on the order of 20 μm or more orthogonally and more than 1 degree off in θ. The best way to check for alignment was to

make the top die's area a bit smaller than the bottom, thus forming a plateau region where registration marks from both substrates can be seen simultaneously by an indepedent optical microscope.

4. **Fix 4: Pyrex Wafer Subsitution and Graphite Insert Monitoring** Foreshadowing to later discussions, the quartz bond glass eventually broke after x-number of bonds. It turns out that a 6" pyrex wafer functioned just as well as the quartz did, and during the pyrex bonding experiments it was determined that the condition of the graphite inserts needs to be monitored closely to enusure the best possible bond quality.

Readers who are interested in the details of these quick-fixes are invited to read Appendix C. But it is suffice to say that these improvements resulted in the successful fabrication of a face-back, die-level integrated 3-D stack. Figure 3-7 is a gallery of a face-back bonded sample that underwent successful acid-encroachement release with a 20 μm Al release layer.

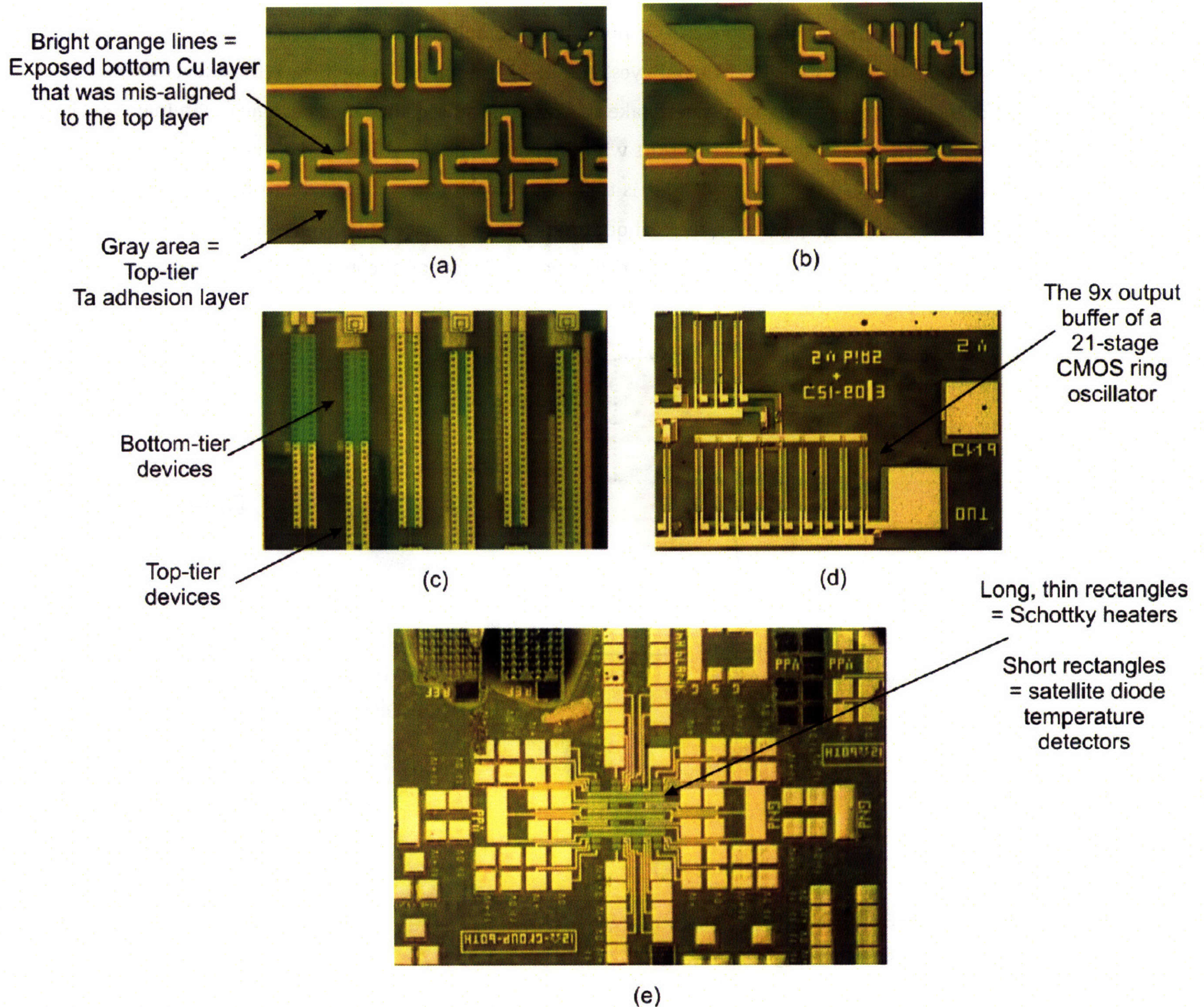


Figure 3-7: A gallery of successful die-level integration. All photos were taken from a face-back, die-bonded sample with a 20 μm Al release layer. (a) and (b) shows that the misalignments between the two wafers were on the order of almost 10 μm. Photo (c) shows is a "looking-glass" structure where in regions devoid of Cu bonding pads, the lower-tier NMOS devices can be simultaneously visualized with the top-tier PMOS devices. Photo (d) shows a section of a completed 21-stage CMOS ring oscillator (when tested later, the output was flatlined but responded to changes in Vdd). Photo (e) shows a two-layer SOI Schottky heaters with satellite temperature- detecting diodes.

3.2 Multi-level 3-D Die Integration

Now that we've "solved" all the problems associated with die-die alignment and bonding, can it actually be used in a 2-level integration process ? Or how about in multi-device level integration processes ? The answers to both questions were fortunately yes, and for the moment, we will leave the 2-level integration results for the next chapter. Right now, let's take an extremely abbreviated look at some preliminary results for multi-level die-level integration. To start with, our vision of a multi-level die integration flow is represented in Figure 3-8, where the basic idea is to first construct a 2-level seed die to which additional die layers can be added onto. However, the real question the reader should be wondering about is: Can our current process actually produce a multi-layer die stack ? The proof of concept can be seen in the 3-layer stack shown in Figure 3-9 and Figure 3-10.

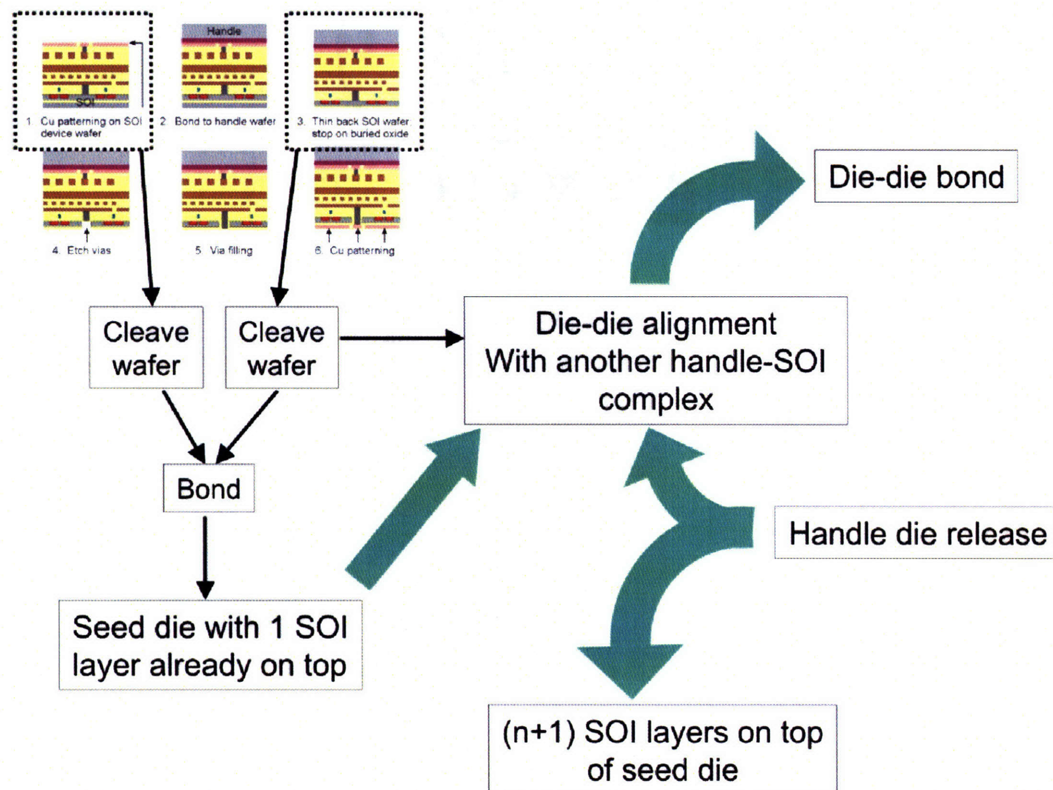


Figure 3-8: The MIT 3-D die-level integration scheme.

Now, can we build the stack even higher ? Absolutely. Figures 3-11 and 3-12 proves the point. shows a 4-layer stack built by three successive Cu bond / acid release loops.

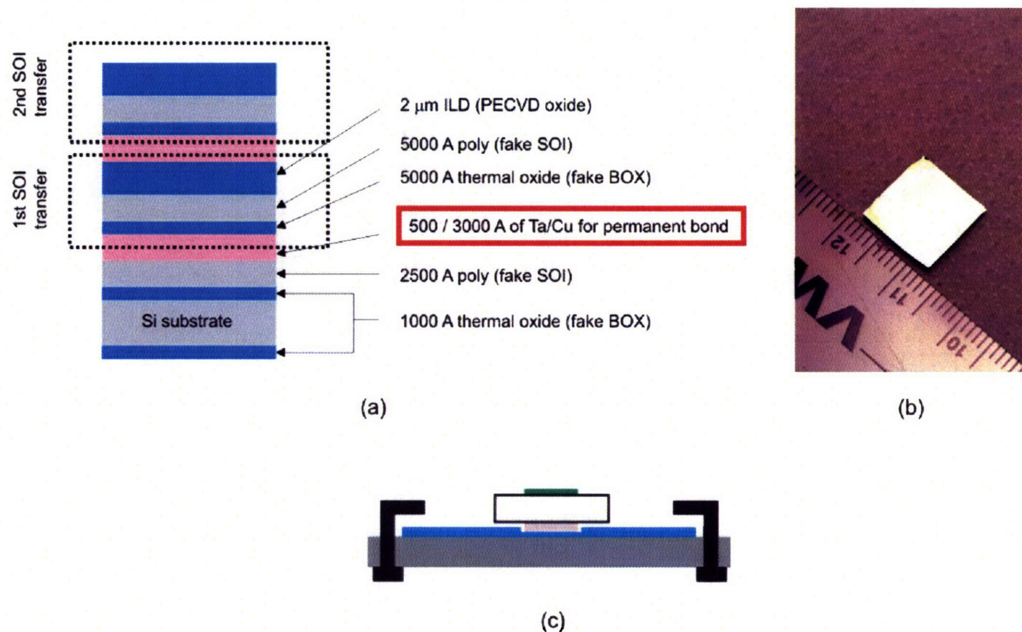


Figure 3-9: Construction of a 3-layer stack. In (a), two separate die-level Cu bonding / acid release loops were created on top of a base SOI substrate, thus forming a 3-layer stack. A proof-of-concept 3-layer blanket stack was made in (b) with the bonding configuration of (c), which was similar to what we had in Figure C-1.

3.2.1 Epilogue

During the course of this entire chapter, we have shown that **within the bounds of our current 3-D process, wafer-level 3-D integration can only work if the devices were face-to-face bonded.** This limitation was due to our inability to successfully release a 6" handle wafer using our current acid encroachment method; that is, we did not construct additional fluidic channels within the handle wafer itself to enhance the acid seepage rate into the Al release layer, of which can be easily done and it is a topic for future work.

On the same token, another topic of interest for future work is using the SmartCut process to facilitate the handle wafer release step - a work started by our former group member C.S. Tan. [24]. Instead of destroying the laminate structure on a macro-scale using acid corrosion or polymer laser ablation (developed by IBM) [25, 26], H^+ atoms can be pre-implanted into the handle wafer prior to the sacrifical handle-SOI bond. Then, at the time of wafer release, a heat treatment of the bond at 400 °C will activate the H^+ implants which evolute into tiny bubbles of H_2 gas. This results in global micro-crack formations on the first 10-100 μm of the Si wafer, thus the handle wafer appears to "peel" off from the 3-D stack during release. It sounds simple, but work has to be done to control the both the thermal cycle of the 3-D process *and any external work done to the handle-SOI bond, i.e. any piston downforces,* after H^+ implants. Preliminary results showed that if a Cu-Cu bond was again chosen to be the handle-SOI bonding medium, the H^2 implant was pre-maturely activated even in a handle-wafer bond at 250 °C and with 2000 N of piston downforce. By stark contrast,

Figure 3-10: SEM photo of the 3-layer stack from sample (b) of Figure 3-9.

C.S. Tan has repeatily shown that if one uses a room-temperature SiO_2-SiO_2 handle wafer bond [1] that was followed by a 3 hr anneal at 300 °C, the H^+ implants remains inactive. This simple example shows that control of the total *work budget* at the handle-SOI interface has to be carefully investigated if SmartCut was to be a viable process.

In terms of die-bonding, we have shown that die-level integration does indeed work and can be repeated numerous times to construct a multi-level structure, albeit a few equipment modifications were required. Nevertheless, in terms of mechanics, a proof of concept was demonstrated in both wafer-level and die-level Cu-Cu bonding. The question now is: Does Cu-Cu 3-D integratio *really* work ? In the next chapter, I hope the reader will agree that the resounding answer to that question is "yes!"

[1] A successful oxide-oxide bond via van der Waals attraction requires, on top of all other parameters, a good microscopic and macroscopic oxide surface planarity. While it's simple to bond a blank, post-CMP oxide wafers with a 0.1 nm RMS local surface roughness, it is much more difficult to bond fully-integrated wafers in which the ILD oxide surface, initially with a +/- 1.5 μm surface topography caused by the underlying MOSFET and Al interconnects, has undergone a professional(Entrepix, Inc. in Arizona) global CMP of which the step heights have been reduced to rolling hills of 200-500 A high. In terms of back-end IC integration, this was pretty much as "flat" as one can get with global CMP, but unfortunately it's not good enough for direct SiO_2-SiO_2 bonding.

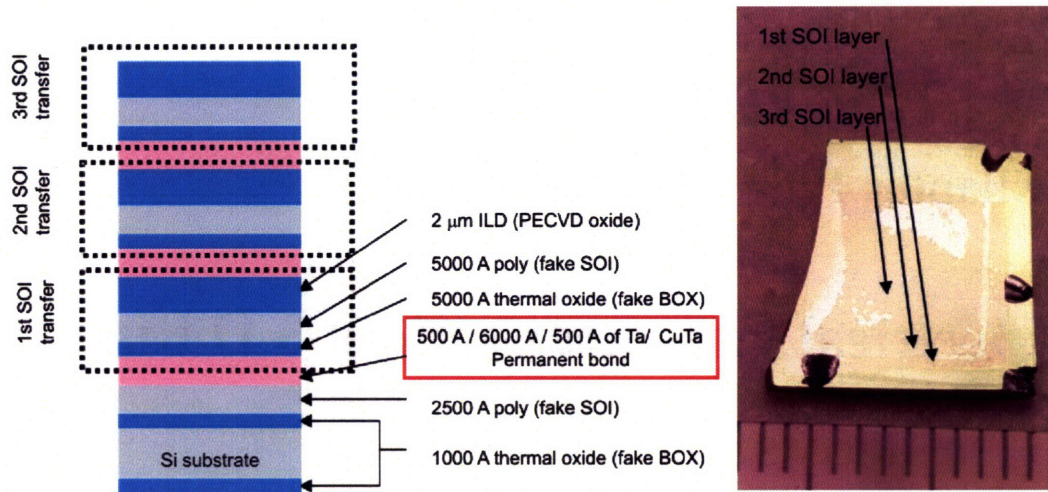


Figure 3-11: Construction of a 3-layer stack. In (a), three separate die-level Cu bonding / acid release loops were created on top of a base SOI substrate, thus forming a 3-layer stack. A proof-of-concept 3-layer blanket stack was made in (b) with the same bonding configuration from (c), of Figure 3-9.

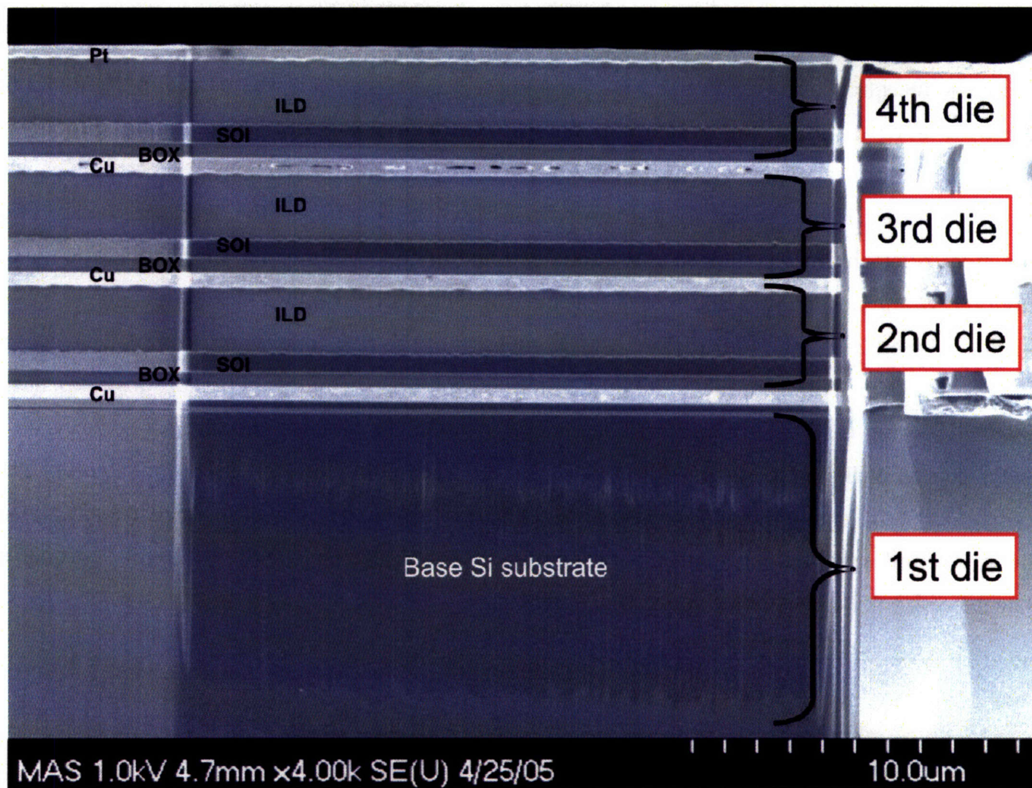


Figure 3-12: SEM photo of the 4-layer stack from sample (b) of Figure 3-11.

Chapter 4

Electrical Characterization: 3-D Ring Oscillators

4.1 Overview of 3-D Ring Oscillator Design

4.1.1 Motivation

From the inception of the MARCO 3-D Interconnect Project, we spent some time in the planning stage deciding on what sort of circuits can best demonstrate the feasibility of 3-D integration. Should we build a simple memory capacitors on top of simple logic ? How about an interconnect-limited FPGA network ? Not before long, it became apparent that a 3-D ring oscillator should be the definitive vehicle of choice - both for its design simplicity (or so as we initially thought) and its fabrication simplicity (ditto the previous notion!). Going deeper, we also decided that the feasibility metric can be "kicked up a notch" by building a CMOS-based 3-D ring, and for each CMOS inverter in the ring, one would completely segregate the PMOS drain-side load from the NMOS transistors by splitting them onto separate substrates. Hence, the 3-D CMOS ring oscillators will ring *if and only if* [1] all Cu-Cu contacts and all PMOS / NMOS devices were aligned well, bonded well, and the devices all work simultaneously within one region of a die. The remainder of this chapter will focus on analyzing the results from our latest lot, which in short, the face-face bonded samples worked but the face-back samples did not. But before one dives into the data analysis, let's take a brief look at the layout design of the oscillators.

[1] That was probably not a good idea to begin with.. putting all the eggs in one basket like that. But I didn't know any better back then!

4.1.2 Overall Structure of the 3-D Ring Oscillator

As previously mentioned, the PMOS and NMOS devices within each circuit were completely isolated from each other by constructing them on separate wafers. To be specific, all NMOS devices will reside on the bottom wafer and will have a choice between a SOI or bulk design, while all PMOS devices will be SOI's and will reside on the top wafer. The actual Cadence layout of a 21-stage CMOS ring oscillator can be seen from Figures 4-1 and 4-2. For this particular example, the width of each NMOS member within the main 21-stage ring was 60 μm, and it corresponding PMOS partner's width was set to 120 μm to offset for the 50% reduction in mobility for holes. Also, each nearest-neighbor inverter within the main ring was oriented in a head-to-tail fashion to minimize the wire length, or equivalently, the parasitic capacitance between the output stage of one inverter and the input stage of the next. A semi-3D schematic of this can be seen in Figure 4-3.

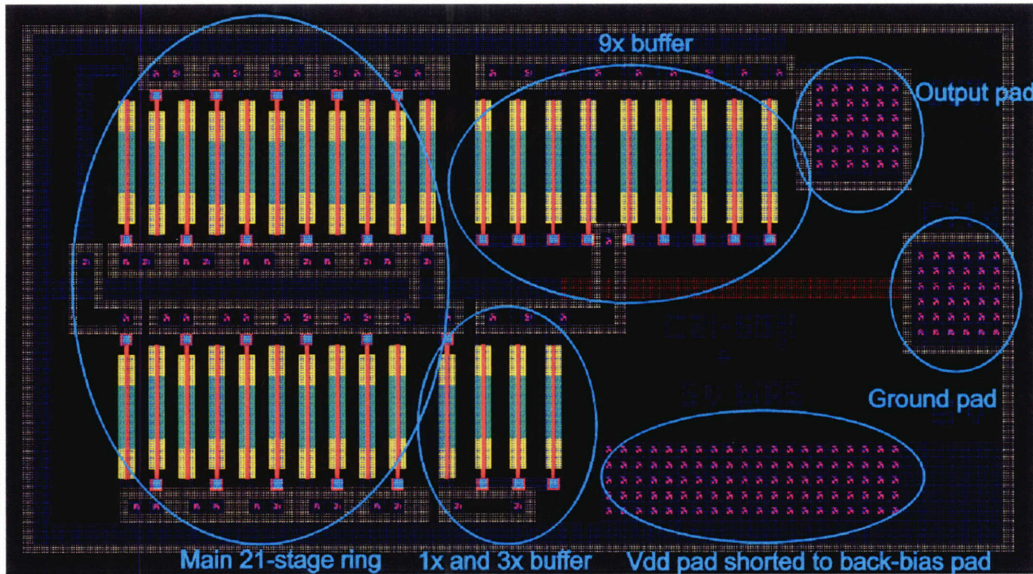


Figure 4-1: Cadence layout of a 21-stage CMOS ring oscillator, W_{NMOS} = 60 μm, L = 1 μm.

If one look a bit closer on the lower right-hand corner of Figure 4-1, there was a giant probe pad labeled "Vdd pad shorted to back-bias pad." This meant that whatever positive Vdd level was applied to the circuit, the PMOS Cu backgates are also going to be at the same positive bias. Physically, this was achieved by shorting the Vdd pad with the auxiilary Cu bond pads that normally would not be electrically active. In fact, during the layout of the Cu bond pads, care was taken such that all electrically-inactive Cu auxiliary pads (i.e. Cu bonding areas that were not in contact with the source / drains / gates of MOSFETS) were shorted together, thus creating a gigantic back-bias plane that could be used to shift the Vt's of the PMOS if be needed to. Furthermore, an additional 10 μm air moat was added to the perimeter of each 3-D ring oscillator cell to completely isolate the backbias gates from inter-cell crosstalk. The Cu-only layout can

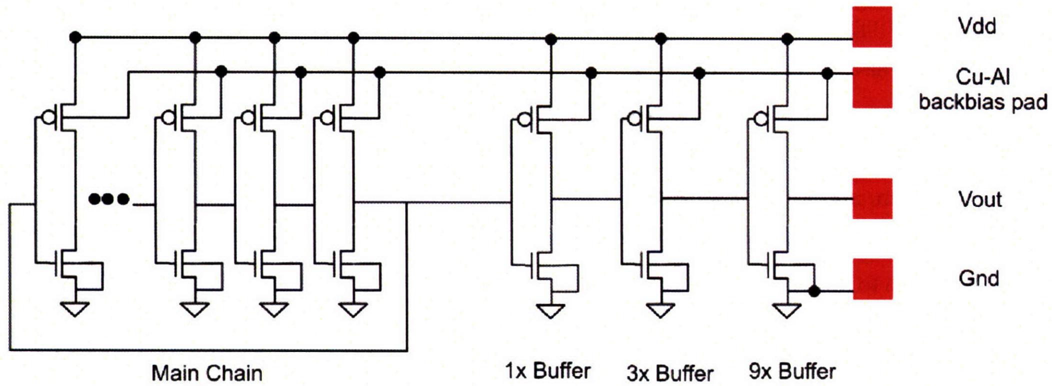


Figure 4-2: Skeleton of a CMOS oscillator schematic, showing all 4 probe pad connections.

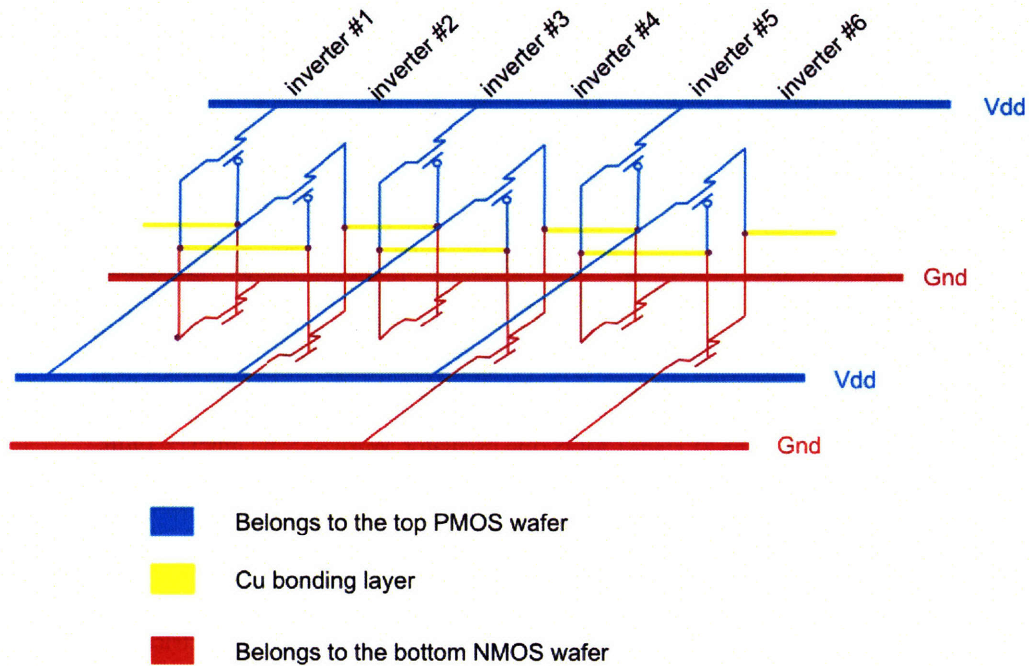


Figure 4-3: A more detailed circuit diagram of a 3-D ring oscillators, showing the head-to-tail connections between the inverter I/O ports and weaving between different wafers.

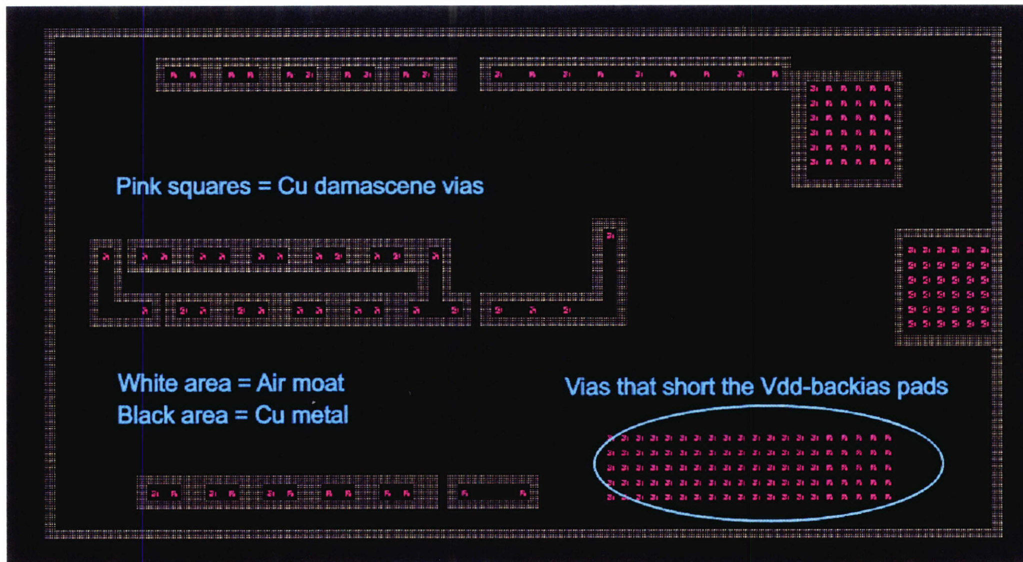


Figure 4-4: A stripped-down layout of the same CMOS ring oscillator, this time showing the Cu bonding layer and the Cu damascene via positions only. Since this is a dark-field mask, the large, black field regions represent Cu metal coverage, whereas the white boxes denote regions of air moat isolations. Also, a group of Cu damascene vias on the lower right shorts the Vdd pad and the Cu backbias region together.

be seen in Figure 4-4. In addition, if one wishes to control the Cu backbias planes independently, all the designer would do is to cut the top Al Vdd pad into two halves - and instantly, you'll have yourself an independent bias pad with the same exact Cu layout as shown in Figure 4-4.

4.2 Preliminary Cadence Simulation of the Ring Oscillators from Real MOSFET Data

4.2.1 Initial Design Considerations

Now we have seen a glimpse of how the 3-D ring oscillators were laid out, let's go briefly into the circuit design side of the layout. To begin with, the entire chip contained two main ring oscillator groups. A quick table of this can be found in Table 4.1.

1. **21 or 43-stage, 2-D NMOS Rings :** All devices associated with these oscillators reside only on the bottom NMOS wafers. The main NMOS switch was either 60 μm or 80 μm wide and all had 1 μm long channels. An active load was used as the load resistor, (these NMOS loads were running in enhancement-mode, where both the gate and the drain were shorted to Vdd), and their widths vary from 20 μm down to 5μm. These 2-D ring oscillators were made as safeguard devices.. just in case if none of the 3-D rings worked, the plan was to pull the 2-D oscillators signals up from the bottom

NMOS wafers to the Al contact pads on the top surface using the huge Cu damascene vias and simple Cu-Cu bond pads as a conduit. If that at least worked, then it was proof that the Cu-Cu contacts were somewhat viable, but of course, not completely viable.

2. **21 or 43-stage, 3-D CMOS Rings:** The NMOS transistor widths in these cells vary were all 60 μm, and the PMOS widths were all twice that, or 120 μm. Within these 2 groups, the MOSFET channel lengths were either 3 μm (my safeguard devices) and 1 μm (more aggressive designs). Further still, each of those groups were divided into cells that could or could not be backbiased independently, and in retrospect, the forced-Vdd bias cells were quite redundant and useless.

NMOS-only rings			CMOS		
	21-stage	43-stage		21-stage	43-stage
	Switch - load	Switch - load	Only NMOS W/L ratio shown; FB = forced backbias, IB = indpendent. backbias		
Inverter #1	60/1 – 10/1	60/1 – 10/1	Inv. #1-2	60/1 – FB	60/1 – FB
Inverter #2	60/1 – 20/1	60/1 – 20/1	Inv #3-4	60/1 –IB	60/1 –IB
Inverter #3	80/1 – 5/1	80/1 – 5/1	Inv #5-6	60/3 – FB	60/3 – FB
Inverter #4	80/1 – 10/1	80/1 – 10/1	Inv #7-8	60/3 – IB	60/3 – IB

Table 4.1: A collection of NMOS-only and CMOS ring oscillator configuratons. Each W/L ratio correspond to the width / length of MOSFETs in microns. In the NMOS-only cells, the first set of W/L corresponds to the switching transistor, while the the second W/L corresponds to the enhancement-mode NMOS active loads. In the CMOS cells, the PMOS transistors' widths doubled their NMOS counterparts. Forced-backbias means that Vdd was shorted to the Cu backbias pad, whereas in independent backbias cells, those two ports are separate.

Before the devices were even built, these ring oscillators were designed with presumed Vt's in mind. Through numerous iterations of Tsuprem simulations and refining the dose of the threshold voltage adjust Vt implants, a threshold voltage Vt of 1.23 V for NMOS and -0.9 V for PMOS provided a starting point for the overall design. For 3 μm long devices without major parasitics, a 21 stage CMOS ring oscillator was initially simulated to ring at arond 120 MHz at 5 V Vdd with rail-to-rail output, which at that time, the design looked somewhat feasible. Such simulated results from Cadence is presented in Figure 4-5. We shall not report the other *a priori* simualted oscillator results here; rather, a more interesting exercise is to obtain the actual measured Vt's from real devices, re-enter them into Cadence and simulate the oscillator's response without parasitics, and then compare it to the measured oscillator results (provided if any of them worked at all in the end).

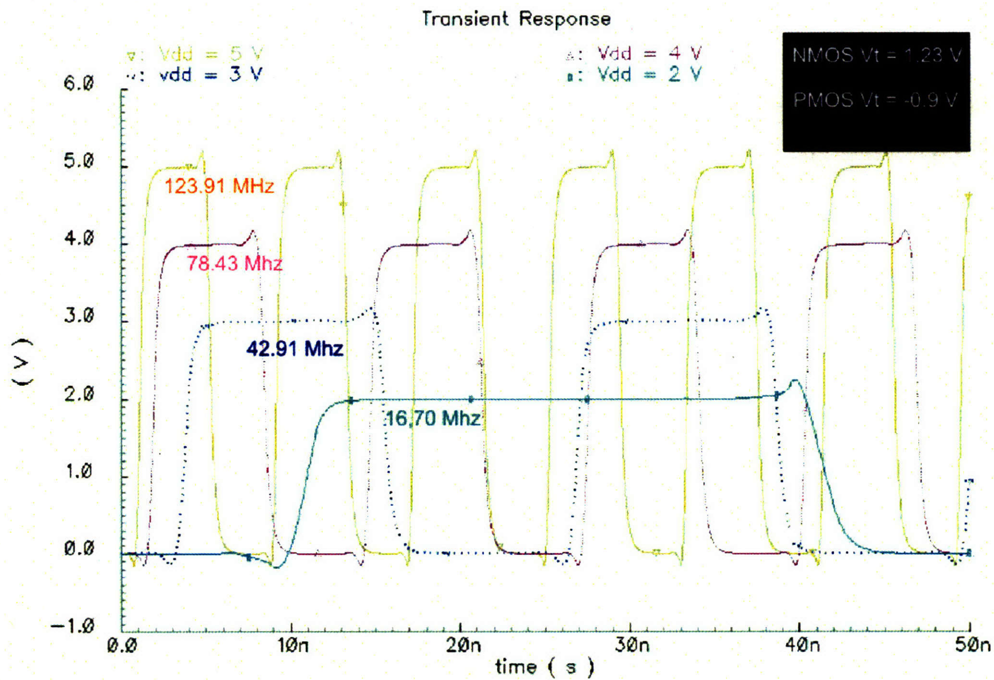


Figure 4-5: The basis simulation from which all other ring oscillators were designed around. Each colored curves represent the output response of a 21-stage CMOS ring oscillator (W/L_{nmos} = 60/3, W/L_{pmos} = 120/3) for a given value of Vdd. The NMOS and PMOS Vt's were assumed to be 1.23 V and -0.9 V, respectively.

4.2.2 Measurement of Fabricated single NMOS / PMOS devices

Upon finishing the design and layout, the final chip contained a total of 12 masks - 2 for PMOS/NMOS active areas on different wafers, 1 for gate poly, 2 for PMOS/NMOS contact cuts, 2 for Al Metal #1 for NMOS and PMOS, 1 for Cu damascene via, 1 for Cu bonding pad, and 3 for the lateral thermal diodes to be discussed in the next chapter. Starting with SOI substrates for both the NMOS and PMOS lots, the NMOS wafers were completed all the way up to its final Cu bonding pad layer and would wait for the PMOS lots to catch up before the final 3-D Cu-Cu bond can commence. On the other hand, the PMOS SOI substrates took a bit longer to go through the entire process; so, upon etching the Metal #1 Al lines, one wafer was taken out of the lot for some quick measurements. Table 4.2 is a collection of unbonded NMOS / PMOS data measured right after the Metal #1 etch. In the NMOS data, the threshold voltage was extracted by plotting the drain current I_d as a function of V_{gs} for at a small V_{ds} of 1.5 V. In other words, from the MOSFET equation in the linear region:

$$I_d = \frac{\mu C_o w}{2L} 2(V_{gs} - V_t)V_{ds} - V_{ds}^2 \tag{4.1}$$

if V_{ds} is somewhat smaller than (V_{gs}-V_t), then the above equation reduces to:

$$I_d = \frac{\mu C_o w}{L}(V_{gs} - V_t)V_{ds} \tag{4.2}$$

and when one then plots I_d as a function of V_{gs} for at a constant small V_{ds}, V_t can be extrapolated from the x-intercept of the graph's least-square linear fit. Moreover, the PMOS table entries also confirmed the expected V_t shift as one applies an increasing negative backbias on an isolated PMOS-SOI device lightly biased at V_{ds} = -1.5 V. Finaly, all pertinent IDVD, IDVG, and the VT least-square extraction graphs can be viewed from Figure D-1 thru Figure D-3 in Appendix D. Within these graphs, two features of the NMOS IDVD plots are worth mentioning. First of all, all 1 μm channel devices (plots (a)-(e) of Figure D-1 suffer from velocity saturation, and this can be deduced at a glance because the spacing between each I-V curves at saturation was linear, not qudratic, with respect to V_g. Simply put, our usual quadratic saturation equation

$$I_d = \frac{\mu C_o w}{2L} 2(V_{gs} - V_t)^2 \tag{4.3}$$

now becomes a linear equation under velocity saturation, with (V_{gs}-V_t)

$$I_d = \frac{\mu C_o w}{2L}(V_{Dsat})^2, \tag{4.4}$$

where

$$V_{Dsat} = \sqrt{2(V_{gs} - V_t)|E_c L|} \tag{4.5}$$

and E_c is the empirical critical electric field parameter at velocity saturation. Therefore, we expect there are some hot electrons within our channel, and as we will see when we probe the actual 3-D ring oscillators,

these hot electrons may be colliding with the lattice with such ferocity that optical phonon interactions could occur [2].

Next, from the IDVD graphs, one can observe that each of the 1 μm devices break down at a higher Vdd as V_{gs} increases. This was also probably due to impact ionization metioned above - a classic case of avalanche breakdown at work. Since we have already established that these NMOS reaches velocity saturation fairly quickly, even at very low gate voltages (like for the V_g = 1 V curve). The channel electrons become "hot" (ie. out of equlibrium with the substrate as they go from the source towards the drain as we sweep V_{ds} from 0 to 4V. This exacerbates the breakdown because the numerous holes generated from the ionization events at the drain, aided by the positive gate electric field, are being swept to and collected at the bulk substrate, or in SOI wafers, charging up the buried oxide-silicon (BOX-Si) interface. Once the BOX-Si interface gets enough positive charge on it, the p-type body and the n^+ source junction becomes forward biased, and then more electrons will be injected into both the body and the drain, causing more impact ionization events.

As V_g increases, the breakdown now happens at a larger V_{ds} because saturation occurs later, thereby delaying the avalanche event. Incidentally, the charging of the BOX-Si interface has a profound effect for the 3-D ring oscillators because as we shall see, a face-to-face bond body-charging event merely changes the Vt of the PMOS and adds a little bit of parasitic capacitance, while in a face-to-back 3-D bond, the charge buildup could become a huge parasitic capacitance between the device layers, although this remains to be proven true. As a last comment, notice that both of these characteristics were less pronounced in the longer 3 μm channel device (Figure D-1, plot (f)).

Unbonded NMOS			Unbonded PMOS			
Type	Vt (V)	ST-slope (mV/decade)	Type	Backbias (V)	Vt (V)	ST-slope (mV/decade)
80/1	0.662	118.35	60/3	Floating	0.103	-77.3
60/1	0.662	125.96	60/3	Grounded	0.105	-77.21
20/1	0.645	150.25	60/3	-5	0.431	Always on
10/1	0.643	208.20	60/3	-10	0.83	Always on
5/1	0.661	293.88	60/3	+10	-0.665	71.16
60/3	0.900	85.82				

Table 4.2: Summary of basic MOSFET parameters for unbonded NMOS and PMOS wafers. Vt = Threshold voltage, and ST-slope = Subthreshold slope

Next, the serires resistance and the effective channel length of the NMOS devices can be extracted if one

[2] To foreshadow the interesting results ahead, the face-to-face ring oscillator's titanium backbias pad (probed at floating bias) had a parasitic oscillation signal drawn from the main-stage MOSFETS that rang at a frequency 6 times faster than the output buffer signals. Furthermore, the peak-to-peak voltage of these parastic oscillations grew to very strong levels as Vdd increases.

plots the on-resistance as a function of the theoretical channel length of each device, for as many values of V_g as possible as long as $V_g > V_t$ is valid. This is because low drain voltages, the channel and the source-drain resistance combined can be given as:

$$R_m = \frac{L_m - 2\delta L}{\mu_n C_{ox} W (V_g - V_t)} + R_x \tag{4.6}$$

where

R_m = the channel resistance + the 2 source/drain resistances (measured linear MOSFET resistance)

L_m = the theoretical gate length of the devices from the maskset

δL = the decrease of channel length L from each side due to process-induced source-drain overlap

R_x = the dual source/drain resistance combination (value independent of channel length)

The data plotted in Figure 4-6 were obtained for the following two devices:

- **NMOS, W/L = 60/3** : V_g = 2 to 5 V
- **NMOS, W/L = 60/1** : V_g = 2 to 5 V

As seen in the graph, if the iso-V_g values were linearly interpolated, the experimental value of R_x can be extracted at the nexus of all 4 curves due to the fact that the source and drain series resistance do not vary with the channel length (notice that we kept the channel width at a constant 60 μm for R-preservation). Thus, a single source *or* drain series resistance was a mere 5/2 = 2.5 Ω, and the channel length shortening due to processed-induced source/drain overlap was about 0.125 = 2 $\cdot\delta$L, or δL = 0.0625 μm total.

Switching gears, the PMOS plots within Figures D-4 thru D-6 of Appendix D show what we expect from an SOI device that has undergone a positive and negative backside bias. The V_t's of the floating body and the grounded body devices were both around +0.1 V, which was quite weird at first glance because even if there was no Vt-adjust implant, the theoretical Vt of a PMOS should always normally be negative. The anomaly on-hand could mean that there was some work-function mismatch between the p^+-doped poly gate and the n-well implant, or if dopant segregation occured within the p^+-doped poly during the source/drain anneal, then Vt shifts could also occur. Nevertheless, Vt can be adjusted easily by applying a backbias on the substrate chuck contact, and this will be a key theme for the 3-D ring oscillator results: If the output ring response is undesirable, just crank up the body voltage - and problem solved !

Notice in plots (c) and (d) of the IDVG figures, the Vt's have been shifted so much in the positve direction that the devices didn't turn off within our V_g sweep. Also, the source / drain series resistance were not extracted from the PMOS devices; instead, we also assumed it to be 2.5 Ω per implanted region. Finally, a critical note for the PMOS lot: **All PMOS-SOI device with channel lengths less than 2 μm had shorted source-drains** [3].

[3] This was due to my mistake in trusting the TSuprem simulations for the source /drain anneal thermal cycles.

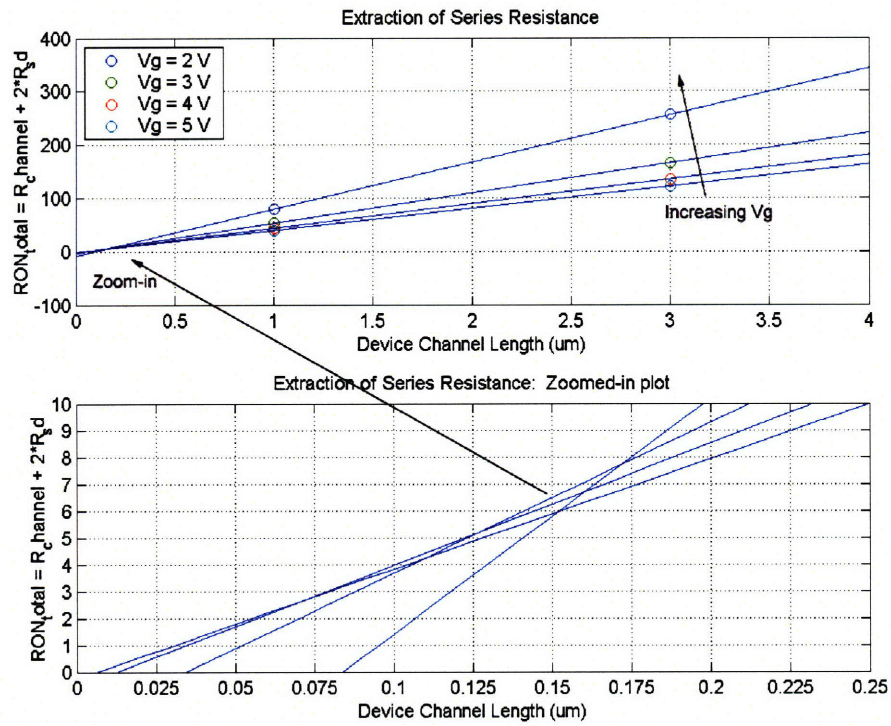


Figure 4-6: Extraction of *unbonded NMOS* series resistance. The on-resistance of each 60 μm-wide NMOS, at Vds = 0.5 V, was plotted as a function of both the gate length L and the gate bias Vg. The total source-drain series resistance $2R_x$ was extrapolated at the approximated to be about 5 Ω, or at intersection of all curves except for the outlier line Vg = 2.

4.2.3 Simulated Oscillator Results: With Measured Vt's and First-order Parastics

Now, once we've found the true Vt's, the corresponding series resistance through measurements, we can re-evaluate our simulated ring oscillator results to get a more realistic expectation of what our real 3-D oscillators would look like [4]. Inserting the corresponding values from Table 4.2 on page 71 into the Cadence circuit models, we can get simulated results such as the the 21-stage, 3 μm-length CMOS ring oscillator shown in Figure 4-7. Again, instead of plotting all these results in separate figures, the ring frequency vs. Vdd for each kind of oscillators were collectively tabulated in Table 4.3.

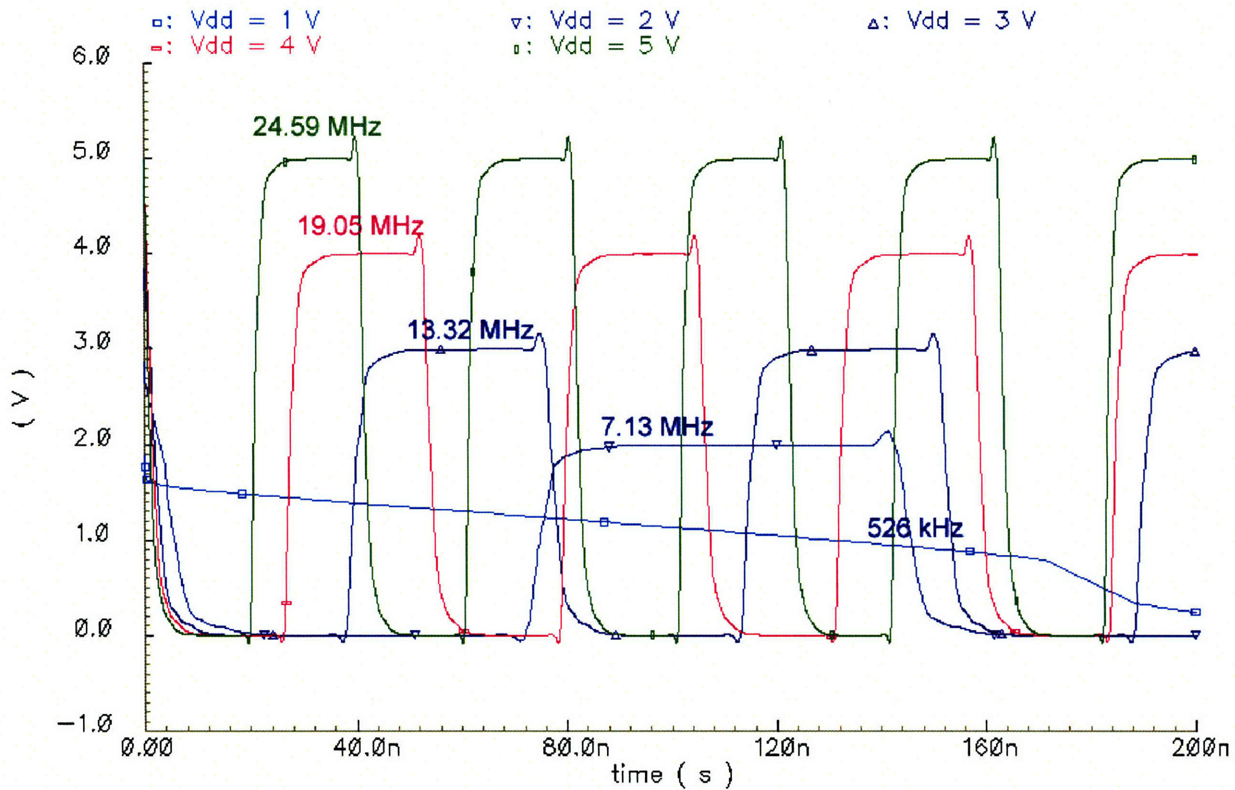


Figure 4-7: Simulated 21-stage, 3 μm-channel CMOS ring, with real NMOS Vt = 0.900 V and a properly-biased PMOS Vt = -0.665 V (taken from value at +10 V backbias for safe measures).

[4]The mobility values and the effective substrate doping (the SOI film was thin enough where a mere Vt implant can change the body doping) for NMOS / PMOS models in Cadence were also changed such that they matched the measured I-V characteristics of individual devices. The combination of these 2 corrections proved to be of upmost imporantance, for if they remained uncorrected, the simulated output frequency of the CMOS oscillator was about 20x faster than the measured output. With the corrections, the simulated output frequency was only 3.7x faster than the measured data.

21-stage CMOS			43-stage CMOS		
	L = 1 μm	L = 3 μm		L = 1 μm	L = 3 μm
Vdd	Freq (MHz)	Freq(MHz)	Vdd	Freq (MHz)	Freq(MHz)
5	82.25	24.59	5	42.19	11.89
4	68.68	19.05	4	33.94	9.25
3	49.50	13.32	3	24.45	6.49
2	27.33	7.13	2	13.49	3.17
1	21.67	0.526	1	1.067	0.260

Table 4.3: Summary of CMOS ring simulations, with real NMOS / PMOS Vt's included in the models. For the all 3 μm long device, Vt_{nmos} = 0.9 V and Vt_{pmos} = -0.665 V (assuming that our Vt can be properly biased to a well-behaved negative number). For the 1 μm long devices, Vt_{nmos} = 0.662V and the PMOS Vt remained the same.

4.3 3-D Ring Oscillator Measurements

The long-awaited results for this entire project are presented in the next couple of sections. Out of the 4 possible CMOS ring oscillator configurations, two of the cells managed to ring: The face-to-face bonded 21-stage and the 43-stage circuits with 3 μm channels. It was no surprise that the 1 μm rings did not work because all 1 μm PMOS transistors became resistors after the overly-aggressive source / drain + poly resistivity reduction anneal. In particular, it was surprising that the 43-stage variety worked at all, since that particular circuit had a combination of 112 FETs scattered bewteen two separate wafers, 102x2 = 204 individual Cu damascene vias, [5] and 92 Cu-Cu bonds on top of those Cu vias to complete the chain. *Every element* mentioned above have to work simultaneously or the chain would have be broken and the outputs would have been flatlined.

Unfortunately, none of the face-to-back CMOS rings worked. This was probably due to a combination of poor bonding, aligning, and most of all, there was probably too much parasitic capacitance between the PMOS Cu back gate and the Cu/Al interconnect layers of the NMOS below. We will go through each of the results and try to make sense and possibly de-embed some of the parasitics involved.

4.3.1 Face-to-Face Ring Oscillators: 21-Stage CMOS, L = 3 μm

Floating Backbias

The grand results of this entire thesis can really be summarized in this and the following section. In brief, this set of devices were made by a hybrid process without the use of a handle wafer. First, the NMOS and

[5] And that's not counting the redundant ones on the big probe pads; these 204 vias were all part of the critical path between the PMOS/NMOS s/d outputs and their respective gate inputs.

PMOS devices were face-to-face bonded with Cu. Then, the Si bulk from the PMOS substrate was removed by a combination of SF_6 plasma etching for 6 hrs and a TMAH etchback that stopped on the PMOS BOX layer. Next, a thin layer of 1000 A Ti was sputtered on top of the PMOS BOX to act as a makeshift backgate. This was done because our "designed Cu backgate" auxiliary bond pads were now sandwiched in-between the Al interconnects and the ILD layers of both PMOS/NMOS layers, or in other words, the backgate was located too far away from the PMOS gates, thus rendering it useless. Last, but not least, an approximately 1.1 μm deep thru-SOI contact hole was etched from the top using both 50:1 HF to etch through the Ti and BOE to etch through the BOX, field oxide, and the LTO passivation oxide that belonged to the PMOS SOI. The exposed Al pads (Vdd, ground, output of oscillator, and useless Cu backbias Al pad) were then directly probed through the vias. Naturally, the Ti metal left over on the top surface will be our makeshift Ti backgate for the PMOS devices. A brief pictorial tour of the above proces is shown in Figure 4-8, and a schematic representation of the face-face circuit is shown in Figure 4-9.

Without further ado, Table 4.4 is a detailed list of the working ring oscillator's frequency and output voltage maxima / minima as a function of Vdd. Also, the measured output signal from the oscillator, the parasitic signal coming the floating Ti backbias pad, and the parasitc signal coming from deep within the "useless" Cu backbias plane (denoted by "Al Bias pad" in the legend) were all plotted as a function of time. For discussion purposes, we will only plot the oscillator response for seleected voltages:

- Vdd = 1 - 2.5 V in Figures 4-10,
- Vdd = 7.0 V and the frequecy vs. Vdd response in Figure 4-11,
- The oscilloscope signal coming from the both the floating Ti backgate and the useless Cu pad within Figure 4-12 in the body of this chapter.

The entire spectrum of this oscillator's response can be found in Appendix D, section D.3.

21-Stage CMOS, L = 3 μm: Backbias = Floating				
Vdd	V-	V+	Vpp	Freq (MHz)
1.0	0.109	0.953	0.844	2.857
1.5	0.094	1.813	1.719	4.310
2.0	0.094	2.313	2.219	4.673
2.5	-0.25	3.063	3.313	4.762
3.0	-0.625	3.625	4.25	5.102
3.5	-0.625	3.438	4.063	5.376
4.0	-0.938	4.0	4.938	5.434
4.5	-0.813	4.50	5.313	5.494
5.0	-1.063	4.938	6.001	5.494
5.5	-1.063	5.188	6.251	5.682
6.0	-1.125	5.813	6.938	5.747
6.5	-1.313	6.563	7.876	5.682
7.0	-1.313	7.50	8.813	5.682

Table 4.4: Output from 21-Stage CMOS ring, L = 3 μm, with floating backbias on the PMOS body

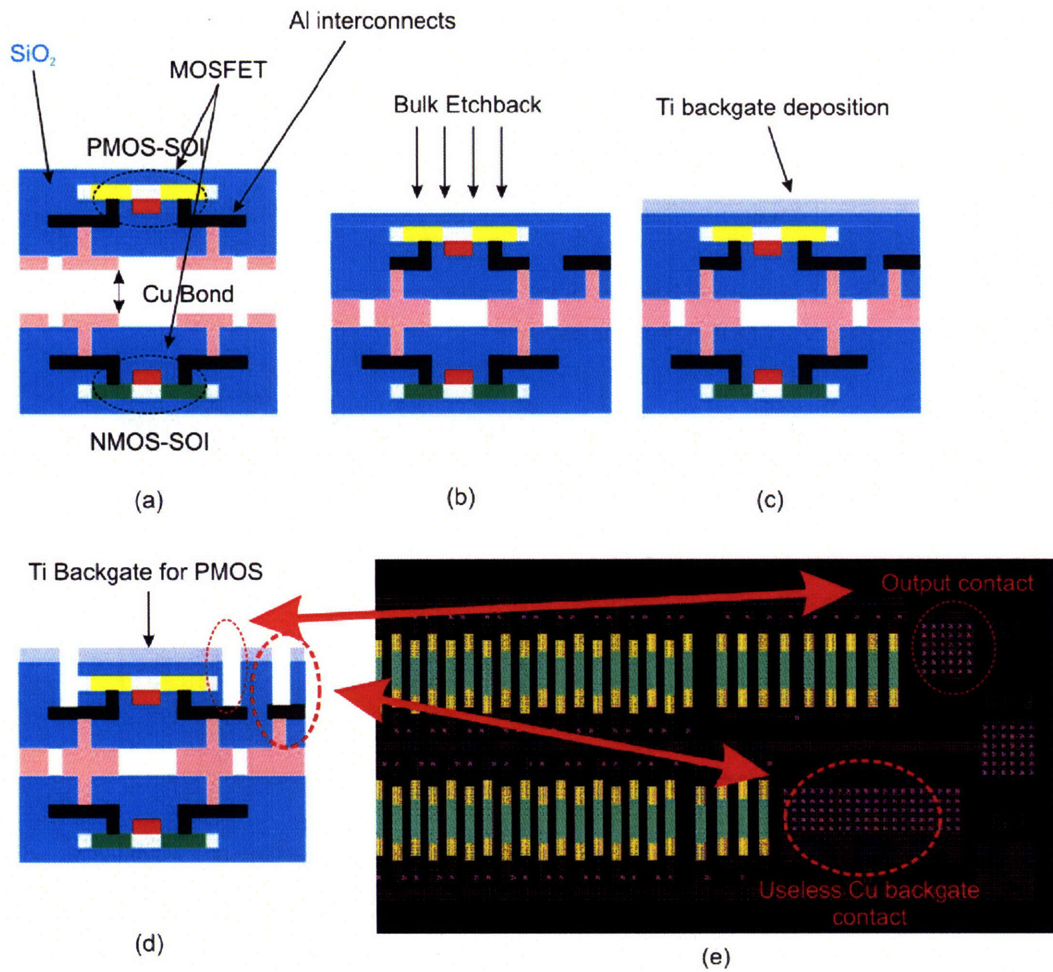


Figure 4-8: The makeshift face-face 3-D process: In (a), the PMOS and the NMOS substrates were bonded with the standard Cu recipe. Then, a SF_6 plasma and a TMAH etch was able to clear the backside of the PMOS substrate in (b). After the Ti backgate depositon (c) and topside via etch (d), the 3-D ring oscillators were ready for probing at the red circles in (e).

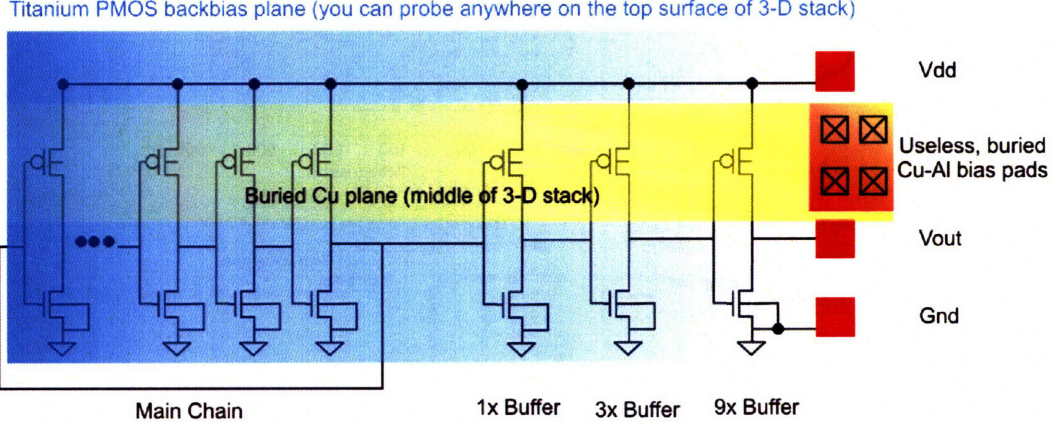


Figure 4-9: The makeshift Ti backgate position relative to all the other probe pads. Notice that the buried Cu-Cu bonding backplane was useless here because the top Ti backgate, sitting right on top of the PMOS BOX, has a more direct access to the PMOS gate than the Cu plane.

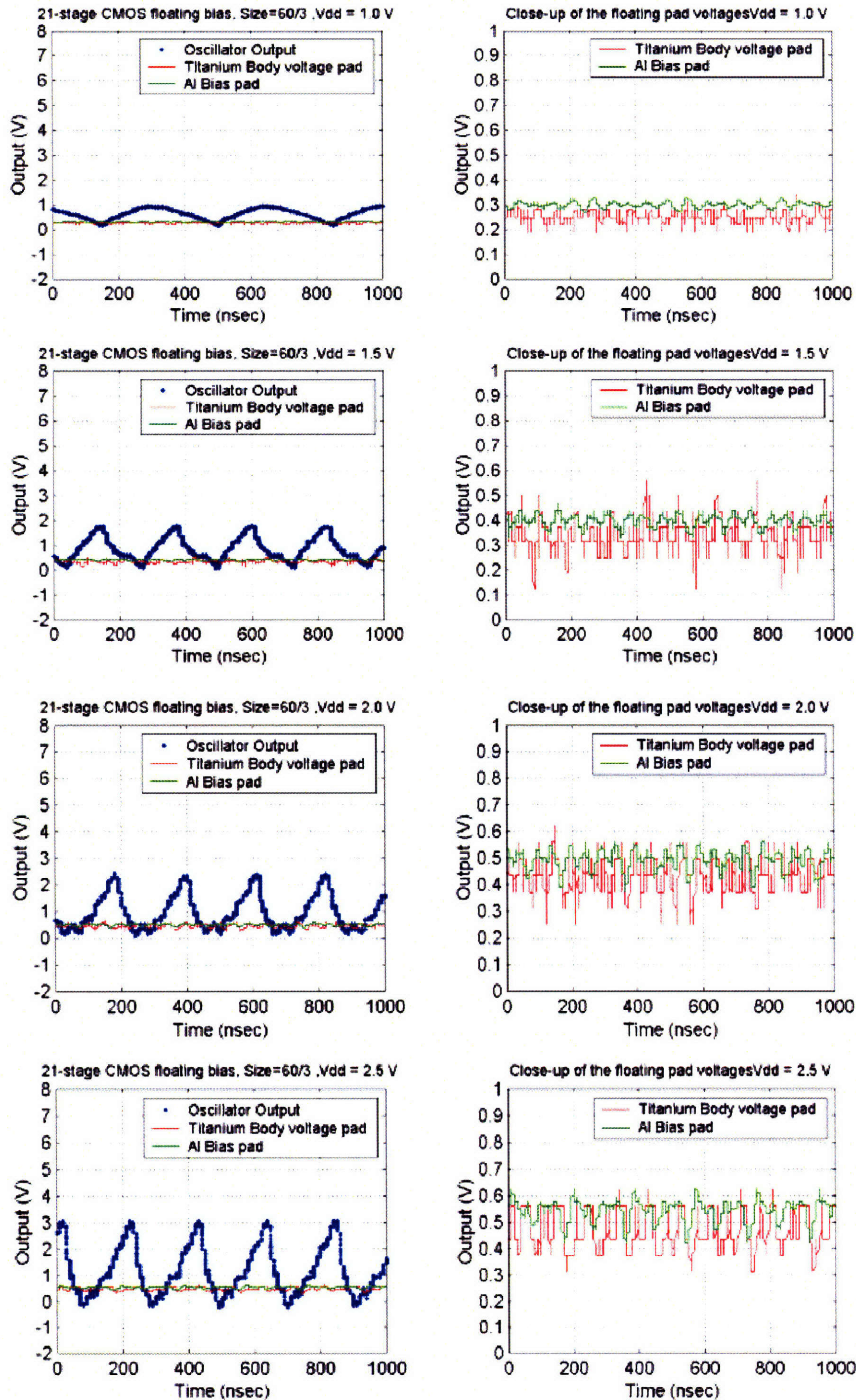
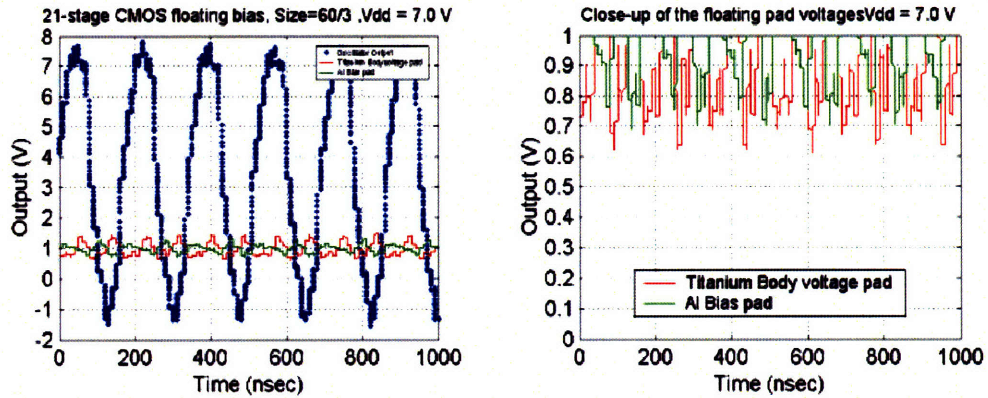
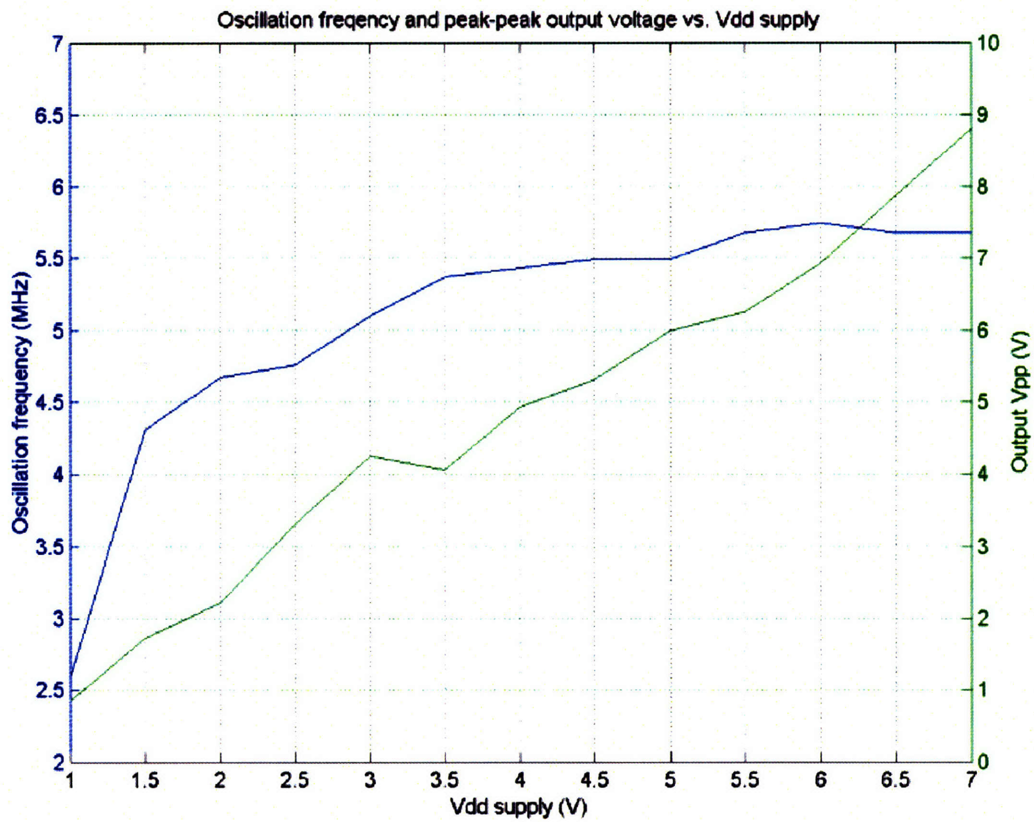


Figure 4-10: 21-Stage CMOS, L = 3 μm: Vdd = 1 thru 2.5 V. Left-column plots are the signals from the 9x output buffer and the tiny traces from probing the floating Ti backgate and the "useless" Cu-Al plane's pad; right-column plots are zoomed-in traces of the aformentioned floating pads.



(a)



(b)

Figure 4-11: 21-Stage CMOS, L = 3 μm: The plot at Vdd = 7.0 is in (a). In (b), the frequency and the peak-to-peak voltage Vpp of the output was plotted as a fuction of Vdd. Note the saturation of the oscillation frequency at high Vdd's.

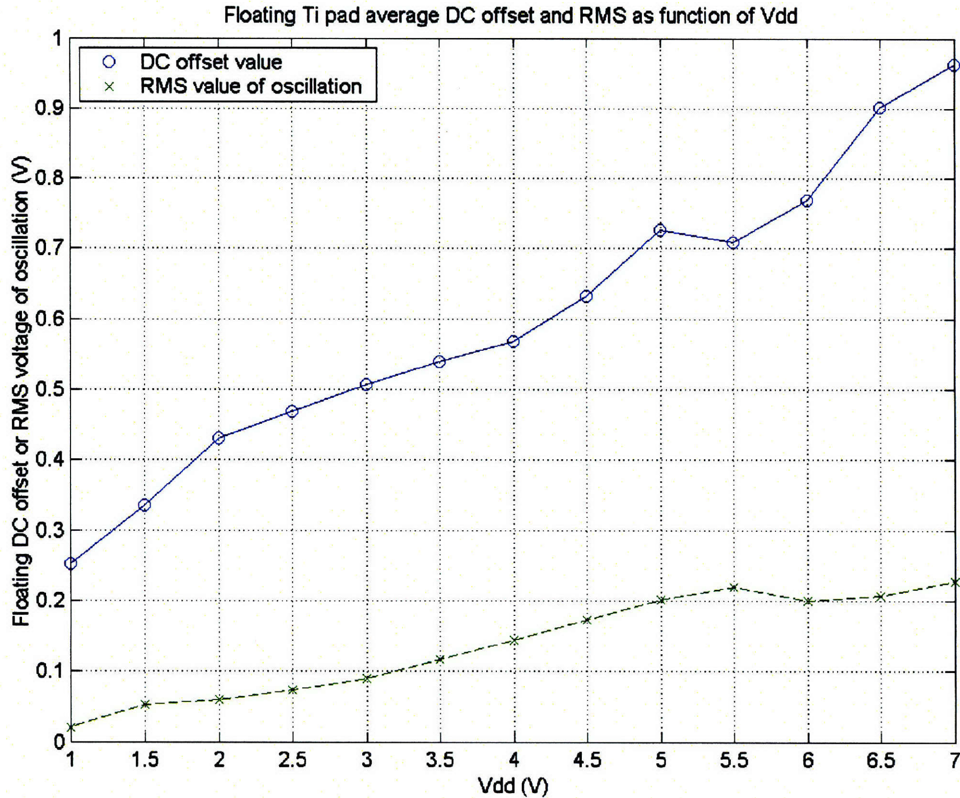


Figure 4-12: 21-Stage CMOS, L = 3 μm: The DC offset and the RMS voltage of the fast oscillations coming off of the floating Ti bckgate and the useless Cu floating pads

As the reader would agree, the preceding figures contains a wealth of information that's tough to swallow in one shot. Let's discuss some of the interesting features in detail.

1. **The Ring Oscillator Speed was SLOWER Compared to the Simulated Results :** One quick comparison between Table 4.4 and Table 4.3, and you will find that our face-face, 21-stage CMOS 3 μm ringer was about *4.5x times slower* [6] than what Cadence told us after we inserted the real Vt's and source-drain series resistance into our model. Evidently, there is a lot of parastics in our Cu-Cu bonded samples that we didn't even account for in our simuation models. As we shall see in a few sections, the two major parasitic contributers in our 3-D circuit are within the Cu damascene vias and the little extraneous capacitances lay littering upon the entire 3-D landscape.

2. **V+/V- of Main Output Exceeds Vdd :** As seen from both Table 4.4 and the left column in every Figure from 4-10 to 4-11, the value of V+ often exceeds over Vdd by a little bit and the value of V-

[6] At Vdd = 5 V, the ratio between the simulated and measured output frequencies is (24.59 MHz / 5.494 MHz) = 4.48

always dipped below the ground rail. Since the PMOS-SOI body voltage was floating, we expect some sort of charging event at the SOI-BOX interface that's raising the effective V+ above the Vdd top rail. "But what about the values of V- dipping below 0 V?' one might ask. A quick answer to this question is that even though the NMOS-SOI body contacts were firmly clamped to zero by the ground chuck, the overall V- voltage can still shift below 0 V by virtue of SOI-BOX charging occuring at the SOI-BOX interface. If one was to explain the above statements in more careful fashion, though, we'll have to take into account our next interesting feature of the oscillator, which was that:

3. **The Floating Titanium Backbias Pad both Rings and Contains a DC Offset** : If one looks at all right-column plots from Figure 4-10 he or she would see that both the floating Ti PMOS bias pads and the "useless" Cu-Al bias pads exhibited some parastic ringer signals that seeped out from the main ring oscillator chains. In addition, these oscillations were about 6x faster than the proper outputs coming out of the 9x-buffer port, and both signals have a DC offset component and an oscillatory V_{rms} component that increases with Vdd [7]. Furthermore, although the the DC offset - Vdd relatioship was linear, the RMS voltage of the fast oscillatory signal saturates at high Vdd (see Figure 4-12).

Now, armed with knowledge #2 and #3, we are ready for a final discourse on what's happening to these face-face bonded 3-D oscillators. To begin with, during the time when the PMOS gates are ON within the oscillator, holes are being injected from the source end (tied at +Vdd) and are collected at the PMOS drains. Assuming that our PMOS reaches velocity saturation very quickly, as in the case when we measured the unbonded PMOS device, the holes drifting towards the drain might acquire enough energy to become hot carriers, thus creating numerous impact ionization events near the drain-channel region. Now, if the impact ionization energy was great enough, generation of new electron-hole pairs will occur. While the excess holes will be re-collected at the drain junction by lateral E-fields, electrons, coaxed by the negative V_{gs} bias during the PMOS ON-stage, would be swept and collected at the either

- Within the SOI body itself
- At the interface traps between the SOI-BOX interface

Remembering that the PMOS BOX neighbors two conductive substrates (Si channel one one side, Ti backgate on the other side), an electron build-up at the Si-BOX interface would cause a buildup of positive charges on the Ti-BOX interface due to charge neutrality. Since the entire Ti backgate is equipotential, the buildup of positive charges on the entire Ti surface is the reason why the measured DC offset was *positive.* Moreover, since the body discharge rate when the PMOS turns off (when the body-to-drain or body-source paths are severely reverse-biased) is much slower than the charging rate, an accumulation of electrons will

[7] In theory, the Ti and Al pad signals should not be shorted together. I think the reason why they're shorting out here was there were some stringers left on the surface after the topside Ti/contact via etch. This can be proved in our 43-CMOS ringer data where indeed a proper etch will electrically isolate these two pads completely

buildup within the body when averaged over time. Since the body and the Ti-gate can store large amounts of charge, as Vdd increases to larger values, the positive voltage on the Ti-gate will continue to increase without bounds. This matches the observation in Figure 4-11, where the DC offset of the Ti-gates keesp on increasing linearly with Vdd. Since this is an process averaged over time, it is considered to be a *slow* process and can be summarized in Figure 4-13.

However, the above does not explain why the RMS voltage of the oscillation saturates at high Vdd. To explain this effect, one also has to consider simultaneous electron accumulation events occuring at the SOI-BOX interface. Given there exists a series of interfacial charge states at the oxide interface, electrons injected into this region from impact ionization will quickly fill these states up when the PMOS is on, and the level of filling depends on the operating voltage, or Vdd. Since these interfaical states (Q_{it}) are few in number compared to the charge capacity within the SOI bulk, they can respond to a much higher frequency stimulus than the aformentioned body charging events. Furthermore, if the value of Vdd exceeds the bandgap, then all Q_{it} sites will be filled, and this was probably why the V_{rms} saturates at high Vdd's. The entire Q_{it} fast-charging scenario was illustrated in Figure 4-13.

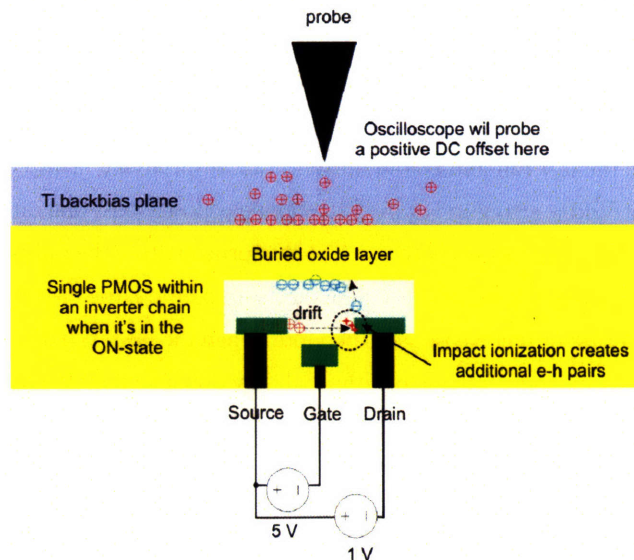


Figure 4-13: Positive DC offset on the face-to-face CMOS floating backgates was caused by the slow body-charging process

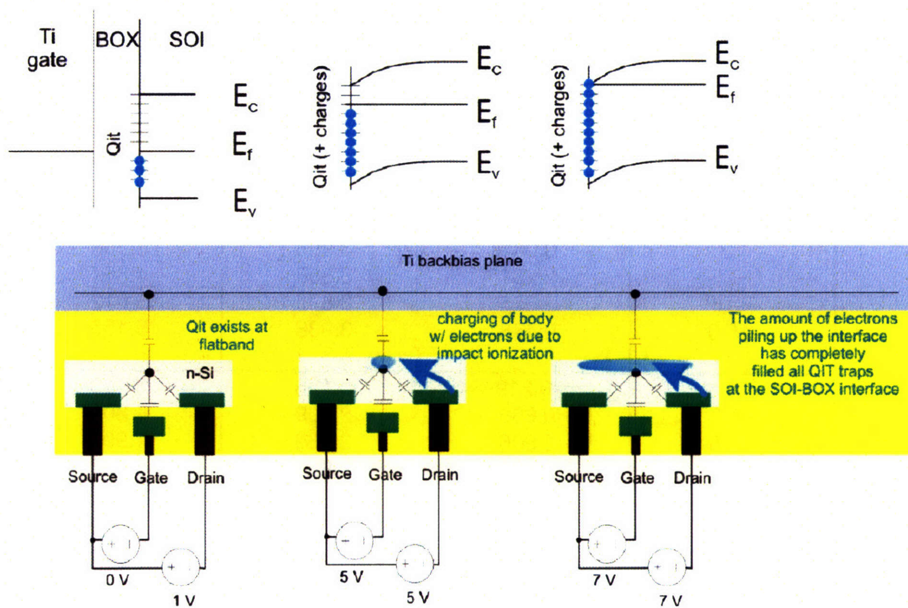


Figure 4-14: Saturation of the V_{rms} occurs because of fast filling / discharging of interficial charge states at the BOX-SOI interface

Pinned Backbias

Having gone through the floating gates results in length, let's look at some results where we forced a backbias onto the Ti backgate. As one would predict, Table 4.5 shows that a postive bias on the PMOS gate can both shift the Vt and limit electron-charging at the Si-BOX interface, thus forcing V-/V+ back to where they belongs - above the 0 V rail and below the Vdd rail, respectively. This effect can be readily seen in measured outputs at Vdd = +5.0 V in Figure 4-15. Taking a closer look, one can see that when the backbias voltage = +5.0 V = Vdd, The output of the oscillator returned within the rail-rail ranges of 0 - 5 V. In addition, as one increases the backbias past +5 V (like +6.5 V in plot (c)), the PMOS tends to remain in the on-state for a much wider input voltage range. Therefore, when the NMOS gates are supposed to be on, the PMOS gates remain on for some fraction of a time, and by doing so there's a period in which both the PMOS and NMOS forms an desirable voltage divider. Thus, this is the reason why the output signal in (c) does not touch the ground rail. A converse argument can be made when when the PMOS Ti-gate is negatively biased, as in the case plotted in (e), where it's the NMOS that fails to turn off properly.

Finally, while only the Vdd = +5.0 response was plotted in this chpater, the aggregate results of backbiasing, from Vdd = +5.0 V down to Vdd = +3.0 V, can be found in Appendix D, starting from page 182.

21-Stage CMOS, L = 3 μm: Backbias = Varied				
Vdd	PMOS backbias	V-	V+	Freq (MHz)
5.0	0	-0.625	4.938	5.682
	+5.0	0.125	4.675	5.376
	+6.5	1.75	5.062	4.95
	+7.5	3.25	4.656	4.629
	-15.0	0.719	4.406	5.319
4.0	0	-0.156	3.563	5.435
	+5.0	0.063	3.906	5.208
	+6.5	2.281	3.656	4.857
	+7.5	2.375	3.875	3.012
	-15.0	0.438	3.938	5.155
3.0	0	-0.219	3.156	5.263
	+5.0	0.656	2.969	4.808
	+6.5	1.906	2.969	1.906
	-14.0	1.563	2.969	5.555

Table 4.5: Output from 21-Stage CMOS ring, L = 3 μm, with varied backbias points on the PMOS body

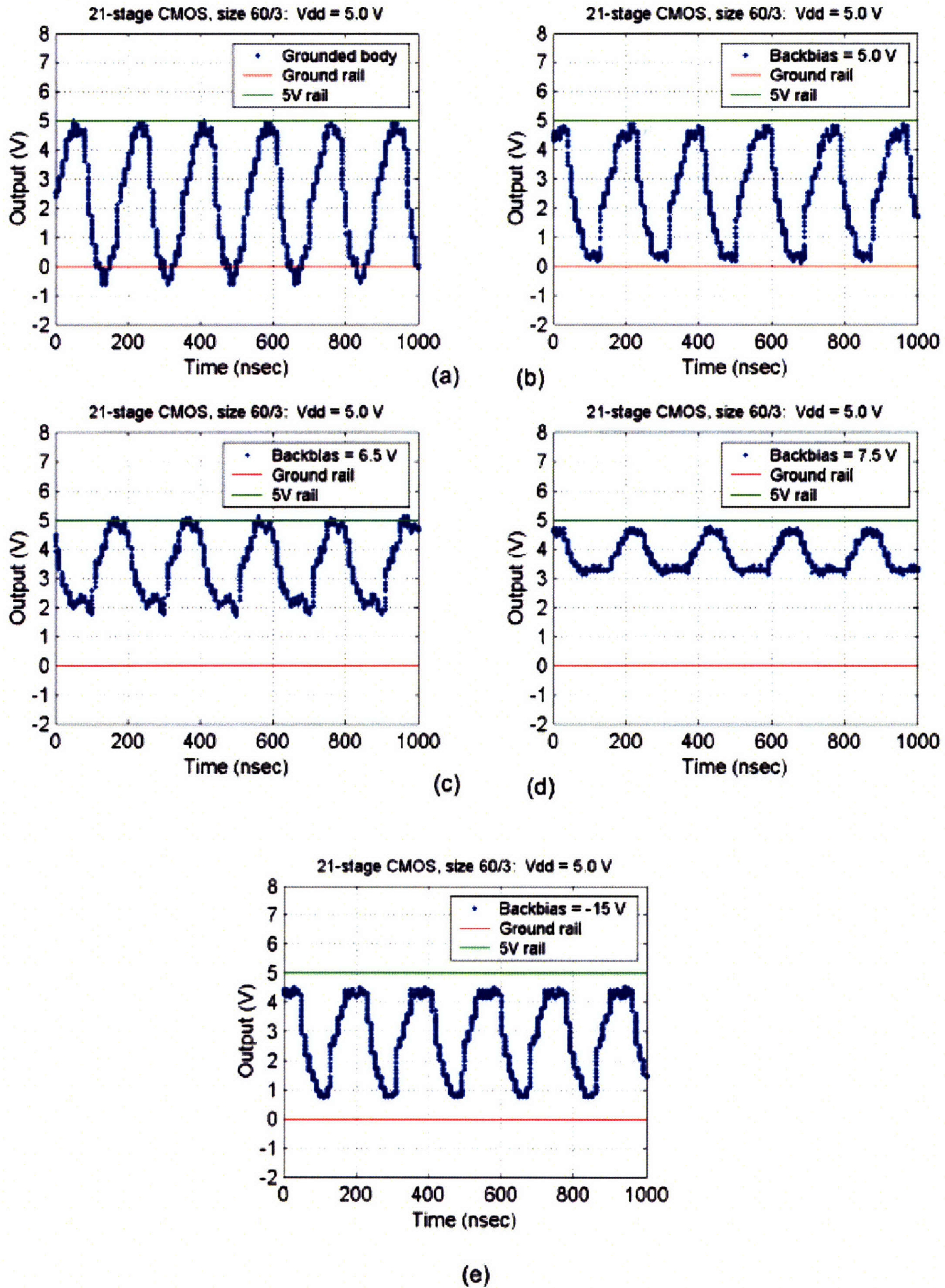


Figure 4-15: 21-Stage CMOS, L = 3 μm, Vdd = +5V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) +7.5 V backbias, (e) -15 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails, and at Vdd = +5 V, V-/V+ almost resided on the rails themselves, albeit with some voltage drop from internal series resistance.

4.3.2 Face-to-Face Ring Oscillators: 43-Stage CMOS, L = 3 μm

Floating Backbias

Without question, the crown jewels of this thesis are the 43-stage CMOS oscillators. It was a miracle that two of these devices actually survived the entire face-face 3-D flow, let alone having all the FETs and the Cu-Cu contacts working simultaneously. Again, we have Table 4.6 displaying the important results in tabular form, and Figures D-15 thru D-18 within Appendix D display the actual output trace from the oscilloscope. What the reader will find is that the 43-stage oscillator can presents the following features even more elegantly than its 21-stage sibling:

- Since the 43-stage oscillator contains around twice the amount of inverters as the 21-stage oscillator, one would expect the oscillation frequency of the longer chain to be half that of the smaller chain. Comparing results between Tables 4.6 and 4.4 (on page 77), this was roughly confirmed.
- Without proper back-biasing, the V-/V+ extrema traveled much farther away from the ground and Vdd rails (V+ was at 12 V for Vdd = 7 V!) as opposed to the 21-stage ringers. This was because the 43-stage chain was larger in area, thus the Ti backgate and the Si-BOX interface could hold more amounts of charging.
- Maximization of the Ti backgate's DC offset from electron charging of the Si-BOX interface was more prominent here because of the larger charging area. Also, the subsequent rise in parasitic oscillation's amplitude due to an increase in avalanche breakdown activity was easier to see in these devices (again, perhaps optical phonon scattering plays a significant role here too).

Again, the backbias situation here differs from the 21-stage oscillators because the Ti backgate covers a much wider area. Therefore, since there are more PMOS devices beneath the Ti surface, this causes a huge amount of electron buildup at the Si-BOX interface and a proportional positive charge buildup on the Ti gate. Table 4.7 shows that it takes an inordinate amount of backbias voltage, +13 V to be exact, to force the V+/V- extrema back to the Vdd and ground rails. One can observe this theme over and over again within Figures D-20 to D-22 plotted within Appendix D.

43-Stage CMOS, L = 3 μm: Backbias = Floating				
Vdd	V-	V+	Vpp	Freq (MHz)
1.0	-0.094	0.766	0.86	1.522
1.5	-0.313	1.563	1.876	2.232
2.0	-0.375	2.00	2.375	2.404
2.5	-0.75	2.594	3.344	2.604
3.0	-1.00	3.094	4.049	2.631
3.5	-0.875	3.813	4.688	2.732
4.0	-1.688	4.188	5.876	2.778
4.5	-1.875	4.4385	6.313	2.778
5.0	-1.688	6.938	8.626	2.778
5.5	-1.875	7.75	9.625	2.841
6.0	-2.063	8.563	10.626	2.809
6.5	-2.375	9.188	11.563	2.874
7.0	-2.563	9.875	12.438	2.874

Table 4.6: Output from 43-Stage CMOS ring, L = 3 μm, with floating backbias on the PMOS body

43-Stage CMOS, L = 3 μm: Backbias = Varied				
Vdd	PMOS backbias	V-	V+	Freq (MHz)
5.0	0	-1.625	7.25	2.8367
	+5	-1.438	6.688	2.778
	+10	-0.938	5.813	2.577
	+13	-0.188	5.375	2.293
	+14	0.125	5.5	2.101
	-10	-0.875	6.375	2.809
4.0	0	-1.188	5.250	2.778
	+5	-0.938	4.563	2.688
	+10	-0.375	4.00	2.359
	+12	0.063	4.00	2.137
	-10	-0.125	5.125	2.632
3.0	0	-0.875	4.0	2.66
	+5	-0.75	3.688	2.577
	+10	-0.188	2.656	1.953
	-10	0.0	3.938	2.358

Table 4.7: Output from 43-Stage CMOS ring, L = 3 μm, with varied backbias points on the PMOS body

4.3.3 Face-to-Back Ring Oscillators: CMOS Failures, NMOS success

Encouraged by the success of the face-face CMOS ring oscillators, tests were also done on the face-to-back bonded CMOS oscillators to see if any of them worked. Summing up all the results in one sentence: All face-face CMOS oscillators had flatline outputs, but its DC offset responds to variations in Vdd. In all three examples listed below, the devices wre constructed from a face-to-back, die-level, 20 μm-Al release layer process. In each figure, the Vout vs. Vdd data points were plotted along with a dashed-green test line to check if at any point Vout was shorted with Vdd. Under a quick examination, Figure 4-16 shows a 21-stage L=3μm CMOS oscillator was a "flatliner" probably because there was too much parasitic capacitance or series resistance linking the PMOS gates, the Cu-Cu bond that doubled up as the PMOS backgate bias plane, and the NMOS interconnect layers that lie immediately underneath the Cu bias plane. Notice also that the metastable output voltage was about half the applied Vdd value (the interpolated data line approximately bisected the angle made between the x=y test line and the x-axis). Please also note that the relationship between Vout and Vdd was approximately a linear one at intermediate Vdd values.

To continue, a quick examination of Figure 4-17 shows that the interpolated data line from this "flatliner" hugged very close to the x=y line, suggesting that this 43-stage CMOS oscillator either had a shorting defect from Cu-Cu mis-alignment, underetched stringers, or from other unknown causes. Again, like its other face-to-back bonded siblings, this oscillator did not ring because it was also in a metastble state of operation. Lastly, even if the oscillator from Figure 4-18 had a floating backbaias, the end results were quite similar to the previous two devices. The linear relationship between Vdd and Vout can be seen across the entire spectrum of Vdd inputs.

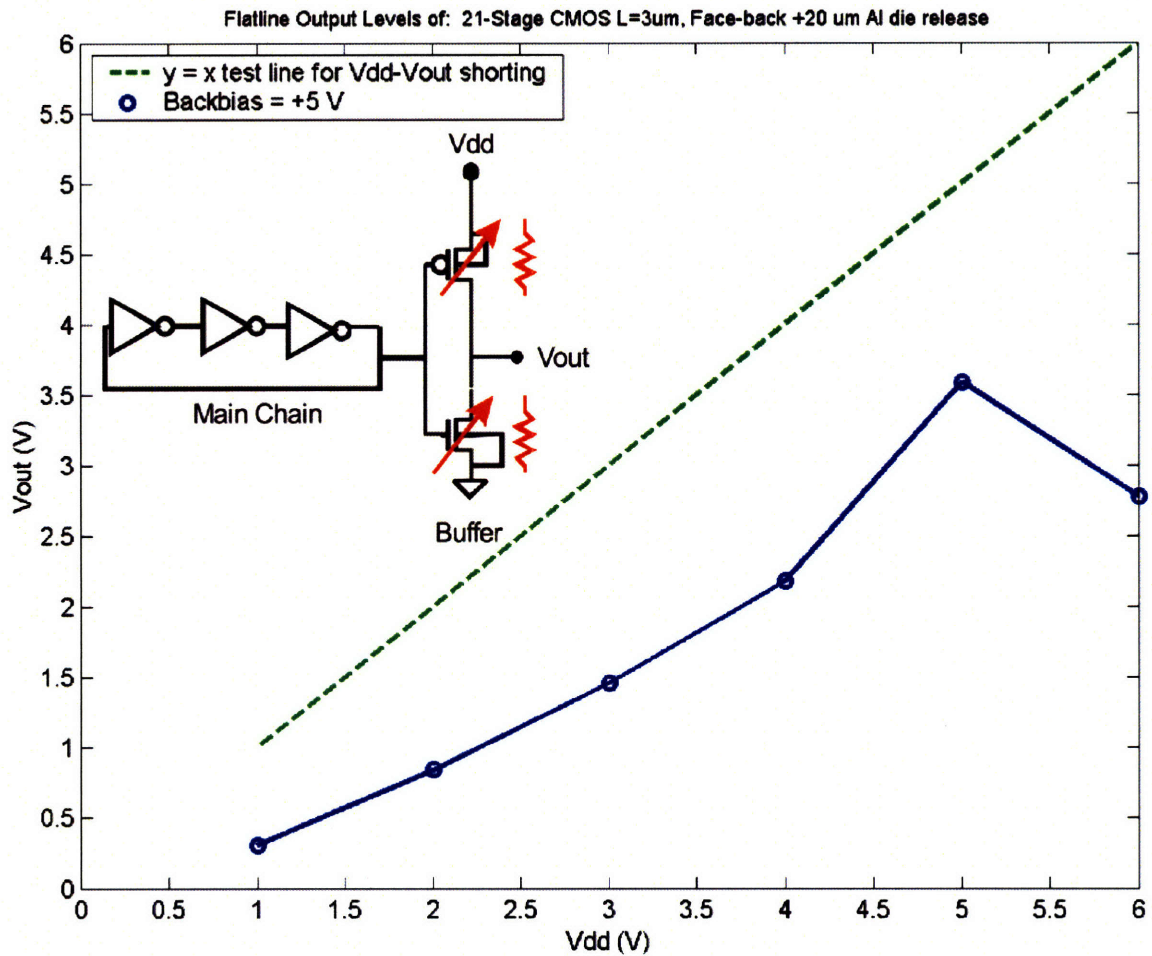


Figure 4-16: 20 μm Al released, Face-back bonded 21-stage CMOS at L = 3 μm with +5 V backbias on the PMOS backgates. Since all output responses were flatlines, the output DC voltage Vout was plotted against the power supply voltage Vdd. This particular circuit had no Vdd-Vout shorts and was in a metastable state that inhibited oscillation.

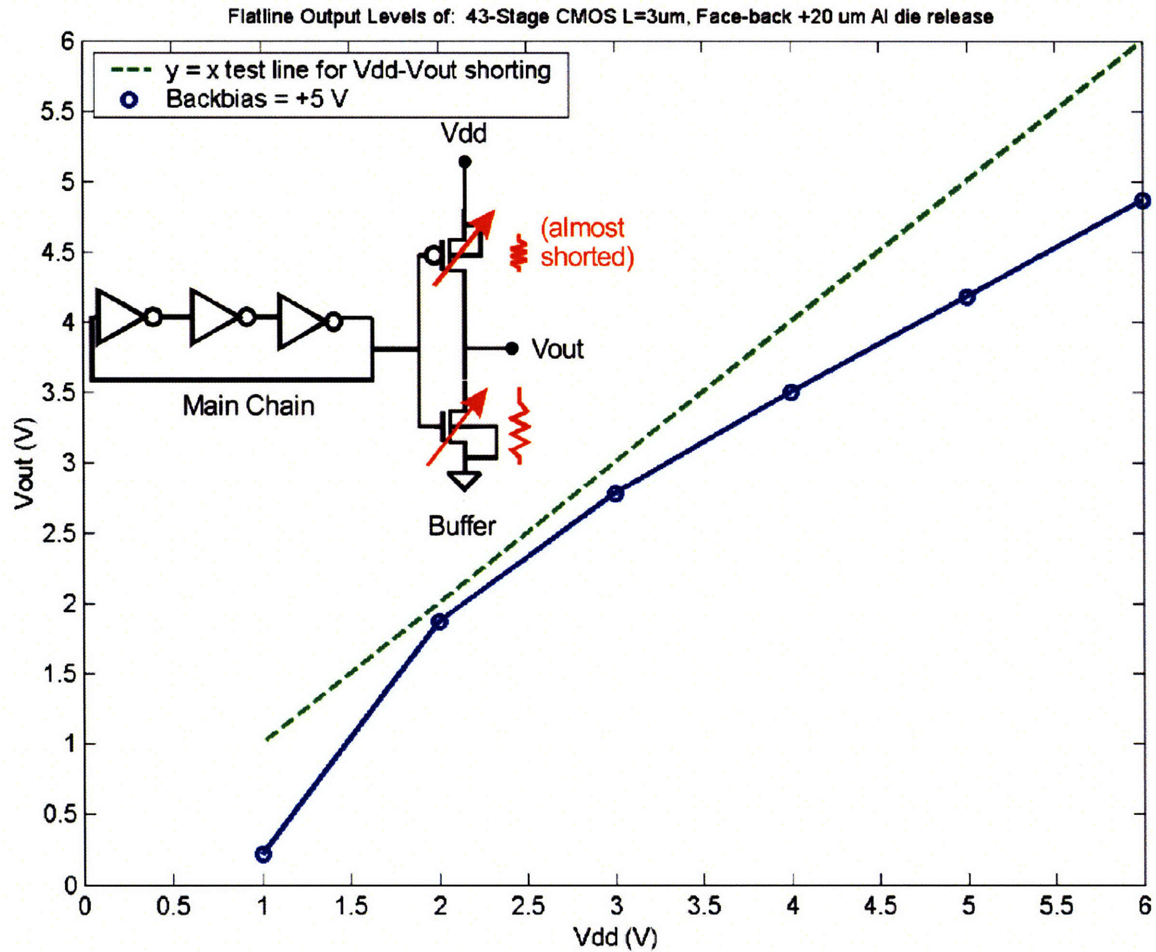


Figure 4-17: 20 μm Al released, Face-back bonded 43-stage CMOS at L = 3 μm with 5+ V backbias on the PMOS backgates. Since all output responses were flatlines, the output DC voltage Vout was plotted against the power supply voltage Vdd. This particular circuit exhibits an almost-shorted Vdd-Vout path near the PMOS transistors, and like the other oscillator, this sample was also in a metastable state that inhibited oscillation.

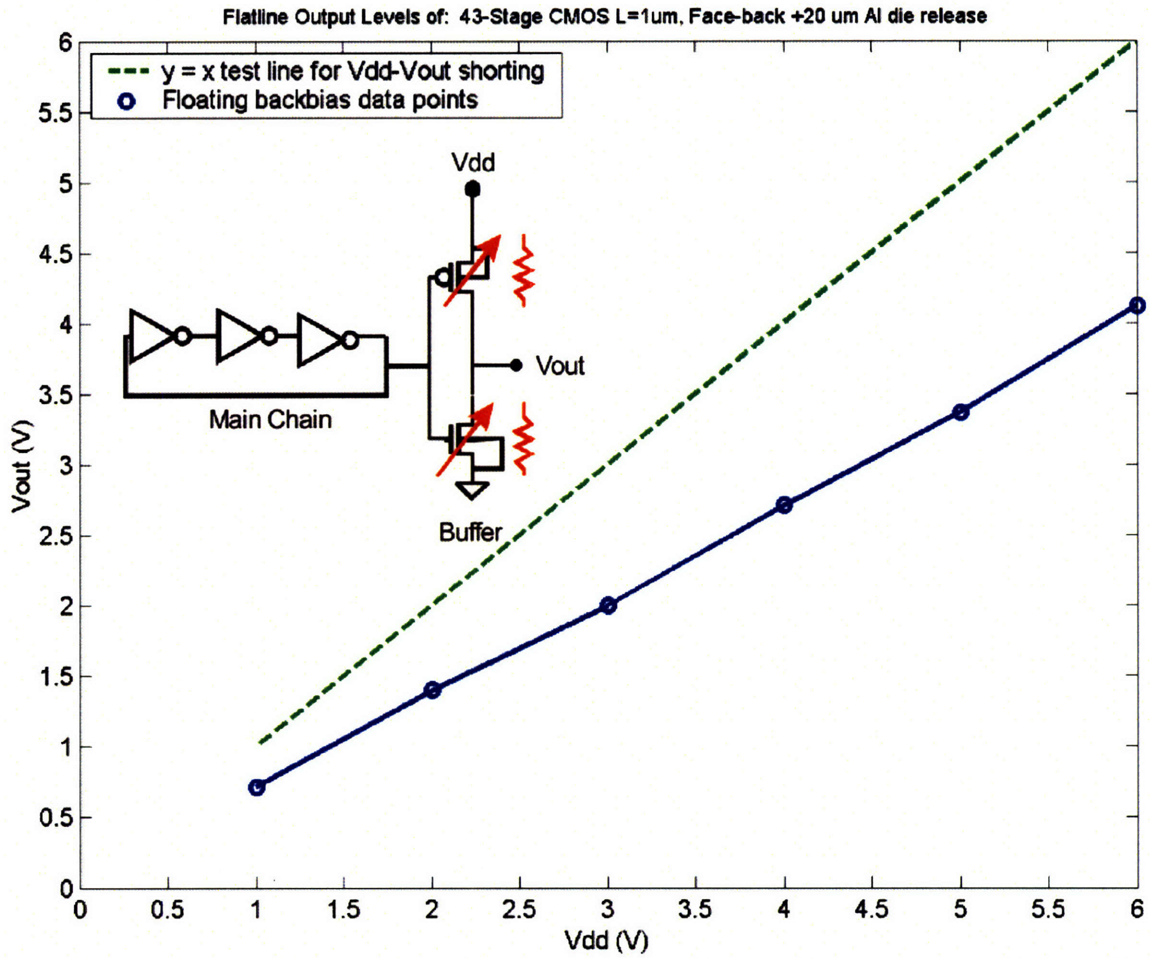


Figure 4-18: 20 μm Al released, Face-back bonded 43-stage CMOS at L = 1 μm with floating PMOS back-gates. Since all output responses were flatlines, the output DC voltage Vout was plotted against the power supply voltage Vdd. This particular circuit also had no Vdd-Vout shorts and was in a metastable state that inhibited oscillation.

4.4 Cu Parasitic Extraction with Bonded MOSFETS and 2-D Ring Oscillators

4.4.1 2-D Ring Oscillator Output From Cu-Bonded Substrates

Although the face-face ring oscillators worked, we still have to investigate the parasitics that contributed to its 4.5x reduction in oscillation frequency when compared to a parastic-free Cadence model. To the first order, these unknown parasitics must involve the Cu-Cu bond and/or planes in terms of either adding stray capacitances or adding series resistance to the oscillators' critical path. One quick way to check for these parasitics is to compare the results of the 2-D, NMOS-only ring oscillators on virgin NMOS-SOI subtrates to the equivalent oscillators in both the face-face and face-back bonded samples. Probing results from the bonded samples can be easily done because we designed a big thru-PMOS via that tunnels down from the stack's top surface down to the NMOS layer's Al Metal #1 pads. A quick schematic of this can be seen in Figure 4-19.

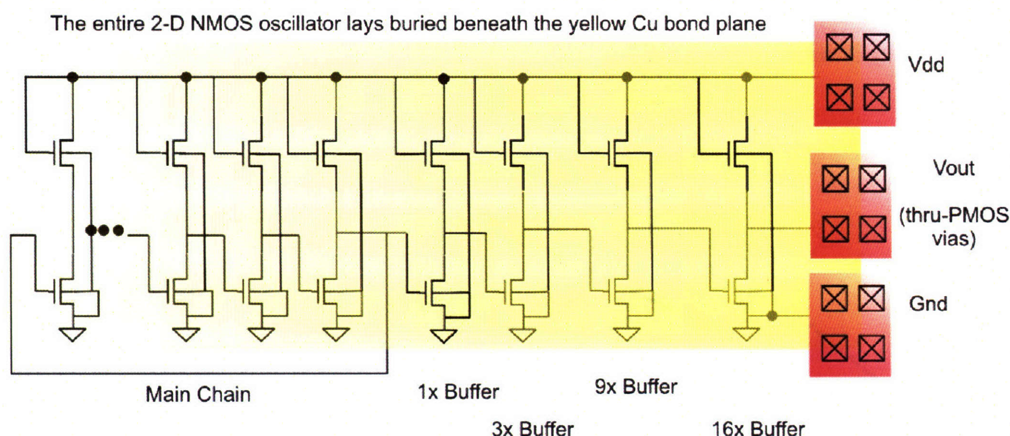


Figure 4-19: The buried 2-D NMOS oscillators can be probed by pulling the signals up from the bottom layer with the Cu-Cu bond pads and the Cu damascene vias.

The litmus test would be based on the following 2-D, 21-stage NMOS oscillators:

- Switch NMOS W/L - Load NMOS W/L = 80/1 - 5/1
- Switch NMOS W/L - Load NMOS W/L = 80/1 - 10/1
- Switch NMOS W/L - Load NMOS W/L = 60/1 - 10/1
- Switch NMOS W/L - Load NMOS W/L = 60/1 - 20/1

The outputs of these 2-D ringers were first measured and recorded from an unbonded NMOS-SOI substrate. Then, those oscillope output traces will be compared with outputs from these 3 samples:

1. Face-back bonded made from a 10 μm Al release layer PMOS-handle complex
2. Face-back bonded made from a 20 μm Al release layer PMOS-handle complex
3. Face-face bonded samples

A sample simulation result of a 2-D ring oscillator, without parastics, is shown in Figure 4-20. Care was made to ensure that the W/L = 80/1, 60/1, 20/1, 10/1, and 5/1 NMOS models approximately matched the measured I-V curves from individual, unbonded devices. Also, a table of comparison between the simulated and experimental results from these 2-D NMOS rings is shown in Table 4.8, where all samples and the simulation itself were biased at Vdd = +4V. This bia was chosen as a merit because it is the maximum sustainable voltage before the on-chip MOSFETS suffer from impact ionization.

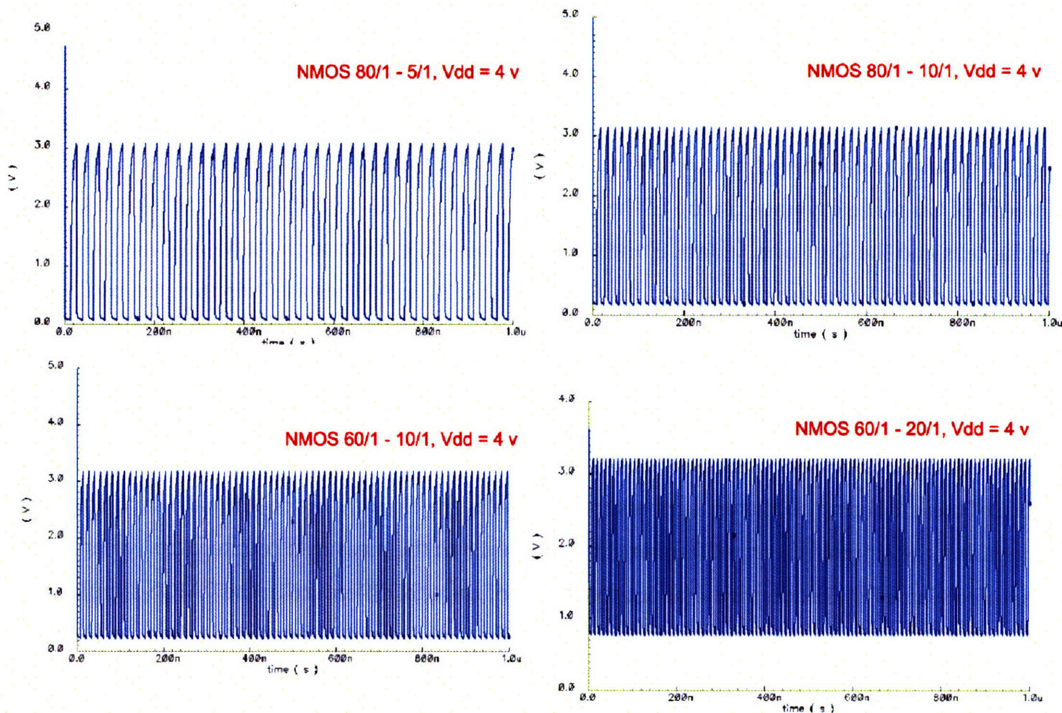


Figure 4-20: Simulated 2-D NMOS rings, all biased at Vdd = +4 V. In each plot, the caption "NMOS W1/L1 - W2/L2" refers to the width/length ratio of the NMOS switch (W1/L1) and the ratio of the active NMOS load (W2/L2).

Furthermore, plotted in Figures 4-21 thru 4-24 were the comparison results for only Vdd = +4 V, and if the if the reader wishes to examine the oscillator output at other bias points, the aggregate results for Vdd = +3, +4, and +5 V can be found from Appendix D.7 to Appendix D.7.4. To begin our short discussion on the 2-D rings, looking only at the measured result in Figure 4-21 (while excluding the simulated plots in Figure 4-20 for the moment), the glaring result was that the unbonded and face-face bonded oscillators were almost

Vdd (V)	2-D NMOS Rings	Simulated (MHz)	Unbonded (MHz)	Face-face (MHz)	Face-back, 10 μm Al release (MHz)	Face-back, 20 μmAl (MHz)
4.0	80/1 – 5/1	39.06	5.154	5.682	5.434	5.882
	80/1 – 10/1	61.72	5.814	6.250	6.098	6.097
	60/1 – 10/1	73.52	8.475	9.259	8.772	X
	60/1 – 20/1	105.26	9.615	10.00	10.00	X

Table 4.8: Summary of simulated and measured rsults from the 2-D NMOS ring oscillators biased at Vdd = +4 V only. The red "X" means that those devices were unavailable for testing.

identical in both operations frequency and V+/V- extrema. On the other hand (only refering to Figure 4-21 here), oscillator output from both face-back bonded varieties match each other, but their performance were inferior to the unbonded / face-face pair. In the face-back pairs, the output V+/V- extrema and their operation frequency suffer from severe series resistance problem. It could not have been predicted that parasitic capacitance mattered as much because *both* the face-face and the face-back samples had the same amount of Cu plane coverage on top of the 2-D NMOS ringer cells and the same vertical separation between the cells and the Cu plane. In fact, if one refers to Table 4.8, even though the V+/V- values in both face-to-back catagories were sub-par, at least for the face-back, 20 μm-Al released circuits, the oscillation frequencies of the 80/1 -5/1 and 80/1 - 10/1 rings were fairly comparable to both the unbonded and face-face moieties. [8].

Although we have just said that the parasitic capacitance between the Cu-Cu bond plane and the Al interconnects of the 2-D NMOS oscillators could be a secondary matter, the layout-dependent parasitic capacitance of these 2-D ring oscillators appears to overwhelm the circuit. One could make such postulate when comparing the simulated output frequencies in Table 4.8 to the measured counterparts, where in all instances, there was a 7x - 11x disparity (about twice that of the 4.5x difference in the CMOS simulation - measurement frequencies) between the Cadence model and the real circuit. In order to refine the Cadence model bridge this parity, a distributed parastic capacitance of around 6 pF - 7.8 pF has to inserted at the inputs of every inverter stage. Physically, this capacitance could probably have been contributed by the long and wide feedback path in the ring oscillator layout itself [9], but it doesn't account for all of the supposed 6 pF worth of extra capacitance per inverter stage. Other sources for this simulation - measurement disparity could also come from inaccurate NMOS Cadence models (despite our efforts to match all the parameters between the model and the experimental MOSFET I-V's). And lastly, of course, the aforementioned parastic

[8] Although some of the oscillope outputs in Figures 4-21 thru 4-24 appears (by eye) to contradict the frequency match arguement made in the main text, after careful replots in MATLAB and going over the raw .CVS data files from the Agilent digital oscilloscope, the frequency results in Table 4.8 were confirmed

[9] Part of the feedback path was made wide at the section where a 2500 A-thick doped-poly local underpass was needed. This was done to decrease the critical path's resistance, but unfortunately it also created a huge poly-substrate capacitance.

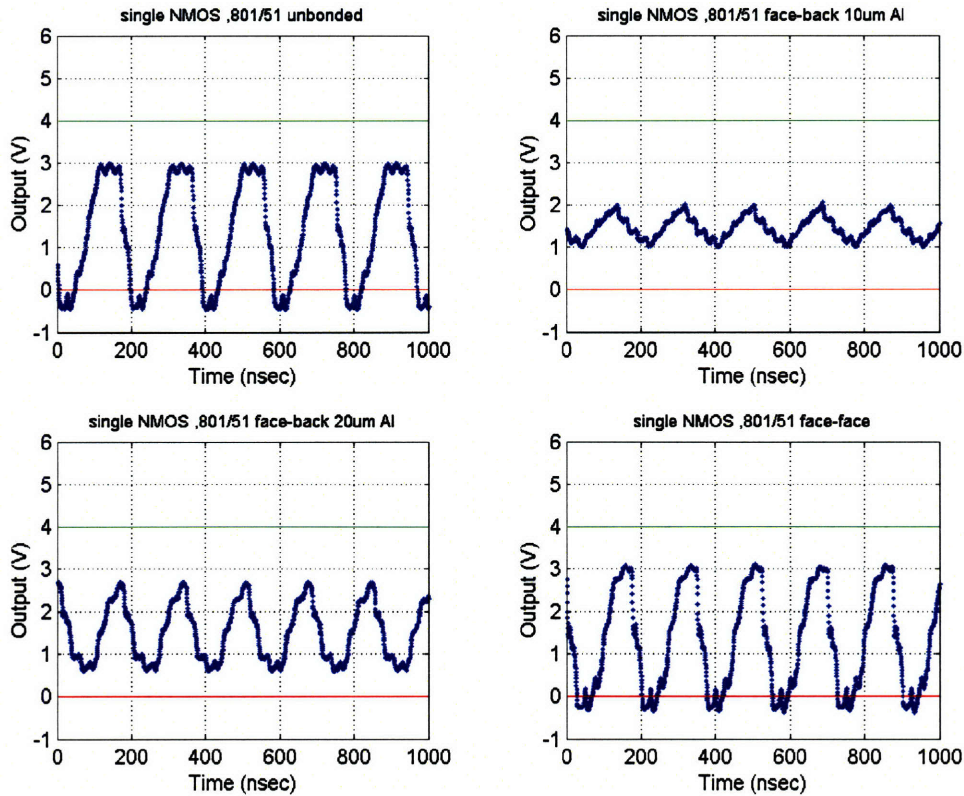


Figure 4-21: 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c), and a face-face bonded sample in (d)

resistance of the layout also needs to be taken into account. It is obvious that more fine-tuning of the simulation model is needed.

With all these facts combined, we conjecture that although the Cu-Cu bonding interface - induced parasitic capacitances do play a huge role in limiting the quality of face-back bonded devices, the series resistance within the Cu damascene vias and within the Cu-Cu bond itself caused a much more observable disturbance in the ring oscillators' output when layout-associated parastic capacitances are excluded. Therefore, the focus of the following section will be on attempting to extracting the parasitic resistance hidden in the critical path. To extract the exact series resistance from the Cu-Cu bond and the Cu damascne vias, it's easier to use results from single NMOS and PMOS I-V's, which we shall take a look at in the next section.

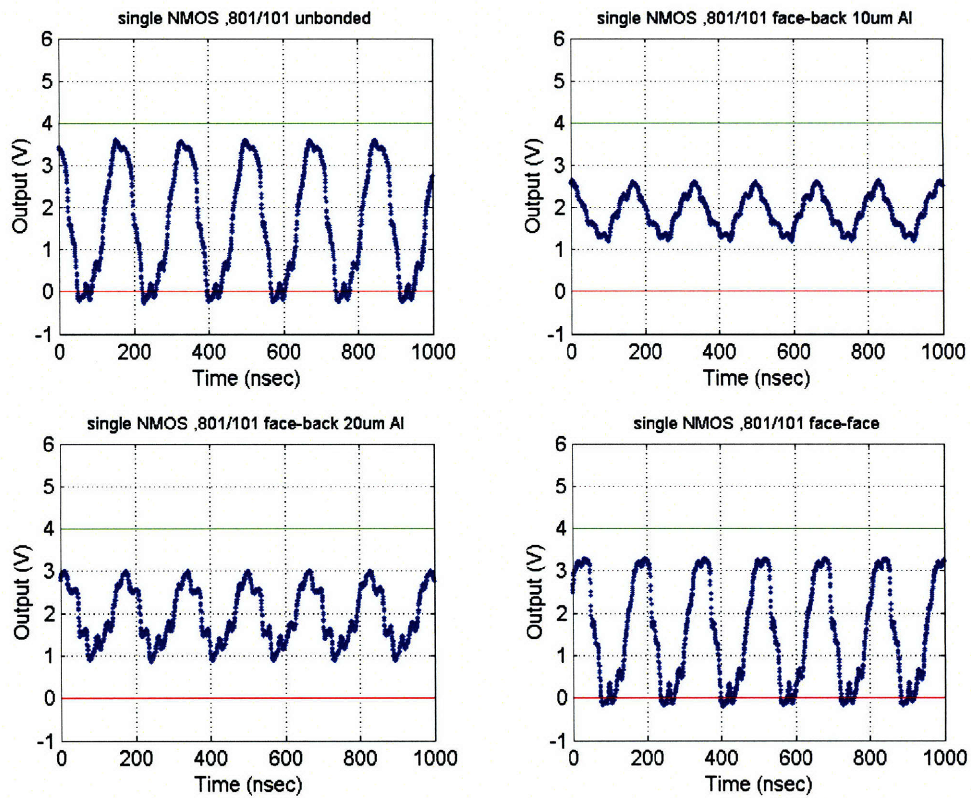


Figure 4-22: 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b) that died during processing, a face-back bonded, 20 μm Al released sample in (c), and a face-face bonded sample in (d)

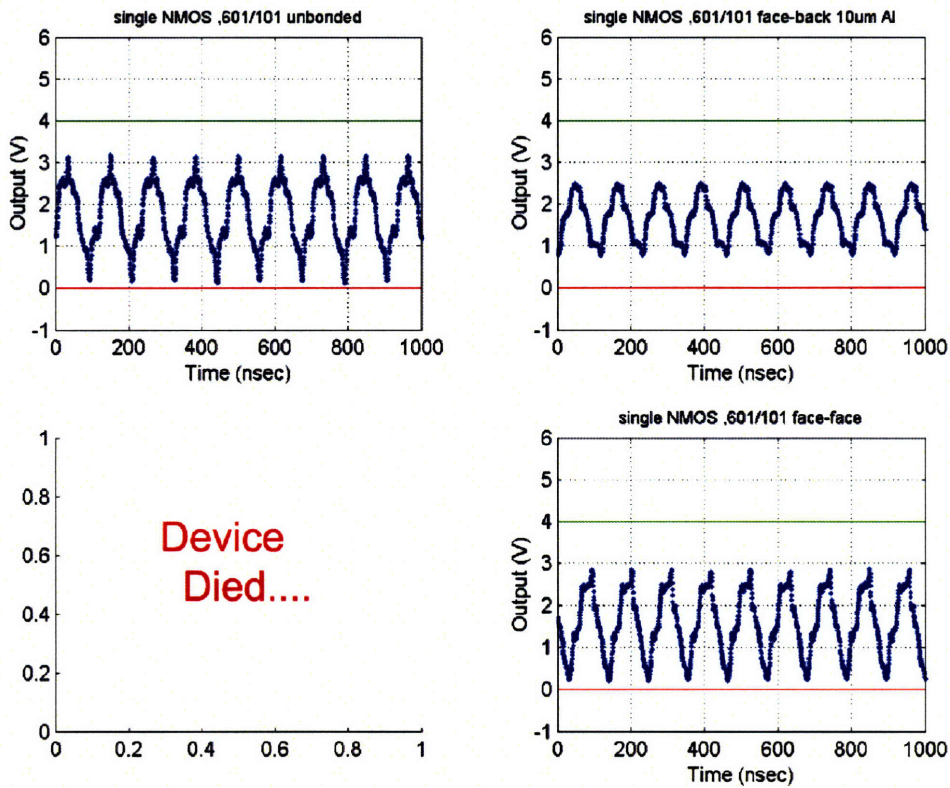


Figure 4-23: 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

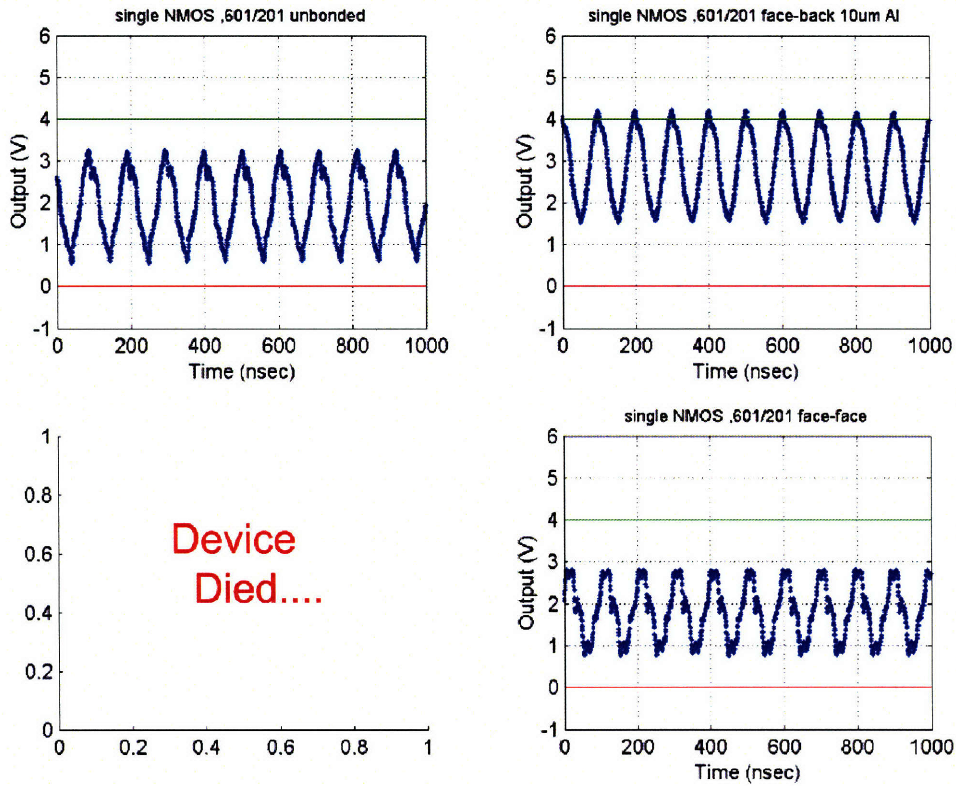


Figure 4-24: 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

4.4.2 Single NMOS / PMOSOutput From Cu-Bonded Substrates

Recalling Figure 4-19 from the last section, we can also probe the buried individual NMOS devices on our bonded samples because our gate-source-drain probe pads have similar thru-SOI vias on them. If one merely repeats the series resistance extraction exercise done as in Figure 4-6 on our new Cu-Cu bonded NMOS samples, then we can compare and extract how much extra series resistance do the Cu damascene vias and the Cu-Cu bond contriubtes from that of the unbonded case. As seen in Figures 4-25 and 4-26, the face-face bonded NMOS have a single source/drain series resistance of about 7.5 Ω, which was 3 times that of the unbonded NMOS wafers. Furthermore, a face-back bonded NMOS device exhibits a single source/drain series resistance of about 127.5 Ω, or about 50 times as much as the unbonded case and about 17 times that of the face-face bonded sample. Since the only processing differences between the face-face and face-back devices were:

- The Cu damascene via quality for the face-face samples was more pristine than the face-back samples because the CMP was smoother
- The Cu-Cu bond for the face-face sample was probably more uniform because it was done on a wafer-wafer level prior to dicing

Could the series resistance factor alone cause the face-back oscillators to fail? Or how about the fact that the face-face oscillators rang slower than the parasitic-free Cadence models? By plugging adding in each series resistance 5, 15, and 255 Ω on the source-end of both PMOS and NMOS devices, the transition of operational frequency decreased only slightly, from the theoretical 24.59 MHz down to about 22.67 MHz, then plunges to about 10.73 MHz in the 21-stage, L = 3 μm cell simulations. To get the simulated frequency down to about 5 MHz (of what we actually measured), either the series resistance has to increase up to about 750 Ω, or if the series resistance stays around 15 Ω as in the face-face bonded sample, then we have to add an inordinate amount of parasitic capacitance across the output of each inverter, which does not make sense physicallly even when we consider the backbias plane's capacitance contribution. More work needs to be done on on de-embedding 3-D structure before one can properly model the entire circuit.

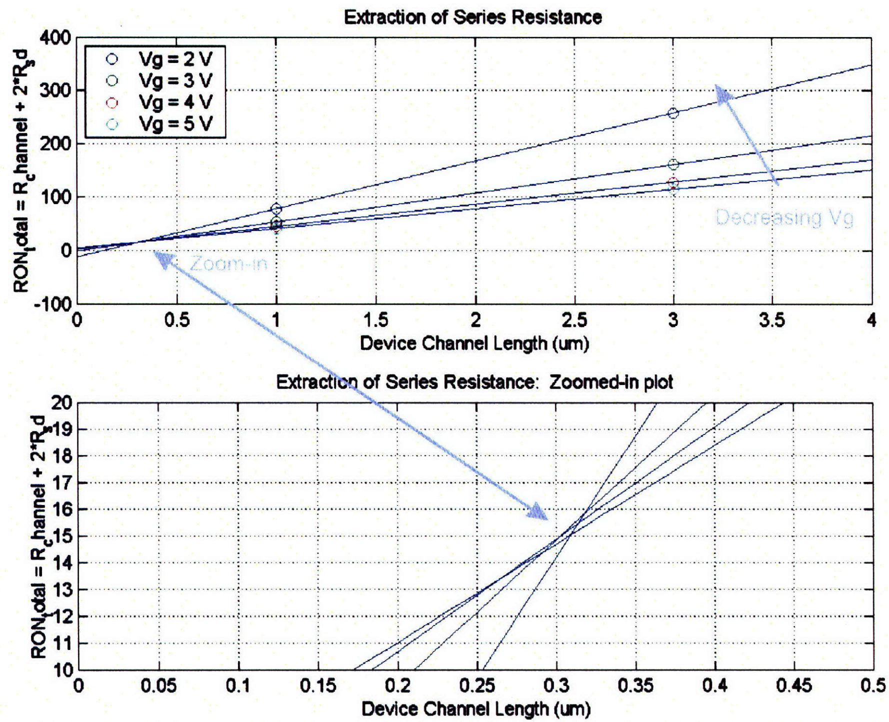


Figure 4-25: Extraction of *face-face bonded NMOS* series resistance. The on-resistance of each 60 μm-wide NMOS, at Vds = 0.5 V, was plotted as a function of both the gate length L and the gate bias Vg. The total source-drain series resistance $2R_x$ was extrapolated at the approximated to be about 15 Ω.

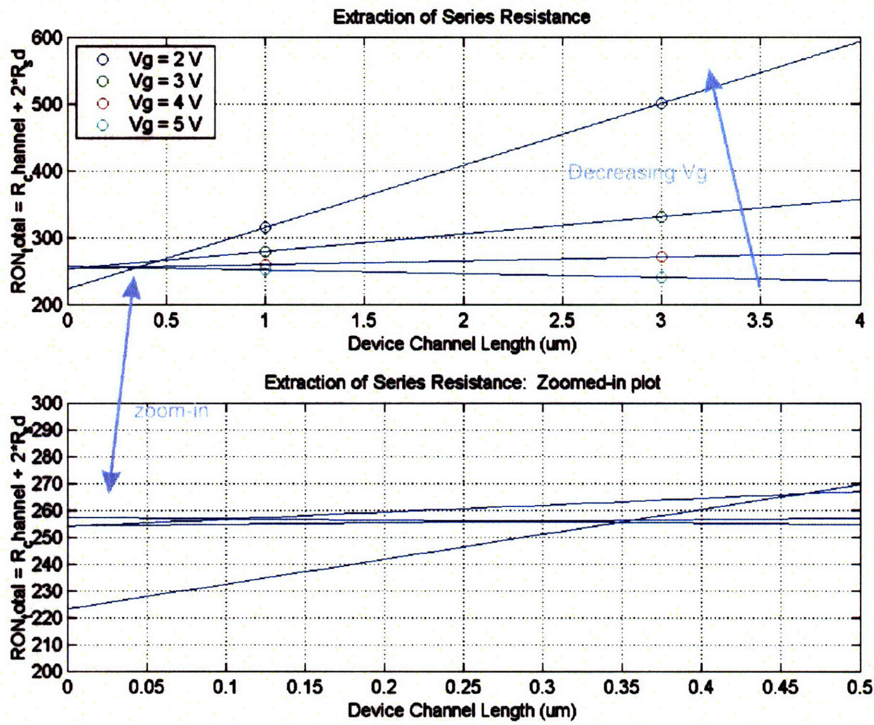


Figure 4-26: Extraction of *face-back, 10 μm Al- bonded NMOS* series resistance. The on-resistance of each 60 μm-wide NMOS, at Vds = 0.5 V, was plotted as a function of both the gate length L and the gate bias Vg. The total source-drain series resistance $2R_x$ was extrapolated at the approximated to be about 255 Ω.

4.5 Summary of Ring Oscillator Results

So, what have we learned by building a 3-D ring oscillator using Cu-Cu bonding ? For starters, a face-to-face configuration was the quickest and most reliable method for achieving good circuit yield because it was:

1. Easier to wafer-align and waferbond than to die-align and die-bond
2. The PMOS wafer surface after Cu via damascene was better when one doesn't have to perform a handle-wafer bond and a substrate etchback before Cu CMP
3. In lieu of #1, the face-face Cu bond yields a lower contact (or series) resistance because the bond quality and uniformity was better on the wafer-level. Having said that, there's no reason why die-die bonding has to be inferior to wafer-level bonding - just as long as we re-design and improve the machinery for the task.
4. The inter-layer parasitic capacitance between the PMOS and the NMOS was definitely lower because the PMOS backbias plane was isolated from the NMOS bulk in the face-face bonding configuration.

However, an asterik has be placed on the last remark. The lower parasitic capacitance for the face-face configuration can only be taken advantage of when the number of device layers is two. If one wishes to add a third device layer onto the 3-D stack, then the managment of stray inter-level capacitance and how or where to apply backbias to the upper device tiers will be an engineering and a design challenge. Going further on the subject of inter-layer biasing, if any top-tier circuitry contains a vast patch of dense SOI devices (such was in the case of the 43-stage, L = 3 μm CMOS ringers), a circuit or device design methodology that decrease the amount of SOI body charging must be take into account. In other words, a method of which one can drain the accumulated charge at the Si-BOX interface would be of great importance any kind of 3-D integration.

Chapter 5

Thermal Characterization: Heat Dissipation in 3-D

5.1 Overview of the Heat Dissipation Problem in 3-D

As previously described in Chapter 1, to maximize the chip performance per unit area, the semicondutor industry's main ice density, lower effective fabrication cost, and higher operating frequencies all come with two major penalties: A higher RC delay due to longer interconnects, and a higher chip power dissipation and density due to Joule heating from fast swtiching times within both gates and interconnects. [27]. Exacerbating this rise power dissipation, SOI devices used in modern circuits exhibit a well-documented problem of "self-heating" [28]. That is, for *medium-doped* Si regions (i.e. inside the channel, where after threshold voltage Vt-adjust implants the doping concentration can be approximately $1x10^{16}$ and $1x10^{17}$ cm^{-3}), both electrons and hole mobilities decrease approximately as inverse-squared of the substrate temperature [29, 30]. Since the resistivity of Si is inversely porportional to the carrier mobility, the resistance inside a SOI channel goes approximately as T^2 [1]. This means that in the self-heating process:

- Current within the MOSFET channel produces Joule heating, or $P = I^2R$, and thereby increases the temperature T inside the channel
- Rise in temperature induces a rise in channel resistance R
- Increase in R further increases the power dissipation P and the overall temperature T

The culmination of these power disspation problems are precisely the reasons why Intel and AMD are shifting their next generation microprocessors to multi-core modules. Instead of having a high-frequency,

[1] "But wait!" one would exclaim, "I thought the equation for TCR (temperature coefficient of resistance) was a linear function of temperature ?" Well, this is normally true if the material is *very conductive*, but for lightly doped n or p-Si, this is not the case. On the other hand, if [dopant] > 10^{18} cm^{-3}, then the mobility is closer to a 1/T relationship, and thus the resistance of highly-doped Si is indeed in the linear form $R = R_o(1+\alpha(T-T_o))$.

high power dissipation chip, one could run two separate, identical cores at lower frequencies and achieve the same or better performance enhancement at the cost of a larger chip area. Although engineers are tempted to utilize 3-D integration in order to reduce the multi-core chip area, the current debate among them is whether going to 3-D will exacerbate the heat dissipation problem even further. At a glance, if a theoretical 3-D stack contains four SOI device layers, how is one going to remove heat effectively from the middle layers ?

Independent of 3-D research, investigators have always been looking into more effiecient heat removal techniques for modern circuits. By far the most popular choice of heat removal is based on forced convection (especially microchannel cooling [31]), and this makes physical sense not only because the heat transfer coefficient in forced convection systems are quite high, it is also a proven technology since within every modern PC or laptop there is a cooling fan installed to carry away heat from the heat sink. However, until advancements in liquid forced convection cooling are mature enough for production, the number one priority should still lie in maximizing the heat conduction efficiency travelling from the device hotspots to the heat sink [2]. Already there are some research groups trying to utilize metal surfaces and thermal vias as microscale heat flux spreaders and conduits, respectively [32], but most researchers believe that if conduction was used as the main heat transfer vehicle, then the design of thermal vias are the keys to success.

It is in the opion of this author, though, that metallic planes can be more useful than thermal vias in reducing the maximum surface temperature of a 3-D stack, and a quick example of this can be made from observing Figure 5-1. Suppose a capacitor was built between Metal # 8 and Metal # 9, and and one of the electrode requires an electrical ground connection. To accomplish this, a circuit designer would simply pull a ground Cu wire from of the substrate, pull it up through the ILD oxide (distance denoted by L_{ox} = 100 μm), and a ground connection can be made with a negligible IR drop because the ratio of the electrical resistacne between the Cu wire and the ILD was near zero, or:

$$\frac{R_{cu}}{R_{ox}} = \frac{0.02}{10^{18}} = 2^{-20} \tag{5.1}$$

where

$$\begin{aligned} \frac{\sigma_{conductor}}{\sigma_{insulator}} &= \frac{\sigma_{cu}}{\sigma_{ox}} = \frac{5 \times 10^{7}}{10^{-14}} \\ &= 10^{21} \end{aligned} \tag{5.2}$$

[2]Real-life example: Why do 8" wafers from Intel undergo backside grinding to remove 550 μm of Si before dicing and packaging ? It's because this can reduce the basal thermal resistance of the chip by 75 %. The Au-Sn solder between the package and the chip acts as both an adhesive and it maintains a good level of thermal condution from the substrate to the heat sink.

Therefore, for every amp of current flowing with through the Cu ground via (suppose there was a *tiny but nonzero* difference in potential bewteen the earth ground and the bottom capacitor electrode), the current flowing through the ILD is basically nil.

Now referring to the thermal analogue on the right-hand side of Figure 5-1, let's say that there is a 100 °C temperature gradient between two metal layers within the circuit, and we require the bottom surface to have an isothermal condition of 0 °C. In the thermal domain, however, one cannot simply connect a Cu wire of a 0 °C ice bath, weave it through the L_{ox} = 100 μm worth of ILD, and expect that bottom metal surface to also be at 0°C. This is because the ratio of the thermal resistance between a good conductor (Cu) and a good insulator (oxide) is no longer near zero. In fact, they are only 3 orders of magnitude apart, or

$$\frac{R_{Tcu}}{R_{Tox}} = \frac{25}{7142} = 3.5 \times 10^{-3} \tag{5.3}$$

where

$$\begin{aligned} \frac{k_{conductor}}{k_{insulator}} &= \frac{k_{cu}}{k_{ox}} = \frac{400}{1.4} \\ &= 285.7 \end{aligned} \tag{5.4}$$

Therefore, for every watt of power transferred through the Cu thermal via, about 3.5 mW of power is also leaking through the thermal-insualating ILD, with a heat flux vector having both vertical *and* horizontal components. The critical point in this exercise is that a good thermal isolator leaks *a lot more* energy when compared to a good electrical isolator, and the disparity here is that:

$$\begin{aligned} \frac{\sigma_{conductor}}{\sigma_{insulator}} &= 10^{21} \\ \frac{k_{conductor}}{k_{insulator}} &= 10^{2} \end{aligned} \tag{5.5}$$

Even if the thermal vias were redesigned to a shorter length, the thermal IR drop (or the thermal gradient $-\nabla T$) across the via while transporting a heat current is still far from negligible because the thermal insulator's R_T scales simultaneously with the thermal conductor. Therefore, instead of concentrating our efforts on maximizing the heat flux magnitude purely in the z-direction, which is inherently a leaky process as was described above, our objective for this chapter will be to use Cu planes in maximizing the *divergence* of the heat flux. In other words, let's use the metallic planes to spread the heat flux such that the heat flow burden can be shared by all three degrees of freedom, thereby reducing thermal gradient along the z-direction. In effect, we'll be maximizing the "leakiness" of thermal insulators and use that to our advantage.

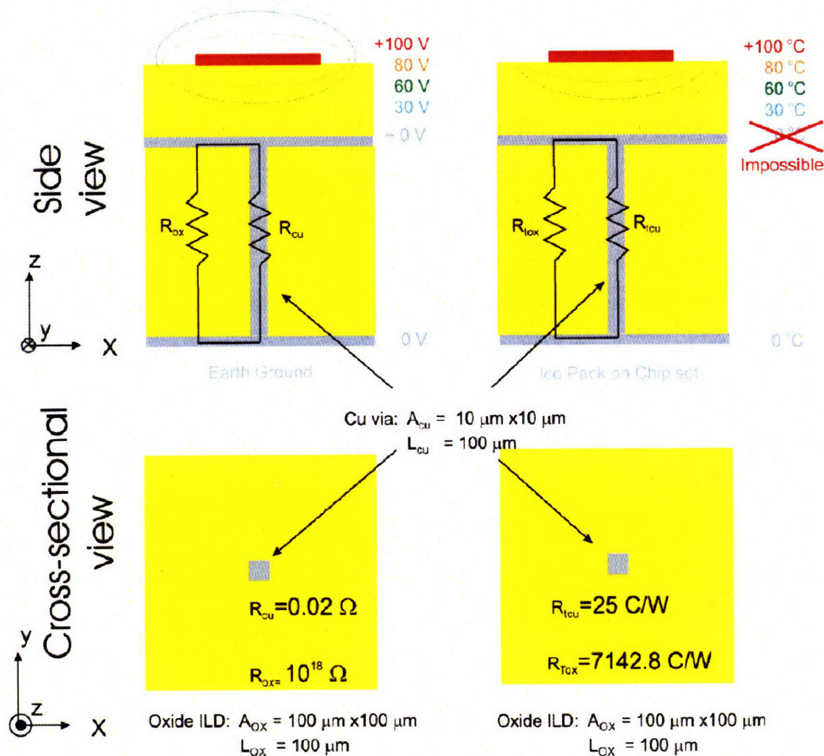


Figure 5-1: Poisson Eq. - Heat Eq. inequality: A quick reasoning as to why thermal vias may not be that useful in decreasing the overall temperature of a chip. See text for details.

5.2 3-D Heat Transfer Simulations

One of the most important aspect of thermal engineering is to design material and physical features that can accurately transport a desired amount of heat from one place to another while maintaining maximum energy efficiency. In systems where conduction is the only medium of heat tranfer, the efficiency of a heat circuit reduces to an exercise of manipulating a system's heat flux direction $k\nabla T$. Moreover, by manipulating the heat flux lines, the two main design specifications in any general thermal network that coule be optimized are:

- The maximum temperature of the system
- The overall isotherm profile of the system

To begin with, the overall system isotherm profile (shape-wise) is a much easier requirement to satisfy than that of the maximum system temperature. Since a system's isotherms are mathematically perpendicular to $k\nabla T$, the isotherm shapes can be engineered if the heat flux lines can coerced into travelling in certain directions using high thermally-conductive conduits. On the other hand, if the maximum temperature

requirement was added, then two design conditions have to be optimized simulatenously:

1. The direction of heat flux $k\nabla$T at a point (x,y,z)
2. The volume integral of $-\nabla\cdot(k\nabla T)$ within the object that the flux $k\nabla$T is passing through

With these two design conditions in mind, let's theoretically compare whether or not Cu planes could be more effective in reducing the maximum temperature of a heat stack. A finite-element model (FEM) simulation is the best course of action here because it's often difficult, if not impossible, to obtain closed-form solutions for the isotherms T(x,y,z) in mutli-layer structures with sharp thermal conductivity discontinuities in 3-D.

5.2.1 Reference Simulation: 2-D SOI Heater

The first reference model simulated was an 1-level SOI (unbonded) wafer containing three rows of n^+-doped SOI self-heating resistors. The choice of the three resistors were based on the fact that we would like to explore the temperature profiles within and outside of a highly-packed, self-heating circuit. By applying a V = 2.5 V bias, the resistors will undergo self-heating process governed by three coupled equations [33]:

$$\begin{aligned} P &= \frac{V^2}{R} \\ C_T\frac{dT}{dt} &= -\frac{T}{R_T}+\frac{V^2}{R} \\ R &= R_o(1+\alpha(T-T_o)) \end{aligned} \quad (5.6)$$

where C_T and R_T were the equivalent heat capacitance and the equivalent thermal resistance of the SOI structure, R_o = 60.75 Ω was the electrical resistance of the heating resistors at the reference temperature T_o = 300 K, R is the total electrial resistance at an arbitrary temperature T, and finally, α = 1161x10^{-6} was thermal coefficient of resistance (TCR) of the Si doped at 10^2 cm^{-3} (which was extracted from mobility calculations). At steady-state, the time-varying term in the second equation approaches zero, which means a single expression for the maximum steady-state tempearture within the structure can be obtained from the coupled equation [33]:

$$T_{ssv}=\frac{R_TV^2/R_o}{1+\alpha R_TV^2/R_o} \quad (5.7)$$

A pre-simulation estimation of T_{ssv} will require us to know the thermal resistance R_T beforehand. Since R_T is a spatially nonlinear function that encompasses the geometry and thermal conductivity of every subdomain in our structure, an *accurate a-priori* estimation for the thermal resistance value is theoretically impossible most of the time. Therefore, we will rely on steady-state FEM simulations for the next few sections to obtain T_{ssv} as a function of geomerty, thus bypassing the need for the value of R_T

The simulated results for the 1-level, self-heating SOI resistor from FEMLAB v3.1 are presented in Figure 5-2, where only half of the structure was simulated because adiabatic symmetry conditions can reduce the amount of meshes in the FEM model, thus dramatically saving memory and computation time. Also, results from the 250 μm x 250 μm adiabatic simulation boundaries could be mirror-imaged and summed over a larger area, such as an 1 cm x 1 cm chip, by again taking advantage of symmetry. Therefore, to simulate a field of dense heating elements on a large chip, all one would decrease adiabatic boundary dimensions and use symmetry to sum up the individual results.

Continuing with the simulation discussion, the input power (excluding self-heating effects) for this structure was solely provided by the 1.5 ($5^2/60.75$) = 617.28 mW of Joule heating using the "one-and-a-half" resistors. The steady-state maximum temperature of 63 ^oC occured at the resistor nearest to the x-axis, which corresponds to the middle of the 3 heating resistors in full geometry. Also, since all 3 heat generators were operating equally at constant power *density* (isoflux condition), the isotherm patterns were fair-shaped ellipsoids and were concentric around the inner-most SOI heater. Since these ellipsoidal isotherms were well-behaved shape-wise and extends all the way towards the bottom heat sink (and the Si substrate has an uniform thermal conductivity k), the heat flux lines -k∇T exhibited a purely radial pattern across most of the structure. Mathmatically, radial flux lines suggest that the volume integral of $-\nabla\cdot(k\nabla T)$ was near its maximum for this FEM model, and unless one drills a direct heat sink going from the resistors to the bottom heat sink, the maximum temperature T_{max} of this structure was very near the optimum point. The physical interpretation of this lies in the fact the Si substrate itself acts as a giant heat flux spreader and no additional thermal conductor planes can improve on their divergence. In terms of thermal circuit design, since the Si substrate is large and thermally conductive, it has effectively decreased the overall thermal resistance R_T by increasing the effective cross-sectional area occupied by the heat flux lines.

5.2.2 Reference Simulation: 2-layer SOI Heater

The second reference FEM simulation performed was that of a double-layered SOI resistors where the top heating elements aligned directly above the bottom heaters, which is the worst-case scenario for self-heating 3-D devices. Once again, the self-heating resistors obey the three coupled equations in Eq. 5.6,the input voltage of these resistors was 2.5 V, and the values of R_o = 60.75 Ω and α = 1161×10^{-6} were also conserved. The two SOI resistor tiers were separated from one another by a 2-μm SiO_2 ILD layer and a thin 2000 A BOX layer that's associated with the top SOI film. Since there are 3 resistors in the equivalent half-circuit, the total input power of the structure, again neglecting self-heating effects to start with, was about $3(5^2/60.75)$ = 1.23 W. Although we have effectively doubled the input power from the 1-layer SOI model, each heating element's power *density* was kept constant bewteen the two models [3]. The overall structure

[3] It seems like we're comparing apples to oranges, but the original thought here was that we wanted to see what happens if we stacked two *identical* structures on top of another, such as the case if one divides a dual-core processor into 2 parts and stacked them into a single column without re-engineering the power dissipation charactersitics of those 2 half-circuits.

has a close resemblance of a face-back bonded 3-D stack with a SiO_2-SiO_2 bonding inteface, and because it lacks any heat flux spreader planes or direct heat conduits, we expect this 3-D stack to heat up immensely. This was the case shown in Figure 5-3 where T_{max} has risen from 63 °C to 154.85 °, which was a + 92 °C or + 245% increase.

Upon closer examination, the isotherm of the SiO_2-SiO_2 bonded 3-D SOI stack in Figure 5-3 were much tighter in the x-y plane than that of the unbonded SOI heaters in Figure 5-2. Specifically, the surface isotherms of the oxide-bonded model were not elliptical but were actually closer to an super-ellipsoid (a rectangular block with rounded edges). This can be seen from a more detailed isotherm plot provided in Figure 5-4, where the heat flux lines $-k\nabla T$ emanating from selected points at the top surface were plotted as blue streamlines, and the thermal gradient $-\nabla T$ were plotted as green streamlines. Since there are no heat spreader or heat conduit structures adjacent the stacked heat sources, heat flux lines near the hot resistors were packed tightly within the xy-plane and were forced to dive straight down towards the heat sink (see Figure 5-5), passing through an SiO_2 ILD layer that's thermally resistive (k = 1.4 W/m-k [34]). In essence, the Heat equation equilibrium

$$-k(T_x^2 + T_y^2 + T_z^2) = \frac{P}{volume} \tag{5.8}$$

was heavily biased towards the z-direction because heat flow in the x-y directions were suppressed by the low thermal conductivity of the ILD [4].

[4] In a close system (5 adiabatic sidewalls and one Dirichlet heat sink constraint), the power generation density (W/m^3) was conserved. Thus, the total heat flow $-\nabla\cdot(k\nabla T)$ must also be conserved, and the isotherms T(x,y,z) are forced to accomodate this flux conservation according to the Heat equation. Therefore, if T(x,x,z) scales in the x-y direction, there has to be a proportional isotherm scaling in the z-direction due to energy conservation

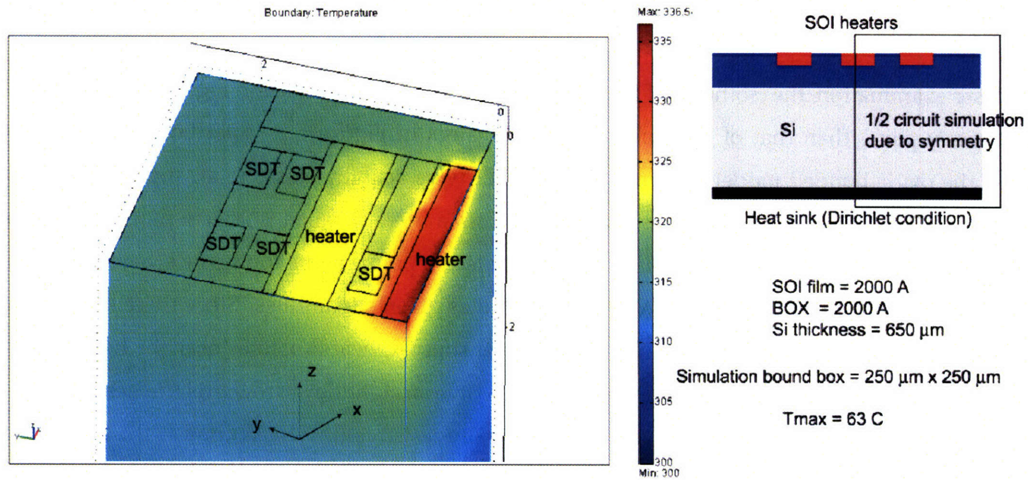


Figure 5-2: Reference FEM simulation # 1: One-level, unbonded SOI heaters. Colorbar temperatures are in Kelvins.

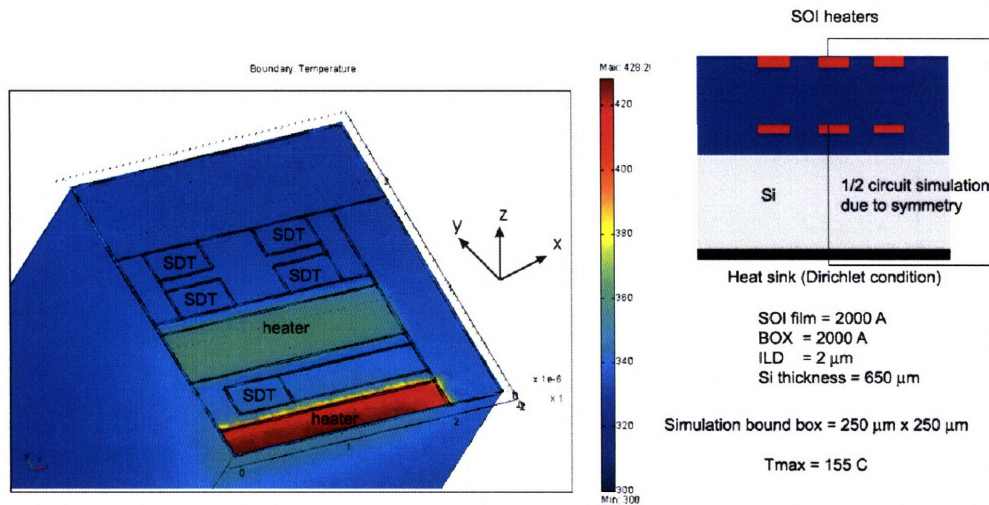


Figure 5-3: Reference FEM simulation # 2: Two-level, oxide-bonded SOI heaters. Colorbar temperatures are in Kelvins.

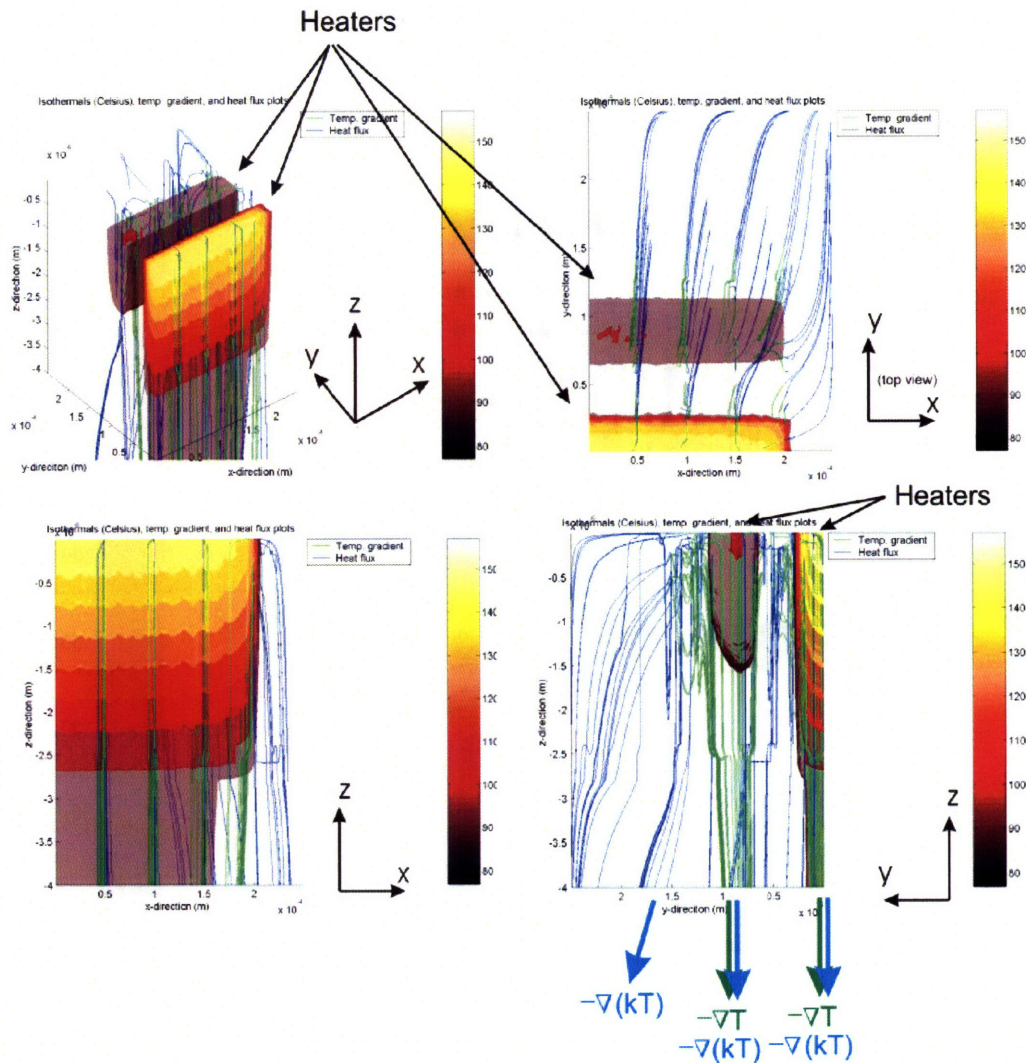


Figure 5-4: Reference FEM simulation # 2: Two-level, oxide-bonded SOI heaters, isotherm plots. Heat flux lines $-\nabla(kT)$ are in blue, and temperature gradient lines $-\nabla T$ are in green. To accentuate the results, the z-axis for each plot was cut off at $z = -4$ μm, or at a depth 0.6 μm below the bottom interface of the BOX from the lower device tier. All colorbar temperatures are in Celsius, and "SDT" stands for "Satellite Diode Temperature" detectors.

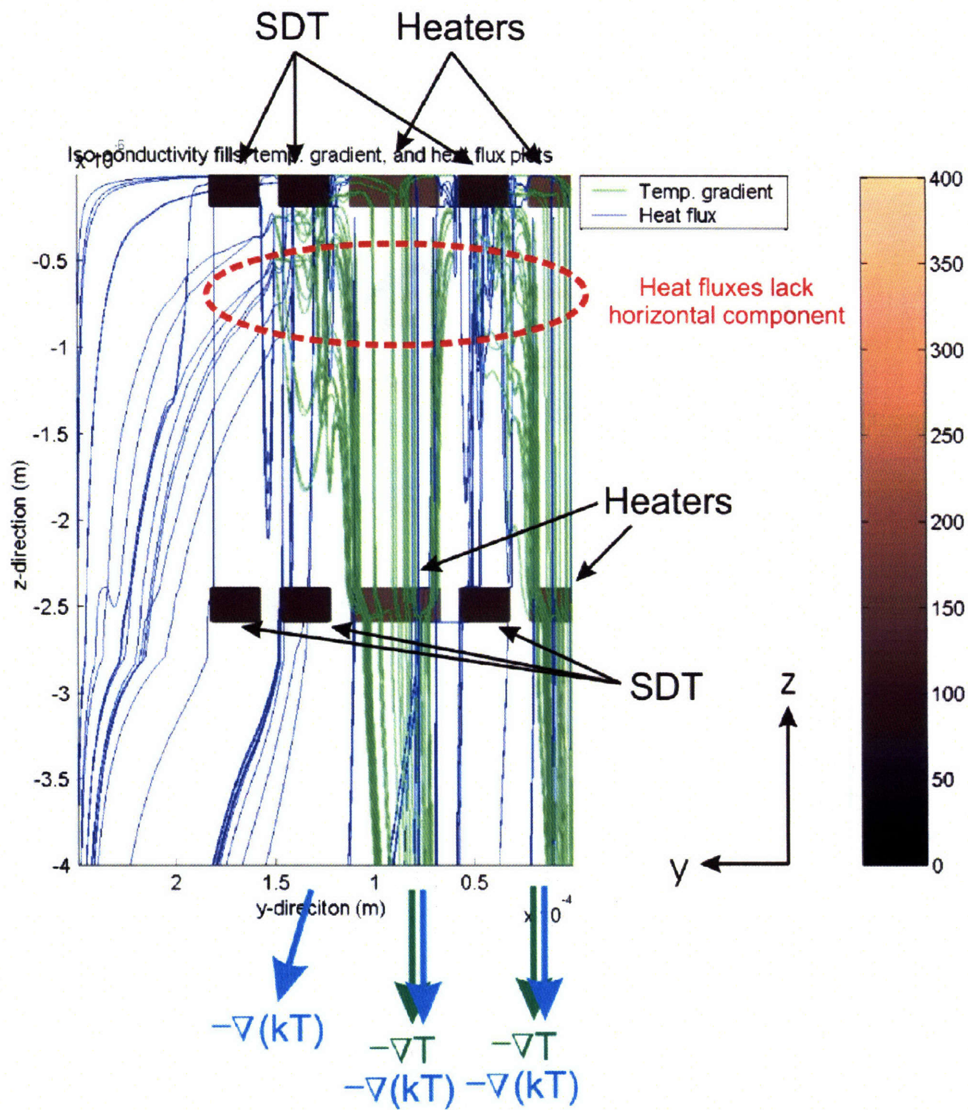


Figure 5-5: Reference FEM simulation # 2: Two-level, oxide-bonded SOI heaters, heat flux plot on a vertical plane. Heat flux lines $-\nabla(kT)$ are in blue, and temperature gradient lines $-\nabla T$ are in green. The values on the colorbar corresponds to the thermal conductivity of the subdomains, in W/m-K.

5.3 Simulation Comparisons: Adding Cu Planes vs. Cu Thermal Vias to the Referenc Structure

In lieu of the previous discussion on the direction of heat flow and conservation of energy in a closed system, one can try to reduce the z-axis temperature gradient by increasing the heat flow along both the x-and-y axes with flux spreaders. The first-order solution to achieve this was to use the Cu bonding plane as the flux diffuser, and as a "thermal profile cleanup" procedure, a second-order solution to reducing the topside temperature was to insert thru-SOI thermal vias. Figure 5-6 displays the cross-sectional and topside views of the reference, 6000 A Cu-bonded with no vias, and 6000 A Cu-bonded with vias structures, and Figure 5-7 shows a direct comparison of FEM simulation results among all three structures. Again, to accentuate the results within the SOI/Cu/ILD/SOI quad layers, the z-axis of each plot was truncated at z = -4 μm, which is located inside the base Si substrate 0.6 μm from the BOX-substrate interface of the lower tier devices.

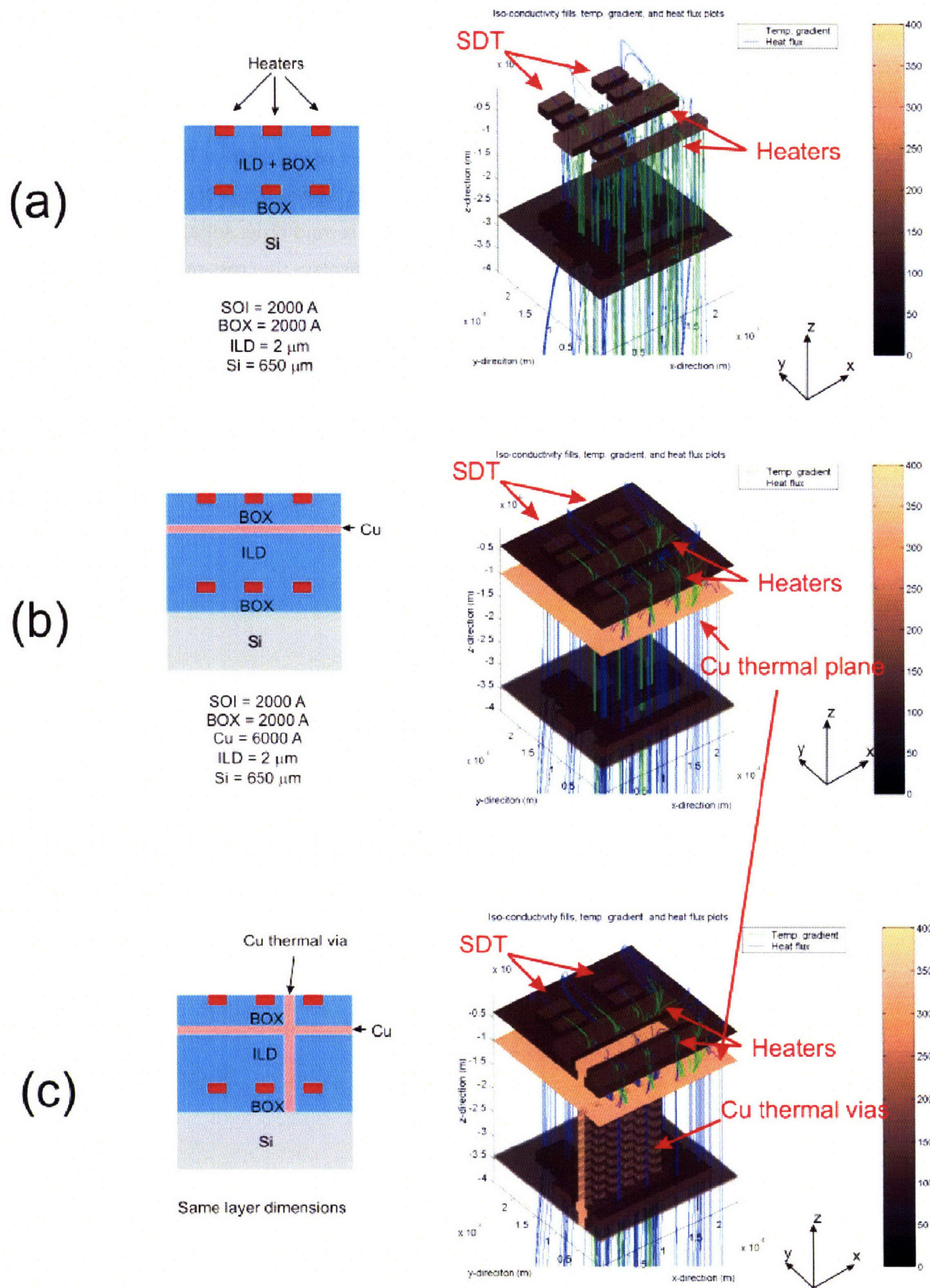


Figure 5-6: Structural between the reference (a), Cu-bonded but no thermal vias (b), and Cu-bonded with thermal vias (c). The colobar coding denotes tge subdomain's thermal conductivity: Pink regions were made of Cu, and brown regions were made Si. Also, SDT stands for "Satellite Diode Temperature" detectors. Also, the heat flux lines were plotted in blue, and the temperature gradient lines were plotted in green.

Discussion - to - Figure Guideline :

Figure 5-7 is an aggregate plot of simulated results for the reference SOI structure, the 6000 A Cu-bonded (no vias) structure, and the 6000 A Cu-bonded (with thermal vias) structure, where the x-y-z, x-y, x-z, and y-z perspectives for each structure were all placed within a given column. Although the discussion of results will mainly be based on the bird's eye-view plot, the reader may wish to resort to the following magnified plots for more details:

- **For COLUMN A:** See Figure 5-4
- **For COLUMN B:** See Figure 5-8
- **For COLUMN C:** See Figure 5-9

X-Y-Z PLOTS :

To start with, from the x-y-z isotherm plot in Figure 5-7 (x-y-z in Columns A, B, and C), the maximum temperature of our double-layer heating structure has decreased from 155 °C to 130.84 °C just by the addition of a single layer of Cu flux diffuser. The inclusion of a thermal via with the Cu plane actually raised the maximum temperature by 1.77 °C, but this was probably a mesh-related artifact of the FEM simulation and for the time being, we will consider T_{max} between the non-via and via structures to be equal.

X-Y PLOTS :

Next, observations from the x-y plots of Figure 5-7 (x-y in Columns A, B, and C), clearly show the decrease in maximum temperature and the overall broadening of the isotherm in the x-y directions when the Cu diffuser plane was added to the reference model. This was also a direct evidence that the addition of the Cu plane has lightened the heat flux burden (referring to Equation 5.8) on the z-axis and transferred some of the heat flow burden onto the x-y axes. As a consequence, the z-axis $-\nabla T$ has been reduced, and this was why the overall maximum temperature has been reduced by more than 24 °C when compared between reference structure to both the Cu-no-via and Cu-with-via structures.

Next, with the addition of thermal vias , the x-y plot of Figure 5-7 - Column C show that warm purple regions between the two heaters have decreased at least 20 °C (outside of the colobar range) due to the via's heat flux siphoning ability. Unfortunately, this did very little to the overall surface isotherm profile and it suggests that thermal vias are probably useful for "heat profile cosmetic surgeries," where random hotspots can be quickly eliminated by placing an adjacent via.

X-Z PLOTS :

Furthermore, focusing on the x-z isotherm plots of Figure 5-7 (x-z in Columns A, B, and C), the spacing between the z-axis isotherms grew when the the Cu diffuser plane was added onto the reference structure and the overall magnitude of the z-axis thermal gradient measuring from the top surface down to the Si substrate has decreased (compare Column A to Column B to see this). This decrease in thermal gradient also signifies a shift of flux balance from z-axis-dominant to a milder z-dependence and stronger x-y dependence. However, an isotherm anomaly occured within the bottom SOI / ILD section when thermal vias

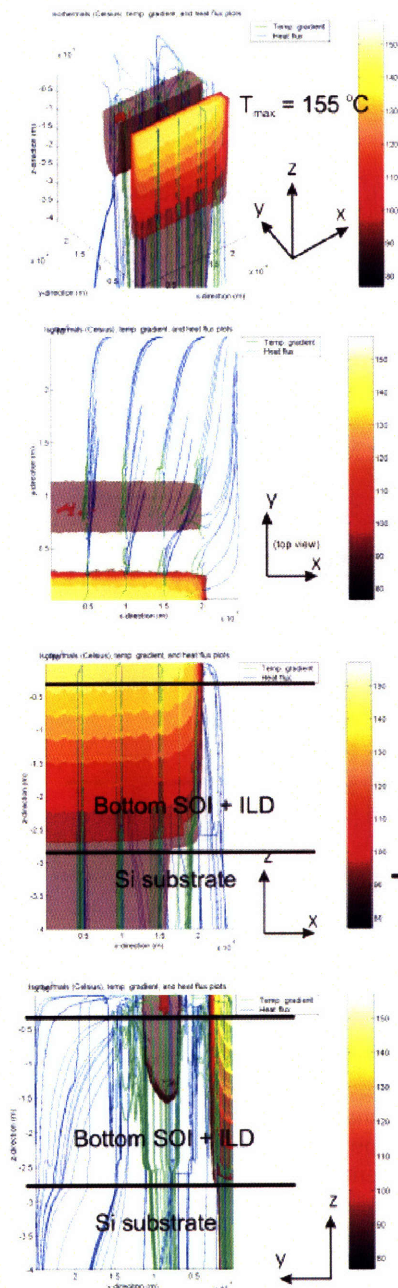
were added to the Cu-bonded model. Although $|-\nabla T|$ continues to decrease slightly, the ILD / Bottom SOI regions below the Cu-Cu bonding layer has become much more isothermal altogether (broader red regions in Column C when compared to Column B), and on average, this region became a bit hotter when compared to the non-thermal via model. This was a major penalty associated with using Cu thermal vias: The walls of the vias have become a thermal Faraday cage, and all regions normal and adjacent to the via sidewalls were forced to reach a common thermal equilibrium that's dictated by the temperature of the Cu via [5]. Since we have already determined that there are no such things as perfect "thermal ground" wires, the Cu via's ambient temperature will be directly proportional to how much heat flux it has siphoned from hot surfaces near the top of the structure. Unfortunately, the more effective the via is in removing surface heat, the hotter it becomes and the negative effects of the Faraday cage becomes worse and worse - regions in the ILD / Bottom SOI can then become hotter than it should.

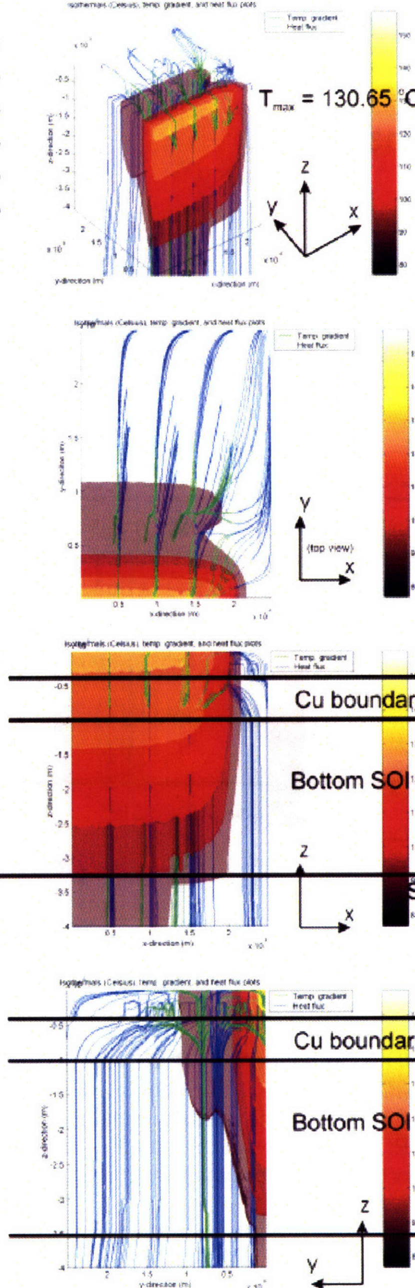
Y-Z PLOTS:

The Faraday cage effect can also be seen in the plots within y-z plane. When thermal vias were added to the system, one can observe that all local heat flux lines were siphoned into the via, thus creating a flux bundle (Column C) that extends all teh way down to the lower SOI tiers. If the thermal vias were overloaded with heat flux, then a hotter isothermal cage will develop.

Conclusion: Cu thermal planes can be used as heat flux diffusers and decrease the overall magnitude of the temperature gradient profile by at least 15%. Cu thermal vias can be used to fine-tune the surface temperature profile and to construct a cold-wall barrier around a hotspot, but care must be taken such that the steady-state via temperature does not interfere with pre-designed cold regions along its major axis.

[5] Mathematically, this is like having an internal isothermal boundary (Dirichlet) condition





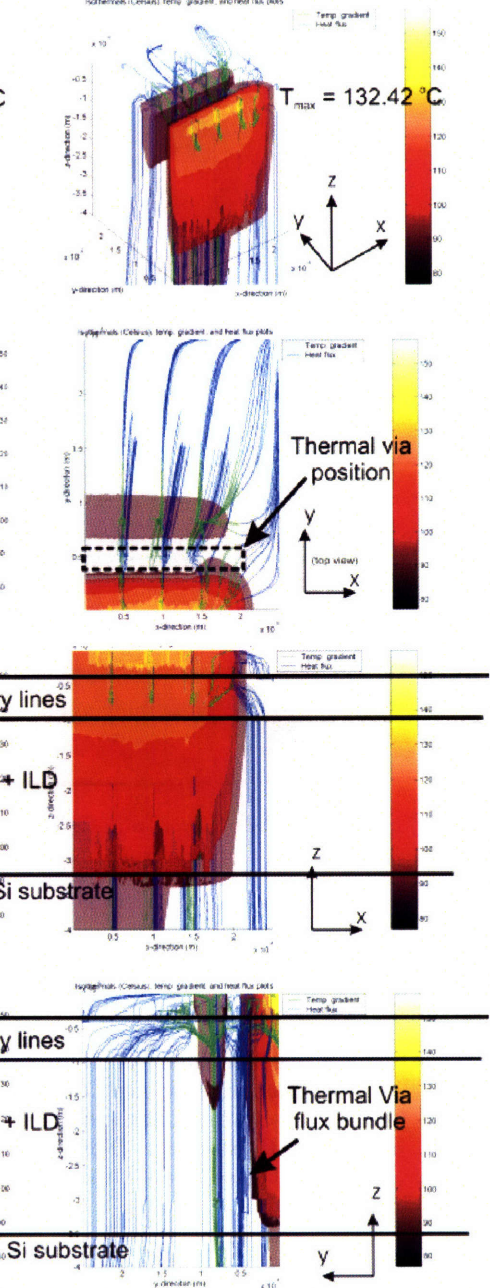


Figure 5-7: Bird's eye view plot: Simulated isotherm comparisons for the reference, the 6000 A Cu-bonded with no vias, and the 6000 A Cu-bonded with thermal vias structures.

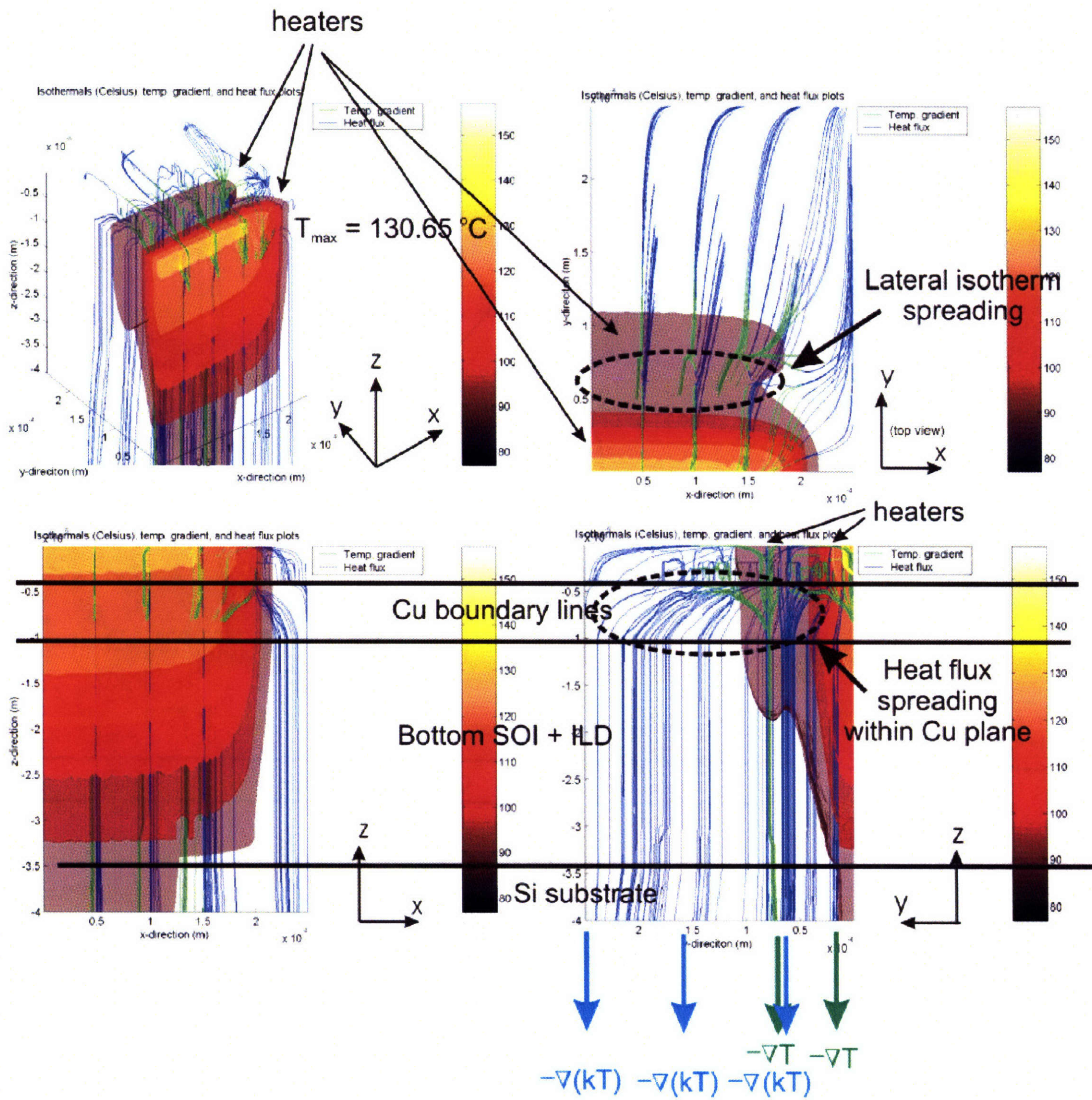


Figure 5-8: Simulated isotherms of 6000 A Cu-bonded structure, with no vias. *These four graphs are the magnified versions of those from COLUMB (B) of Figure 5-7.* The simulated results suggests that the Cu thermal plane does spreads out the heat flux and reduces the z-axis thermal gradient when compared to the reference, non-Cu structure. DISCLAIMER: The outlines of the satellite diodes do not readily show up on the isothermal plots because their neighboring inter-subdomain temperature gradients were small.

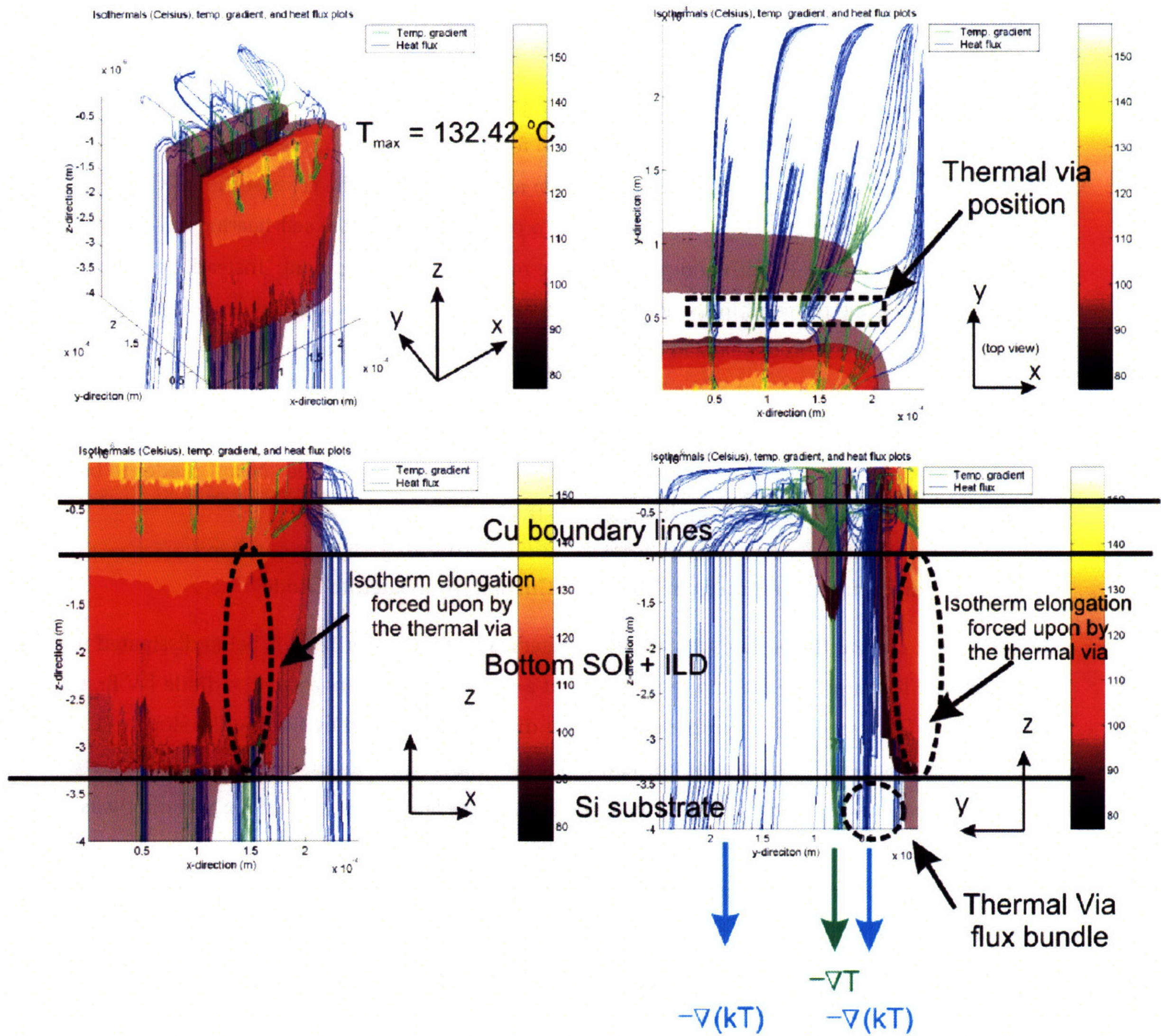


Figure 5-9: Simulated isotherms of 6000 A Cu-bonded structure, with thermal vias. *These four graphs are the magnified versions of those from COLUMB (C) of Figure 5-7.* The simulated results suggests that the Cu thermal vias do siphon the heat flux from its local surroundings, but they do not decrease the maximum temperature of the structure. Also, the thermal Faraday cage effect (elongation of the isotherm) can be seen from both the x-z and y-z plots. DISCLAIMER: The outlines of the satellite diodes do not readily show up on the isothermal plots because their neighboring inter-subdomain temperature gradients were small.

5.4 Measured Results from Real Heat Structures

5.4.1 Reference Measurement and Thermal Calibration: Cu-bonded SOI Heaters, Excluding Thermal Vias

Structure Schematics

In an attempt to verify the simulated results, an aligned triple SOI heater structure was fabricated using face-face Cu bonding. The original design called for highly-doped SOI resistors inside the three long trenches depicted in the layout within Figure 5-10(b) and also seen in the cross-sectional view in (c), but those 10 Ω resistors did not absorb enough power to create an observable self-heating effect because the device current exceed the 100 mA compliance of the measurement equipment. Instead, the SOI trenches were kept intrinsic and the resulting Schottky diodes were used as self-heating power generators operating in constant current mode, as depicted in Figure 5-10(a). Since the Schottky devices were quite resistive, self-heating was observable up to either the 2-W equipment power compliance or the 40 V voltage compliance, where measurements were stopped upon hitting either of the compliance points. Figure 5-10(a) also shows the two groups of satellite diodes measured; obviously, one would expect the "Near" diodes to be hotter than the "Far" diodes solely based on distance from the Schottky heaters alone.

Method of Temperature Extration from Measured Current from Satellite Diode

For each input current forced into the Schottky heaters, a certain power level is being dissipated into the test structure. To extract the on-chip temperature at a particular position, a satellite diode I-V first has to be measured. Theoretically, the temperature of that diode can be found by finding the slope q/kT of the resultant line when one plots ln(subthreshold diode current) vs. the input voltage. Unfortunately, since the Schottky heaters did not heat up to an appreciable temperature, the q/kT slope variation was barely distinguishable as one varies the Schottky power dissipation level. On the other hand, it was found that the reverse saturation current I_o in the ideal diode equation:

$$I_d = I_o \cdot (e^{qV/kT} - 1) \tag{5.9}$$

exhibited an observable upshift as the Schottky forcing current increase. Therefore, the temperature extraction from the measured diode I-V will be based on the shift in I_o. Specifically, the extraction routine was as follows:

1. Stabilize the measurement chuck temperature T_{chuck} to a value
2. With no current forcing in the Schottky diode, measure the satellite diode I-V and plot the data as $\ln(I_d)$ vs. V on the 4155C.

3. Now, varying the Schottky forcing current, re-measure the satellite diodes and superimpose the I-V curves within the same screen

4. The aggregate data of step # 2 step # 3, plotted as $\ln(I_d)$ vs. V, shows a gradual vertical shift in $\ln(I_o)$ as the Schottky heater's power dissipation increases. Since this vertical shift was a constant for a wide range of V's in the subthreshold regime, we took the I_d values at different Schottky power dissipation levels for V = 600 mV only. When this data set was plotted against the Schottky input power at T_{chuck} = 25 °C, one would get either the 'O' or 'X' dark blue curve in Figure 5-11.

5. Repeat steps 1-4 for a different T_{chuck}

After the data collection was completed (ie. The entire dataset is within Figure 5-11) , the rest of the task will purely consist of number-crunching. First, since ordinates of Figure 5-11 was exponentially related to the abscissa, Figure 5-11 will be re-plotted in semilog ordinates and the result is shown in Figure 5-12. Now, our task will be to try to use the current data at different T_{chuck} values at Power = 0 W (denoted by "Region # 1") as the absolute correspondence between I_d vs. T_{chuck}, and also, pick a higher higher Schottky power (denoted by "Region # 2") and also use the data at different T_{chuck} as a correspondence between I_d vs. δT, where $\delta T = T - T_{chuck}$. Using least-square fits for both Regions # 1 and # 2 (and remembering that the resuting slope and y-intercept are exponentiated because of the semilog ordinates), and then combining the Results from Regions 1 and 2, we get two identities that relates to the measured I_d for a given chuck temperature at either the "Near" or "Far" diode positions:

$$\begin{aligned} T_{near} &= (Region2_{calibration}) + (Region12_{bridge}) + T_{chuck} \\ &= \frac{1}{0.056} \cdot ln(\frac{i_{meas}}{i_{o2}}) + \frac{1}{0.0615} \cdot ln(\frac{i_{meas}}{i_{o1}}) + T_{chuck} \\ &= ln[(\frac{i_{meas}}{i_{o2}})^{17.85} \cdot (\frac{i_{meas}}{i_{o1}})^{16.26}] + T_{chuck} \end{aligned} \tag{5.10}$$

and

$$\begin{aligned} T_{far} &= (Region2_{calibration}) + (Region12_{bridge}) + T_{chuck} \\ &= \frac{1}{0.0533} \cdot ln(\frac{i_{meas}}{i_{o2}}) + \frac{1}{0.0621} \cdot ln(\frac{i_{meas}}{i_{o1}}) + T_{chuck} \\ &= ln[(\frac{i_{meas}}{i_{o2}})^{18.76} \cdot (\frac{i_{meas}}{i_{o1}})^{16.1}] + T_{chuck} \end{aligned} \tag{5.11}$$

where

- T_{chuck} = The refence chuck temperature in Celsius

- i_{o2} = The measured diode current at the lowest input power to the Schottky heater in Region # 2, in amps
- i_{o1} = The measured diode current when the Schottky heaters are off, in amps
- $i_{measured}$ = The measured diode current, in amps
- T_{near} = The extrapolated temperature for "Near" diodes, in Celsius
- T_{far} = The extrapolated temperature for "Far" diodes, in Celsius

Finally, one can directly use the above equations and convert the measured currents in Figure 5-11 into the absolute temperature in Figure 5-13

A temperature difference of less than 5 °C exists between the "Near" and "Far" diodes when the Cu-bonded Schottky heater was dissipating power ranging from 0 to 2 W - an obvious deviation from the simulated results in previous sections because our on-chip heaters were Schottky diodes, not doped resistors. Also, one can only conjecture that areas adjacent to the heaters, whether near or far, were fairly isothermal in nature, which was consistent with the simulated results for Cu-bonded strucutures in Figure 5-7. Also, the satellite diodes may not be as hot as those from simulations because we have neither adiabatic boundaries nor arrayed heaters on our chip to act as thermal insulators. So, as of now, this result can only be taken as a reference point because we did not have any devices that were of the SiO_2-SiO_2 bonded variety.

5.4.2 Comparison: Cu-bonded SOI Heaters, with and without Thermal Vias

Our next task is to see whether an addition of thermal vias can lower the local temperature at a point surrounded by heat sources. A Cu-bonded structure with Cu thermal vias was constructed on the same face-face bonded sample, and its layout was presented in Figure 5-14(a). A direct temperature comparison between the "Near" diodes in Figure 5-14 (b) and in Figure 5-10(a) was valid because they were identical in film composition, positional coordinates, and their zero-power subthreshold i-V curves. Thus, the current-to-temperature conversion was performed with the same parameters as in Equation 5.10 and the results were plotted in Figure 5-14(c). Hence, despite that the "Near" diode was surrounded by heat sources on two sides, the addition of a nearby thermal via has decreased the diode temperature by about 12 °C, or in relative terms, a linear -16.6 % temperature decrease at input power ranging from 0.2 W to 1.2 W. The percentage change was very similar to what was simulated in x-y isotherm plots of Figure 5-7, where the insertion of a thermal via has led to at least a 20 °C reduction in adjacent areas that was about 90 °C (or -22.2 %) to begin with. [6].

[6] Again, the purple isotherm disappeared because its temperature was below the colorbar minimum.

5.5 Summary of Thermal Results

In the end, the main conclusion from FEM simulation is that 3-D structures built with Cu planes have an inherent advantage over SiO_2-SiO_2 bonded (or polymer-bonded) analogues because one obtains a free metal layer in which heat flux from the top device tiers could be diffused over a larger area. In addition, that same Cu thermal diffuser plane doubles as a possible backbias plane to control the V_t's of top-tier SOI devices, thus further increasing the value of Cu-3D integration. If the effectiveness of the Cu heat flux spreader was locally maximized within small cluser of intense 3-D heat generators, thermal vias could then be used to selectively target undesirable hotspots and decrease the thermal gradient around those pinpoint locations. Cu thermal vias, however, will never be more efficient than Cu thermal planes in heat removal in 3D-IC's, and the main reason is the cost of via real estate. Since these vias will not even electrically active, the value of adding thermal vias will only be worth it if it *really* decreases the local temperature gradient by a substantial amount.

Next are some comments about the validity of the simulated and experimental results presented in this chapter. Comparing the FEM simulations to the rudimentary temperature extraction of a fabricated 3-D SOI heating circuit, we have some experimental proof that thermal vias can drastically decrease the local isotherm gradients by as much as 16 % when Cu flux diffuser planes were also present in the 3-D stack, although consistency of the experimental data remains in question. Any fabrication mishaps, Cu-Cu via misalignments, inadequate Cu grain growth during bonding, and the quality of the satellite / Schottky diodes themselves could possibly change the effective thermal resistance of the entire structure, thus causing a basal shift in the measured diode current and making the subsequent current-temperature conversion inaccurate. Since we also did not have a reference SiO_2-SiO_2 3-D stack for measurement comparisons becuase of fabrication mishaps in the lab, it was unclear how much improvement did the Cu-plane-only reference structure offered to begin with.

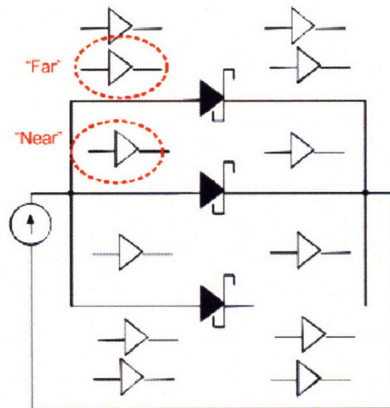
Furthermore, one has to be wary when comparing FEM thermal simulation results from different publications because heat transfer characterstics of a chip are highly dependent layout, film compositions, and where the Neumann / Dirichlet conditions are located in the model. For example, referring to simulation results between the oxide-oxide reference model and Cu-plane-only structure, our simulation result of a -15 % decrease in the overall T_{max} from FEMLAB seemed to be much more optimistic than the results presented by Banerjee et. al.[32], where the introduction of a copper plane only reduced T_{max} from 385 °C to 380 °C (-1.3 %) when compared to a polymer-bonded 3-D stack ($k_{polyimide}/k_{oxide}$ = 0.1). However, upon closer examination, their FEM structure differ greatly in that their model contained:

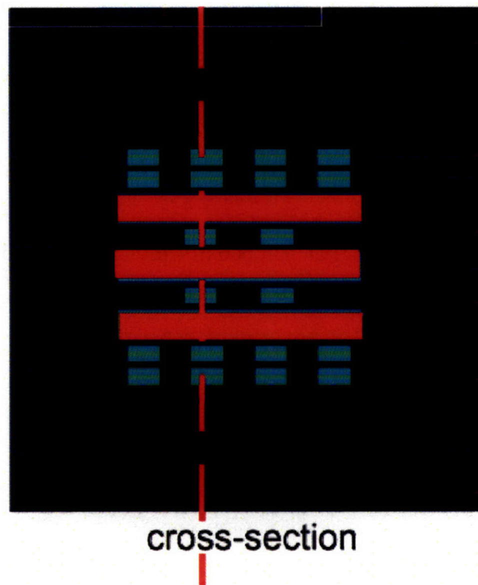
- A 10 μm-thick ILD versus our 2 μm
- A 1.4 μm-thick Cu diffuser versus the course of action for the past 10 years has been to push scaling to its limits. The benefits of high devour 0.6 μm
- A pre-set temperature of 320 °C at the lower tier Si-ILD interface to simulate a power dissipation of

0.615 W/mm^2 on the lower device layer, and then they added another 0.615 W/mm^2 of input power on the top tier surface and let the FEM simulation determine the steady-state maximum temperature. In our simulations, the lower tier Si-ILD temperature of around 80 °C was simulated based on a power dissipation of 100 mW from each SOI resistor located in different (x,y,z) coordinates and a Dirichlet condition of T_{ref} at the bottom surface of our 650 μm-thick Si substrate

- Their adiabatic boundaries were macro-scaled (more than 500 mm^2 of chip area), whereas our adiabatic boundaries were micro-scaled (6.25x10^{-2} mm^2) because we wanted to closely examine how layout can influence heat transfer in more detail.

In the end, it is very difficult to compare our simulation results to theirs (or to most other 3-D thermal papers, for that matter) unless all FEM parameters were *fully* reported and were *scalable* from one model to another. Therefore, while our thermal model may not necessarily follow the trend of others, the fact that we were able to show and match an approximate 16-20 % decrease in local temperature using Cu thermal vias should give some credence to our model's validity.





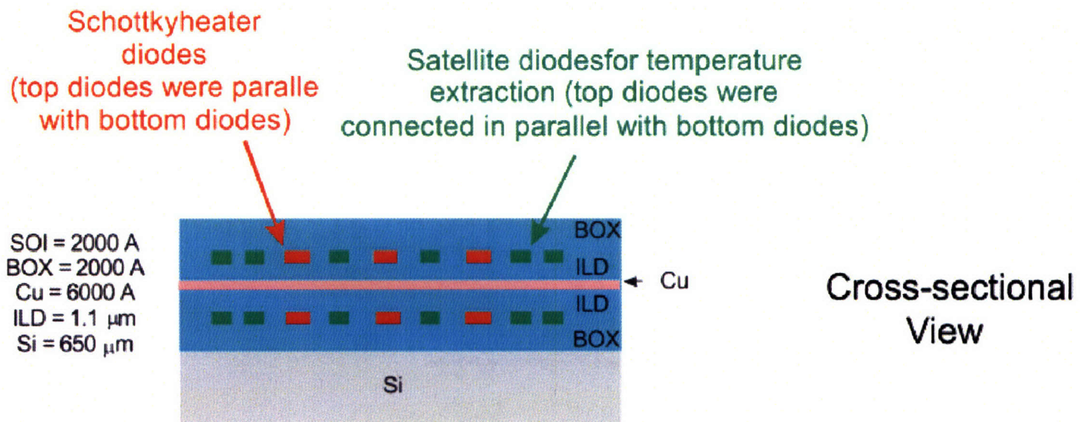


Figure 5-10: (a) Schematic of Face-face bonded Triple SOI heaters. There are 3 *pairs* of current-driven Schottky heaters (3 diodes on top tier, 3 unseen diodes stacked underneath) wired in parallel, and each individual satellite diodes in the schematic are also stacked top/bottom tier pairs. (b) A layout of the reference structure to be tested. (c) A cross-sectional view of the face-face bonded reference structure

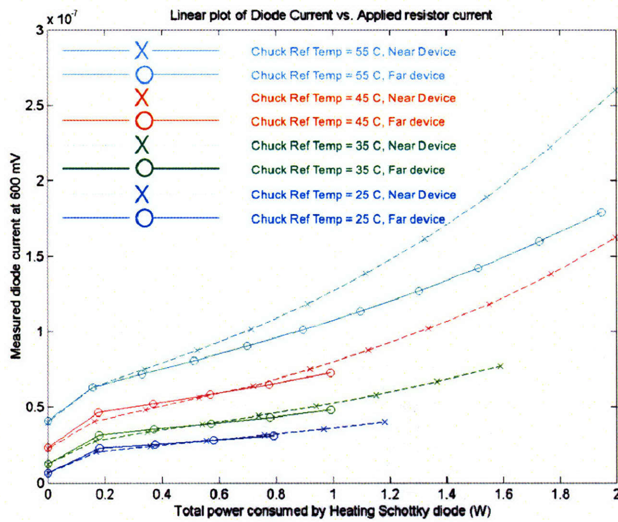


Figure 5-11: Measured satellite diode current ("Near" or "Far") at 600 mV plotted as function of the Schottky heater power dissipation at different chuck temperatures

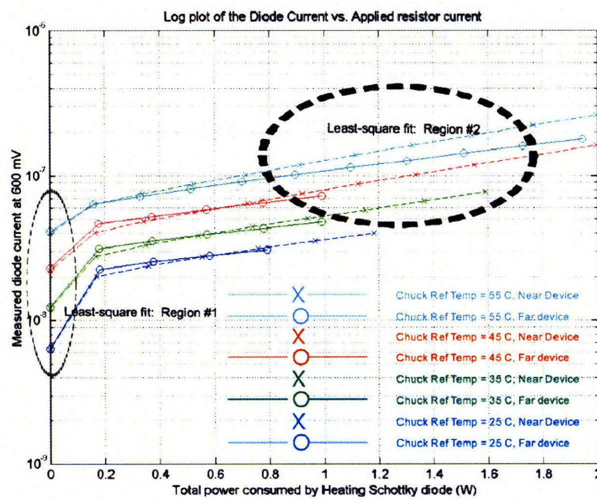


Figure 5-12: Same plot as Figure 5-11, but now in semilog y-axis.

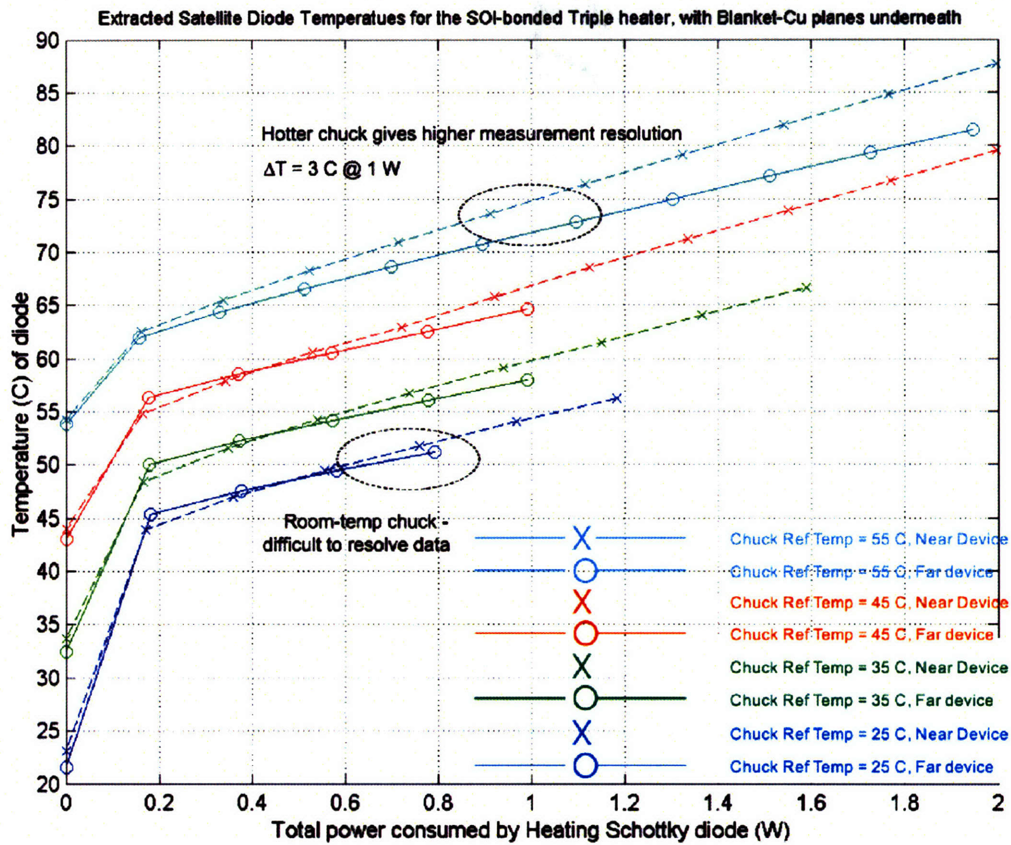
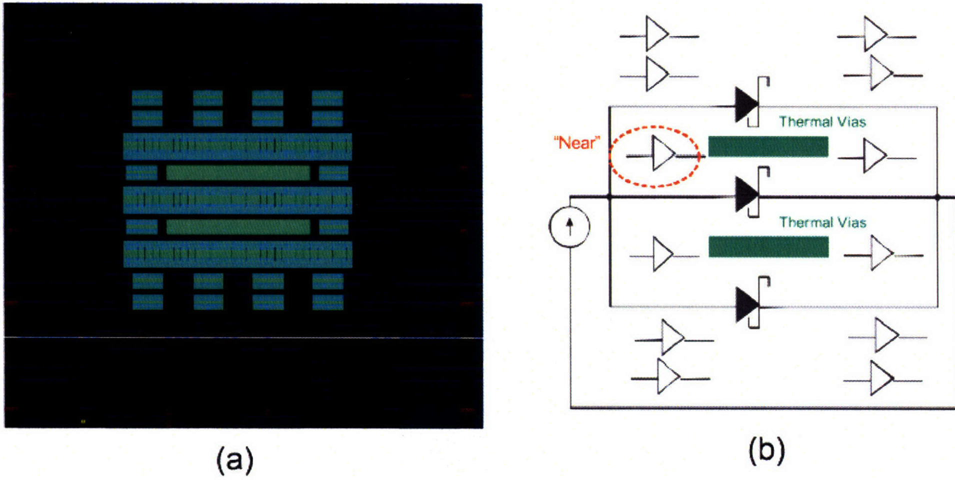


Figure 5-13: Extracted temperatures from data collected in Figure 5-11.



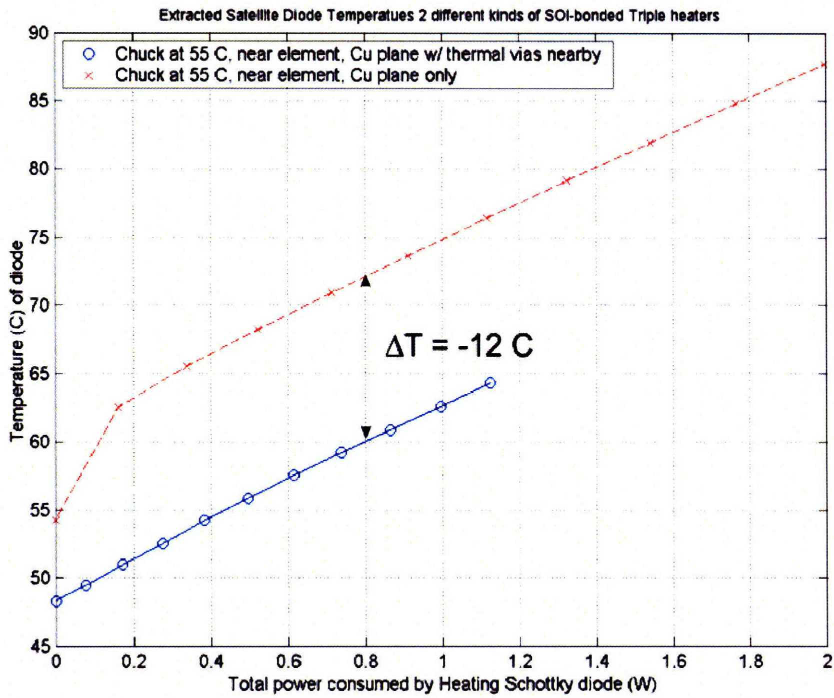


Figure 5-14: Reference Structure Measurement: (a) Face-face bonded Triple SOI heaters layout with thermal vias sandwiched between the heating elements. (b) The tempearture of the "Near" was measured and the results were compared (c) with the non-via analogue measured in Section 5.4.1. The addition of a Cu thermal via decreased the regional temperature by 12 °C.

Chapter 6

RF Sprial Inductor Integration and Magnetic Shielding Studies

6.1 Introduction

In the age of modern telecommunication, no battle have been more fierce than the evolution of the cell phone, because whichever telecommunication corporation that can back up their claim, "We produce the smallest, most powerful, and most energy-efficient cell phone" rules the wireless world. As the miniturization of cell phones continues, the RF components on and off-chip inside the casing also has to scale accordingly, both in size and in power efficiency. In terms of size reduction, one of the most stubborn passive component on the RF chip that resists dimensional scaling is the monolithic RF spiral inductor. In modern RF circuits, spiral inductors are routinely used in RF tank circuits in voltage-controlled oscillators (VCO) and as impedance-matching components in both low-noise amplfiers (LNA) and passive filters [35, 36]. Since the typical inductance used in these applications are usually around 1 - 20 nH, the size of a spiral can vary from a 3 nH, 5-loop, 150 μm x 150 μm inductor, to a giant 20 nH, 8-loop, 530 μm x 530 μm inductor [37]. The bulkiness in their lateral dimensions stems from the fact that magnetic coupling between *multiple* loops are needed to create an appreciable amount of inductance, and winding arms of opposing currents needs to be separated with an adequate distance (i.e. the air core region in the center of the spiral) to reduce negative coupling within the inductor. Moreover, the metal trace widths needs to be wide enough to decrease series resistance, [1] and at the same time, there is a finite lower limit for the wire pitch because the internal parasitic capacitance could degrade device performance [38, 39, 40] The shear size of spiral inductors can be seen from a sample RF chip photo in Figure 6-1, where the 5 inductors cover at least 15 % of the left cell's area.

[1] The overall inductance, however, is not a strong function of metal width because magnetic coupling depends on the pitch between wires.

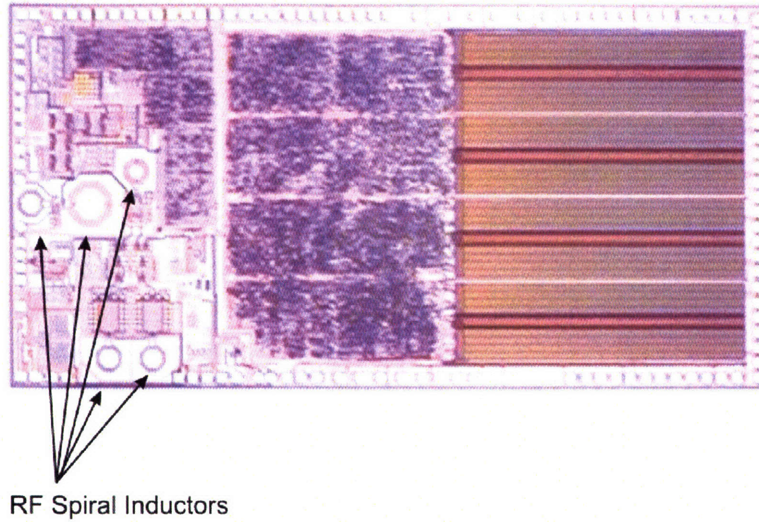


Figure 6-1: Sample RF chip photo, showing five real estate-consuming RF spiral inductors. Photo taken from http://www.techonline.com

Furthermore, in addition to bulky lateral dimensions, the vertical dimension constraints of spiral inductors result in significant Si real estate occupation. Although spiral inductors are usually fabricated on the two uppermost interconnect layers of a chip, RF layout rules prohibit device cell placement underneath any spirals because the substrate eddy currents induced by the inductor's magnetic field can be deleterious to any underlying circuits. Therefore, when one places an inductor on a layout, a substantial chunk of valuable Si area (and volume) is usually sacrificed just for the sake of temporary magnetic energy storage. Taking into account all the spatial constraints associated with monolithic spiral inductor, could one invent a more efficent way to integrate these passives onto a chip ? A passive-on-active 3-D stack immediately comes to mind, and various research groups have approached this problem not out of area concerns, but their initial goal was to increase the Q of the inductor.

Although many articles have been written about the use of patterned ground shields (PGS) to reduce the eddy current losses due to Lenz's Law, [35, 41, 42], the Q of a monolithic RF spirals implemented on Si rarely exceeds 10, and this is a major limiting factor in certain RF applicaitons, such as maximizing the transducer power gain (G_t)in current bipolar LNA design [36]. To push the inductor's Q to over 10, radical innovations such as double-level spirals operating in differetial mode [43], free-standing MEMS meandering inductors [44], suspended MEMS spiral bridges, [45], 3-D air core solenoids [46], spirals built on silicon-on-sapphire substrates [47], and much more are being persued. With the exception of the double-level spiral, all other aforementioned alternative inductors will require some form of integration to go along with an existing RF bipolar or RF-LDMOS circuitry [48]. Natually, 3-D can be the answer to these integration challenges, where a high-Q inductor can be fabricated separately from the main circuit and the two could then be integrated using Cu-Cu wafer or die-level bonding. Even then, the main circuitry still has to be magnetically

decoupled from the inductor or spurrious eddy currents can wreack havoc on the entire RF system. Thus, the last chapter of this thesis work will focus on finding an IC-compatible magnetic shielding configuration that can potentially be used in both 2-D and 3-D applications - with hopes that one can someday integrate magnetically-active components on top of existing circuity with ease *and* be able to save layout area on the main chip. The material of choice for the EMI (electromagnetic interference) shield is cobalt (Co), and it was chosen due its high permeability (μ_r(initial) = 110, μ_r(max) = 600) [49] and its acceptance as a CMOS-compatible metal.

6.2 Cobalt Magnetic Shielding Measurements

6.2.1 Reference Stucture: Al Spiral, No Shielding

The fabrication and testing of cobalt magnetic shielding reported below are for detecting RF substrate crosstalk between neighboring inductors. The base structure to be simulated was a 4.5 turn, 2.6 nH inductor (approximate value taken from [37]) with these flowing characteristics:

- Spiral lateral dimensions: 270 μm x 270 μm
- Spiral wire width = 10 μm, height = 1 μm, material = Al
- Spiral winding pitch = 10 μm
- Spiral underpass via depth = 1 μm
- Underpass metal height = 1 μm

The inductor was fabricated on a 650 μm Si subtrate with an 0.5 μm of top surface thermal oxide. The main crosstalk test structure consists of two side-by-side inductors with their GSG (ground-signal-ground) terminals facing opposite directions to minimize any air flux linkage of the two signal pads during RF test. The layout of the reference structure with no Co shielding is shown in Figure 6-2. Upon SOLT (short-open-load-thru) calibration, an input RF power of +4 dBm (2.51 mW) was swept from 500 MHz to 20 GHz at the GSG terminal of one inductor, and any power transfer to the neighboring inductor through substrate crosstalk was observed by measuring the value of |S21| at the other GSG probe. In this setup, the dynamic range of the power measurements was already at its maximum because the PNA-L's RF generators became unbalanced if one attemps to increase the output power beyone +4 dBm.

Since the load-side terminal of each inductors was shorted to ground, we expect to see a high |S11| and |S22| because of full reflections off of the loads, and if there were any substrate crosstalk, one would see some |S21| or |S12| signals, although they should remain low because the Si substrate was only doped to 10^{15} cm^{-3}. We also expect that as the RF input frequency increases, the substrate crosstalk will grow because the capacitive coupling through the 5000 A thermal oxide becomes a dominating factor over the

inductance. Even though the inductor is a linear and symmetric element (Port 1 response should be identical if applied to Port 2), the layout of a spiral inductor is inherently asymmetrical due to the Al underpass. Therefore, we ought to see some parity in the reflected powers of |S11| and |S22| because power reflections are directly related to the position of impedance mismatches. Specifically, there will be more capactive coupling among the substrate, underpass, and the main spiral at Port 2's entrance instead of Port 1. However, this asymmetry should not affect the thru values S12 and S21, though, because power transfer for thru-waves are independent of the position of impedance changes [2].

In summary, if one gazes at the measured results in Figure 6-2 and 6-3, we see all of the expected results mentioned above.

[2]This statement is true here because the total spiral length of 3.7 mm was still only at $\lambda/2$, and the skin depth for the 1μm-thick Al spirals is about 0.58 μm at 20 GHz (λ = 7.59 mm with ϵ_r = 3.9 for SiO_2). Therefore, propagation loss within the metal and the dielectric was negligible. For the case when length $>>$ λ, as in the case for long microstrip lines operating at 50 GHz or more, then this statement would be false because ohmic loss due to both length and skin effect during wave propagation will no longer be negligible

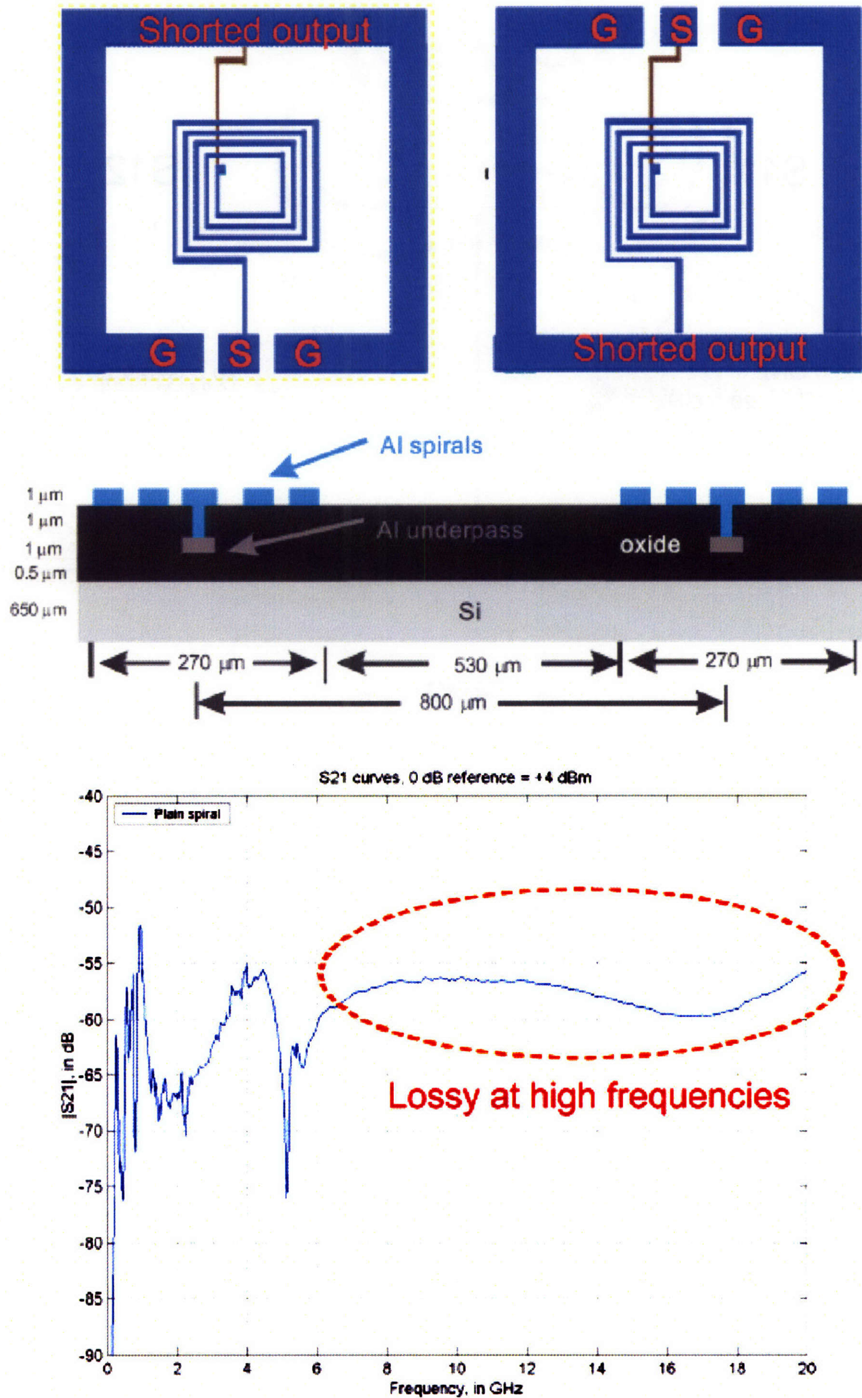


Figure 6-2: S21 measurement of reference Si crosstalk structure. Notice the substrate crosstalk increases with frequency because of capacitive coupling.

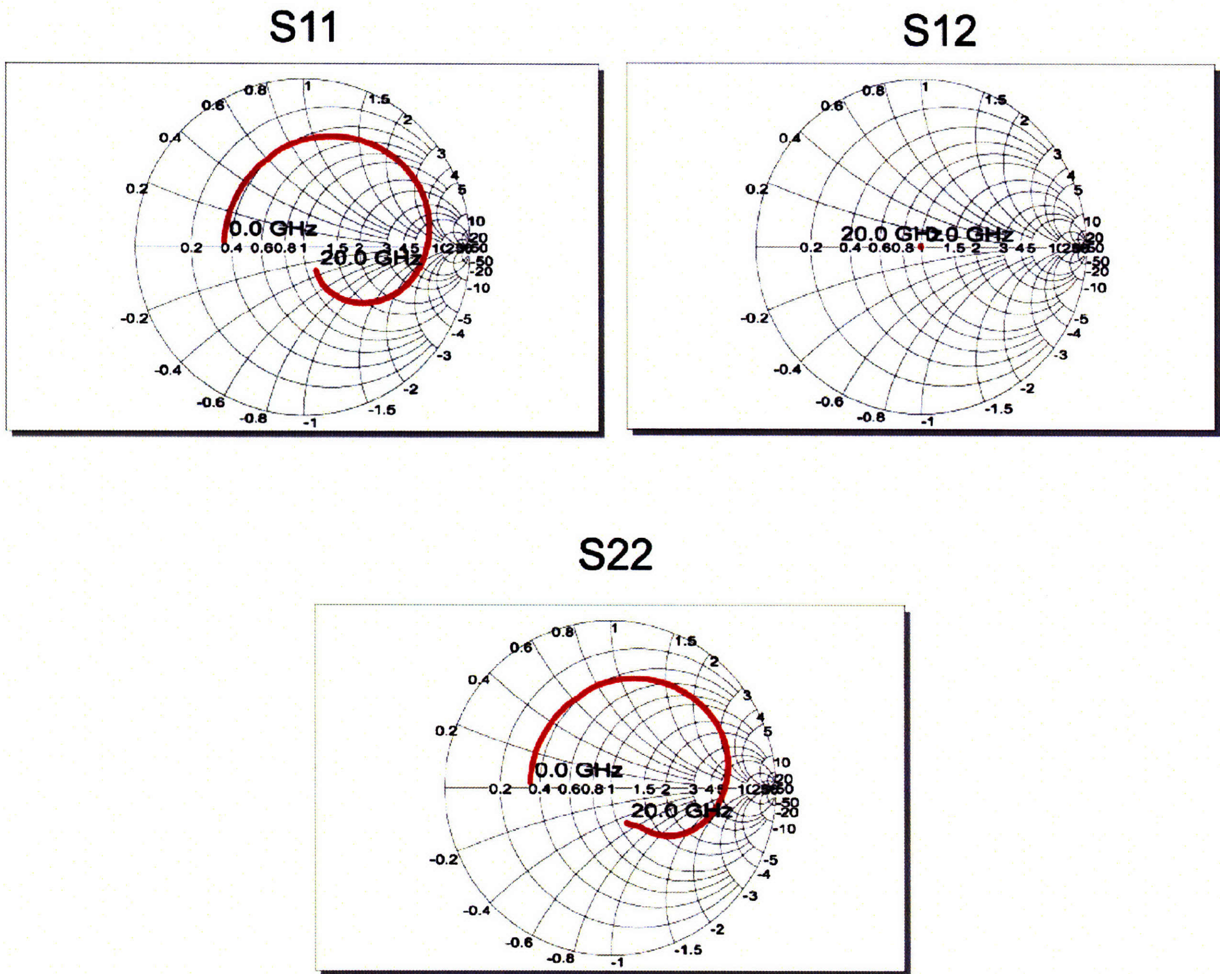


Figure 6-3: S11, S21, and S22 measurement of reference structure in Smith Chart form. Notice that the inductor ports were slightly asymmetric because Port 2's entrance geometry was very different than Port 1.

6.2.2 Comparison of Refence Structure to Solid Co Shields

Next, the effects of a rudimentary solid Co shield was tested by adding a 4000 A sputtered Co plane in between the spiral and the Si subtrate, shown as the green layer in the layout within Figure 6-4 [3]. With a relative permeability that varies between 100 to 600, one would expect the Co film to block most B-field-induced substrate currents in lieu of Lenz's Law; and instead of power dissipation occuring in the substrate, the image current will dissipate onto the top surface of the Co shield with a skin depth of around (taking μ_r = 350 as an average, and σ_{co} = 1.66x10^7 S/m) [50]:

$$\begin{aligned} \delta &= \frac{1}{\alpha} \\ &= \sqrt{\pi f \sigma \mu} \\ &= 93.2nm \end{aligned} \tag{6.1}$$

at 5 GHz. Comparing this skin depth with our 400 nm-thick Co film, one can be confident that these Co shields should provide ample magnetic isolation between both spirals and the substrate. Indeed, the |S21| thru-measurements in Figure 6-4 show that the Co shields, compared with respect to the aformentioned reference structure (fabricated on the same wafer), provided an isolation of around -10 dB at frequencies lower than 5 GHz. As input frequency increases past 6 GHz, the isolation improves steadily until it peaks at 13 GHz with a -24 dB of isolation. Even though the isolation diminishes as the frequency ramps up above 13 GHz, a respectable isolation of around -10 dB still exists at 20 GHz. This gradual degradation of isolation exists in any shielding materials at frequencies of 20 GHz (and above) because induced currents within the shield are unable to respond to the switching speed of external magnetic fields. Nevertheless, in common RF applications where the frequency of operation seldomly passes 20 GHz, these Co shields seem to be a promising tool for on-chip EMI protection.

Finally, an important item to note in the |S21| plot of Figure 6-4 is the common dip of the isolation curve at 5 GHz for both the reference and the solid cobalt shield structures. This dip is probably due to an overall structural resonance where (most probably) a parasitic capacitance was offset by the inductance of both inductors at that frequency, thus temporarily deleting the substrate crosstalk path and hence contributing to the observed increase in isolation.

[3] The choice of 4000 A Co was based on numerous experimental failues, in which Co layers thicker than 0.4 μm exhibited too much stress and will generally peel off even with an adhesion layer present. In fact, thicker Co will be a waste of metal anyways because the skin depth at frequencies higher than 5 GHz will be on the order of 100 nm or less.

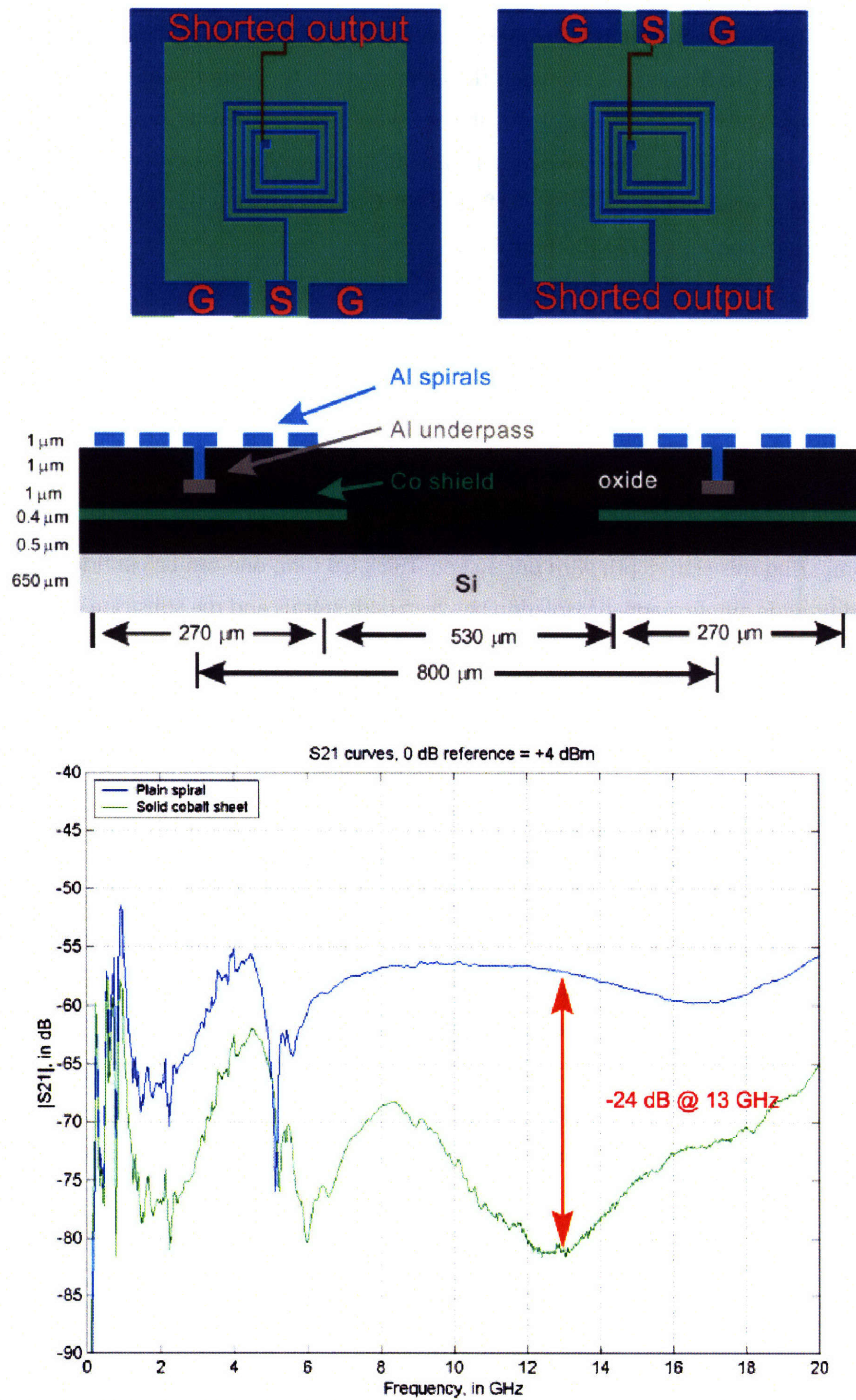
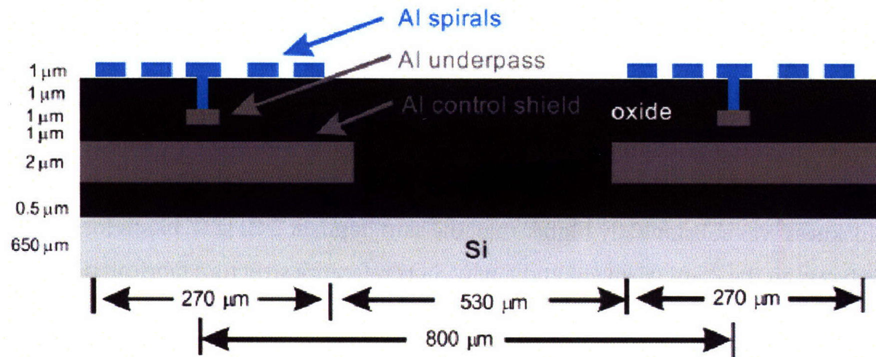


Figure 6-4: S21 measurement of reference Si and of a solid cobalt shield. The blue curve represents the reference crosstalk structure with no shields, and the green curve represnts the attenuation of crosstalk with a solid Co magnetic shield.

6.2.3 Comparison of Solid Co Shielding to Solid Al Shielding

Since we have established that Co shieldings do in fact work, one can now compare the effectiveness of Co shields to shields made from other materials. By default, since the RF sprials themselves were made of Al, a 2 μm layer of Al was used in place of the 400 nm Co as the new shield. The Al shield thickness was made to be 5 times the thickness of 400 nm Co shields because we would like to match the situation where the metal thickness was substantially higher than the skin depth at 5 GHz (1.16 μm for Al). Comparing the |S21| values between the 2 μm Al shield and a new SiO_2 reference structure fabricated on that same wafer, the isolation provided by the Al shield was surprisingly low. As seen from Figure 6-5, even though the thick Al shield offered very good isolation below 4 GHz, at frequencies above 6 GHz, the isolation effects became spurrious at best, with a mid-range frequency peak of only -8.75 dB and a high-range frequency peak of -17.5 dB. Therefore, it appears that when the shields under investigation were compared with the same reference structure built on their own respective substrates, the thin 400 nm Co magnetic shield outperforms the 2 μm Al ohmic shield at mid-range RF frequencies between 6 GHz to 20 GHz.

With regards to the measurements made in the Al shielded wafers in Figure 6-5, one should note that the reference structure's isolation was already about -10 dB better than the reference isolation in Figure 6-4. One could argue that we underestimated the Al shield's isolation because the |S21| measurements could be approaching the noise floor of the setup. This notion can be disproved by the fact that during SOLT calibrations, the "open" isolation values were on the order of -100 dB or more at low frequencies and about -85 dB at 20 GHz, and these calibrations were done with the same reference 0 dB level being the +4 dBm of input power. In all, the dynamic range of these measurements should have been more than adequate to detect small changes in power levels within structures of high thru-atteneuation.



(Al control Shield was solid with no cavities)

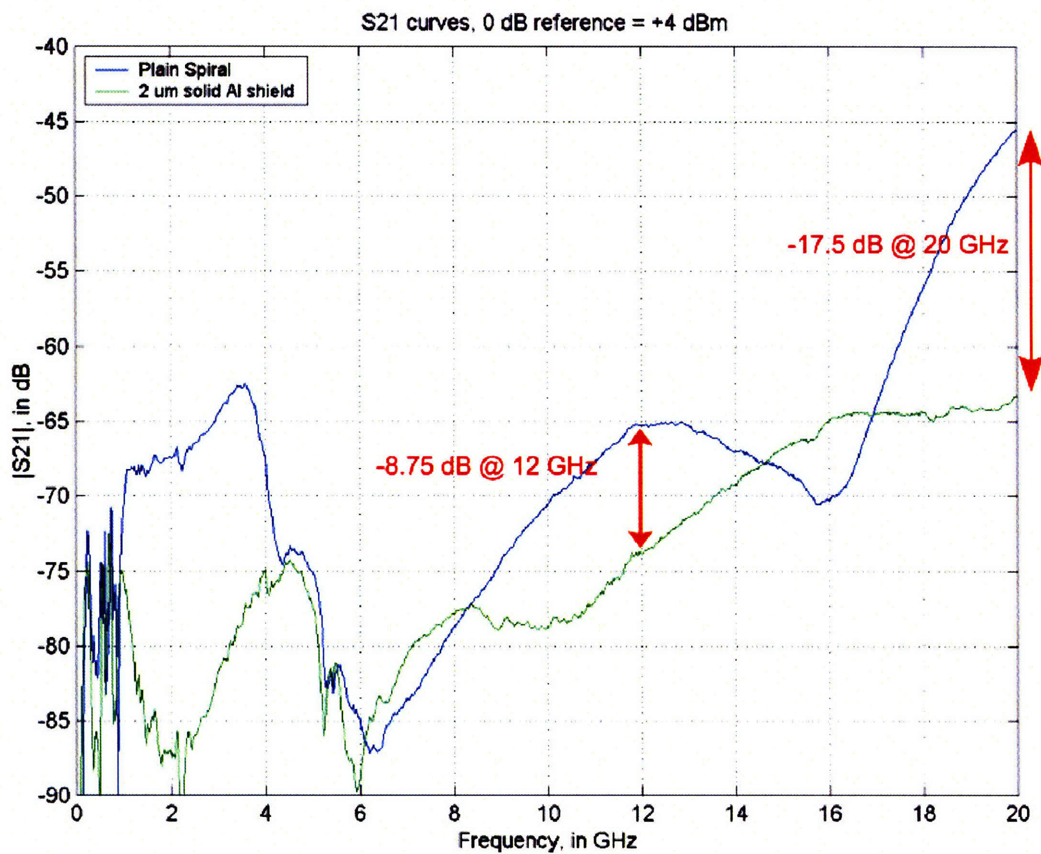


Figure 6-5: S21 measurement with new reference (blue trace) and a 2 um Al shield (green trace).

6.2.4 Comparison between Different Co Shield Configurations

Once we have established that the Co shields were even more effective than an Al shield 5 times as thick, one can now vary the Co shield configurations and see which shield geometrty will

- Have the best isolation performance
- Have the least power dissipation in order to maintain a good "Q."

There are three main groups of shields within the test matrix, and their layouts are shown in Figure 6-6. To begin with, we will take the data from the already-tested solid cobalt shield (which was electrically grounded) and the non-shielded reference structures and use them as the first and secondary standards for comparison, respectively. Next, since it is generally accepted that a patterned-ground-shield (PGS) can improve the Q of an inductor by both creating a low-resistance bypass of the substrate capacitance / resistance *and* lowering the power dissipation within the shield by creating discontinuties in the circularly-induced current, Co PGS of two different linewidths were included in the test matrix. Futhermore, according to research done by Chang, et. al [51], a distrbuted network of floating shields can further decrease the amount of power dissipation within the shield due to perimeter restriction in induced loop currents. By creating these floating shields, though, one pays a penalty of increasing the effective spiral-substrate capacitance. We also decided to include three types of distributed floating shields within our test matrix, each with different Co particle sizes.

Figure 6-7 shows that the solid cobalt shield gives the best magnetic shielding configuration almost within the entire 500 MHz - 20 GHz spectrum. This was expected because the shield did not contain cavities through which B-field penetration could occur. In descending order of isolation effectiveness, the 5 μm-wide and 10 μm-wide Co PGS's provided adequate isolation up to 12 GHz, but all of the distributed floating Co shields offered very little or no isolation in comparison to the non-shielded reference structure. Focusing our discussion on the Co PGS results first, even though magnetic fields penetrated these cavity-filled PGS's with relative ease, good shielding characterics were actually obtained at low-to-medium frequencies because the magnitude of the induced currents [4] was still large enough to generate a sizable opposing magnetic field. At higher frequencies, however, each leg of the Co PGS probably became very resistive due to the skin effect, and thus beyond 12 GHz, one can see that both Co PGS's did not function function as effective magnetic shields.

As the input frequency approaches 20 GHz, an anomaly occured where both Co-PGS samples somehow initiated a highly effective crosstalk pathway that propelled the |S21| values above the values of the reference structures. Taking into account the following passive parasitic in these devices:

- Capacitive coupling between top spiral - underpass and spiral - PGS
- Capacitive coupling between underpass - PGS

[4] Because of the perpendicular slots within the PGS, these current loops are now elongated ellipses that trace each leg of the PGS.

- Capacitive coupling between underpass - Si substrate
- Resistive coupling beteewn substrate underneath both inductors,

there were no indications as to why any of these parameters would suddenly change and provide a +16 dB increase in crosstalk when compared with the reference structure at 20 GHz. However, when one considers *active* RF processes, then an explanation for the increase in crosstalk could be provided. In short, since our spiral inductor lies only 3 μm above the PGS (most inductors fabricated in the industry have a spiral-PGS separation of at least twice that, if not more), the capacitve coupling can become so dominant such that the entire inductor-PGS structure looks like a transmission line dipole. And at the end of that dipole, RF radiation can occur from charge oscillations between the spiral and PGS terminals. In essence, our double-inductor crosstalk detector have evolved into two antennae [50, 52] that can transmit and receive RF energy by pathways independent of the substrate parastics. Again, playing devil's advocate, one can argue that this phenomenon should also be happening in the case of a solid cobalt shield, and indeed it could. However, since the metal-metal surface overlap between the spiral and a solid shield is much larger than the overlap between a spiral-PGS pair, the radiation efficiency in the spiral-solid shield structure could have been suppressed to nominal levels because the total power dissipated on the metal surfaces is higher. This scenario is very similar to that of the microstrip patch antennas, where the broadside transmission / reception efficiencies are inherently low because the metal surface overlap betwen the antenna patch and the ground plane consumes too much RF power [52].

Finally, turning our attention to the floating cobalt structures, these distributed shields did very little in isolating the substrate from the RF magnetic fields. The reason for this could be that the effective area for induced currents to flow was reduced so much that shielding failed because the induced B-field was not strong enough to counteract the forcing B-fields.

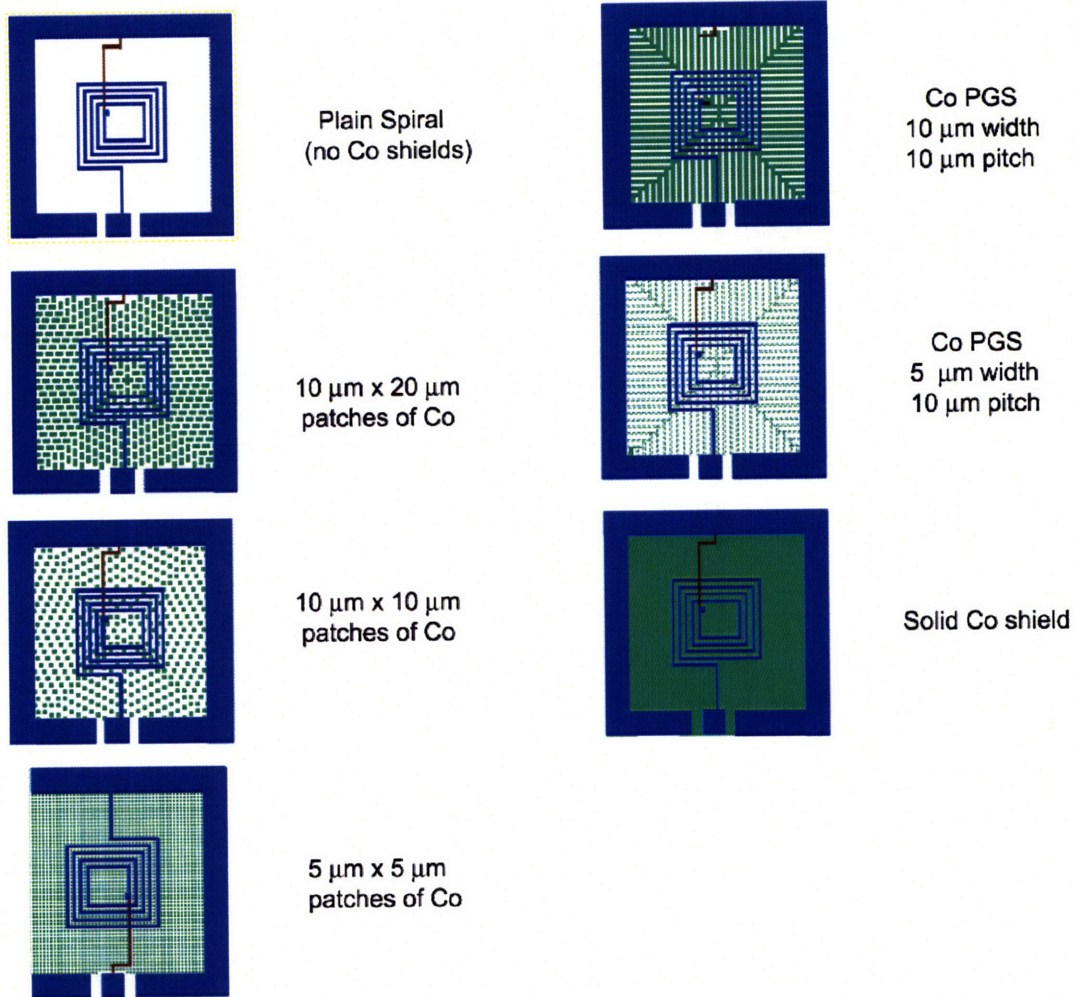


Figure 6-6: A schematic list of all the Co shield configurations tested. The rectangular and square "patched" Co shields were electrically floating, while in all other shield configurations the Co metal was shorted to the ground test pads. This list is in the same order as the legend in Figure 6-7.

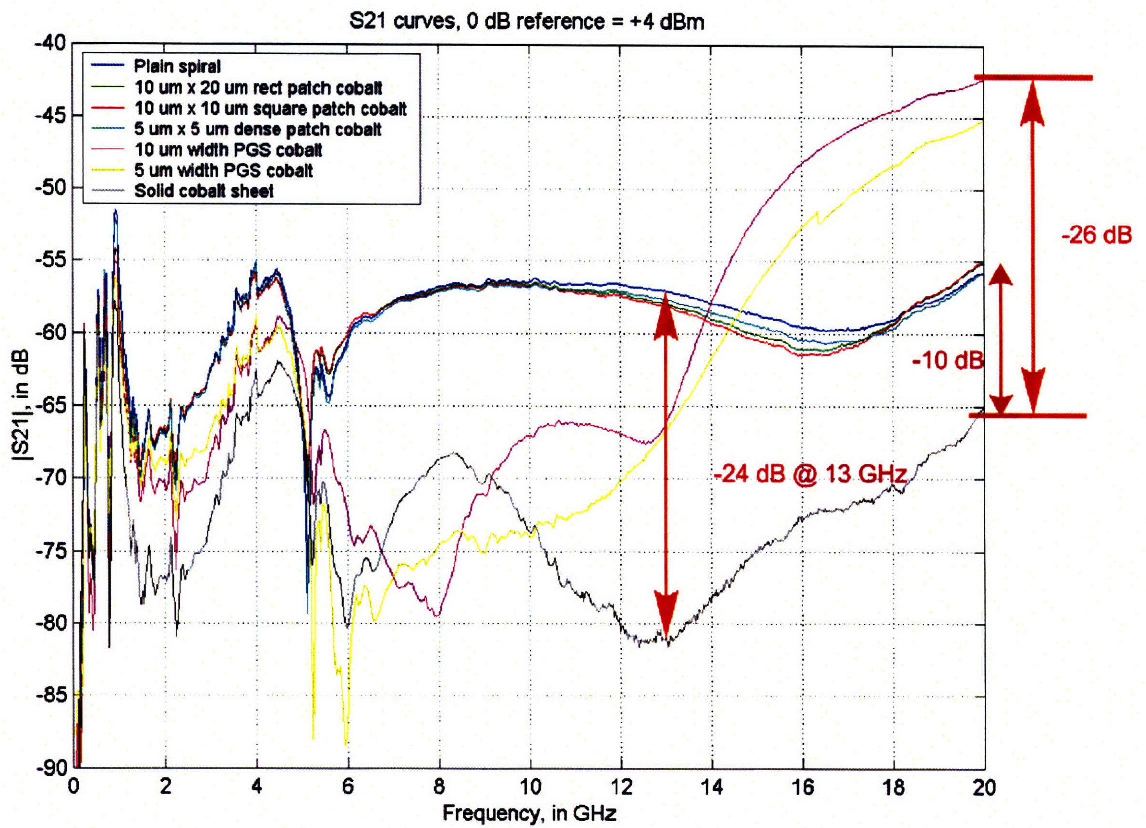


Figure 6-7: S21 measurement of reference Si and various configurations of cobalt shields. The solid cobalt shield again offered superior magnetic isolation when compared to all other Co shield configurations when considering the entire span of frequency range.

6.3 Summary of Inductor Results

In summary, we have shown that Co magnetic shielding do function as advertised, and when compared to a solid Al ohmic shield with 5 times the thickness, the solid Co shield was still the superior B-field isolator. From the surface, one could say that Co magnetic shields are indeed feasible tools for future applications such as providing EMI beween an integrated, high-Q MEMS-based RF components stacked on top of RF CMOS circuitry using 3-D. In reality, though, there are some remaining questions that needs to be tackled before any of these dream structurs could be realized. To start with, even though the theoretical permeability of Co is around 100 - 600, we were not certain of its exact value for our 400 nm sputtered Co films. Also, these Co wafers actually underwent an annealing step at 400 °C for 1 hr in forming gas (normal sintering recipe for MOSFETs) in attempt to improve the metallurgical junction quality between the Co and the Ti/Al contacts. We do not yet know if this also affected its magnetic properties, but on the first order, the effects from this anneal was probably minimal since the Curie temperature of solid Co metal is about 1115 °C, one of highest known in nature [53]. The most it could have done was probably to increase the overall grain size and the electrical conductivity of the film.

Furthermore, since Co is inherently magnetically hard (the coercive force H_c is very large), it may or may not be suitable for extremely high-frequency applications where the shield will be asked to respond to ultra-rapid changes in B-field. Conversely, this apparent deficiency may possibly become an advantage in some cases, for example, if one wishes to shield a low-frequency parasitic carrier signals from areas with high-frequency circuitry. Last, but not least, any usage of shields in conjunction with RF spiral inductors should involve some discussion on the shield's impact on the value of Q. With cobalt being more resistive than Al or Cu, the ohmic loss within the shield could be of major concern if one places the shield too close to the spiral inductor. This ohmic loss can be exacerbated if one uses a high-permeability material that decreases the skin depth, and hence increasing the effective RF resistance of the metal. Also, in terms of maximizing Q, a solid cobalt shield is worse than a PGS because the induced circular currents are not physically impeded by integrated perpendicular slots. With that in mind, Co magnetic shields could have great potentials as an EMI shield and area-saving gadget for RF appliacations if they are:

- with low - Q RF spiral inductors, similar to those used in matching the BJT's base impedance in LNA's
- with high -Q RF MEMS spiral or solenoid inductors, with 3-D integration as an enabling technology
- as an on-chip low-frequency or DC B-field shield

Chapter 7

Concluding Remarks

7.1 Summary of Accomplishments and General Conclusions

In conclusion, the salient features of this thesis work encompass the following:

1. Demonstrating the feasiblity of wafer-level 3-D integation by utilizing Cu-Cu bonding with a face-face configuration
2. Demonstrating the feasiblity of die-level, mutli-layer 3-D integation by utilizing Cu-Cu bonding in both face-face and face-back configurations
3. Successful implemenation of two varieties of 3-D CMOS ring oscillators, thereby demonstrating the feasiblity of the Cu 3-D process flow
4. Thermal modeling and rudimentary experimentation on utilizing Cu thermal planes and vias as heat-managing components in 3-D integrated circuits
5. Experimentation on utilizing thin-film cobalt as an inter-layer magnetic shield in RF 3-D integration.

While 3-D is conceptually an enticing technology for future microprocessors, the dicussions throughout this thesis work (expressed in the sole opinion of the author !) suggest that there are still major technological roadblocks that need to be resolved before 3-D becomes a viable technology. Fundamentally, however, this thesis work has shown that Cu-Cu bonding, whether on a wafer-level or on a die-level, can be a feasible technology if one is not limited by equipment limitations. The success of the 43-stage CMOS ring oscillator demonstrates that if every on-chip component was working properly and if all fabrication processes flow according to plan *prior-to and during bonding*, then there are no theoretical reasons why Cu-Cu 3-D integration would physically ever fail. Thus, the choice of whether or not 3-D should in one's design toolbox will purely be application-driven. And if 3-D was determined to be benefical for the application under consideration, then it is in the author's opinion that 3-D integration with Cu bonding should be the definitive chioce over all other existing 3-D technologies.

"Why," one may ask. Consider the following: First of all, by using thermal FEM simulations and a subsequent quick test using SOI Schottky heaters, and combining the successful implementation of the 3-D ring oscillator, we have shown that Cu 3-D integration has an inherent advantage over all existing 3-D techniques (oxide-based, polymer-based, TFT-based) because one receives a free metal layer that can be utilized :

- As an free interconnect layer that can be used in routing, as demonstrated by the Cu-Cu input-output local interconnects between neighboring inverters within the CMOS ring oscillators
- As a psedudo-backgate for Vt - adjustment in 3-D SOI devices, which was also demonstrated by the ring oscillators
- As a RF or DC ground or Vdd planes
- As a an effective heat flux diffuser

Furthermore, if the face-back MTI 3-D flow was used, it can offer the highest level of 3-D integration when the metric is the vertical via density [1] This is because a shallower inter-level via, created after a handle-bond / substrate etchback combination *prior* to bonding, will inherently have a smaller aspect ratio than a deeper inter-level via created *after* the completion of of a 3-D bond [2]. Hence, a smaller aspect-ratio via's width can be scaled more aggressively, and when combined with an an accurate wafer or die aligner with sub-micron registration accuracy, one can pack more vertical vias per unit area with our process flow than any other integration schemes.

Finally, Cu-Cu 3-D integration is superior to that of polymer or oxide-based integration because of the ease in which Cu films can bond to each other. Though in-situ sputter cleans can improve the theoretical quality of the Cu-Cu bond, the neglible Cu-Cu series resistance extracted from our face-face bonded MOS-FETs showed that barring from equipment-related limitations, a well-bonded Cu-Cu interface can be made without any special surface cleaning treatments. Combined this with the fact that Cu deposition and film adhesion can be easily be performed on many different types of substrates (the same cannot be said for either oxide nor polymers), heterogeneous integration between devices fabricated from different technologies can be realized without too much material compatibility issues. Last, but not least, if one chooses to perform hybrid MEMS or CMOS RF components onto pre-existing BJT or CMOS circuitry, the thesis work have also shown that cobalt magnetic shielding can be used an effective inter-layer EMI shield.

[1] Theoretically, TFT epitaxy-based 3-D circuit would have the highest inter-level via density becasue direct contact between source /drains from different level can be made with a front-end process. However, such structures have yet to be integrated into any kind of real circuitry, like those of a ring oscillator.

[2] such as the case in both polymer and oxide-based bonding.

7.2 Future Work

7.2.1 Handle Wafer Release Optimization

From experimentations involving wafer-level 3-D integration, it was determined that the limiting factor in the process flow was the efficacy and effiency of the handle wafer release. The proposed method in this thesis work - utilizing acid-encroachment to destroy the handle-SOI lamniate structure, was obviously sub-optimal in terms of efficiency, but above all, it just didn't work very well. If one was to approach the handle wafer release problem using chemical corrosion of interlaced films as a basis for the solution, then it is in the author's opinion (and the opinion of everyone that've been investigating this problem) that integrated microfludic channels must be used to facilitate the release process. By drilling holes through the handle substrate or creating pre-existing micro-channel grooves within the handle-SOI laminate structure, one can enhance mass transport of the releasing reagent and speed up the overall erosion of the desired layers between the handle wafer and its bonded SOI counterpart.

The choice of the releasing reagent, whether it be an aqueous acid / base solution or an acid / base vapor (HF vapor, XeF_2 gas, etc.), depends on the material selection of the to-be-eroded film. Even though a thick layer of Al was chosen as the release layer in this thesis work, it is easily conceivable that one can use SiO_2, polysilicon, or other metallic or organic film systems as the release layer of choice. However, most of these proposed materials often resists corrosion unless HF or TMAH was used, and this can be a problem because most ILD materials (like oxide, BCB, etc.) or Si-based materials (the Si substrate itself, gate-poly, etc.) will also quickly succomb to aqueous / vapor-phase fluoride or aquesous hydroxide attack. Therefore, the solution to this problem cannot be solved overnight, and most likely it'll require a comprehensive chemical corrosion analysis before one can optimize the choices between:

- The designated material to be corroded
- The corrosion reagent and its compatiblity with pre-existing films on chip
- Additional on-chip physical structures required to facilitate proper and timely corrosion

On the other hand, one does not have to limit him/herself in a corrosion-based release process. Wafer delamination methods such as SmartCut, or sublimation films such as poly-formaldehyde (in Prof. Karen Gleason's lab), laser ablation of other polymer films [26], or any other un-perceived methods can also be used as a release mechanism. Or, if one chooses *not to release the handle wafer at all*, then one can always resort to the old-fashioned method of Si grindback / TMAH etchback of the handle wafer. This is probably possible if the 3-D stack was no more than 3 device-layer thick. If more than 3 or 4 handle grindback / etchbacks were performed on a growing 3-D stack (as in the case of a face-face bonded seed stack), then the overall structural integrity of the base Si substrate will be an issue.

7.2.2 Die-Die Bonding and Alignment: Equipment Optimization

In Chapter 3, we have examined how one can get around some of the equipment-related issues in die-die bonding by building home-made jigs that improved the quality of the Cu-Cu die bond. However, in order to maximize the potential of die-die bonding, one needs a reliable setup that can provide a good die-die alignment accuracy. In short, it is suffice to say that one cannot guarantee a post-bonding alignment accuracy on the order of 0.5 μm between two substrates without some kind of real-time verification of the alignment prior to bonding. The easiest method to realize this capability is to acquire an IR aligner with a complete set of IR-transparent bond chucks. Thus, one would have definitive proof that the substrates are indeed aligned prior to both substate clamping and substrate bonding.

Furthermore, a good quality die-die bonding setup should also have the capability to bond and align small dies with dimensions around 5 x 5 mm^2. **For an university research environment, this could be of great help** because in most graduate circuit design projects, the tape-out chip size from TMSC, UMC, or National are usually on the order of 0.6 x 0.6 cm^2. If an university fab can obtain a reliable die aligner / bonder that can handle these chip sizes, then one can create great synergy between the circuit design and fab research groups - and in the long run, it will hopefully energize and propel microelectronics research in a university to the next level. In my opinion, this can be especially beneficial at MIT, where MTL have a world-class faculty in both circuit design and in device physics.

Appendix A

The 3-D Process Flow: Detailed Explanations

In this section, we will go through each bullet depicted in Figures A-1 and A-2 slowly and describe the processes in detail.

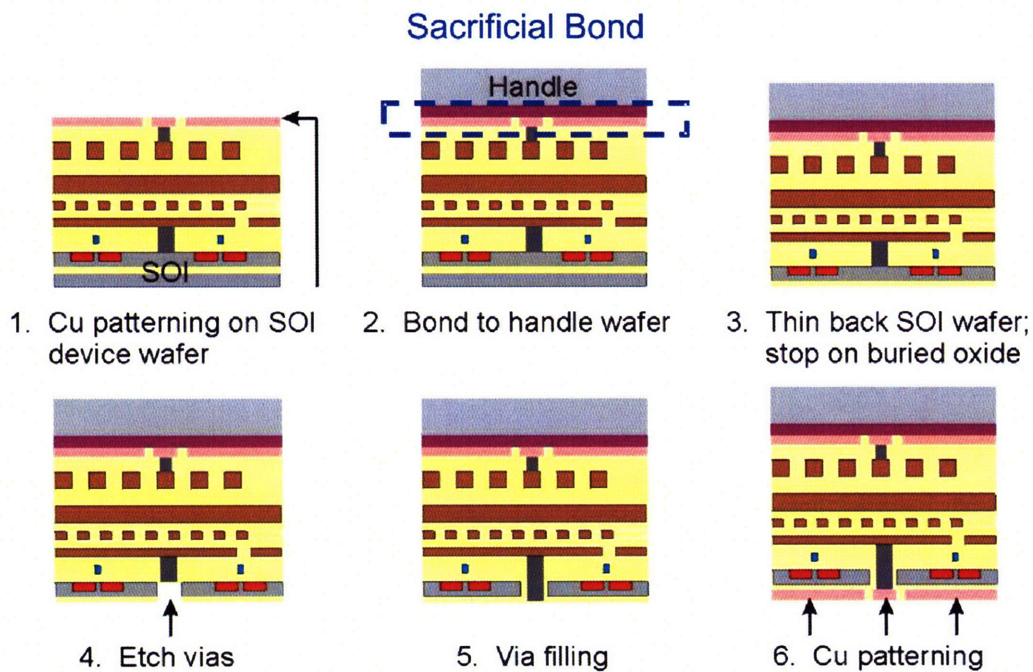


Figure A-1: MIT 3-D process flow, part 1

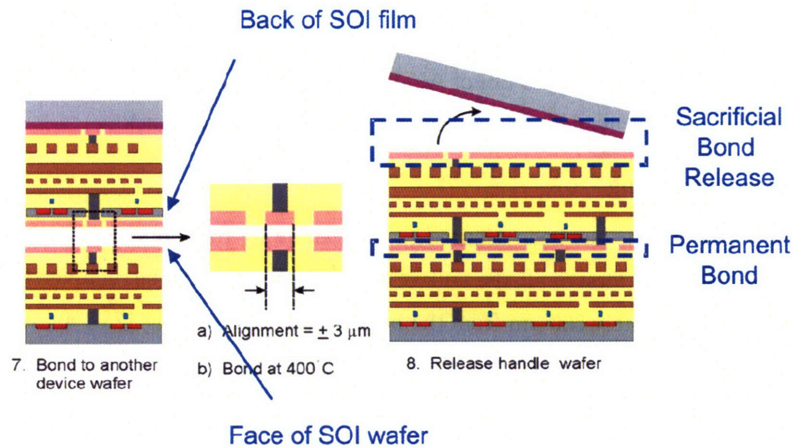


Figure A-2: MIT 3-D process flow, part 2

1. **Cu patterning on top surface**

   The first step, as denoted in Figure A-1, was to create the necessary Cu bonding pads for future bonding steps. Using standard lift-off techniques, a bilayer of 500 A Ta and 3000 A Cu was deposited by e-beam evaporation and then patterned. For the thesis device wafers, the SOI thickness was 2000 A and the buried oxide thickness (BOX) was also 2000 A. *Note: Step 1's Cu patterning was eliminated entirely in the thesis device wafers because we had no plans to bond a "Device Layer 3" onto our 3-D stack.*

2. **Handle wafer bonding**

   To prepare the device wafer for Si backside thinning, one must attach a support wafer to the non-grinding side of the device substrate in order to provide mecahnical stability during grindback. First, the entire surface of the SOI device substrate was passivated by a thick PECVD oxide (about 5 μm, and then the top surface has to be globally planarized by CMP. Since the underlying MOSFET / Metal-1 topography combines to about +/- 1.5 μm, a 5 μm oxide overlayer should provide an ample amount of sacrificial material for a smooth finish. After CMP, the target distance between the top surface of the top-most metal layer and the air-PECVD oxide interface should be about 1.5 μm.

   Next, a combination of layers dubbed the ***laminate structure*** (the term used in our patents, denoted by the purple layer in Figure A-1) has to be deposited on the bonding surface of the device and dummy handle wafers. The laminate structure serves two functions:

   - Facilitates a temporary bond between the SOI and the handle wafer that is stable enough to withstand mechanical grindback (shear stress), aqueous alkaline attack, and heat stress.
   - Facilitates a pre-determined weak point such that when attacked by a selective reagent other than the aforementioned stress factors, the laminate structure will be destroyed.

Foreshadowing to material discussed later in the thesis, the current laminate structure was composed of a penta-layer sandwich of 20 μm Al, a bilayer 500 / 3000 A of Ta/Cu bonded to a mirror-image bilayer of Cu/Ta, and yet another 20 μm of Al. The thick Al layers were completely covered by the Ta/Cu bilayers for protection against subsequent aqueous hydroxide attack.

3. **Backside substrate thinning**

   Once the SOI device wafer has been secured to the handle by the laminate structure, backside Si removal can then proceed. First of all, the bulk of the 6" Si substrate can be removed by mechanically grinding away approximately 500 μm from the backside. The remaining 125-150 μm of the leftover Si can be removed by either wet or dry etching. The preferred wet etching method is to use a 1:1 volumetic mixture of wafer and 25% wt tetramethylammonium hydroxide (1.79 M TMAH final) at 90^oC, and the etch time varied from 3 hours to 4 hours depending on the water evaporation rate (higher hydroxide content yields a slower etch). Since this mixture has a 20,000:1 selectivity to SiO_2, the SOI wafer's BOX is a natural etch stop for TMAH. Moreover, a 25% wt potassium hydroxide solution (5.84 M KOH) at 80^oC will also suffice as a substitute, albeit its selectivity to oxide is only around 800:1.

   On the other hand, if one wishes not to use hot alkaline solutions for whatever reason, the remaining 125-150 μm of Si can also be removed using a non-passivating SF_6 recipe at 600 W forward RF power in an inductive-coupled plama (ICP) etcher. The blanket Si etch rate in MTL's STS etcher is around 2.5 μm per minute - quite fast indeed, but it has minimal, or if any, selectivity to oxide. If one wishes to achieve a full-stop on the BOX, you would have to switch to a etch/passivation multiplex mode on the STS. Only then can you achieve the 150:1 Si:oxide selectivity that the Bosch process guarantees, albeit with an overall slower Si etch rate.

   Last, but not least, one can choose to remove the Si by using XeF_2 gas - the *most* selective and the *softest* thin-film Si removal method available to date. Since it contains no ion bombardment and is a passive vapor etch based on mass transport only, XeF_2 gas theoretically has an infinite selectivity to SiO_2 and for almost all metals, provided that the water vapor content is minimized during etching. The penalty for such selectivity is its slow and vapor pressure-dependent etch rate.

4. **Backside via etch**

   After completing the backside substrate removal, inter-layer vias were then etched from the backside. The via positions were designed so that they lie directly within the LOCOS field oxide regions and were placed directly below the Metal-1 Al contact pads. Theoretically, these oxide vias should be plasma etched to retain its rectangular fidelity, but MTL does not have a 6" Au-contaminated plasma etcher that's plumbed with CF_4 or CHF_3. Instead, the thesis device wafers' inter-layer vias were wet etched in Silox Vapox III, which is a mixture of acetic acid (CH_3COOH) and hydrofluoric acid (HF) that has a limited selectivity between oxide and Al provided that the water content inside the

acid solution was kept to a minimum. Visual cue for etch completion is the roughened Al-Si surface created by acid corrosion during DI water rinse, and the resulting via profile is an top-overblown cylinder with a depth of approximately 1.1 μm and width of about 5 - 10 μm.

5. **Backside via Fill**

   The backside vias can be filled by a variety of methods, the most conformal being W CVD, intermediate conformality with sputtering, and the least conformal with e-beam evaporation. Cu sputtering was chosen to be the filling method because W CVD was unavailable and e-beam evaporation had a throughput problem. To be specific, since the vias were around 1 μm deep, an adhesion layer of 500 A Ti and the main 2 μm Cu film were sputtered in order to provide the usual 1.5 - 2x overfill height needed for Cu CMP. Also, since a 2 μm layer of Cu exhibits a very high tensile stress, waferbow compensation was done by sputtering a bilayer of 500 A Ti / 1 μm Cu on the other side of the handle wafer. Without it, the tensile waferbow is too much for the Cu CMP-head's vacuum ports to compensate. Finally, the filled inter-layer vias were damascened using Cu CMP to create a flush back surface. The quality of this damascene step can make or break the quality final Cu-Cu 3-D bond, and the unforseen importance of this was actually quite surprising. More on this topic later.

6. **Backside Cu pad patterning**

   Upon Cu CMP, the backside BOX surface needs to be prepared for the permanent Cu-Cu bond. In the thesis device wafers, the dimensions of the electrically-active Cu pads (those that sit directly on top of inter-layer vias) were maximized to avoid wafer-wafer alignment troubles. The most important dimension here, however, is the air-gap distance between the electrically active vs. inactive Cu pads. If the air moat was too narrow, then any significant wafer-wafer misalignment will cause a global electrical short across the entire wafer, but too big of a gap will decrease the bond strength of the overall structure. Thus, any electrically-active Cu pads in our wafers will have at least a 10 μm airgap moat surrounding it.

7. **Wafer-wafer alignment and the permanent Cu-Cu bond**

   After Step 6, the thinned-down SOI device layer will be ready for the final Cu-Cu bond. One would first take a desired base wafer from the production lot and pattern Cu bond pads that are mirror images of the ones from Step 6. Next, the base wafer was aligned to the SOI-handle complex inside a non-infrared optical aligner. Assuming a gross alignment tolerance of +/- 3 μm, the 10 μm air moat around each electrically-active Cu pads offer more than ample protection from mis-alignment shorts. Once aligned, the wafers were bonded by Cu-Cu thermocompression for 30 min at 400^oC, 10^{-3} torr vacuum, and with the piston down-force at 10,000 N for 6" wafers. Afterwards, a post-bonding at 400 ^oC for 30 min in an N_2 purged furnace gives the pair extra bonding strength by promoting larger grain growth.

8. **Handle wafer release**

Upon completion of the permanent Cu-Cu bond, the handle wafer needs to be released from the multi-layer structure if one wishes to bond additional layers on top. Handle-wafer release was performed by soaking the entire wafer stack in hydrocholoric (HCl) acid, in which it dissolving the Al cladding layers in the SOI-handle laminate and hastens the destruction of the entire laminate structure. Then, the handle wafer simply floats off from the 3-D structure, and further bonding can be continued since the the resulting substrate can be seen as our new "base wafer." In essence, once a two-layer base is built, a n-layer stack can be constructed by merely repeating Steps 7 and 8, but each time bonding with a different pre-made SOI-handle complex.

As one can imagine, there are serveral obstacles that need to be surmounted if this process flow were to be successful. Techniques and the accuracy of wafer-wafer alignment is a major one, as well as how exactly can one make a laminate structure that can withstand numerous stress patterns and yet be targeted for destruction by a single stress method. And what about the overall quality and uniformity of the Cu-Cu bond ? These are some of the question that this thesis will address in the following sections. Once we've introduced the pertinent variables associated with wafer-level bonding, we will revisit those same variables in the context of die-die bonding - with much less introductory material and more direct report of results.

Appendix B

Handle Wafer Release Mechanism

Once the laminate structure survives grindback, etchback, CMP, and the permanent Cu-Cu bonding steps, it will be ready for destruction with the application of our release reagent, the hot HCl. The basic concepts behind the wafer release mechanism are simple: One has to expose the edges of the wafer so that acid can start to attack the Al release layer. Then, if the Al release layer thickness was large enough to overcome surface tension effects, the hot acid solution can continue to encroach deeper and deeperinto the release layer, dissolving away the Al along the way and leaving all other structures intact; in essence, the hot acid solution behaves just like a drill at an oil well - the deeper it goes, the more difficult it is to drill because mass tranport and surface tension effects within the liquid column will become ever-growing problems. For more details about surface tension, interfacial phenomenon, and some miscellaneous experiements on liquid penetrance, I invite the reader to browse through some excellent literature on these subjects [54, 55, 56, 57, 58].

But for now, let's take a real brief look at what's happening during acid release. Figure B-1 represents the simplest perspective in the release process, and the story only has three parts: Chemical dissolution of the release layer, mechanical agitation to assist mass transport of debris, and finally the separation of the substrates.

Well, if only life can be that simple. The second-order effects that accompany Figure B-1 are the real show-stoppers in the effectiveness of our acid release scheme. With brevity, it will suffice to say that as the HCl solution enters deeper into the release cavity, it becomes more and more difficult to transport the by-products of Al corrosion, the H_2 bubbles, and the reactants (acids from the bulk solution) in or out of the liquid column. A cartoon of this can be seen in Figure B-2.

Why do we have a bubble traffic jam anyways? There are actually 3 second-order variables that can affect the system's mass transport characteristics.

1. **Ionic Strength of Solution:** A solution with high acid concentrations (or more accurately, a higher ionic strength) will generally increase the surface tension between the air-liquid inteface to a value

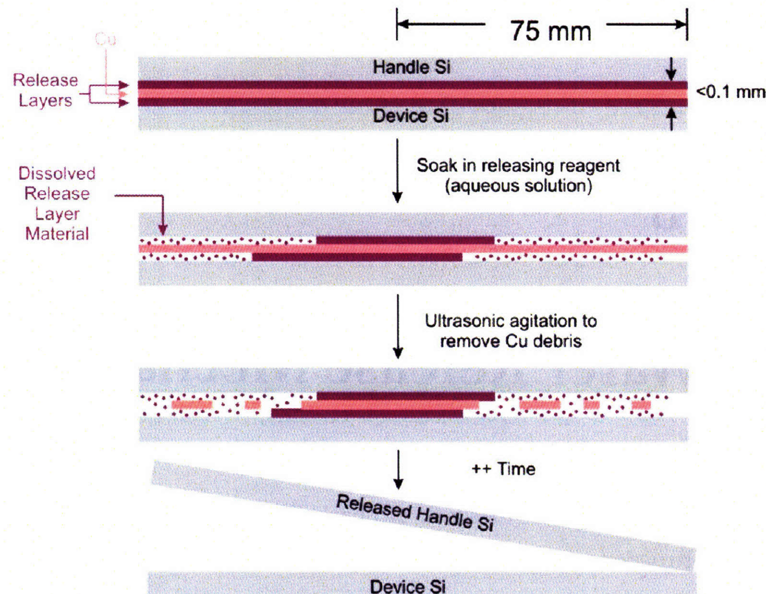


Figure B-1: Extremely simplified overview of the handle wafer release process: Dissolution, agitation, and separation

higher than the normal 72 mN/m value between pure water and air at 25 °C.

2. **Temperature:** In general, elevated temperatures will decrease the surface tension between a liquid-gas interface because the inter-molecular hydrogen bonds between water molecules are disrupted by an increase of their kinetic energy. An increase in temperature will also elevate the rate of Al dissolution in HCl, therefore also controlling the generation rate of H_2 bubbles and it's the **first knob in controlling the transient response of the acid encroachment rate.**

3. **Wetting Angle:** In addition, the local surface roughness and surface chemistry among the acid-Al-Ta-Cu interfaces can change the wetting angle at the liquid-gas-solid intersection points, thereby influencing the effects of surface tension by changing the direction of the tension vector.

The end results from the combination of these factors are the following: If the gas-liquid surface tension is too high and the capillary sidewall wetting is too low, a H_2 bubble can get "stuck" on the sidwall and cause partial blockage to the entire capillary. This will hinder the mass transport of liquid from the bulk solution to the reaction site located deep within the column -the place where HCl liquid front is dissolving more Al metal and creating more H_2 bubbles. Now, two things will then happen: These newly created H_2 bubbles from the reaction site have nowhere else to go but to coalesce with the old H_2 bubble blockage, thereby creating an even bigger bubble traffic jam. Also, the acid concentration at the reaction site will gradually decrease to a level where the Al corrosion rate becomes very slow. The final result is that the traffic jam will stop the acid encroachment altogether.

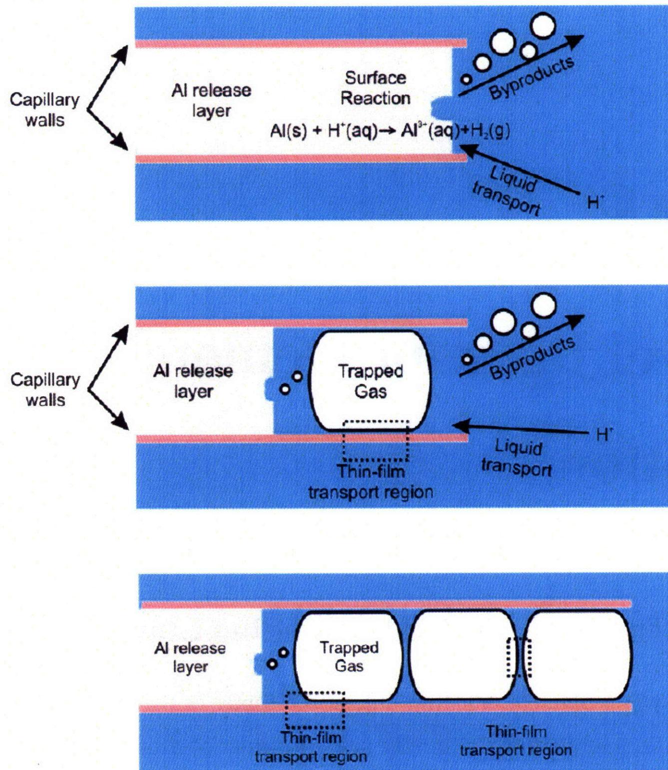


Figure B-2: An extremely simplified overview of the liquid penetrance problem: A traffic jam of trapped H_2 bubbles henders mass transport of acid and the bubbles to and from the capillary

If one really wants to complicate the situation, one can add surfactants to the acid solution in order to tailor the wetting angle and the surface tension within the entire system. It sounds like a great idea, but now we're introducing a tertiary system parameter - phase transformation of the bulk, which is directly related to the *viscosity* of the bulk solution [1]

And if one is keen in fluid mechanics, viscosity is directly related to shear force and fluid velocity; or in other words, a tertiary effect of changing surfactant concentration in acid soluions is to induce a phase transformation, which in turn induces a change in viscosity, which in turn serves as a **second knob in controlling the transient response of the acid encroachment rate** [2]

[1] In general, the phase of the bulk solution is an implicit function of two variables: Surfactant concentration (at phase-transformation boundaries associtaed with breakpoints such as the ***critical micelle concentration***, the ***Krafft eutectic point***, etc.) and temperature. Also, viscosity within a particular phase is usually constant, and only when a phase transformation occurs (due to temperature or surfactant concentration variations) do bulk viscosities change.

[2] It is actually very difficult to predict the exact transient behavior of surfactant solutions as a function of viscosity (or equivalently within a particular phase). This is because in most cases, the bulk solution will exhibit a non-Newtonian viscoelastic response; that is, mass transport of the acid-surfactant solution involves both a non-linear dissipative shear force ***and*** an elastic deformation restoring force that's difficult to model (ie. It cannot be modeled as a simple RC equivalent).

Appendix C

Die-Level Jig Improvements: A Detailed Description

C.1 Fix # 1: The Homemade Die-Bonding Chuck

To begin with, the 4 mm diameter vacuum hole at the center of the die-bonding chuck (see Figure 3-2) seemed to create havoc on bonding uniformity near the center of either a 1x1 cm^2 or a 1x1 in^2 Cu die pair. More specifically, as the Si subtrate decreases in size, the extra pressure exerted by the piston (assuming downforce was kept constant) now has less area for energy dispersion, and theoretically this is good because the Cu-Cu bonding inteface can now absorb more energy and a better bond should be the result. Unfortunately, the Si die is compliant enough such that the extra external energy applied at the center of the die is now transferred into local deformation of the substrate at the interface between the bottom Si die and the bonding chuck near the vacuum port. The result of this was that in all dies bonded with the mesa chuck, the center Cu regions had subpar contact with each other. Figure C-1 shows the progression as one changes from the mesa chuck to a home-made die-bonding chuck made by bonding two 4" Si wafer together, with the top wafer having a pre-etched square hole in the middle. No bond glass was used here because we wanted to isolate each mechanical problem separately; hence, each die pair was grossly aligned by eye and was bonded with the regular 400 °C Cu recipe. Then, the handle die was subsequently released in hot HCl in 6 hrs, thereby also **showing the feasibility of acid encroachment release on the die-level.** As seen from the photos, by switching from the mesa chuck to the homemade chuck, the improvement of the bonding quality was unmistakable.

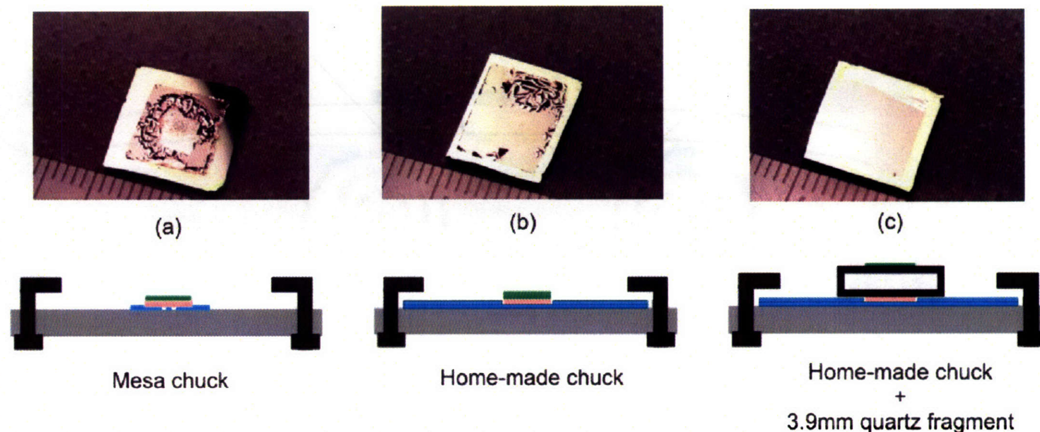


Figure C-1: The effecs of chuck flatness on the die-bonding quality. (a) Left photo shows the mesa chuck result, (b) middle photo shows the home-made chuck result, and (c) right phot shows the combination of home-made chuck and a 3.9 mm quartz fragment used as a makeshift bond glass.

C.2 Fix # 2: Bond Glass Bow Reduction by Overpressing

As mentioned previously, when two Si dies were clamped by the bond glass, there is a natural bow that forms right over the substrates due to its 4 rigid support points - two of them were located at the bond glass clamps, and the other two are located at the contact points between the glass and the outer corners of the top die. Theoretically, one can calculate the air gap distance between the top die surface and the natural bow's vertex, both depicted in Figure C-2, using a combination of geometry and the material parameters of the quartz glass. Once we know the theoretical bow curvature, all one needs is to dial in the necessary overpress distance on the piston backstop collar and we're done.. *right?*

Unfortunately that's probably the most inefficient way to solve this problem because a robust theoretical bow calculation has to take into account a not-so-tiny lateral variation of the support points, and if either die had an offset from the geometric center of the bond glass (could easily be a millimeter off or more laterally in any given run), then the entire calculation becomes invalid because the bow geometry is now asymmetrical and the load point (piston coordinate) is now off-axis. Furthremore, the more one uses the quartz glass, the more it deforms because of stress buildup from mutliple thermal cycles.

Therefore, the entire problem was solved empirically by bonding numerous die pairs varying only the piston backstop collar height and the piston downforce. The starting points for the test matrix were the "theoretical limits of the jig" given by the machine manufacturers, of which at those values the dies didn't bond very well to begin with (see photo (a) in Figure C-3):

- **Backstop collar height** = Total stack height, where the stack include the base wafer thickness of the home-made chuck, the thickness of the two to-be-bonded dies, the of the quartz bond glass in the non-flexed state, and the graphite insert.

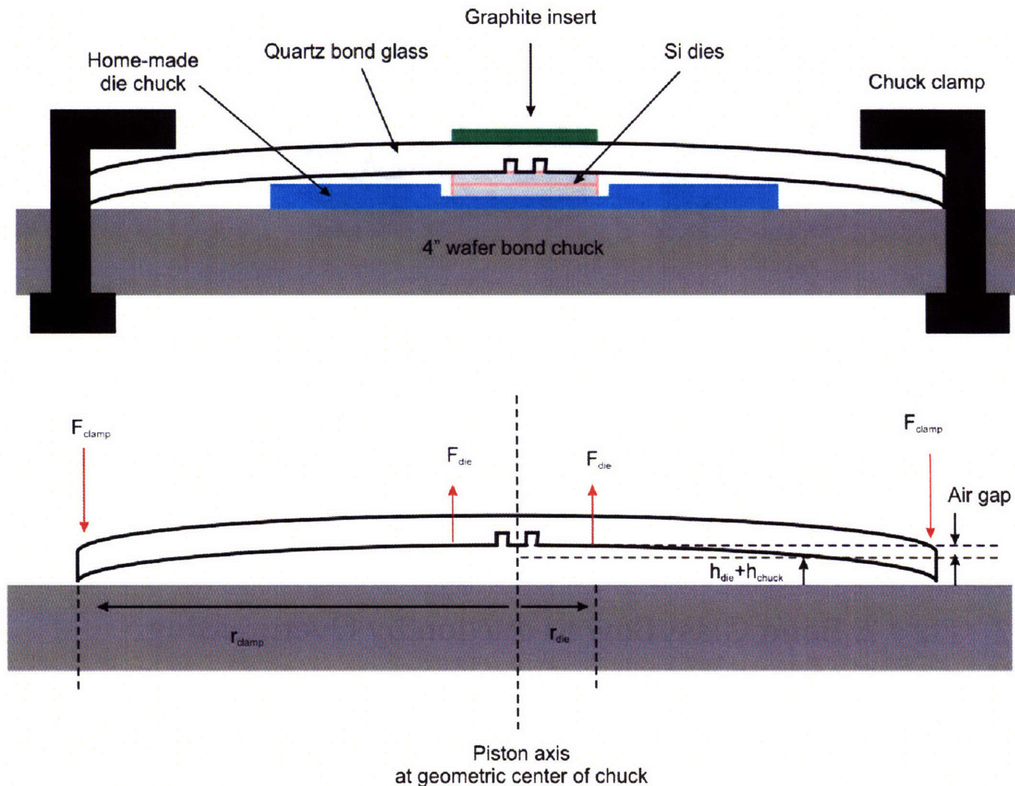


Figure C-2: Bond glass bow: The setup and its corresponding force diagram

- **Piston downforce** = 300 N, where we were warned that the risk of cracking the quartz bond glass was high if this force value was surpassed.

The litmus test for the experiments were the quality of the Cu-Cu bond upon wafer release and checking for signs of bond glass fatigue after each bonding run. Instead of reporting every possible force / distance combinations used in this study, a bond quality comparison between an un-optimized and optimized parameters was shown in Figure C-3 below. For our die-bonding system, the optimized mechanical parameter values which eliminated the quartz plate's bow were:

- **Backstop collar height** = (*Total stack height*) - (500 μm)
- **Piston downforce** = 1500 N.

C.3 Fix # 3: Plateau Dies and Post-alignment Check

One of the many lacking aspects of the MIT 3-D integration flow is the following: When the laminate structure's layer stack was composed of un-patterned Al, Ta, and Cu layers, IR microscopy cannot be used in

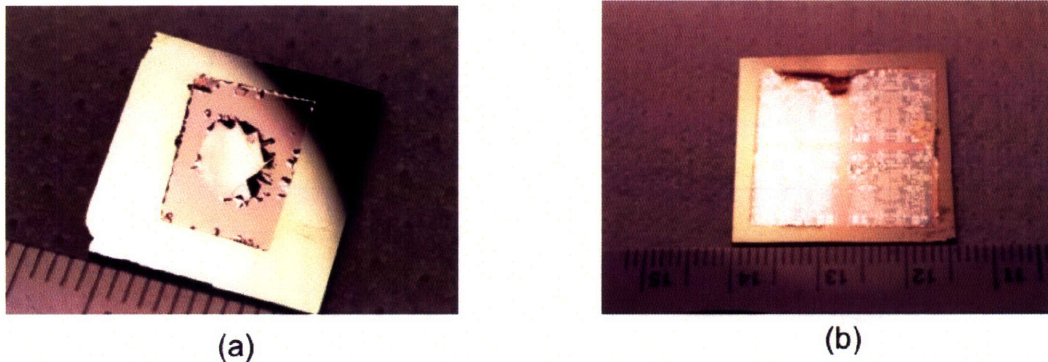


Figure C-3: The left photo (a) displays the bonding quality before backstop collar and downforce optimization, and the right photo (b) displays the optimized force parameters. For both cases, the home-made chuck elevation was h_{chuck} = 0.46 mm, the 2-die combination thickness was h_{die} = 1.24 mm, the quartz bond glass was 2.303 mm, and a *new* graphite insert thickness was 1.0 mm. Each ruler tick mark in the photos correspond to 1 mm.

either the wafer-wafer or the die-die alignment process because the laminate is impervious to IR penetration. On the same token, once the two dies were aligned and clamped by the quartz bond glass, visual verification of the alignment accuracy is very difficult because patterns on at the Cu-Cu interface will be buried. In theory, this issue would be moot if one can guarantee the double- sided registration alignment on the base die and the actual die-die alignment procedures were both perfect. But in reality, without post-alignment verification, all bonded dies tend to have at least a +/- 20 μm in both x and y-directions and with at least 1 degree or more off in θ. The presumed enemy here is the inherent difficulty in squaring-up the two dies inside the aligner. And the reason: **Bad angular alignment**.

To begin with, although the wide angle objective provided a wide enough field of view so that one could see the center of a die, the objective could only traverse 25% past that center point before the eyepiece is stopped by an internal safeguard. This created a major problem: To have accurate alignment between two substrates in the θ direction, the apparent arc length-per-angular deviation ratio has to be maximized; moreover, if one cannot optically observe the arc length deviation from ***opposite sides of the die***, then the effective die-die alignment accuracy will decrease significantly. To quickly explain this, let's take a look at Figure C-4.

In cartoon (a), if two substrates that were misaligned by a given angle θ, one can visually detect this by looking at the apparent arc length displacement S_1 or S_2 of an alignment mark placed initially placed at a distance r_1 or r_2 away from the pivot point. In other words, an imperceptible error in θ can be made into an optically observable distance S provided that the "magnification factor r" is large enough . Obviously, one would like to maximize the S per θ ratio because it's easier to correct for the larger error S_2 as opposed to S_1, and the ubiquitous fab trick of maximizing the moment arm r is to place at least 2 wafer or die alignment marks as far away from the geometric center of the substrate as possible. This was the situation depicted in

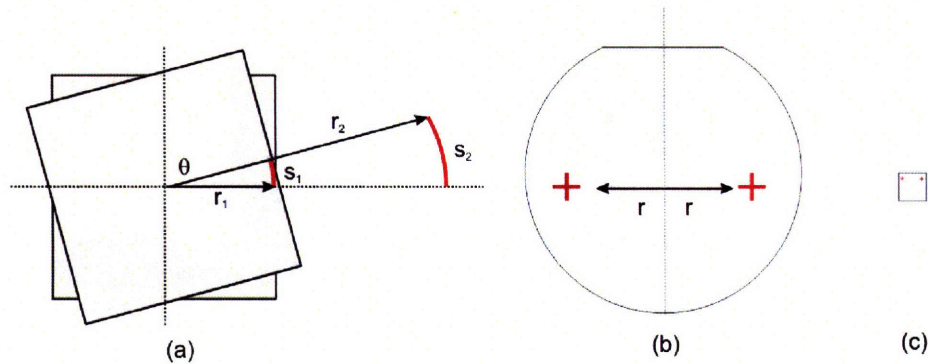


Figure C-4: Essence of die-die θ-misalignments. In (a), given an angular misalignment of θ between the two substrates, the size of the projected misalignment arc lengths S_1 and S_2 ar related to the moment arm dimensions r_1 and r_2 by the equation $S = r\theta$. In (b), the larger r-value on a 6" wafer registration set creates a larger S. In (c), a small r-value on an 1" die results in an imperceptiable value of S.

cartoons (b) and (c). Now, problems with this theory arise when:

1. The moment arm r for die-die alignment is only 1 cm long, and
2. When your microscope objective can only travel a distance 75% of 2r.

Thus, the two factors mentioned above create a fatal flaw in our die-die alignment system: It is impossible to properly orthogonalize an 1x1 in^2 die because the wide-angle objective can only see past a distance of 0.5r past the geometric center of the die. As seen from Figure C-5, a dark zone exist between the two microscope objectives because the wide-angle scope cannot travel too far beyond the meridian of the bond chuck. This means the maximum effective r has been reduced from 1.2 cm to about 0.6 cm, [1] and if we consider r = 0.6 cm as our system's metric, then a 0.5 degree ($8.73\text{x}10^{-3}$ radians) misalignment in θ corresponds to a small-angle approximation arc of $s = r\theta = 52.36$ μm, which is equivalent to an 1 μm shift in one orthogonal direction and a 52 μm shift in the other orthogonal direction. It should be clear to the reader that the alignment tolerance is not sufficiently accurate enough for implemting high-density 3-D integration.

In tackling the entire alignment issue, our current solution involves a repeated, brute-force alignment verification that sacrifices some Si area on one of the two dies, and the highlights of this can be seen from Figure C-6.

First, the dies were rotated 90° such that the alignment marks were not truncated by the microscope dark zones. Next, the second die to be inserted into the aligner has to have an area smaller than the first die. By cleaving away some Si area directly near the alignment mark positions, we have created an exposed plateau region where alignment marks from both substrates can be seen simultaneously, albeit it required a bit of re-focusing during the die alignment procedures. An additional benefit to the Si plateau scheme is

[1] Please remember that maxiumum r can be only considered from the mechanical pivot point of the system, not just the mere midpoint between 2 opposing alignment marks.

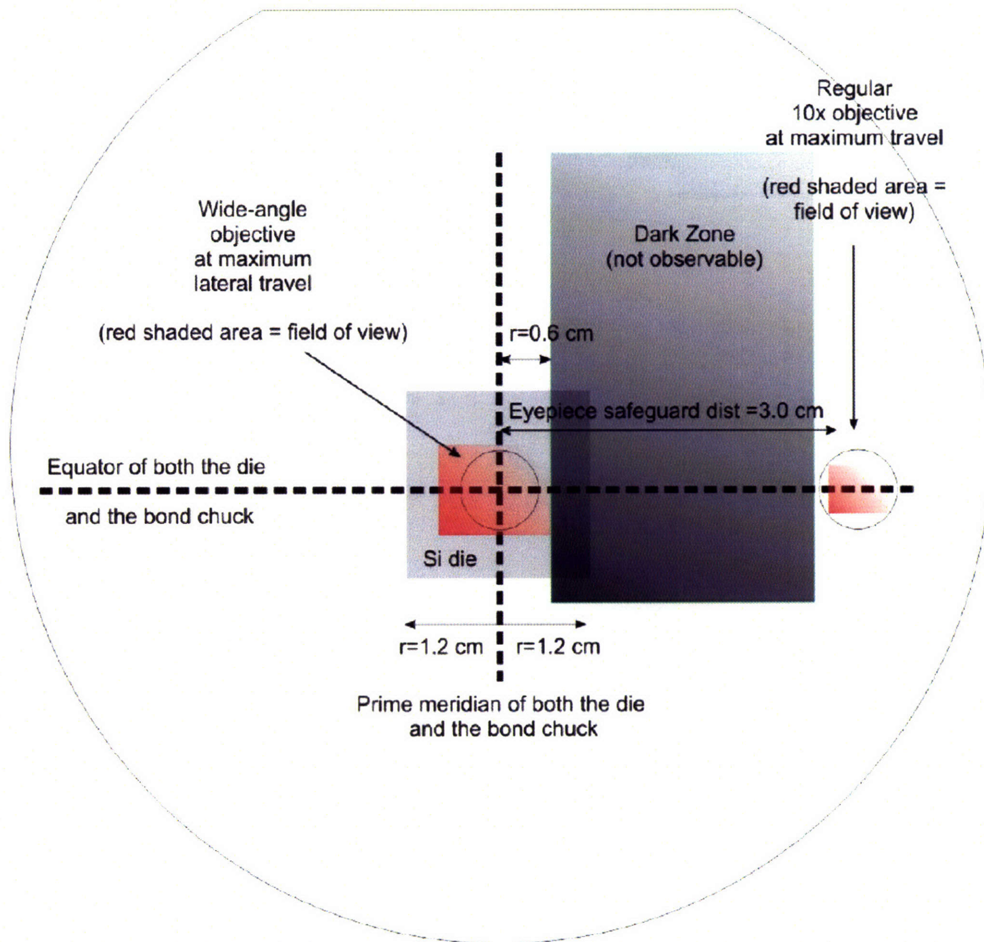


Figure C-5: Mechanical travel limits of the microscope objective and their small field of views both createa dark zone that limits the accuracy of die-to-die θ-alignment. The dimension of the outer "wafer-like" outline was not drawn to scale.

that once the dies were aligned and clamped by the quartz bond glass, a non-IR visual confirmation of the alignment can be performed by placing the entire bond chuck setup under a normal microscope [2]. If the alignment was off, then the entire aligning process was repeated. This solution actually worked quite well and was indeed a repeatable process.

[2]In reality, we actually took apart the mask holder / wafer chuck combination on KS2 aligner in TRL, placed the entire bond chuck assembly onto its sliding loader, and used the KS2's microscope setup for visual examination of the die alignment. This was done because the vertical travelling distance of every stand-alone microscope in TRL was too short

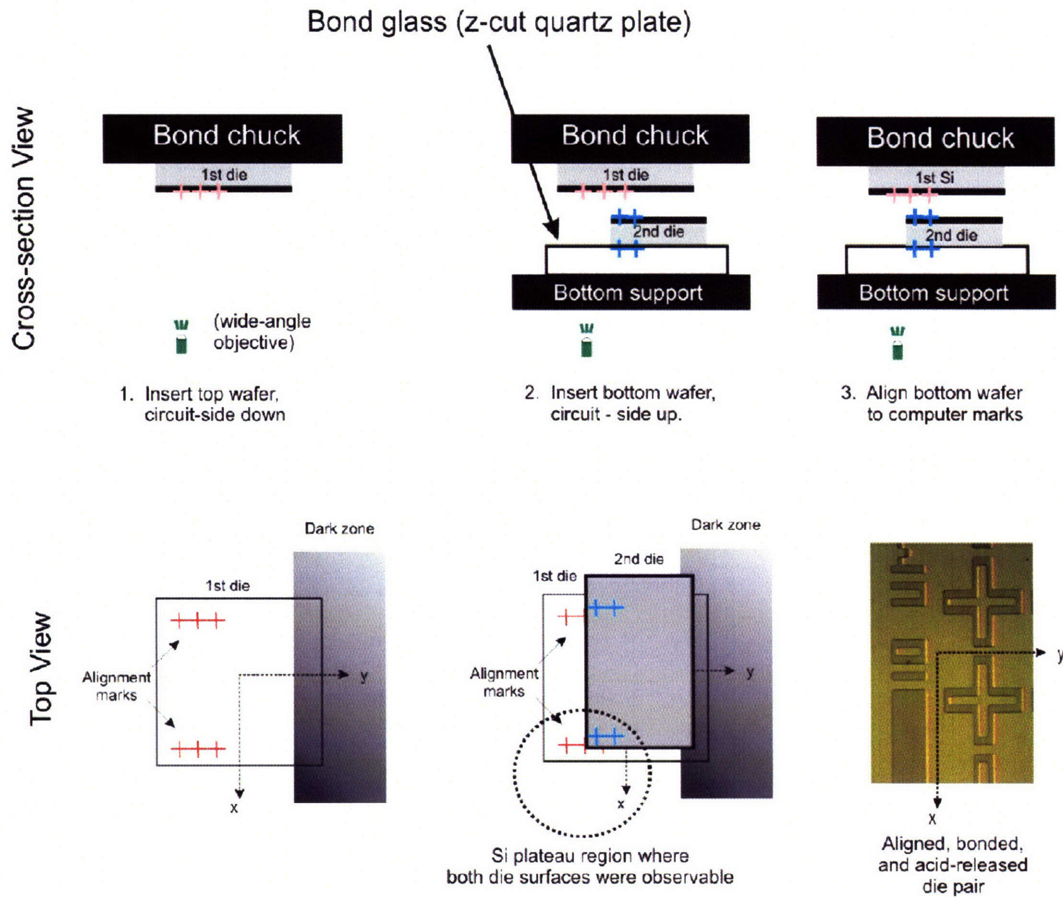


Figure C-6: Sacrificing some die area, the plateau region created by area reduction of the second die increases the accuracy of die-die alignment and faciliates a easy method of post-alignment verification.

C.4 Fix # 4: Graphite Insert Monitoring and Pyrex Wafer Substitution

As previously alluded to, one can try to calculate and compensate for the bond glass bow, check and double-check the alignment prior to bonding, making sure the quartz bond glass's condition is optimal, and etc, and you can *still* get a bad quality die bond. Out of all available die-bonding variables, this thesis has not touched much on the mysterious graphite insert and its role in bonding. Simply put, the physical condition of the graphite insert is the *most important* factor in determining whether or not the top PMOS-SOI film will delaminate from the NMOS base die after acid-encroachment handle wafer release.

To start the discussion, after placing the aligned dies inside the bonder, an 1"x1" graphite insert was placed on top of the quartz bond glass. Since this insert material was extremely compliant, theory suggests that the graphite cutout will absorb the energy from the piston and re-distribute it evenly onto the quartz glass plate, thus achieving a more uniform bond. Using the mechanical parameters listed on page 160, six

die pairs were consecutively bonded over a course of 2 days during mid-April 2006 [3]. After each run, the quartz bond glass was also checked for signs of fatigue after cool-down, and although the graphite insert thickness remained 1 mm after each run, the insert material itself became much harder with each run. In fact, it almsot looked like if the graphite was being carbonized, but since all other mechanical parameters stayed constant during each run, the graphite insert was re-used in trying to match each bonding conditions. Figure C-7 shows the bonding result after each die pair underwent approximately 5 hrs of passive (no-ultrasonic agitation) acid release [4]

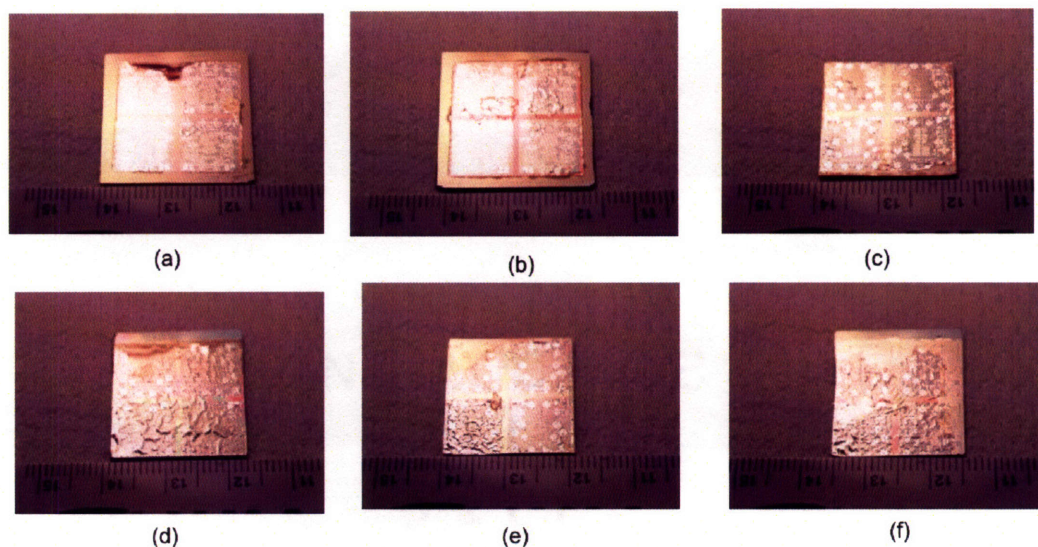


Figure C-7: A set of 6 die-bonding pairs made in mid-April, 2006. The released dies in (a)-(f) were bonded in succession with identical mechanical bonding parameters indicated within the main text. The thickness of the die stack prior to bonding matches those mentioned in the captions for C-3, and each ruler tick in the photos correspond to 1 mm in length.

As one can see from Figure C-7, the first two 1" dies in (a)-(c) bonded extremely well and only exhibited a few apparent "wrinkles" on the PMOS SOI surface. These wrinkles represent regions where the Cu-Cu bond delaminated after some amounts hot HCl encroached into pre-existing defects and corroded away the Al pads / interconnects, the Ti adhesion layer of the Cu damascene vias, and oxidized portions of the Cu metal itself. Nevertheless, the important data to note is that overall bond quality grew progressively worse as the bond sessions continued, as dies (d) and (f) both have about a 99% bonding area failure. As previously mentioned, every mechanical parameter stayed constant after each bond with the exception of the graphite insert's hardness.

[3] Samples were bonded with the usual recipe of 400 °C for 30 min in vacuum. Also, in-between each run, the piston backstop collar was re-zeroed by testing the waferbow pin's contact to the surface of a wafer-less, standard 6" teflon chuck used in Si-Si direct bonding.

[4] For reference, the bottom dies in (a) and (b) were made of blanket Cu films (the top dies had circuitry), while in pairs (c)-(f), both bottom and top dies had real NMOS / PMOS devices on its surface.

To verify that the graphite insert's hardened condition was indeed the bad-bond culprit, a second set of die bonds was performed two weeks after the ones from Figure C-7. Again, the mechanical bonding parameters remained the same, and a fresh graphite insert was provided for the first bond pair only. Figure C-8 shows the result of the second round of bonding. Once again, Figure C-8 shows the degradation of

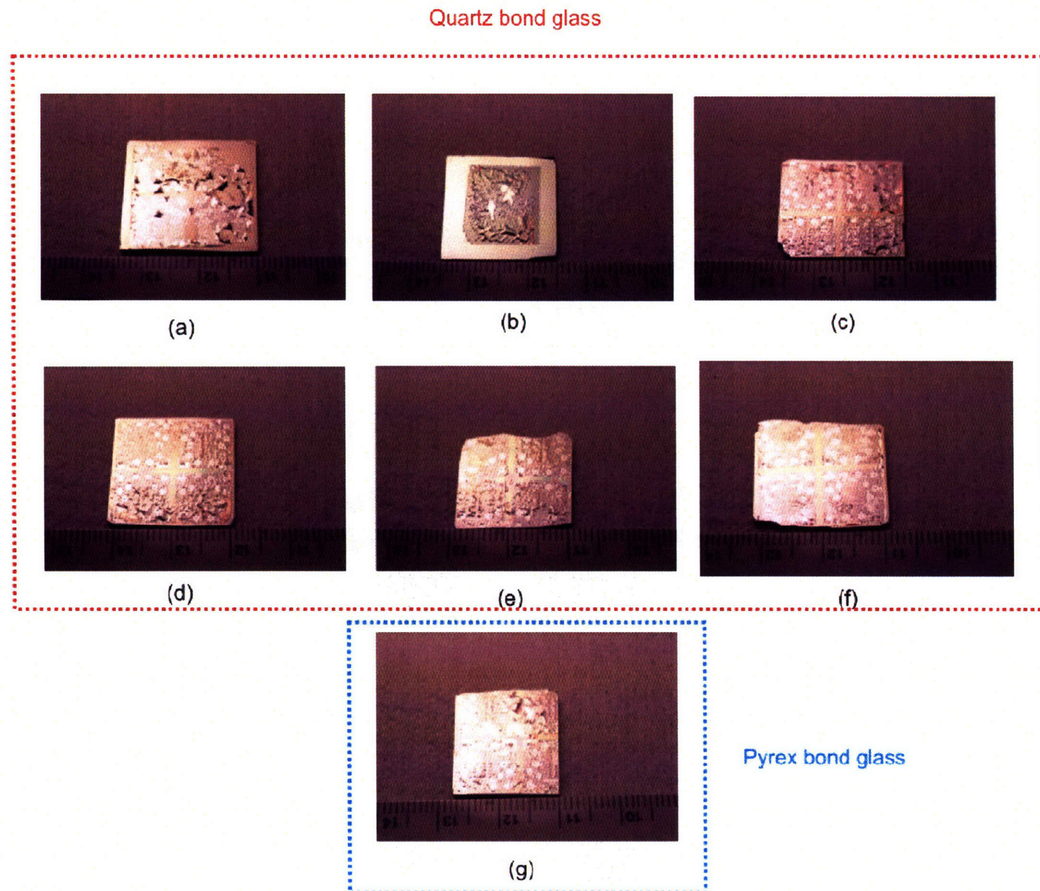


Figure C-8: A second set of 7 die-bonding pairs made two weeks after the first set. Upon breaking the quartz bonding glass after pair (f), a 2.00 mm pyrex glass subsitute was used in pair (g), in which the pyrex glass also broke.

bond quality as one continues to use the same graphite insert, but this time the quality of the initial pair, seen in (a), was much worse than that of (b) or (c) in Figure C-7. Nevertheless, the southwest quadrant of pair (a) exhibited a pristine PMOS surface after acid release, thereby showing the possibility that the center of the die was probably shifted from the geometric center of the chuck during aligning. After the 6th die pair bond, however, the manufacturer's quartz plate finally succombed to fatigue and split into two after chamber cool-down. In place of the quartz glass, a 2.00 mm pyrex wafer was used as the new bond

glass with the graphite insert remaining in its place [5]. Although the pyrex wafer combination was known to have worked before, it also shattered after the 7th die bond. Hence, a new die pair was immediately bonded with a new pyrex wafer and with a new 1 mm graphite insert was used on top of the pyrex glass. The result immediately reverted back to the likes of (a) in Figure C-7. Thus, **these two expensive mishaps were testaments proving that the condition of the graphite insert can make or break the die bonding quality**.

[5] Since the pyrex wafer contains no vacuum grooves, the die was kept stationary during the alignment process from stiction produced by a partially-dried pool of 2-propanol sprayed on the pyrex surface. This stiction was stable enough to last 3 rounds of die-die re-alignment, if necessary.

Appendix D

Supplemental Results and Graphs from the Ring Oscillators

D.1 Unbonded Single NMOS Devices

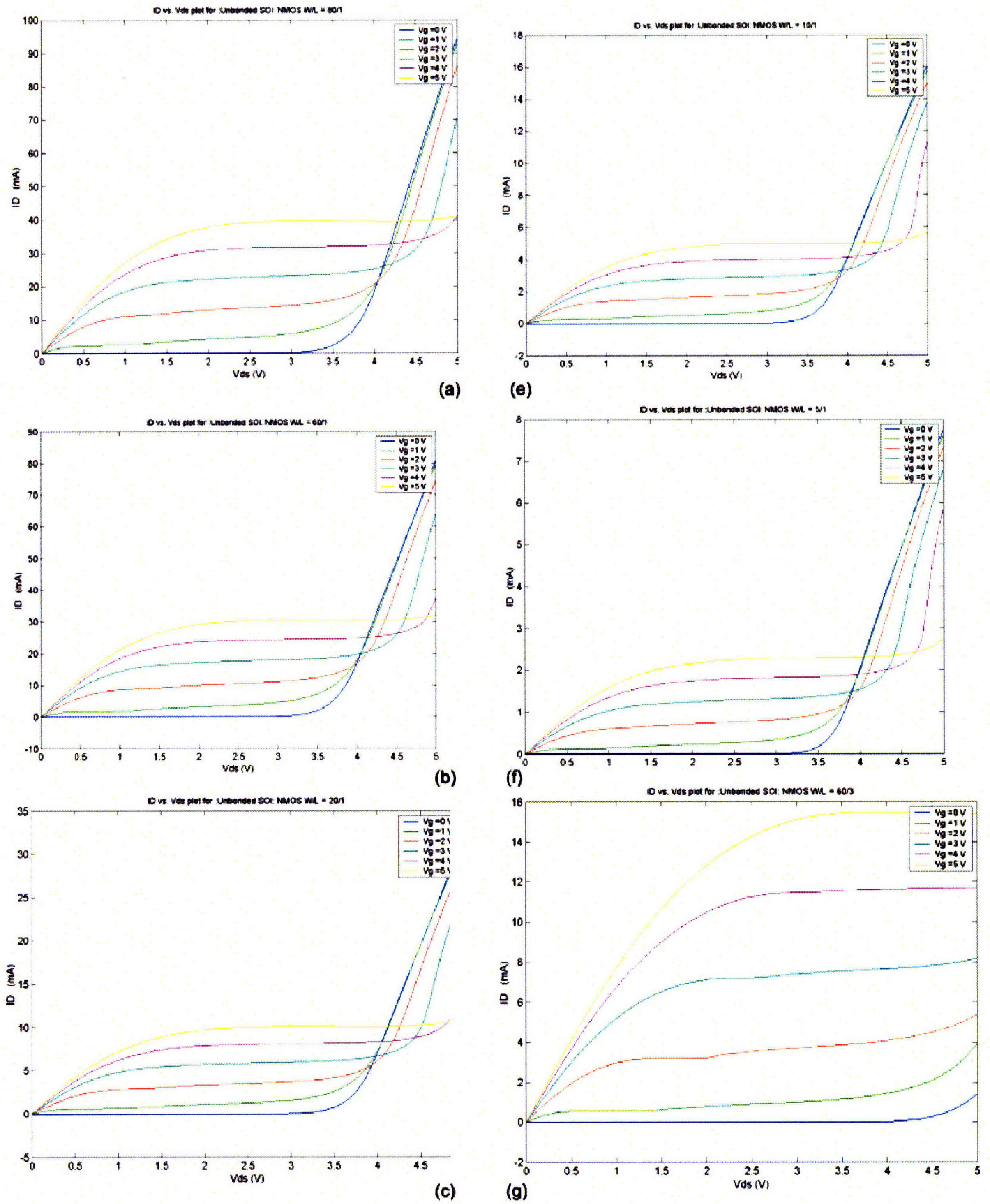


Figure D-1: Unbonded NMOS Id-Vd plots. The width/length ratio in microns for each NMOS were: (a) 80/1, (b) 60/1, (c) 20/1, (d) 10/1, (e) 5/1, (f) 60/3.

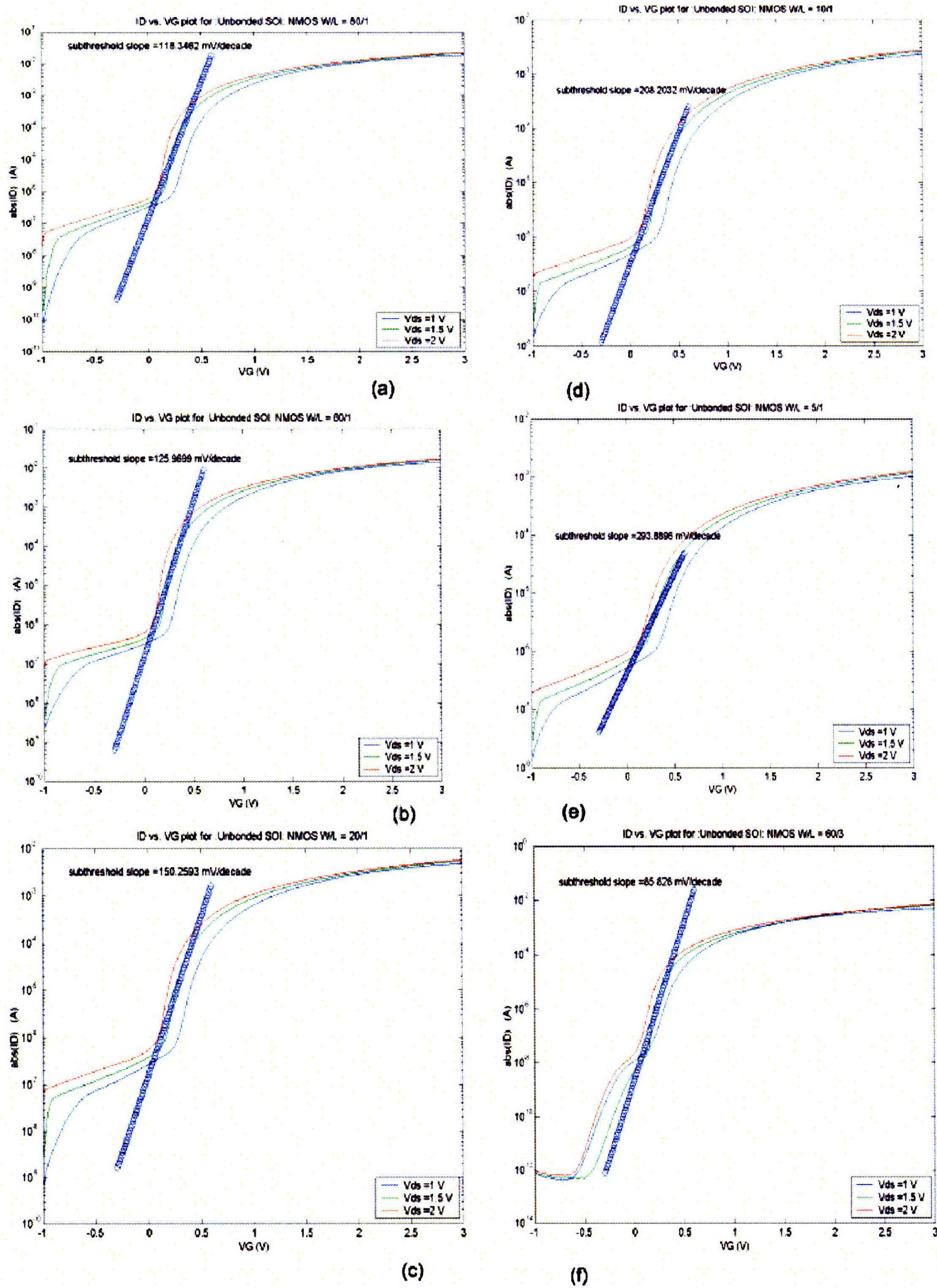
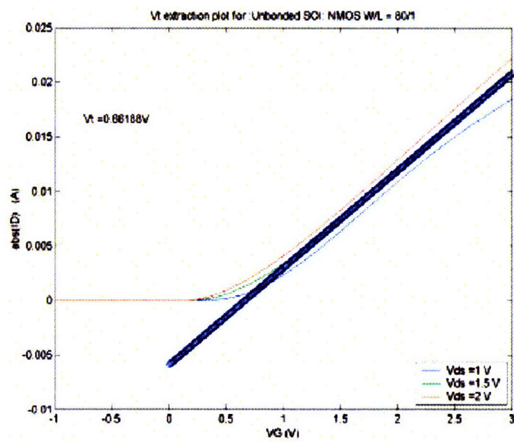
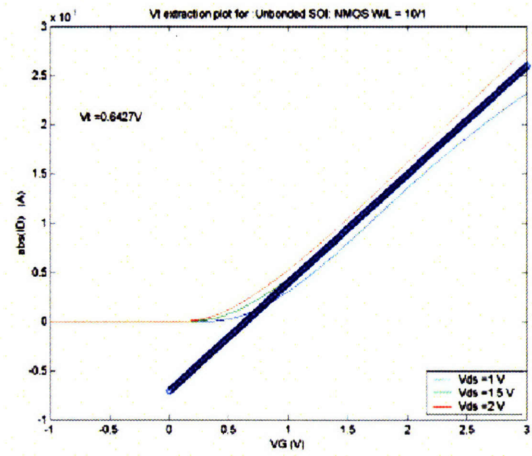


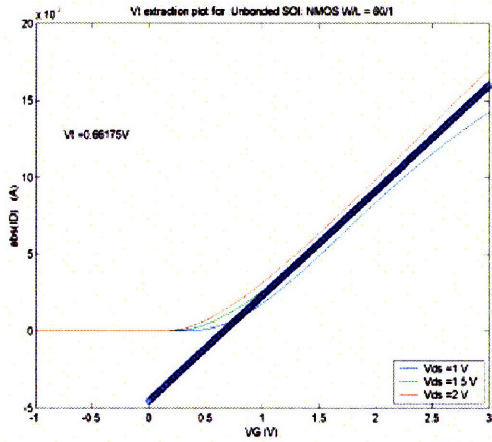
Figure D-2: Unbonded NMOS Id-Vg plots, with subthreshold slope extraction. The width/length ratio in microns for each NMOS were: (a) 80/1, (b) 60/1, (c) 20/1, (d) 10/1, (e) 5/1, (f) 60/3. As seen from the subthreshold slope, these devices do not turn off as well as they should. We probably have some serious edge leakage problems as well as a double-Vt hump due to the much-smaller Vt caused by our layout's **poly gate extension**.



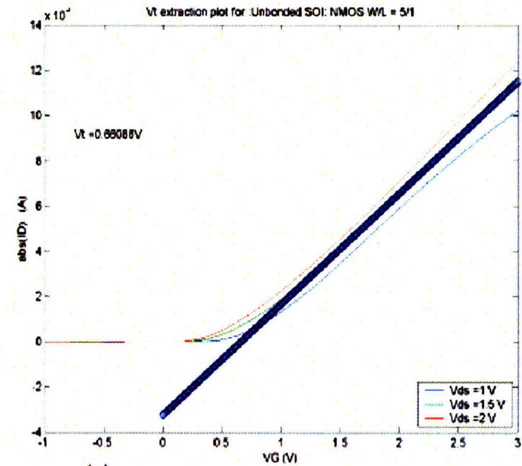
(a)



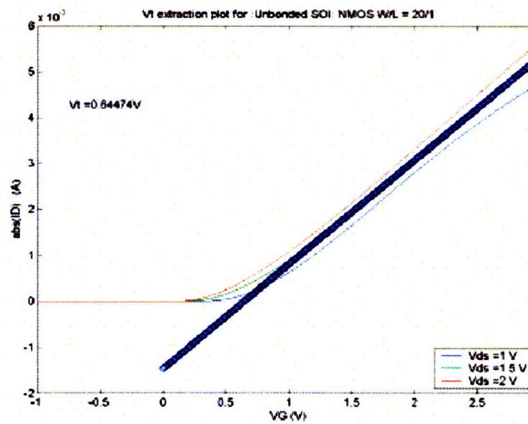
(d)



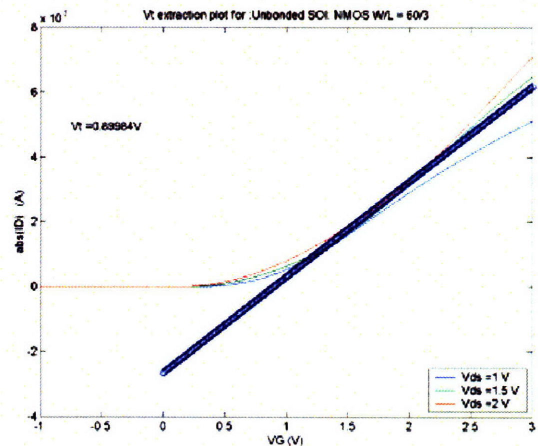
(b)



(e)



(c)



(f)

Figure D-3: Unbonded NMOS Vt-extraction plots. The width/length ratio in microns for each NMOS were: (a) 80/1, (b) 60/1, (c) 20/1, (d) 10/1, (e) 5/1, (f) 60/3.

D.2 Unbonded Single PMOS Devices

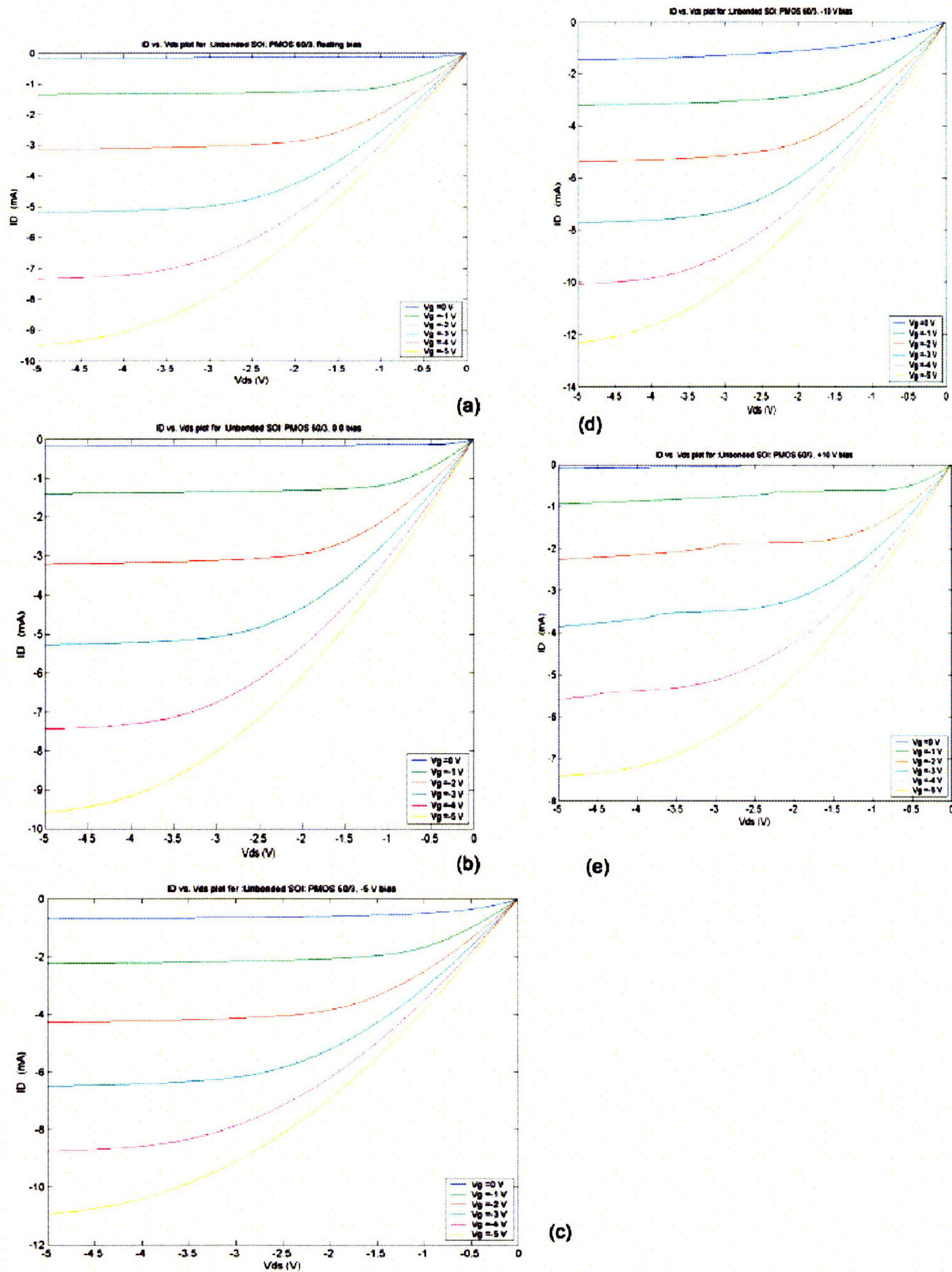


Figure D-4: Unbonded NMOS Id-Vd plots. The width/length ratio in microns for each PMOS were all 60/3, and the difference between the plots were the backbias voltages: (a) floating, (b) grounded, (c) -5 V, (d) -10 V, (e) +10 V

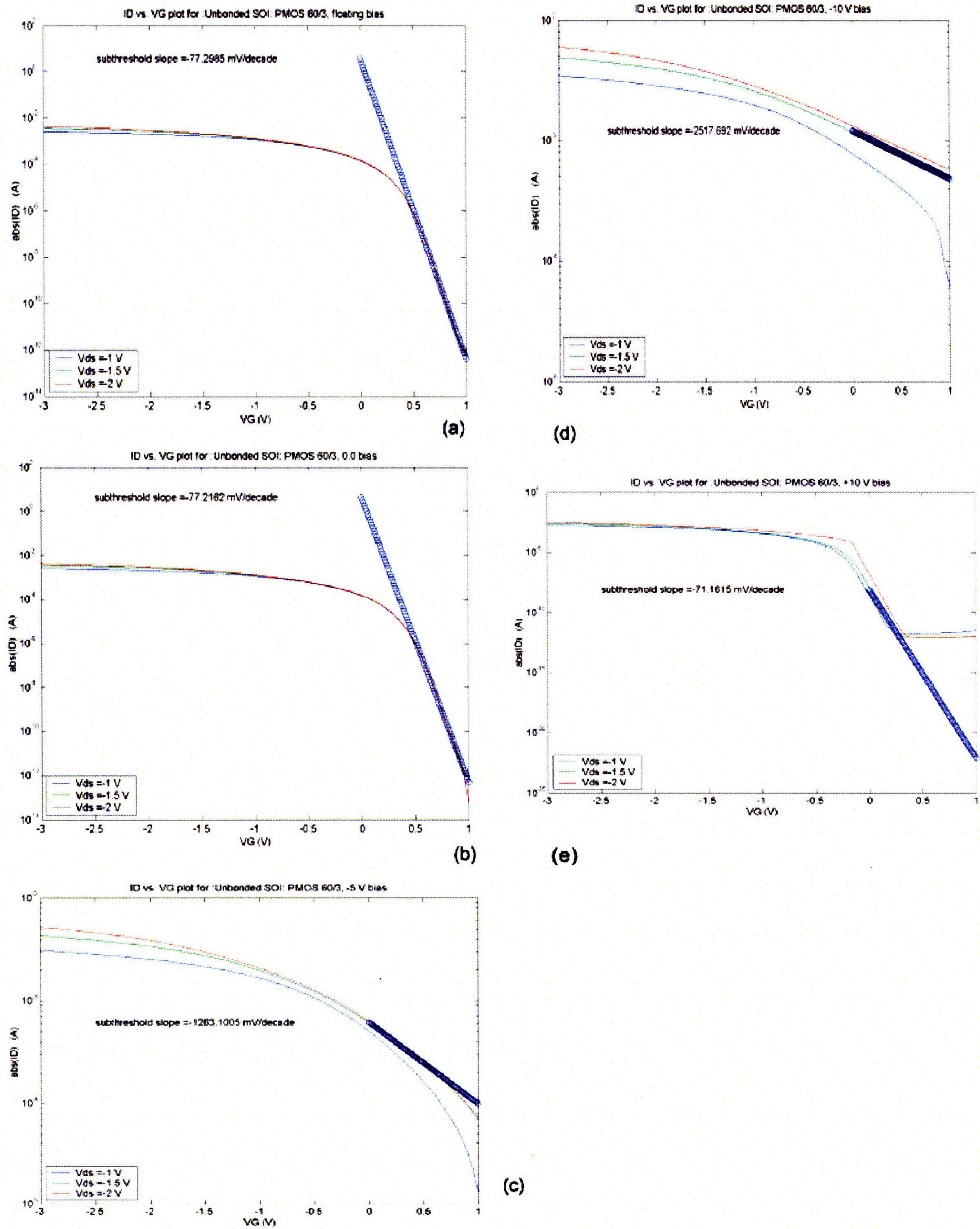


Figure D-5: Unbonded NMOS Id-Vg plots. The width/length ratio in microns for each PMOS were all 60/3, and the difference between the plots were the backbias voltages: (a) floating, (b) grounded, (c) -5 V, (d) -10 V, (e) +10 V

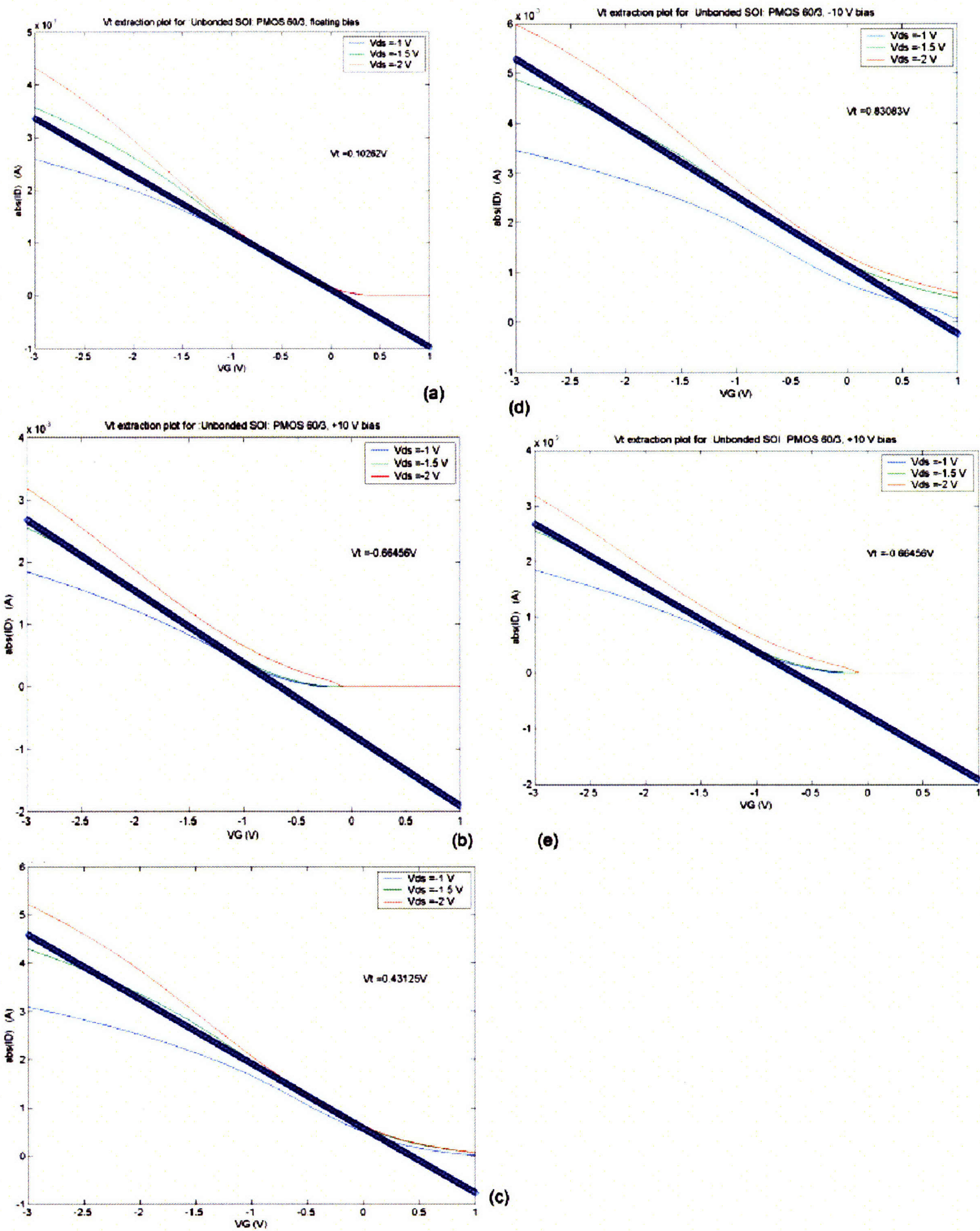


Figure D-6: Unbonded NMOS Vt plots. The width/length ratio in microns for each PMOS were all 60/3, and the difference between the plots were the backbias voltages: (a) floating, (b) grounded, (c) -5 V, (d) -10 V, (e) +10 V.

D.3 Face-Face, 21-Stage CMOS Ring Oscillators, Floating-body

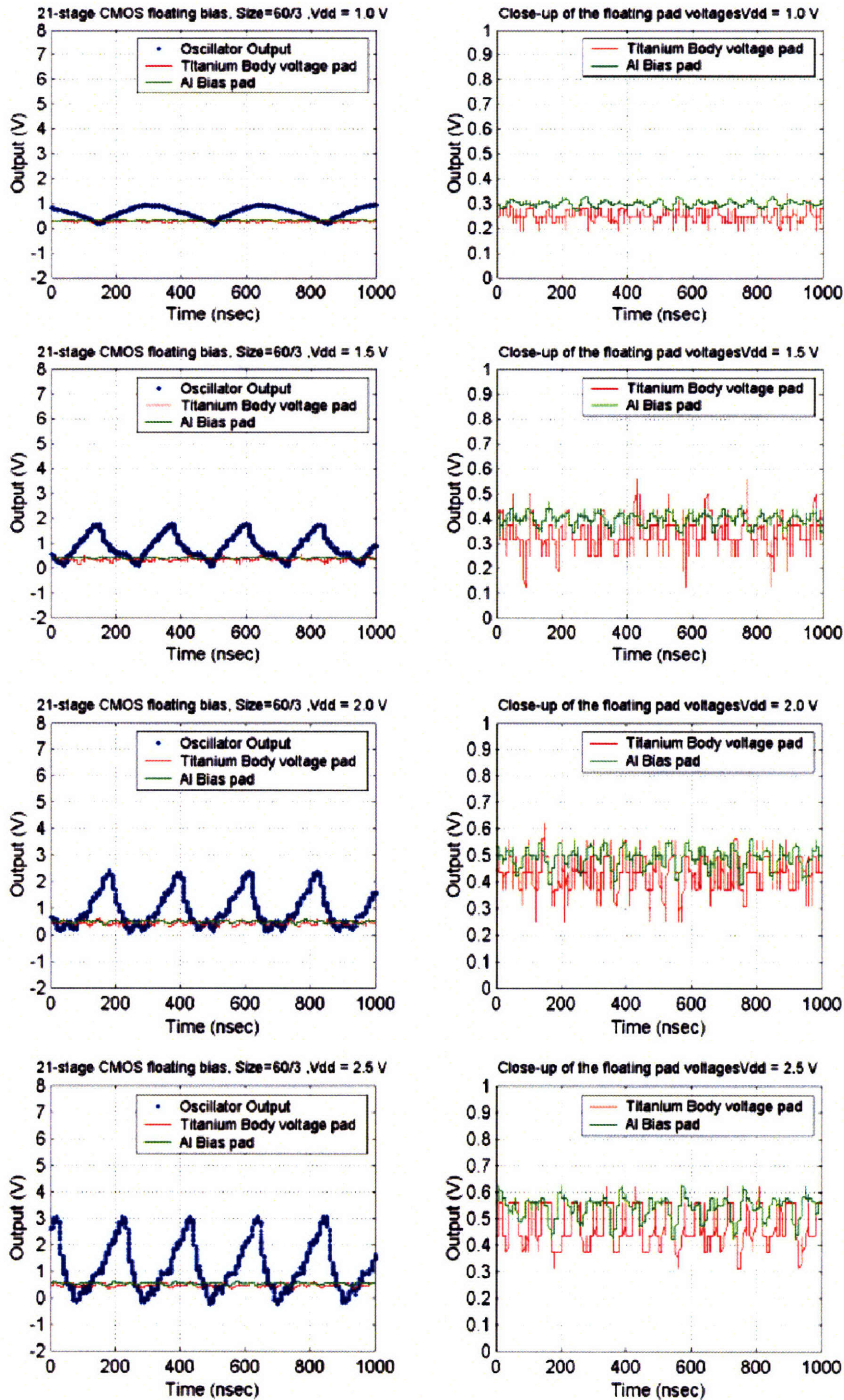


Figure D-7: 21-Stage CMOS, L = 3 μm: Vdd = 1 thru 2.5 V. Left-column plots are the signals from the 9**x** output buffer and the tiny traces from probing the floating Ti backgate and the "useless" Cu-Al plane's pad; right-column plots are zoomed-in traces of the aformentioned floating pads.

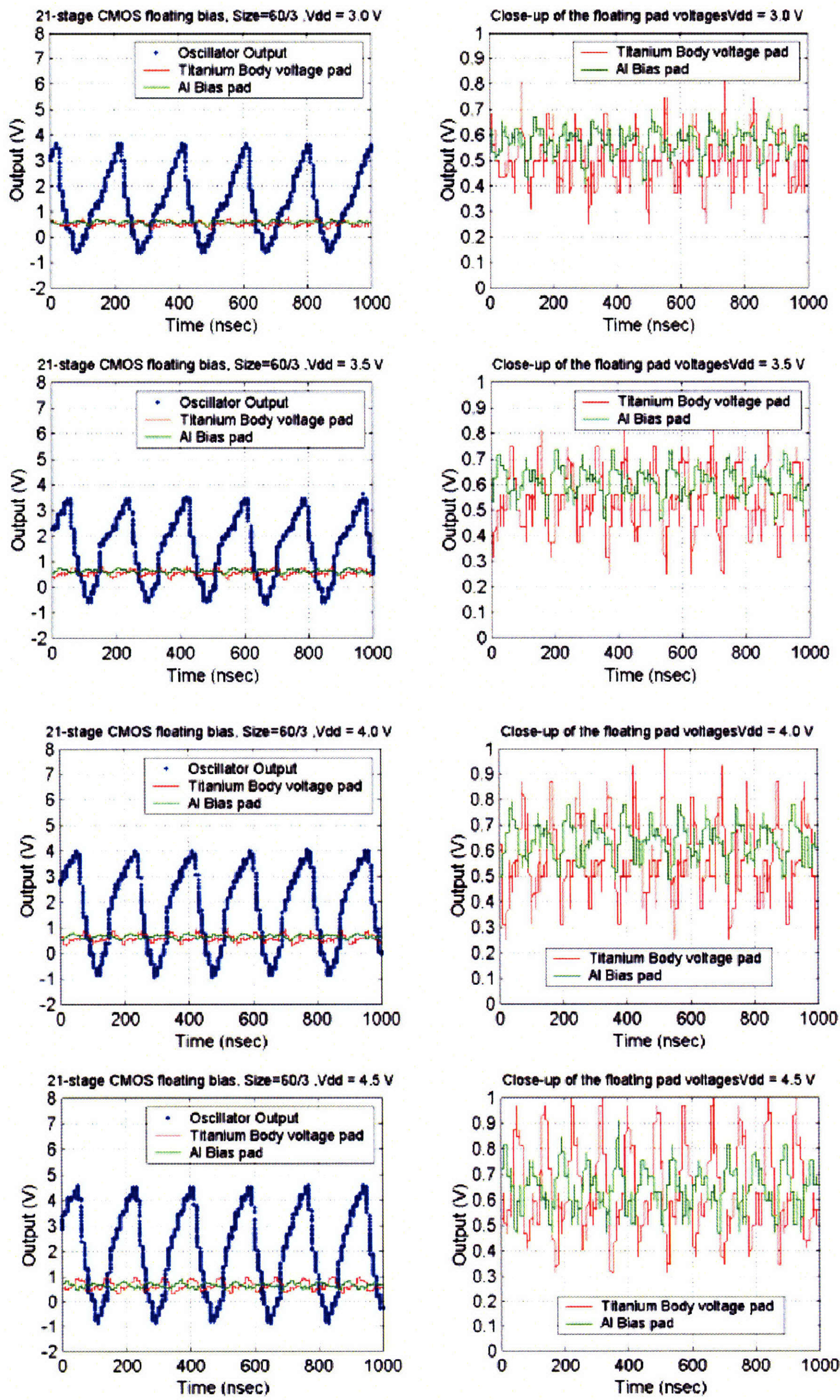


Figure D-8: 21-Stage CMOS, L = 3 μm: Vdd = 3 thru 4.5 V.

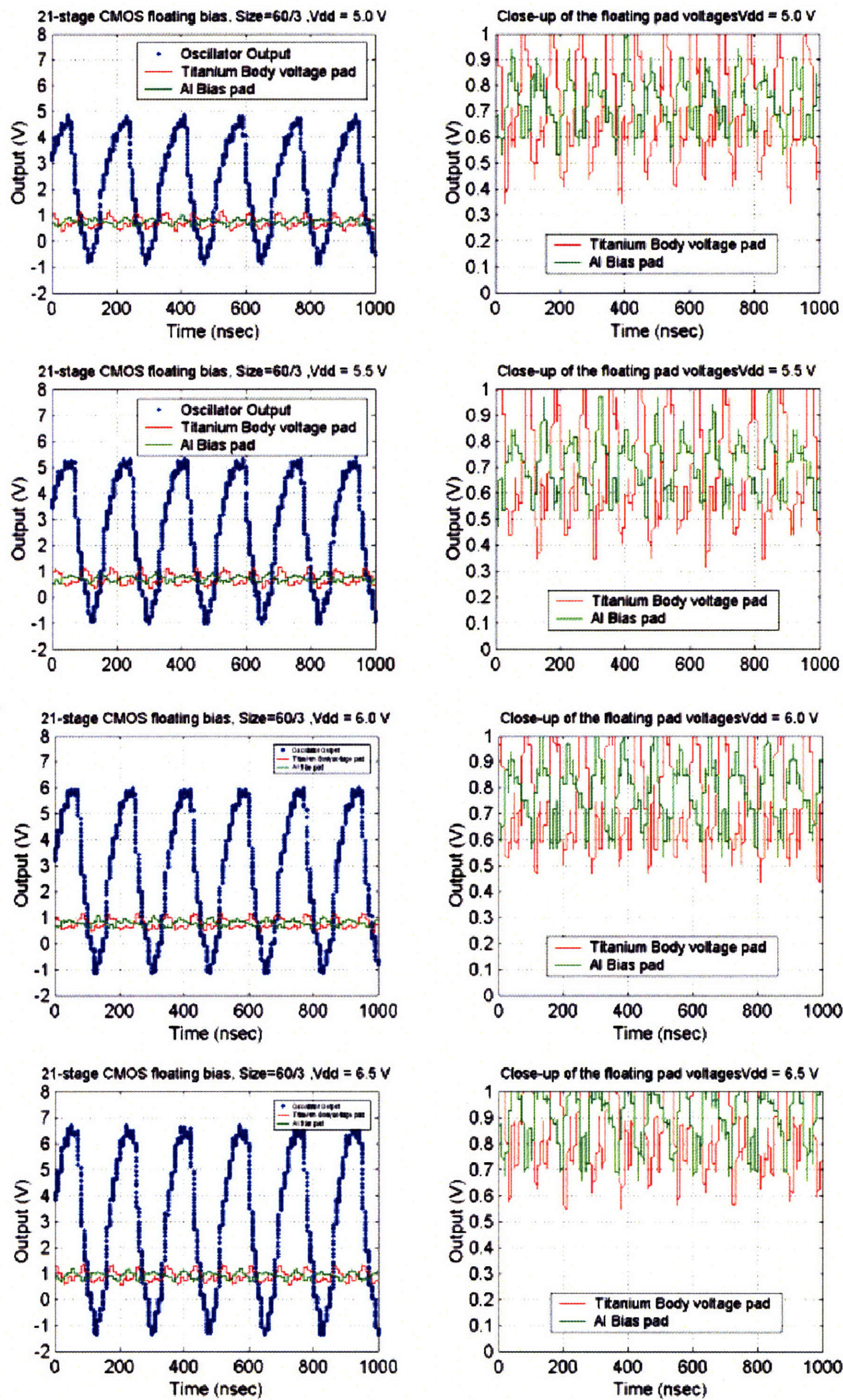
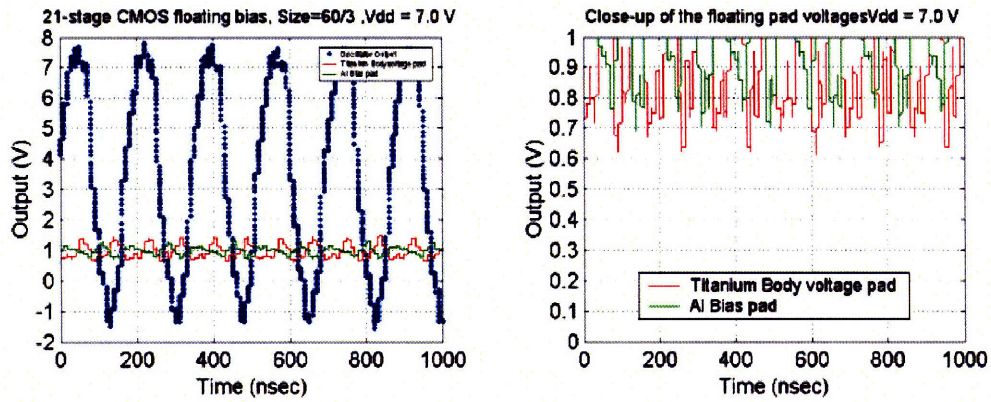
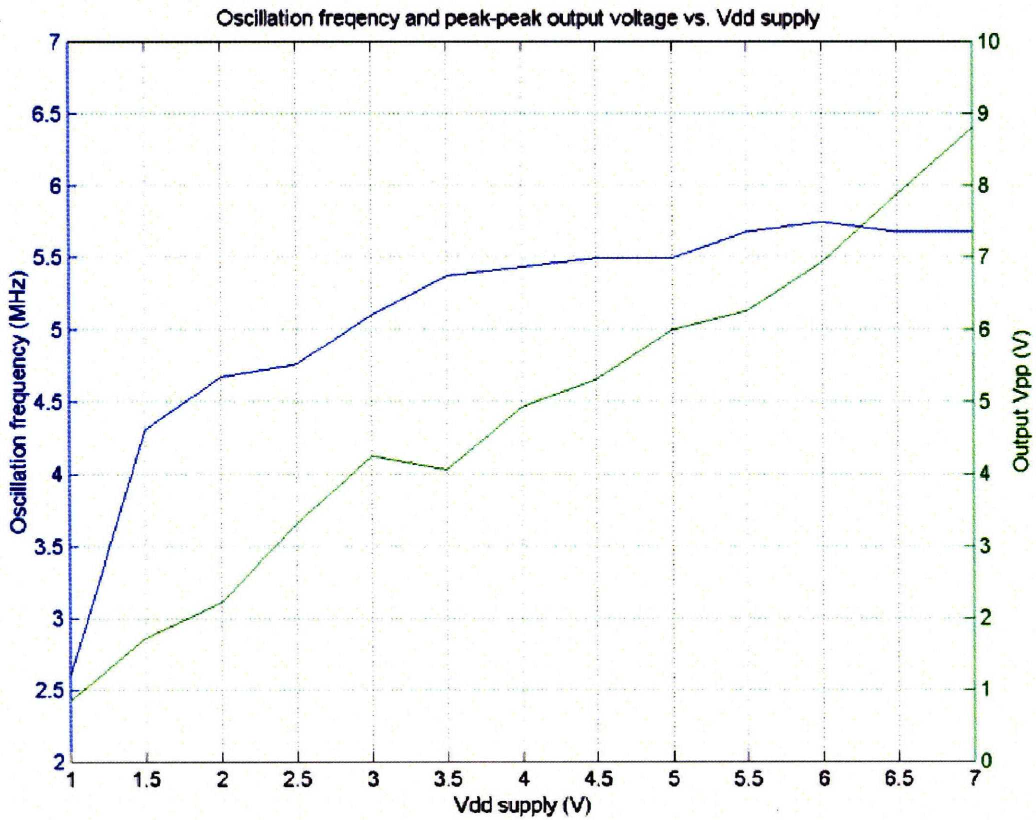


Figure D-9: 21-Stage CMOS, L = 3 μm: Vdd = 5 thru 6.5 V.



(a)



(b)

Figure D-10: 21-Stage CMOS, L = 3 μm: The plot at Vdd = 7.0 is in (a). In (b), the frequency and the peak-to-peak voltage Vpp of the output was plotted as a fuction of Vdd. Note the saturation of the oscillation frequency at high Vdd's.

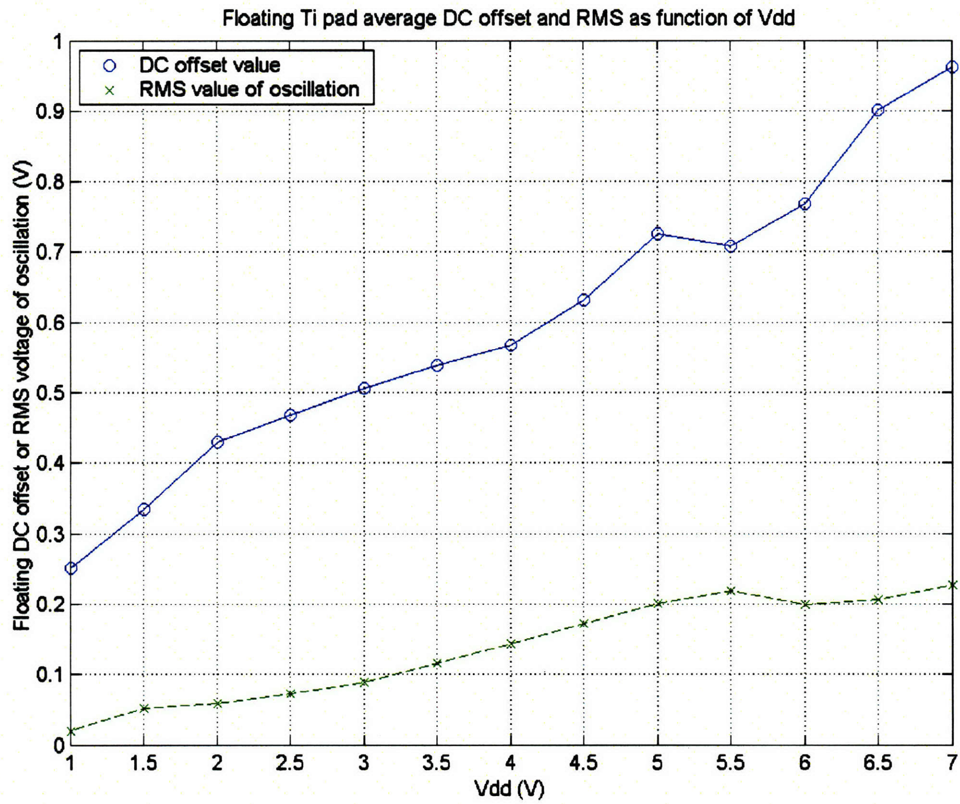


Figure D-11: 21-Stage CMOS, L = 3 μm: The DC offset and the RMS voltage of the fast oscillations coming off of the floating Ti bckgate and the useless Cu floating pads

D.4 Face-Face, 21-Stage CMOS Ring Oscillator, with Backbiasing

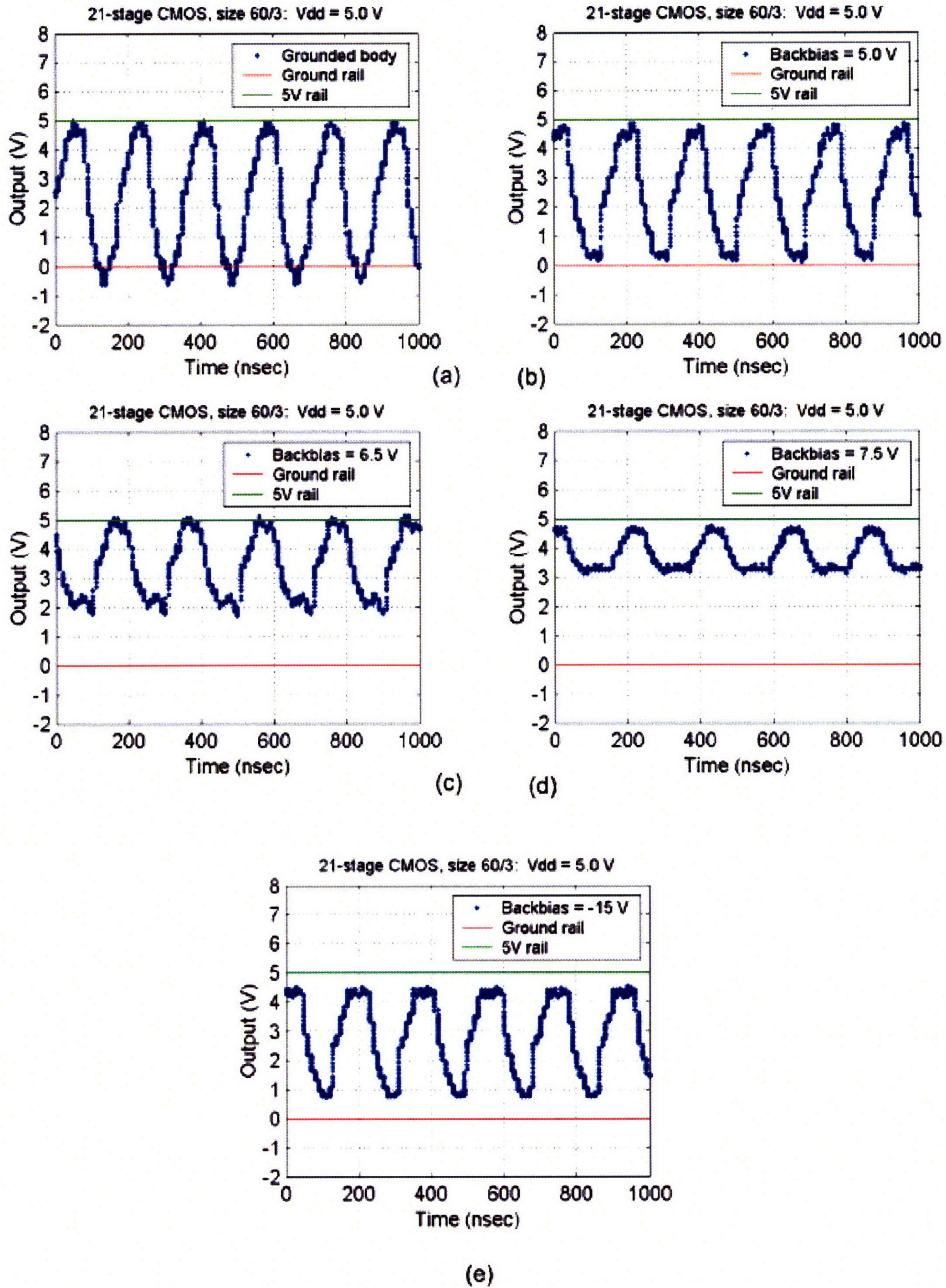


Figure D-12: 21-Stage CMOS, L = 3 μm, Vdd = +5V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) +7.5 V backbias, (e) -15 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails, and at Vdd = +5 V, V-/V+ almost resided on the rails themselves, albeit with some voltage drop from internal series resistance.

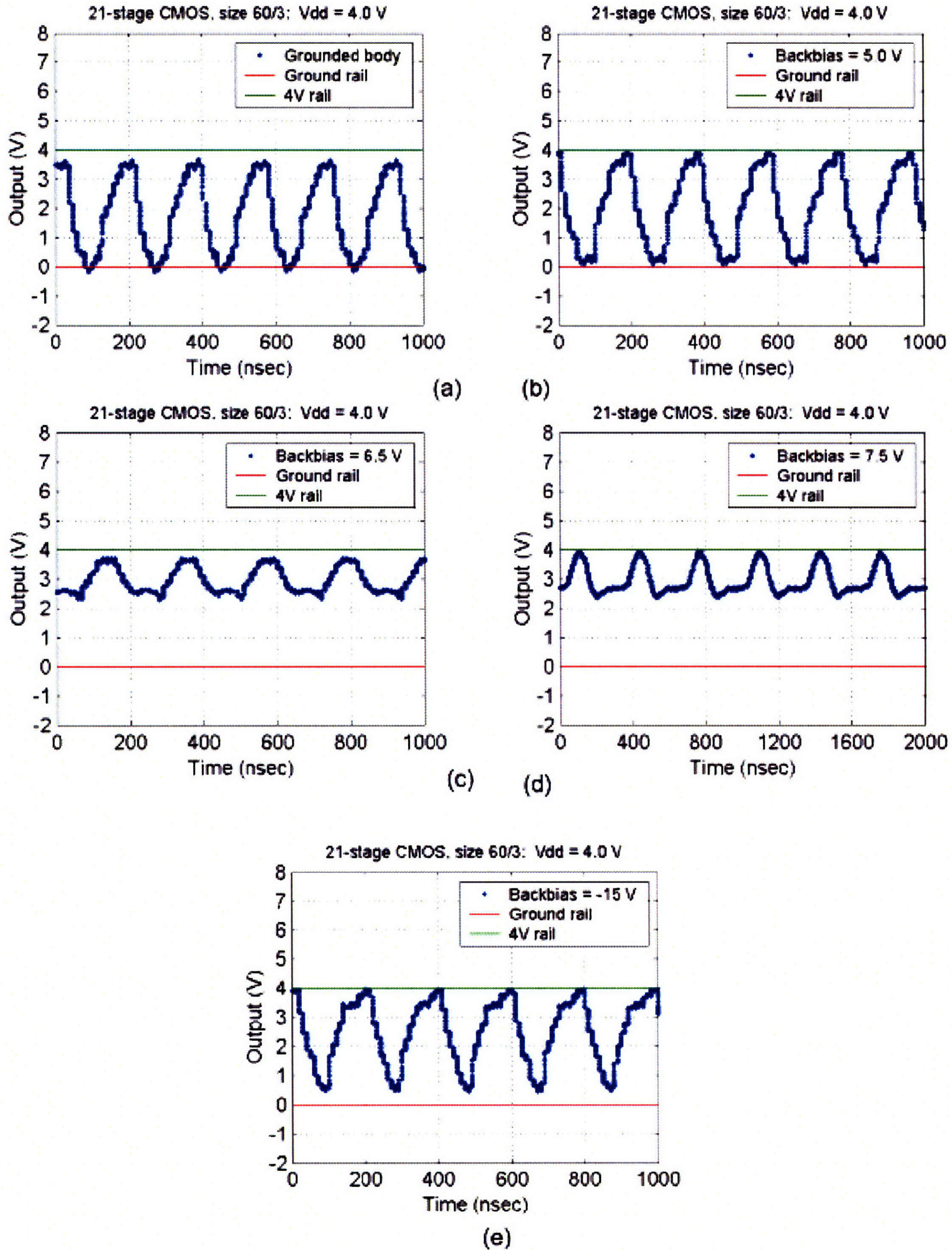


Figure D-13: 21-Stage CMOS, L = 3 μm, Vdd = +4V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) +7.5 V backbias, (e) -15 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails.

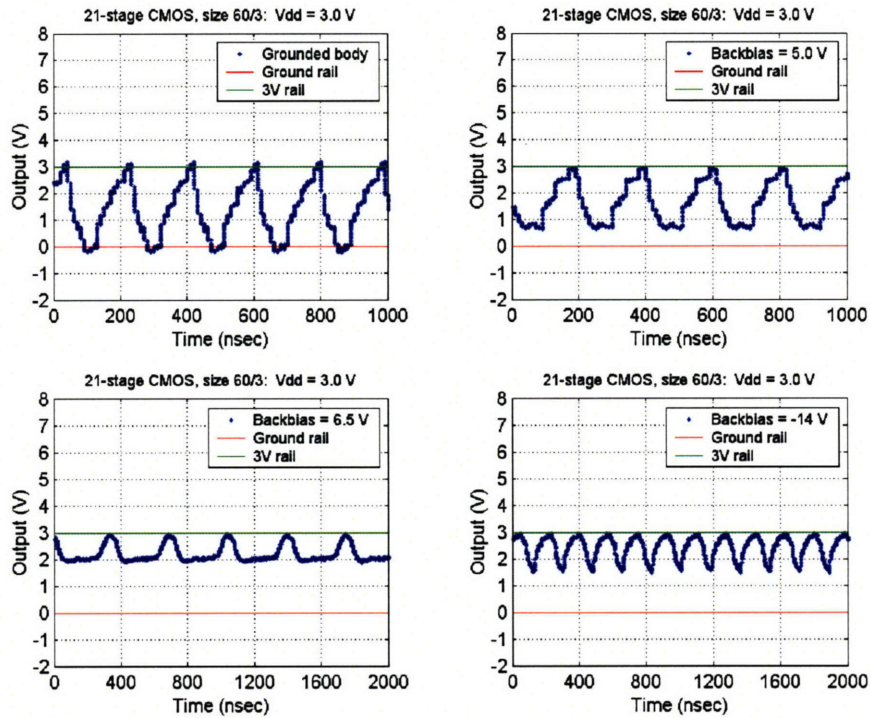


Figure D-14: 21-Stage CMOS, L = 3 μm, Vdd = +3V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +6.5 V backbias, (d) -15 V backbias. Note that at Vdd = +3 V, the PMOS devices can no longer tolerate a backbias of more than 6.5 V without severe degradation to the output signal.

D.5 Face-Face, 43-Stage CMOS Ring Oscillators, Floating-body

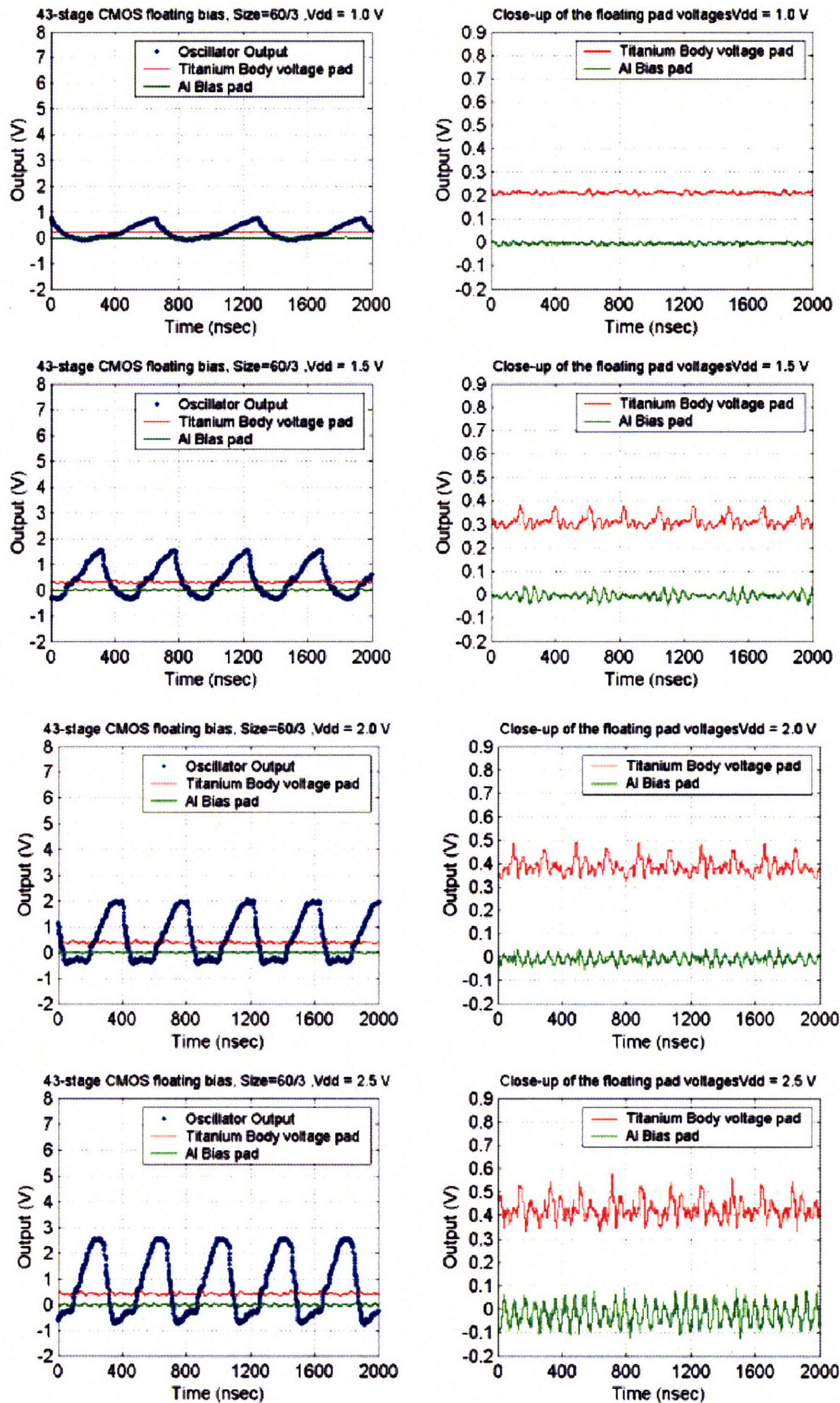


Figure D-15: 43-Stage CMOS, L = 3 μm: Vdd = 1 thru 2.5 V. Left-column plots are the signals from the 9x output buffer and the tiny traces from probing the floating Ti backgate and the "useless" Cu-Al plane's pad; right-column plots are zoomed-in traces of the aforementioned floating pads. Notice now that the Ti and Al pads were now electrically separated, thus supporting the previously mentioned notion that the 21-stage oscillator's shorted pads were probably from leftover stringers.

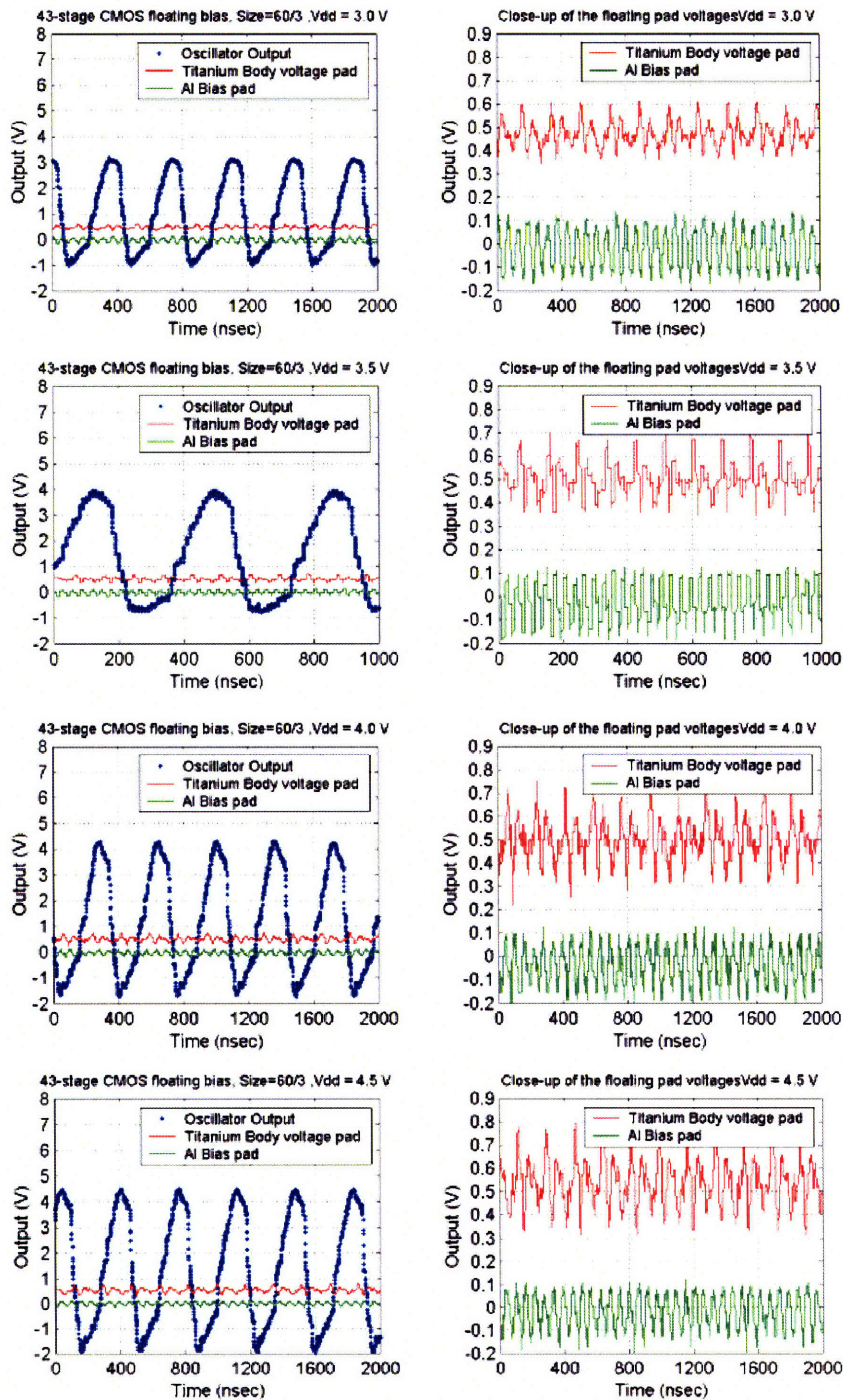


Figure D-16: 43-Stage CMOS, L = 3 μm: Vdd = 3 thru 4.5 V.

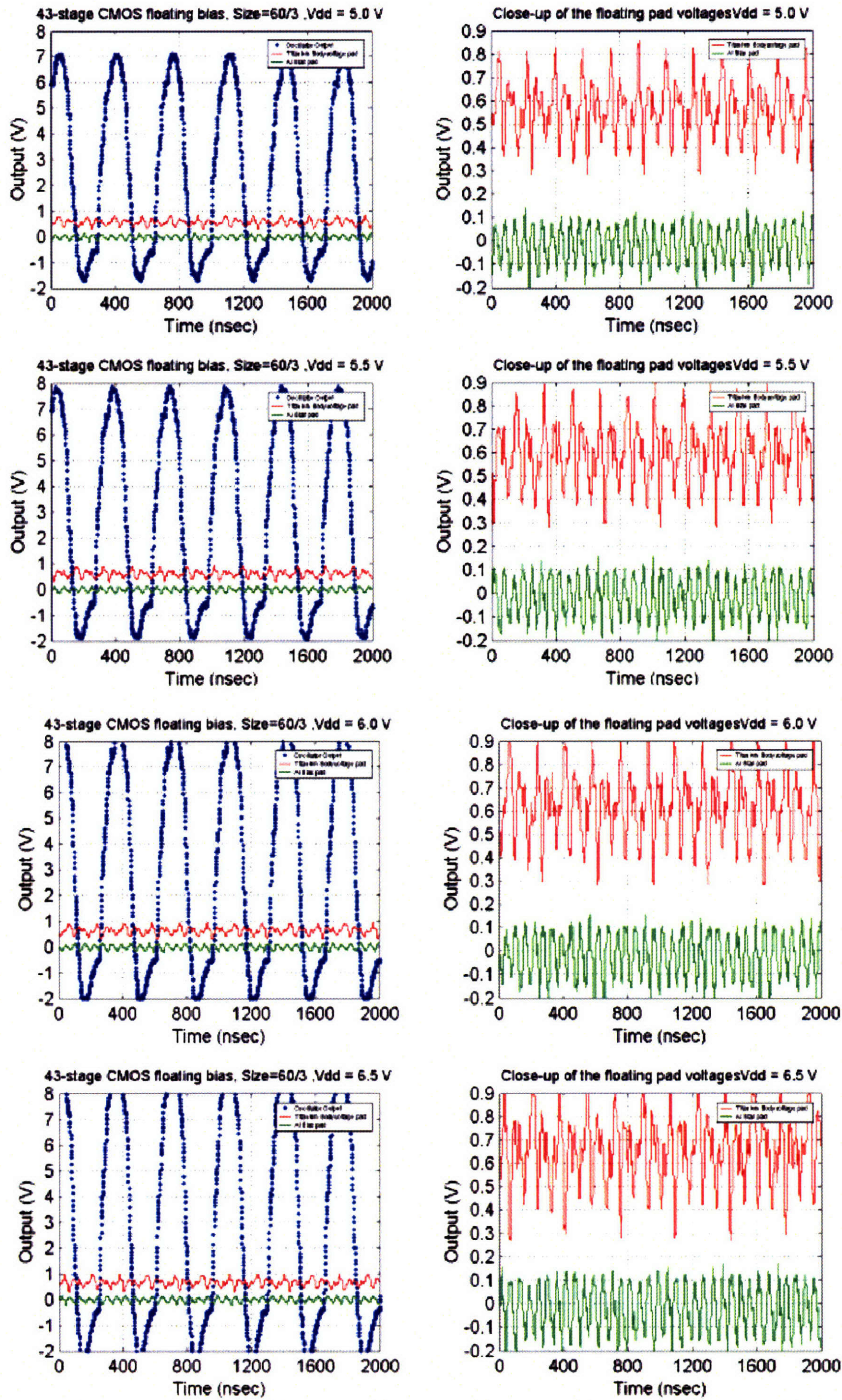
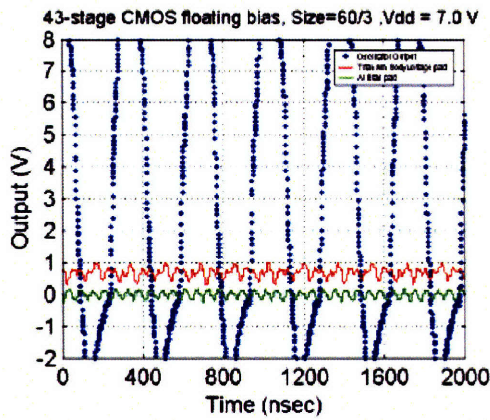
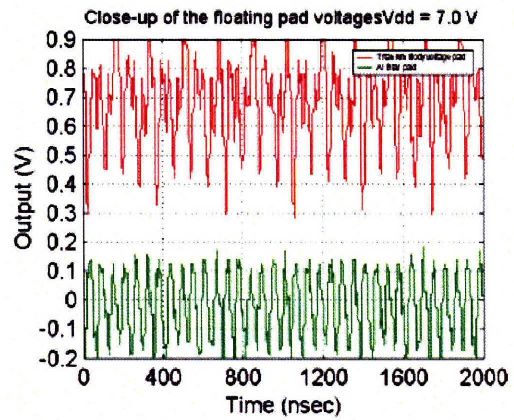
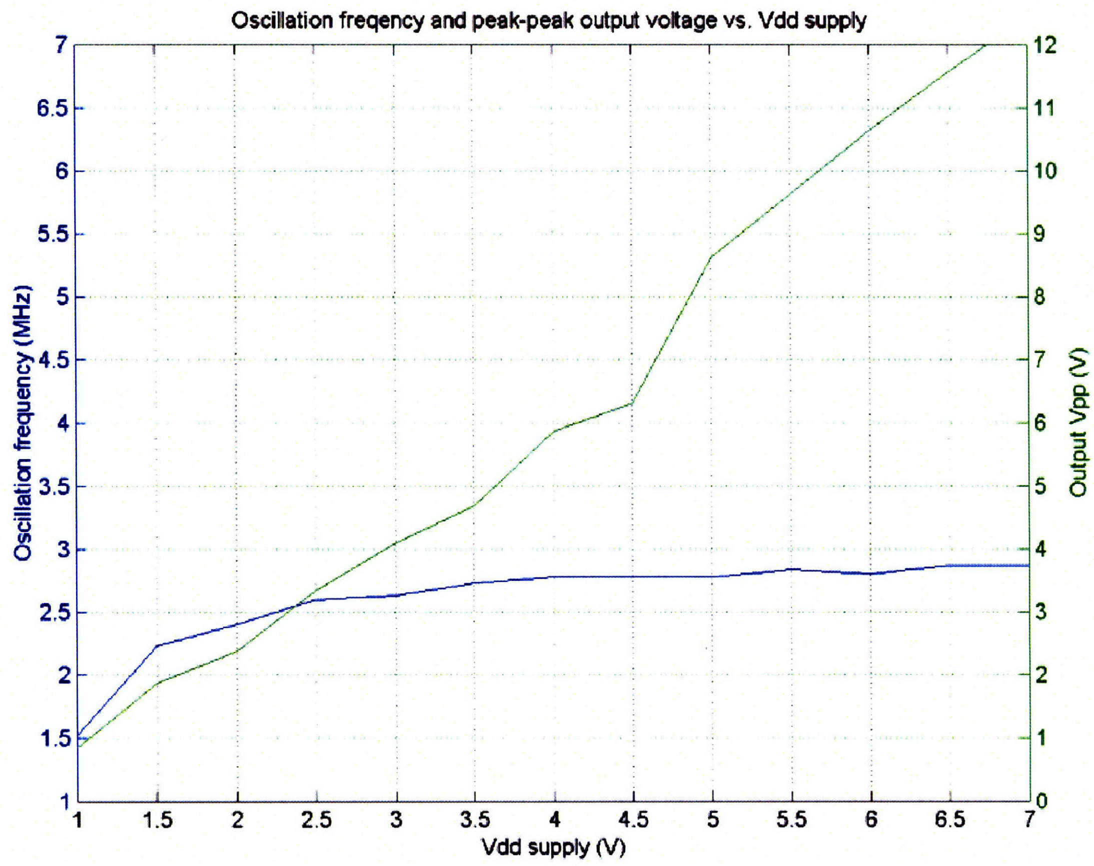


Figure D-17: 43-Stage CMOS, L = 3 μm: Vdd = 3 thru 4.5 V.





(a) (b)



(c)

Figure D-18: 43-Stage CMOS, L = 3 μm: The plot at Vdd = 7.0 is in (a). In (b), the frequency and the peak-to-peak voltage Vpp of the output was plotted as a fuction of Vdd. Note the saturation of the oscillation frequency at high Vdd's. Also note this oscillator rang approximately half the speed of the 21-stage oscillator data shown on 81.

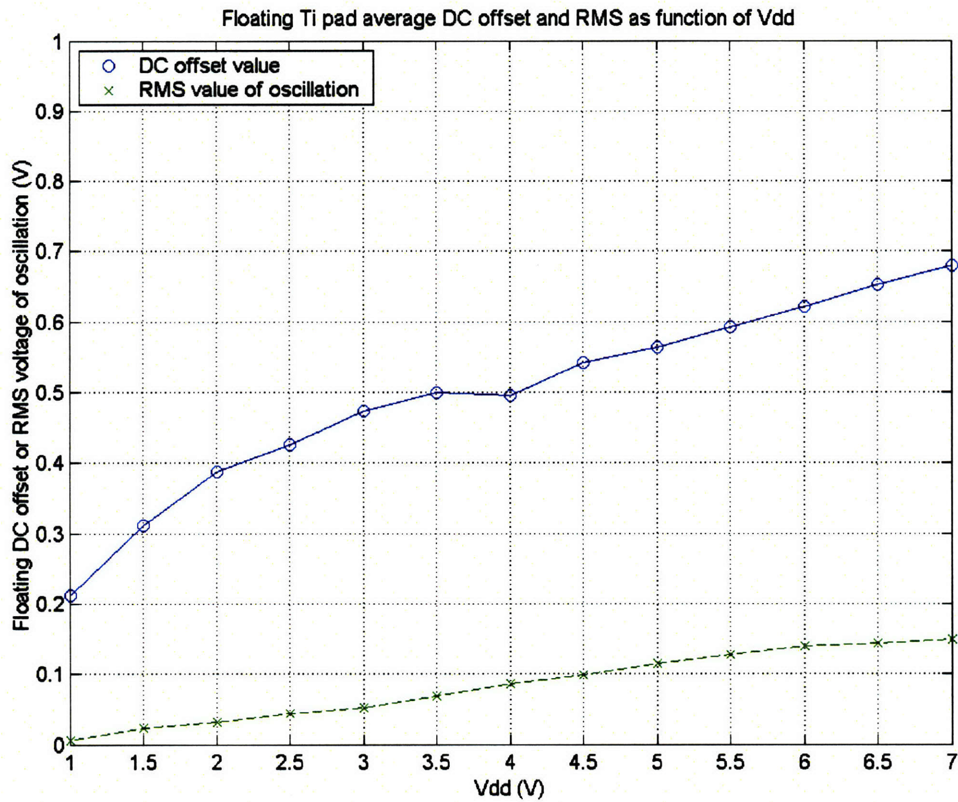


Figure D-19: 43-Stage CMOS, L = 3 μm: The DC offset and the RMS voltage of the fast oscillations coming off of the floating Ti bckgate and the useless Cu floating pads

D.6 Face-Face, 43-Stage CMOS Ring Oscillator, with Backbiasing

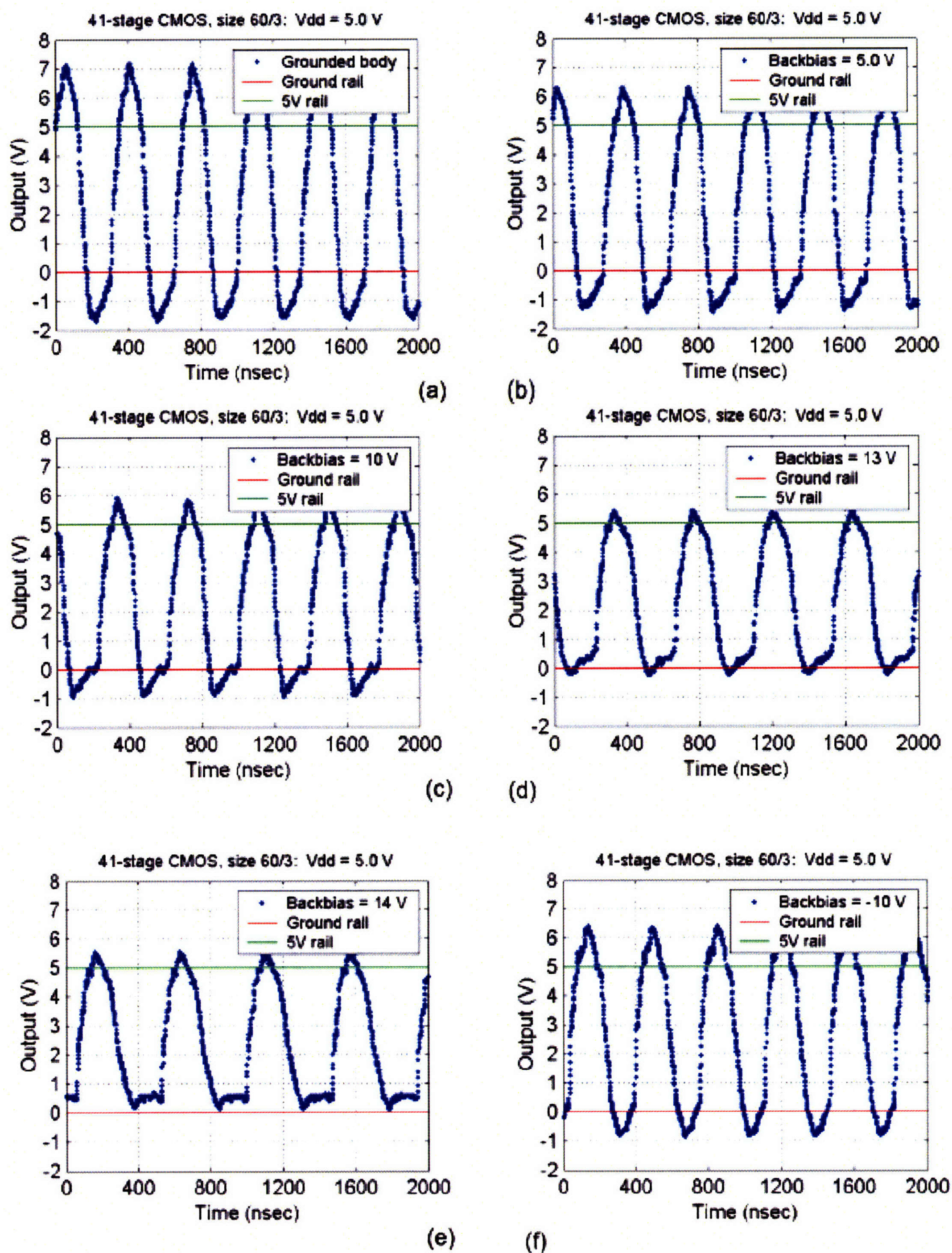


Figure D-20: 43-Stage CMOS, L = 3 μm, Vdd = +5V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +10 V backbias, (d) +13 V backbias, (e) +14 V backbias, and (f) -10 V backbias. Note that as the positive backbias increases, the V- and V+ values crawls back within the bound ground and +Vdd rails. However, unlike the 21-stage variety, it tkea +13 V of backbias to regain control of the V+/V- extrema.

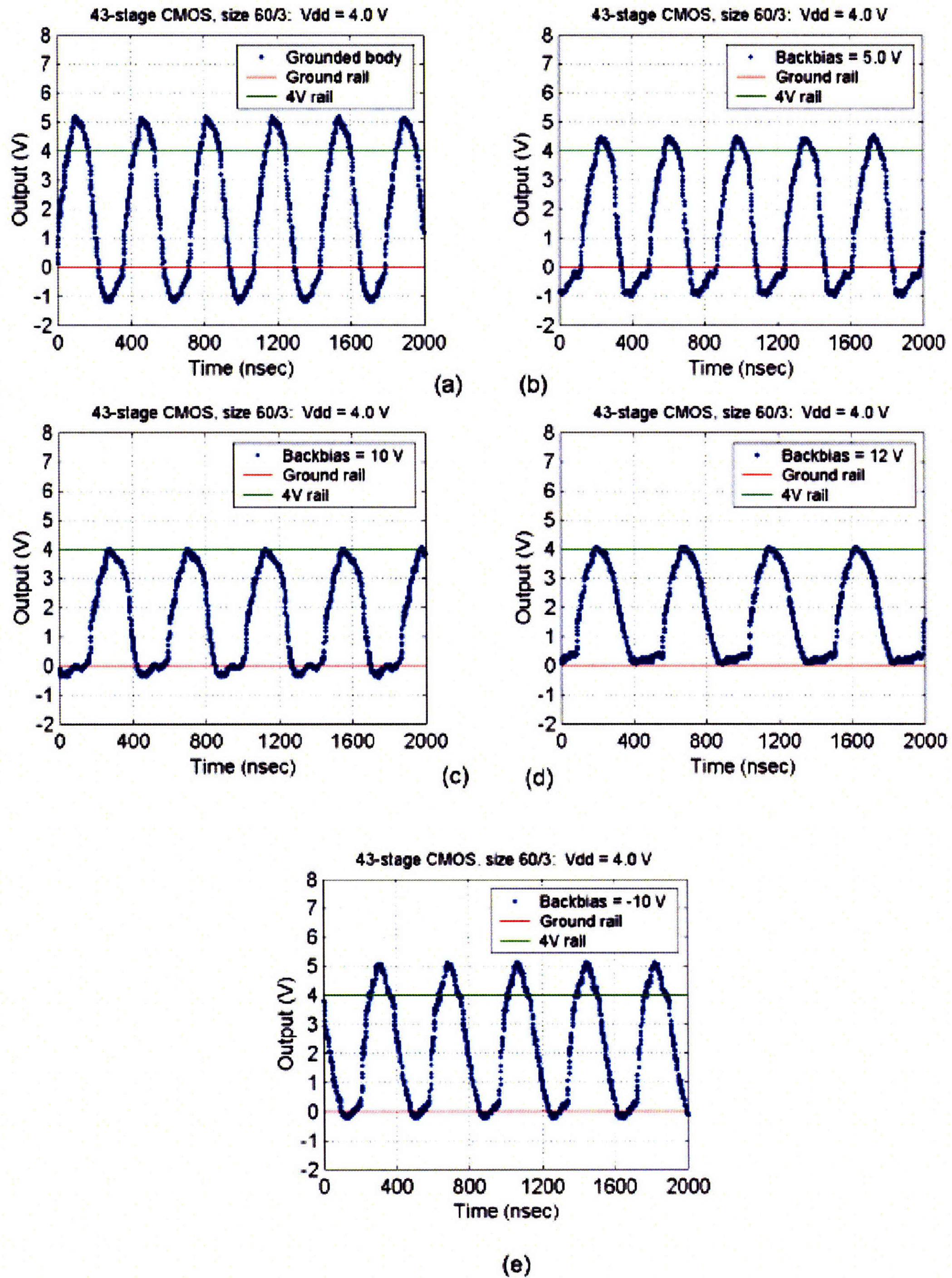


Figure D-21: 43-Stage CMOS, L = 3 μm, Vdd = +4V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +10 V backbias, (d) +12 V backbias, and (e) -10 V backbias.

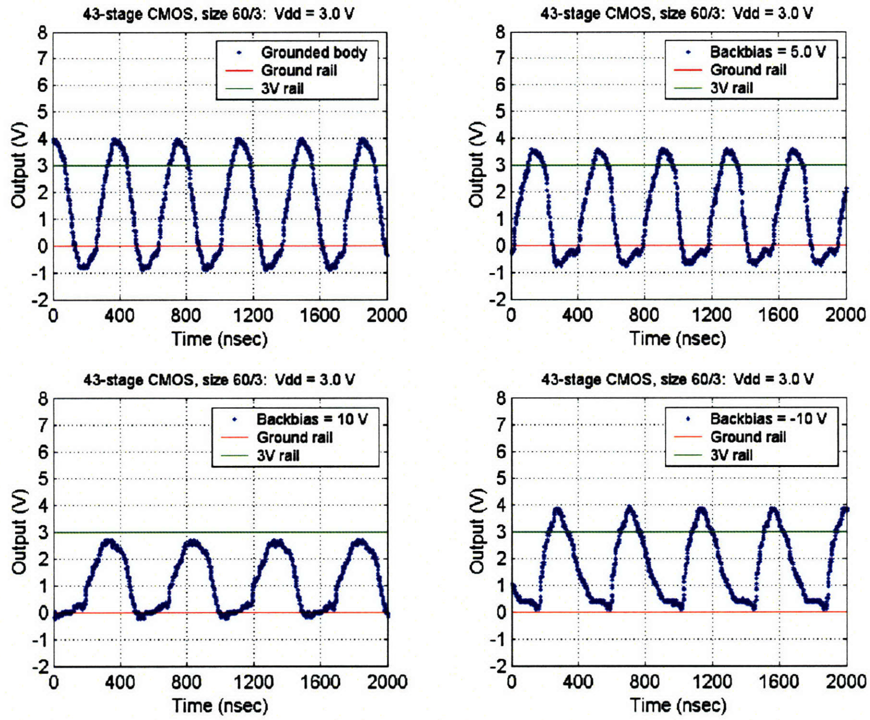


Figure D-22: 43-Stage CMOS, L = 3 μm, Vdd = +3V, with: (a) Grounded backgate, (b) +5 V backbias, (c) +10 V backbias, and (d) -10 V backbias.

D.7 2-D NMOS-only Ring Oscillators

D.7.1 Table of Results

Data considered within main text →

Vdd (V)	2-D NMOS Rings	Simulated (MHz)	Unbonded (MHz)	Face-face (MHz)	Face-back, 10 μm Al release (MHz)	Face-back, 20 μmAl (MHz)
5.0	80/1 – 5/1	49.77	5.263	5.319	5.376	5.682
	80/1 – 10/1	76.86	5.814	6.098	6.024	5.952
	60/1 – 10/1	91.32	8.929	9.259	8.928	X
	60/1 – 20/1	130.03	10.00	10.00	10.00	X
4.0	80/1 – 5/1	39.06	5.154	5.682	5.434	5.882
	80/1 – 10/1	61.72	5.814	6.250	6.098	6.097
	60/1 – 10/1	73.52	8.475	9.259	8.772	X
	60/1 – 20/1	105.26	9.615	10.00	10.00	X
3.0	80/1 – 5/1	25.84	5.319	6.098	5.681	3.144
	80/1 – 10/1	41.56	6.024	6.250	6.250	6.329
	60/1 – 10/1	49.87	8.620	9.259	8.772	X
	60/1 – 20/1	74.23	9.615	10.20	10.00	X

Figure D-23: A complete table of results for teh 2-D NMOS-only oscillator biased at Vdd = +3, +4, and +5 V.

D.7.2 2-D NMOS-only Ring Oscillators, Vdd = 3 V

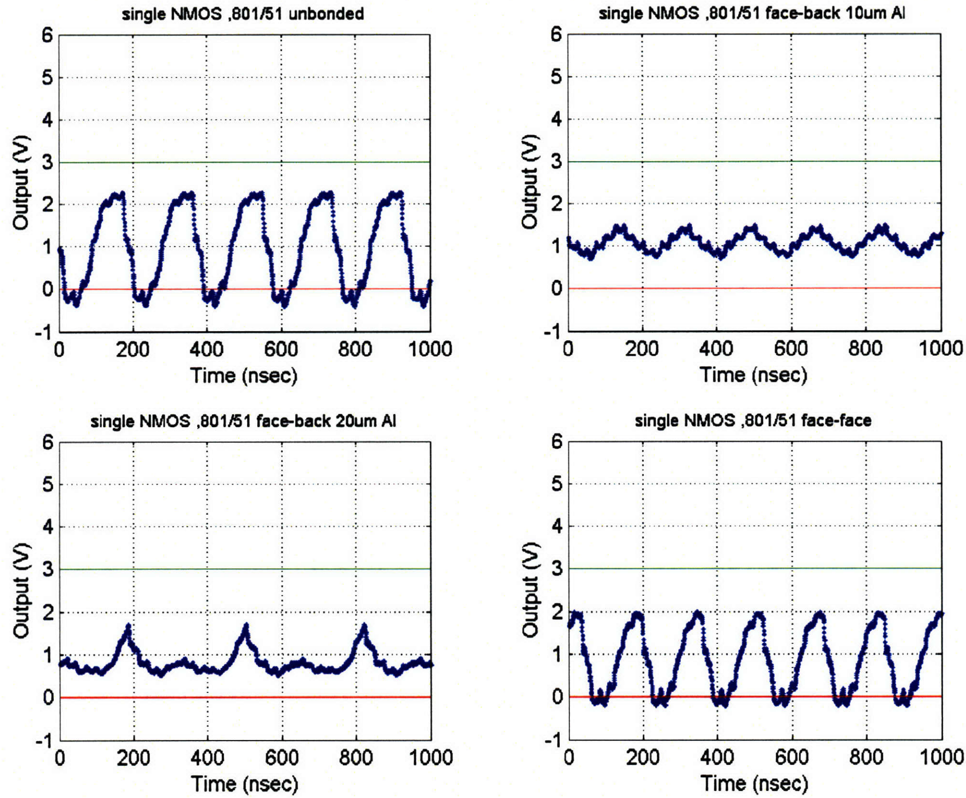


Figure D-24: 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

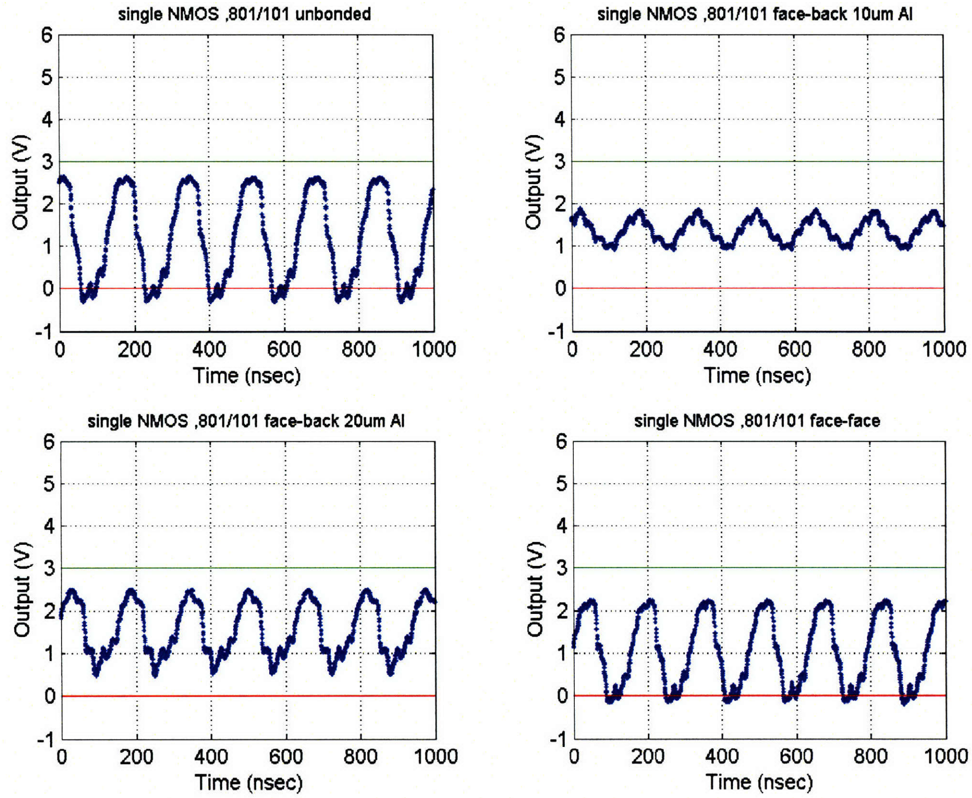


Figure D-25: 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

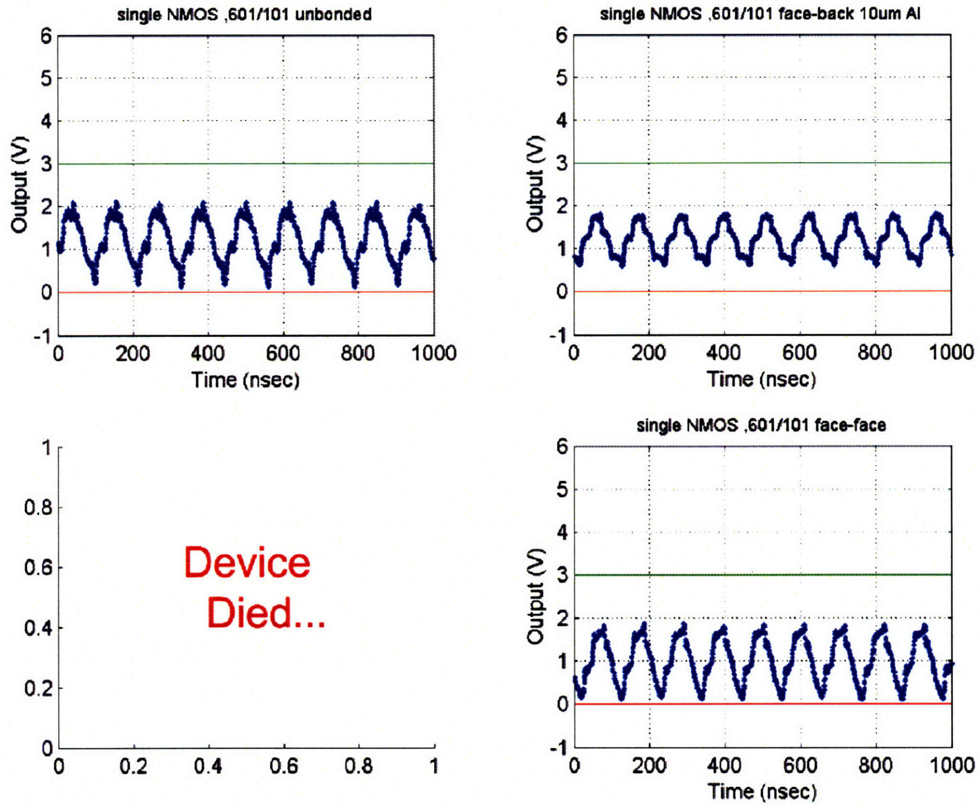


Figure D-26: 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

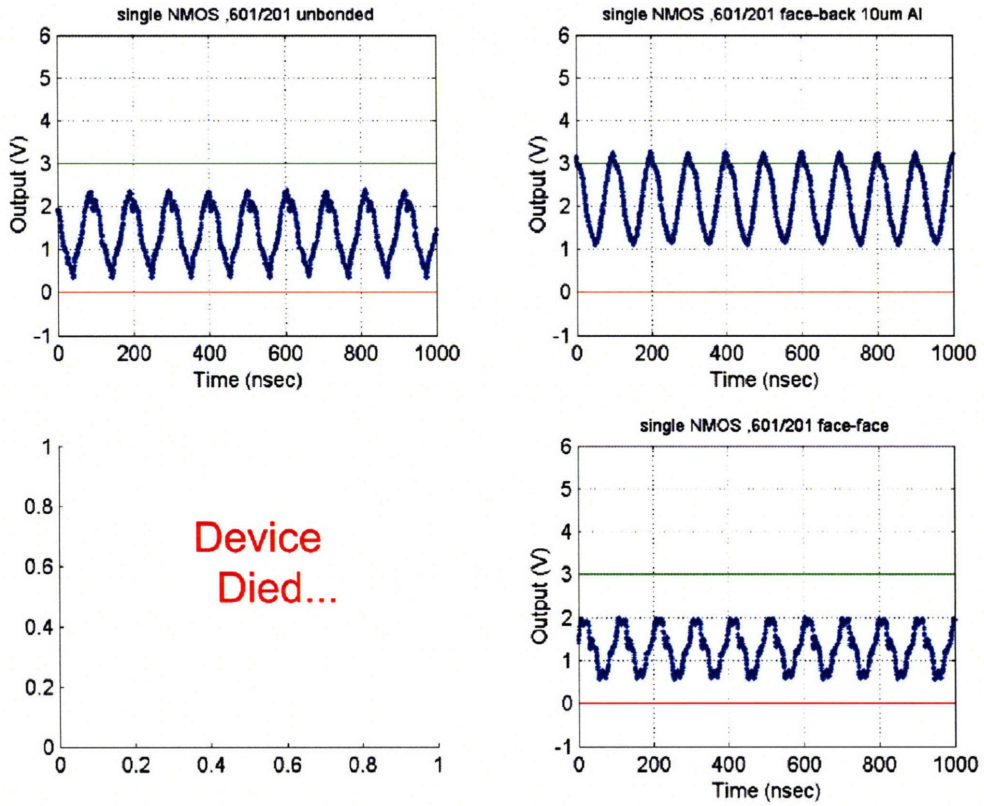


Figure D-27: 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +3 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

D.7.3 2-D NMOS-only Ring Oscillators, Vdd = 4 V

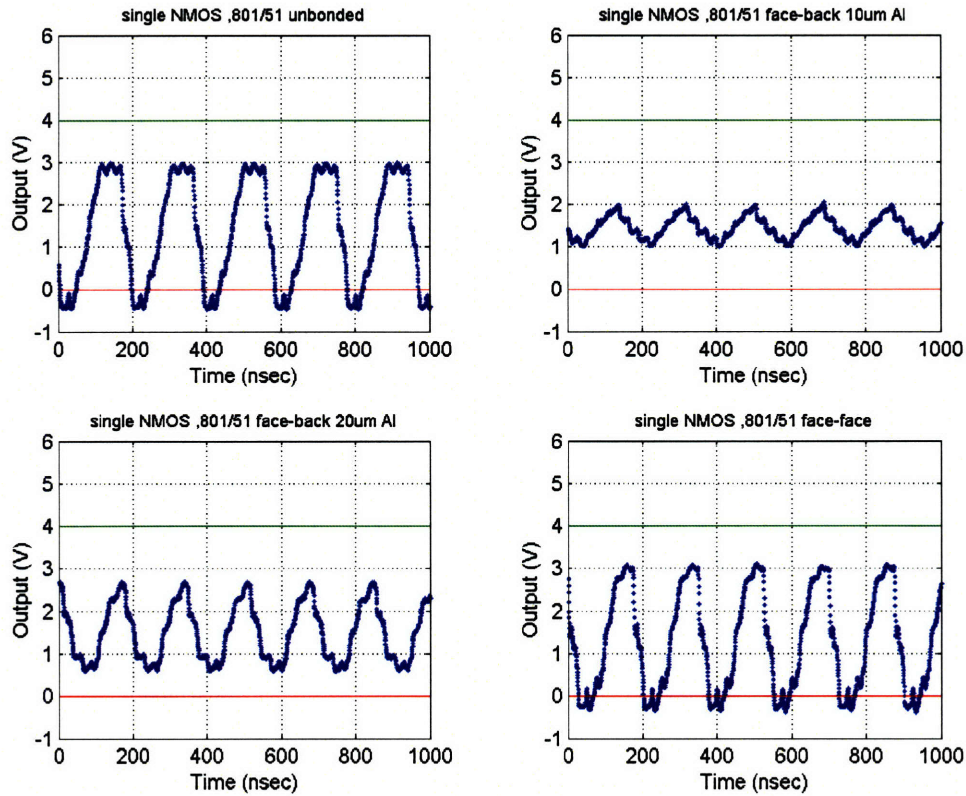


Figure D-28: 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

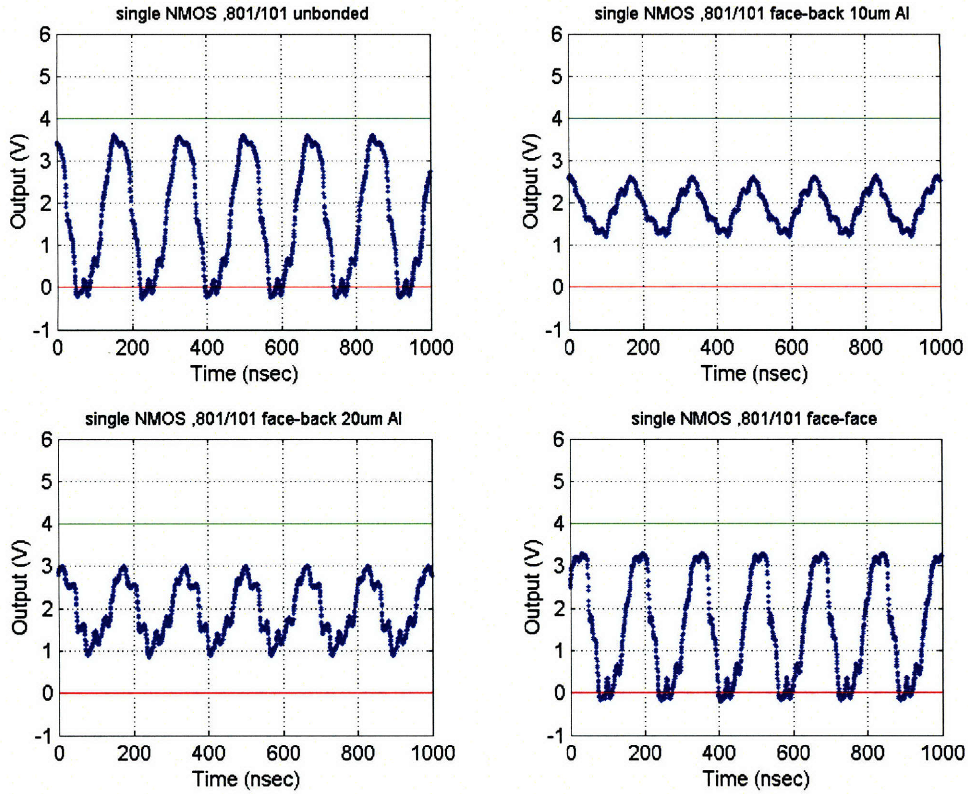


Figure D-29: 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

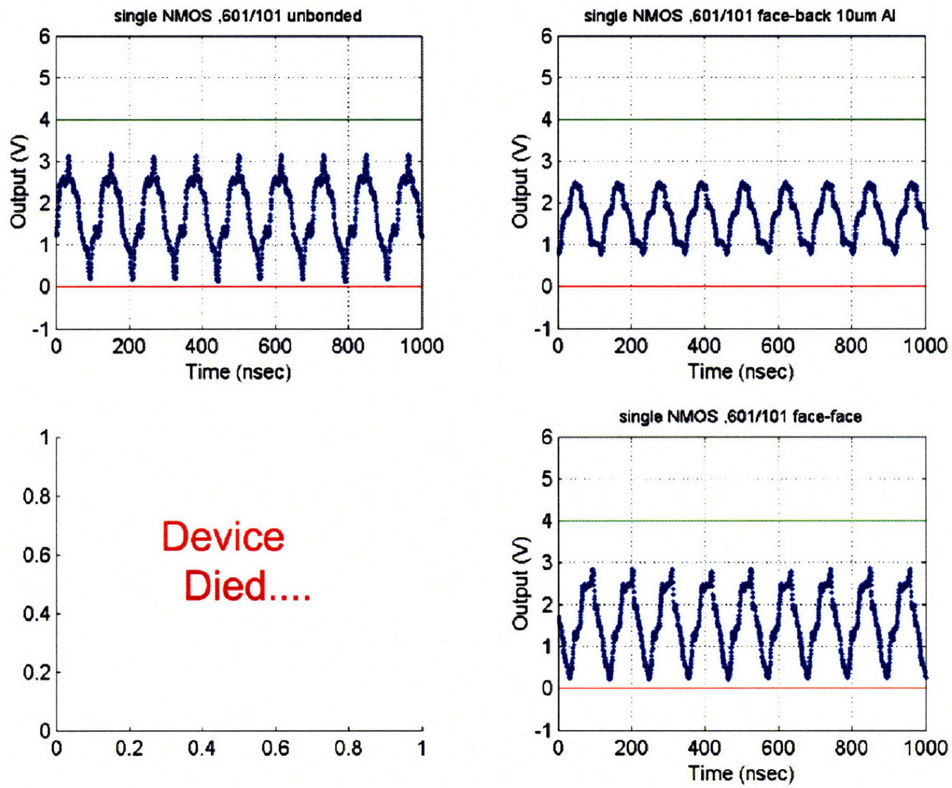


Figure D-30: 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

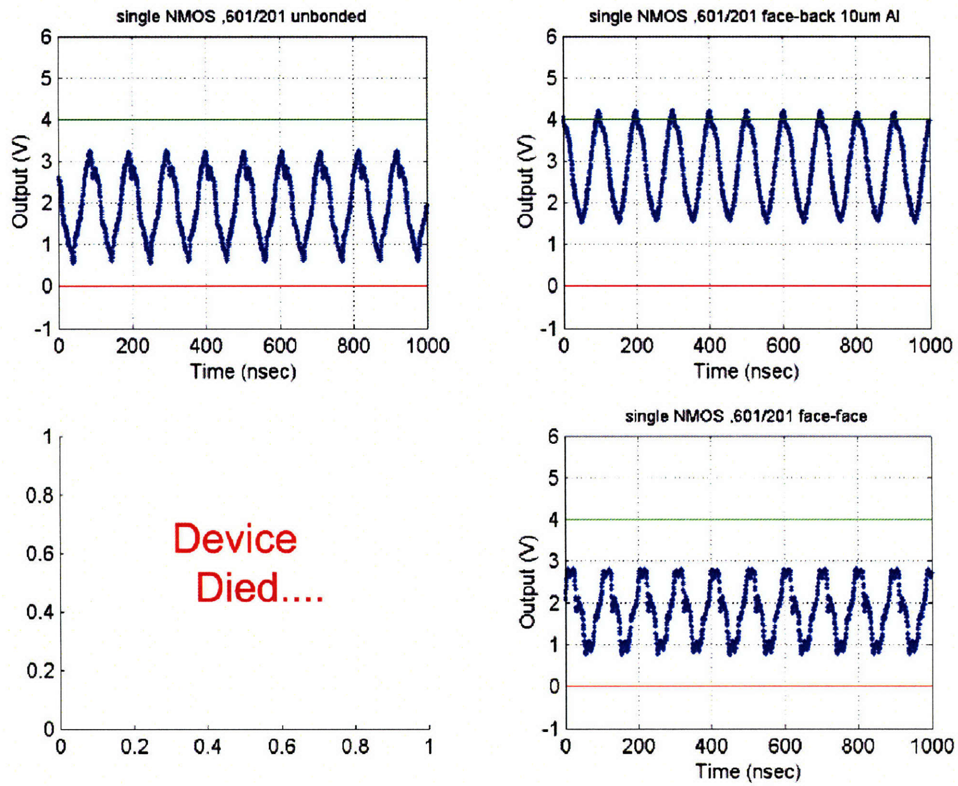


Figure D-31: 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +4 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

D.7.4 2-D NMOS-only Ring Oscillators, Vdd = 5 V

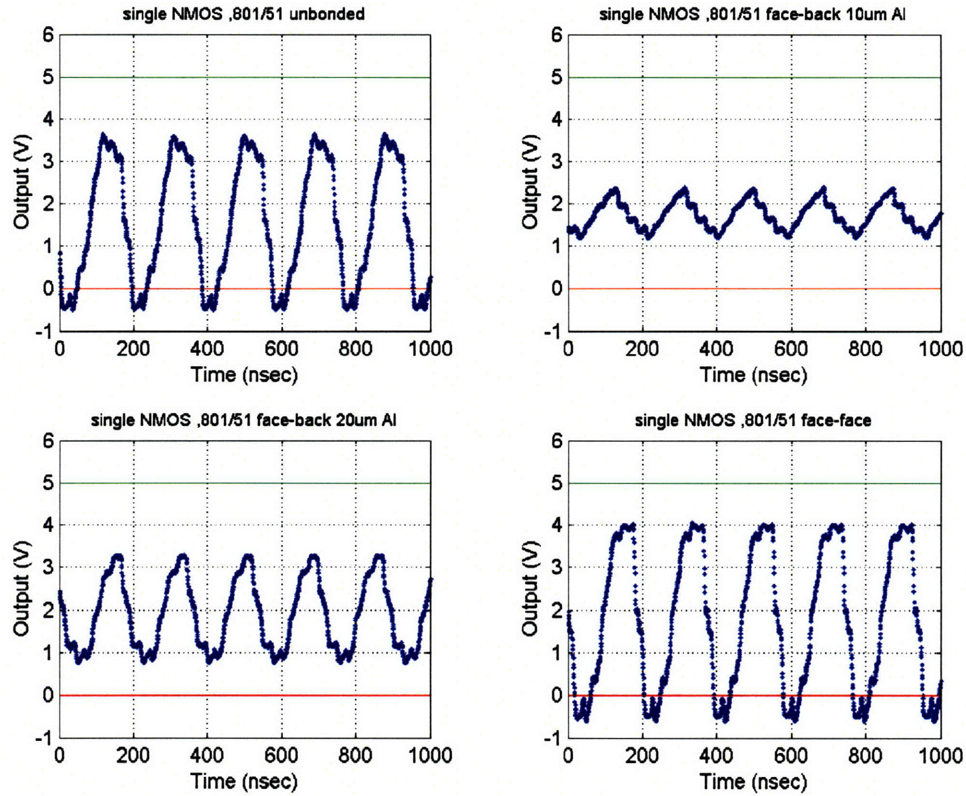


Figure D-32: 2-D NMOS-only, 80/1 - 5/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

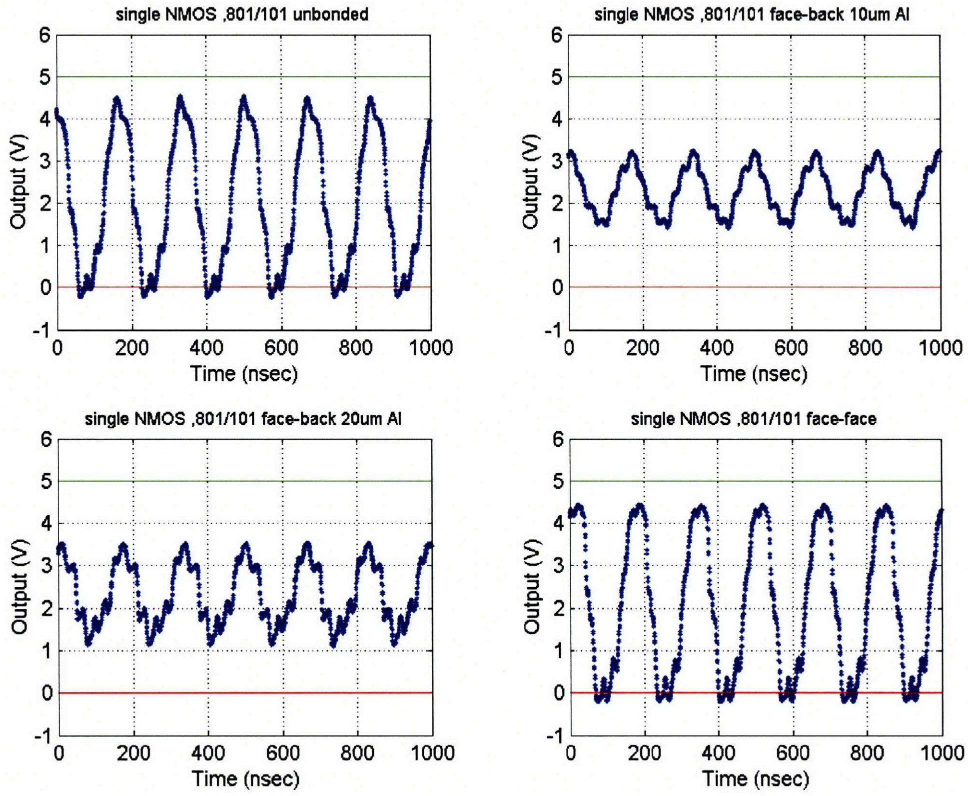


Figure D-33: 2-D NMOS-only, 80/1 - 10/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

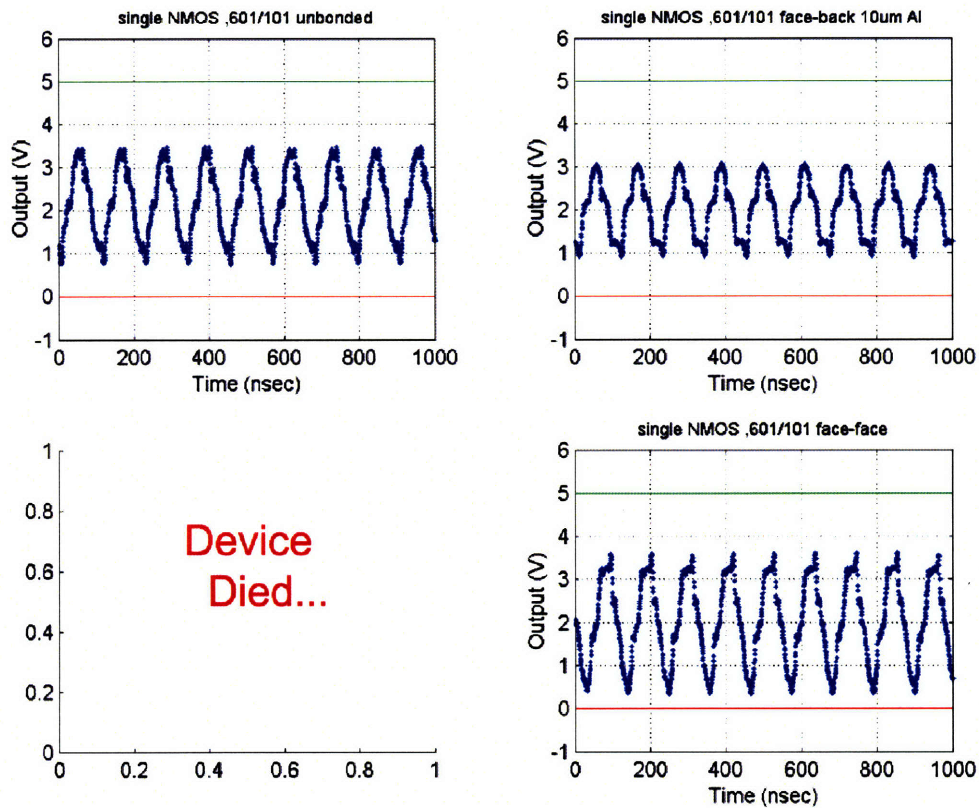


Figure D-34: 2-D NMOS-only, 60/1 - 10/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

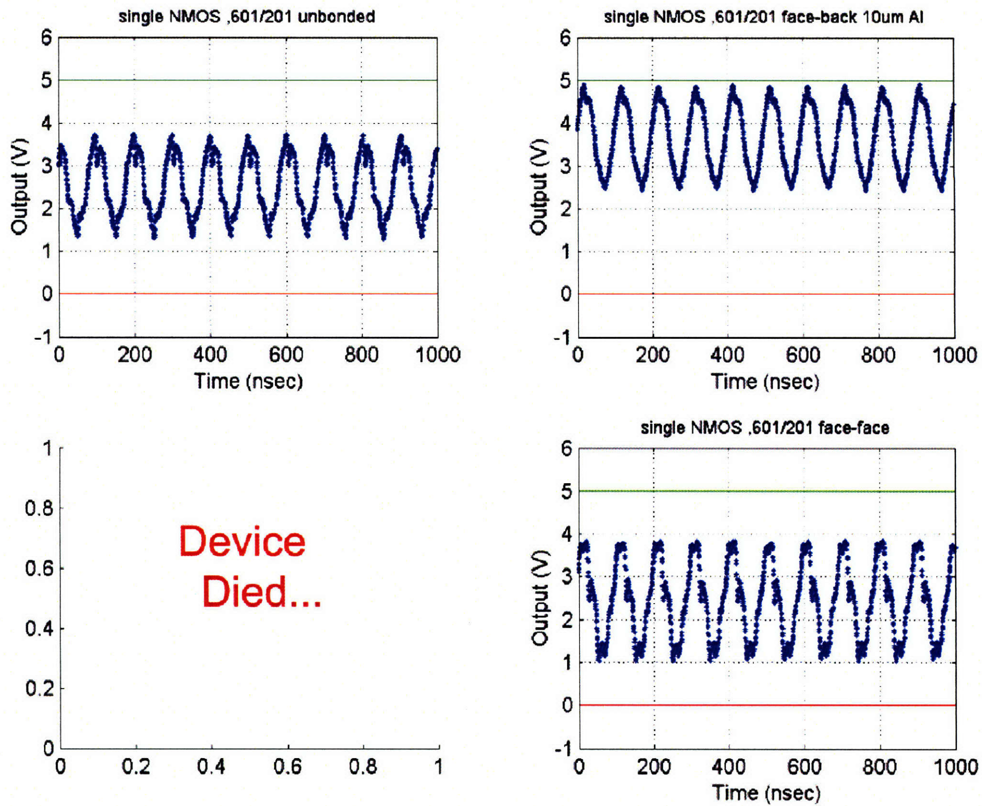


Figure D-35: 2-D NMOS-only, 60/1 - 20/1 ring oscillator powered at Vdd = +5 V, from an unbonded NMOS-SOI wafer (a), a face-back bonded, 10 μm Al released sample in (b), a face-back bonded, 20 μm Al released sample in (c) that died during processing, and a face-face bonded sample in (d)

Appendix E

A Heat Transfer Primer for EE's

E.1 Thermal-to-Electrical Duality: The Heat Equation vs. the Poisson Equation

To gain a more in-depth understanding of the heat transfer discussions in later sections, it is important for the reader (assuming an EE background) to get re-acquainted with the Heat Equation and the fundamentals of heat transfer. The best way to approach this exercise is to appreciate the duality between the Heat Equation and the Poisson Equation in EE. We will use Figure E-1 extensively in the discussions below.

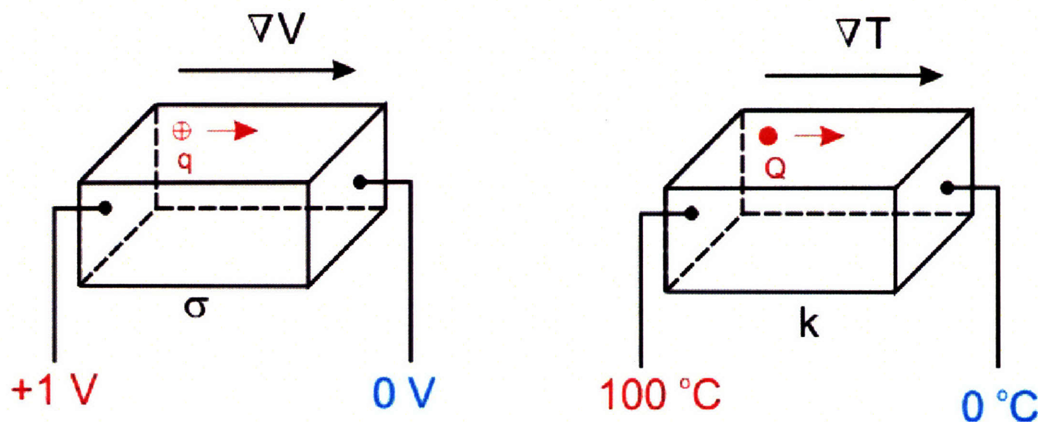


Figure E-1: Duality between electricity and heat. The left cartoon shows a charge q moving through a slab of material with electrical conduct under a potential difference of V. The right cartoon shows an unit of heat Q moving through a slab with a thermal conductivity of k under a temperature difference of 100 °C.

E.2 Ohm's Law and its Thermal Duality: Fourier's Law

To start with, the Poisson Equation can be built from Ohm's Law, or I = V/R. Removing the cross-sectional and length dependence from from the left cartoon in Figure E-1, we can rewrite Ohm's law in the continuity formulation:

$$\mathbf{J_e} = \sigma \mathbf{E} \tag{E.1}$$

where bold-faced variables denote vectorized quantities, $\mathbf{J}_e$ is the electric current density, σ is the electrical conductivity of the material, and **E** is the electric field. For pedagogical reasons, let's again rewrite and rename the variables of this equation using $\mathbf{E} = -\nabla V$:

$$\mathbf{J_e} = -\sigma \nabla \mathbf{V} \tag{E.2}$$

where $\mathbf{J}_e$ is now called the ***charge flux vector*** and ∇V is the ***electric potential gradient vector***. The physical interpretation of Equation E.2 is that an electric charge q will move if an electric potential difference V is present somewhere in space, and in addition, the direction of movement will concide with the ***steepest gradient*** of the potential difference [1]. Furthermore, the magnitude of the charge movement depends on the magnitude of the electrical conductivity σ of the material and the magnitude of the gradient ∇V.

On the same token, heat transfer can also be explained in the same manner as in Ohm's Law, and the corresponding equation in the thermal domain is called Fourier's Law. To begin, let's refer back to the right cartoon in Figure E-1. If a block of material was to have a 100 °C difference between two opposing surfaces, then units of heat "Q' (in Joules, not Watts) will travel in the direction of the steepest temperature gradient with a magnitude proportional to both ∇T and the thermal conductivity of the material k. Therefore, Ohm's Law in Eq. E.2 can be rewritten into Fourier's Law in the thermal domain:

$$\mathbf{J_Q} = -\mathbf{k} \nabla \mathbf{T} \tag{E.3}$$

where $\mathbf{J}_Q$ is the heat current density (thermal energy rate per unit area, or [Joules/sec/area], or [Watts/m^2], or properly known as the *heat flux vector*), k is thermal conductivity in [°C/W-m], and ∇T is the temperature gradient in [°C/m]. By comparing Ohm's Law and Fourier's Law, the duality relationships between them are:

[1] This is true only if the electrical conductivity was constant for all directions. If a non-elementary conductivity tensor exists, then the direction of charge travel will obvisously deviate from the direction of ∇V.

$$Duality = \begin{cases} q \leftrightarrow Q, & \text{electric charge (Coulombs): heat (Joules)} \\ i = \dot{q} \leftrightarrow \dot{Q}, & \text{charge flow (amps): heat flow (watts)} \\ \sigma \leftrightarrow k, & \text{electrical conductivity tensor : thermal conductivity tensor} \\ \frac{l}{\sigma A} \leftrightarrow \frac{l}{kA}, & \text{electrical resistance : thermal resistance} \\ V \leftrightarrow T, & \text{equipotential contours : isothermal contours} \\ \mathbf{E} = -\nabla \mathbf{V} \leftrightarrow -\nabla \mathbf{T}, & \text{electric field lines : temperature gradient lines} \\ \mathbf{J_e} \leftrightarrow \mathbf{J_Q}, & \text{charge flux lines : heat flux lines} \end{cases}$$

E.3 Poisson equation and its duality: The Heat equation

To derive Poisson's Equation, we start from Gauss's Law - the governing equation that relates the electric field to its source charge, and plug in the relationship $\mathbf{E} = -\nabla\mathrm{V}$:

$$\begin{aligned} \nabla \cdot (\epsilon \mathbf{E}) &= \rho \\ \nabla \cdot (-\epsilon \nabla V) &= \rho \end{aligned} \tag{E.4}$$

If the permitivity of the material is isotropic in nature, we can take ϵ as a constant and move it outside of the divergence operator. The resulting equation is Poisson's Equation:

$$\begin{aligned} -\epsilon \nabla^2 V &= \rho \\ &= SourceDensity_{E-field} \end{aligned} \tag{E.5}$$

For the Heat equation, one can do the same exercise starting from Fourier's Law:

$$\begin{aligned} \nabla \cdot (\mathbf{J_Q}) &= SourceDensity_{heat} \\ \nabla \cdot (-k \nabla T) &= SourceDensity_{heat} \end{aligned} \tag{E.6}$$

If the thermal conductivity of the material is isotropic in nature, we can take k as a constant and move it outisde of the divergence operator. The resulting equation is the Heat equation ***at steady-state***:

$$-k \nabla^2 T = SourceDensity_{heat} \tag{E.7}$$

where the heat source density is usually consist of work being done on the system, and in the case of microelectronic circuits, $SourceDensity_{heat}$ = Work density done on system = Power dissipation density from Joule heating, or $I^2R/(spatialmetric)$. Comparing the Heat equation and Poisson's equation, one can reaffirm that fact that if one was given a specific charge distribution $\rho(\mathbf{r})$ in space and a geometrically identical power source distribution q(r), then the resulting E-field lines derived from Poisson's equation should ge-

ometrically coincide with the heat flux lines $k\nabla T$ if ϵ and k tensors were isomorphic. The significance here is that we can solve a Heat equation problem with relative ease if we already know the Poisson solution of a geometrically equivalent problem. An exaggerated case of this can be seen in Figure E-2.

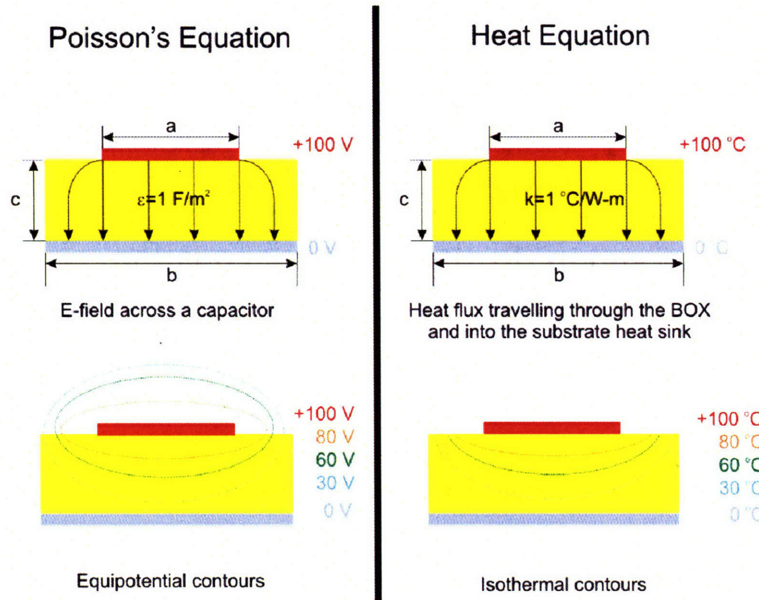


Figure E-2: Equivalence between Poisson and Heat equations. If the dimensions a, b, and c were all equal, then the electric field **E** and the heat flux lines $k\nabla T$ would be geometrically and numerically be identical if $k = \sigma = 1$. Moreover, the equipotential and isothermal contours would also be equal.

Now that we have introduced the duality between the Heat and Poisson's equations, it's time for us to dive into an in-depth analysis into heat transfer design in microelectronic circuits.

E.4 Spreading Resistance

E.4.1 Mathematical Introduction and Physical Interpretation

Within the thermal engineering chapter of the thesis body, there was a lengthy dialogue (through FEM simulations) on convincing the reader that the insertion of a Cu thermal plane into a field of on-chip hotspots will increase the divergence of the heat flux and thus reduce the magnitude of the thermal gradient. Basically, the action of spreading the heat flux has effectively increased the thermal cross-sectional area associated with the flux lines, and because of it, the overall thermal resistance and the maximum temperature gradient of the 3-D structure was decreased. The cause-and-effect of "heat flux spreading" and "thermal resistance reduction" was collectively coined as the *spreading resistance* by thermal engineers as early as the 1970s [59]. Mathematically, following the treatment of Yovanovich, et. al. [60] and Ellison [61], consider the situation in Figure E-3, where we have:

- A heat source of arbitrary shape with a cross-sectional area of A_s, with zero thickness, a power density generation of $\dot{q}$(x,y) [W/m^2], and a total power dissipation of $\dot{Q}$(x,y) = $\dot{q}A_s$ [W]
- A material of conduction with thermal conductivity k, thickness of t, and a conduction cross-sectional area of A = ab.

In lieu of Fourier's law, one can write a relationship between the total power dissipation $\dot{Q}$ and the total system temperature gradient ∇T :

$$\begin{aligned} \dot{Q} &= -\frac{\nabla T}{R_{total}} \\ &= \frac{T_{source}(x,y,z) - T_{sink}(x,y,z)}{R_{total}} \end{aligned} \tag{E.8}$$

where the total thermal resistance R_{total} can be written as the sum of the 1-D rectangular resistance R_{1D} and the the spreading resistance R_s:

$$R_{total} = R_s + R_{1D} \tag{E.9}$$

where

$$R_{1D} = \frac{t}{kA} \tag{E.10}$$

The physical interpretation here is that if the heat source can be made smal when compared to the area in which heat conduction occurs ($A_s << A$), then the total resistance R_{total} can be made small. Furthermore, through FEM simulations, we saw that if one adds a Cu plane within the conduction region, then it has the effect of drastically reducing the value of R_s (by flux spreading) while adding only a tiny increase in R_{1D}; hence, by adding a Cu thermal plane, we decrease the overall thermal resistance R_{total} even further.

E.4.2 Further Mathematical Treatment

Now, continuing with our mathematical treatment, if the only input of the system were the heating element area A_s and the power generation density q, and also if the output variable was the spreading resistance R_s, then Equation E.8 can be rewritten as (for a simpler case where we set impose a Dirichlet condition on the heat sink, or T_{sink}(x,y,z) = T_o:

$$R_s = \frac{T_{source}(x,y,z) - T_o}{\dot{q}(x,y) \cdot A_s} - R_{1D} \tag{E.11}$$

where the spatial dependence of the source temperature isotherm T_{source} can be solved if the spatial dependce of $\dot{q}$(x,y) was plugged into the source-free heat equation (aka. Laplace's Eq):

$$-k\nabla^2 T(x,y,z) = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = 0 \tag{E.12}$$

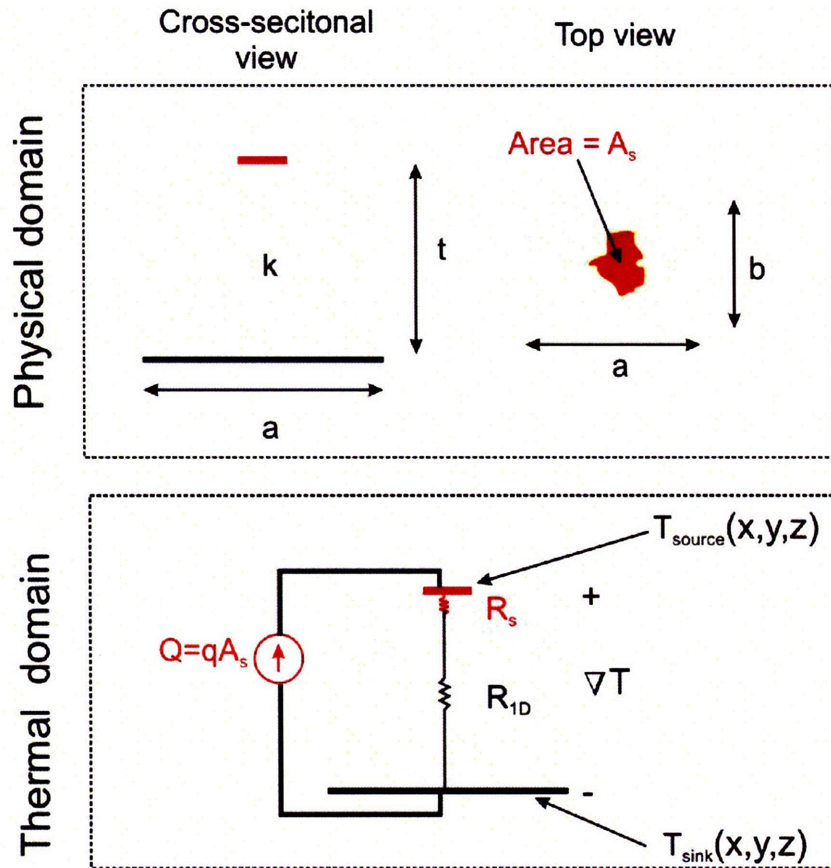


Figure E-3: Setup of the spreading resistance problem. Material of conduction has conductivity k, thickness t, and cross-sectional area of A. Material of heating has cross-sectional area A_s, effective thickness of zero, power density generation of $\dot{q}$(x,y), and power dissipation generation of $\dot{Q} = \dot{q}(x,y) \cdot A_s$.

with an upper isoflux boundary condition of:

$$k\frac{\partial T}{\partial z} = -\dot{q}(x,y) \tag{E.13}$$

So, Equation E.11 is the general form for the spreading resistance inside a rectangular block, composing of a single material with an isothermal heat sink temperature of T_o [2] It may appear innocuous, but I assure you it's not, for if the heat source shape is neither rectangular nor circular, in other words, if $\dot{q}$(x,y) is a complicated function, then both the Laplace equation solution for T_{source} and the expression for R_s will be very nonlinear. Numerous literature have dealt with the thermal resistance when $\dot{q}$(x,y) are of different shapes [62], general shape factors within the contex of Laplace's Equation [63], and thermal resistance where the conduction medium was multi-layered and the geometry of $\dot{q}$(x,y) was complicated [60, 61, 64,

[2] As a sanity check, if the heat source area A_s equals the entire cross-sectional area A, then A_s = A = ab, and the math will work out such that $R_s = 0$, which is what we expect.

65], and I invite the reader to indulge on this subject there. The basic notion gathered from all the references was that if a system to be analyzed has more than three or more conduction layers and if 2 out of 3 degrees of freedom were unbounded (i.e. 2 semi- infinite boundaries, as in the case for an isolated cell of self-heating MOSFETs, located on top of a thick 3-D stack with no neighboring clusters), then a closed-form solution of for the total thermal resistance, R_{total}, is impossible to obtain. This is precisely the reason why most thermal engineering soluions in microelectronics are always solved based on a simulation-first approach, and the nonlinear feedback due to self-heating only exacerbates the problem.

Bibliography

[1] Map of Athens, circa 500 BC, from http://www.cultralresources/images/maps/AthensBig.jpg.

[2] Map of Rome, circa 300 AD, from http://garyb.0catch.com/rome-map/ancient-city-by-type.html.

[3] Map of modern-day Boston, from http://www.mapquest.com.

[4] C. Y. Chang and S. M. Sze. *ULSI Technology*. McGraw-Hill, New York, 1996.

[5] J. A. Davis, Vivek K. De, and J. Meindl. A stochastic wire-length distrubtion for gigascale integration (gsi)-part i: Derivation and validation. *IEEE Trans. Electron Devices*, 45(3):580–589, 1998.

[6] J. A. Davis, Vivek K. De, and J. Meindl. A stochastic wire-length distrubtion for gigascale integration (gsi)-part ii: Applications to clock frequency, power dissipation, and chip size estimation. *IEEE Trans. Electron Devices*, 45(3):590–597, 1998.

[7] ITRS 2005 Edition Online, from http://www.itrs.net/Common/2005ITRS/Home2005.htm.

[8] B. Herring. The secrets of roman concrete. *Constructor*, page 16, Sept 2002.

[9] R. P. Zingg, J. A. Friedrich, G. W. Neudeck, and B. Hofflinger. Three-dimensional stacked mos transistors by localized silicon epitaxial overgrowth. *IEEE Trans. Electron Devices*, 37(7):1452, 1990.

[10] V. W. C. Chan, P. C. H. Chan, and M. Chan. Three dimensional cmos integrated circutis on large grain polysilicon films. *IEEE IEDM*, pages 161–164, 2000.

[11] R. A. Fillion. A forecast on the future of hybrid wafer scale integration technology. *IEEE Trans. Components, Hybrids, and Manufacturing Technology*, 16(7):615, 1993.

[12] M. L. Cambell and S. T. Toborg. *IEEE Trans. Components, Hybrids, and Manufacturing Technology*, 16(7):646, 1993.

[13] S. A. Kuhn, M. Kleiner, P. Ramm, and W. Weber. Thermal analysis of verticlaly integrated circuits. *IEEE IEDM Tech. Digest*, pages 487–490, 1995.

[14] et. al. P. Ramm. *Microelectronic Engineering*, 37–38:39, 1997.

[15] P. M. Sailer, P. Singhal, J. Hopwood, D. R. Kaeli, P. M. Zavracky, K. Warner, and D. P. Vu. Creating 3-d circuits using transferred films. *IEEE Circuits and Devices*, page 27, Nov 1997.

[16] K. Warner, J. Burns, C. Keast, R. Kunz, D. Lennon, A. Loomis, W. Mowers, and D. Yost. Low temperature oxide bonded three-dimensional integrated circuits. *Proc. IEEE Int. SOI Conf.*, pages 123–125, 2002.

[17] T. Matsumoto, M. Satoh, K. Sakuma, H. Kurino, N. Miyakawa, H. Itani, and M. Koyanagi. Three-dimensinal integration technology for real time micro-vision system. *Jpn. J. Appl. Phys. Part 1*, 37(3b):1217, 1998.

[18] H. Takagi, R. Maeda, T. R. Chung, and T. Suga. *1997 International Conference on Solid-State Sensors and Acuators. Digest of Tech. Papers*, 1:657, 1997.

[19] Y. Hayashi, K. Oyama, S. Takahashi, S. Wada, K. Kajiyana, R. Koh, and T. Kunio. Three dimensional ics, having four stacked active device layers. *IEEE IEDM Tech. Digest*, page 657, 1991.

[20] T. Suga and F. Yuuki adn N. Hosoda. A new approach to cu-cu direct bump bonding. *Proc. 1997 IEEE IEMT/IMC*, page 146, 1997.

[21] S. Bengtsson. Semiconductor wafer bonding: A reviewof interfacial properties and applicaitons. *J. Electronic Materials*, 21(8):841, 1992.

[22] M. A. Schmidt. Wafer-to-wafer bonding for microstructure formation. *Proc. IEEE*, 86(8):1575–1585, Aug 1998.

[23] K. N. Chen, A. Fan, and R. Reif. *Journal of Material Science*, 37:3441, 2002.

[24] C. S. Tan, A. Fan, K. N. Chen, and R. Reif. Low temperature oxide to plasma enhanced chemical vapor deposition oxide wafer bonding for thin film transfer application. *Applied Phys. Lett.*, 82(16):2649–2651, 2003.

[25] et. al K. W. Guarini. Electrical integrity of state-of-the-art 0.13 μm soi cmos devices and circuits transferred for three-dimesional integrated circuit fabrication. *IEEE IEDM*, pages 943–945, Dec 2002.

[26] M. Despont, U. Dreschler, R. Yu, H. B. Pogge, and P. Vettiger. Wafer-scale microdevice transfer / interconnect: Its application in an afm-based data storage system. *J. Microelectromechanical Systems*, 13(6):895–901, Dec 2004.

[27] K. Banerjee and A. Mehrotra. Global interconnect warming. *Circuits and Devices*, pages 16–32, Sept. 2001.

[28] F. Yu, M. C. Cheng, P. Habitz, and G. Ahmadi. Modeling of thermal behavior in soi structures. *IEEE Trans. on Elec. Dev.*, 51(1):83–91, Jan. 2004.

[29] R. F. Pierret. *Semiconductor Device Fundamentals.* Addison-Wesley, Reading, MA, 1996.

[30] R. S. MUller and T. I. Kamins. *Device Electronics for Integrated Circuits, 2nd ed.* Wiley and Sons, Palo Alto, CA, 1986.

[31] J. M. Koo, S. Im, L. Jiang, and K. E. Goodson. Integrated microchannel cooling for three dimensional electronic circuit architectures. *J. of Heat Transfer*, 127:49–58, Jan. 2005.

[32] S. Im and K. Banerjee. Full chip thermal analysis of planar (2d) and vertically integrated(3d) high performance ics. *IEEE IEDM*, pages 727–730, Dec. 2000.

[33] S. Senturia. *Microsystems Design.* Kluwer Academic Publishers, Norwell, MA, 2001.

[34] Chang. *Ibid.*

[35] B. Razavi. *RF Microelectronics.* Prentice-Hall, Upper Saddle River, NJ, 1998.

[36] G. Gonzalez. *Microwave Transistor Amplifiers: Analysis and Design.* Prentice-Hall, Upper Saddle River, NJ, second edition, 1997.

[37] S. S. Mohan, M. Hershenson, S. P. Boyd, and T. H. Lee. Simple accurate expressions for planar spiral inductances. *IEEE J. Solid State Circuits*, 34(10):1419–1424, Oct. 1999.

[38] C. P. Yue and S. Wong. Physical modeling of spiral inductors on silicon. *IEEE Trans. Elec. Devices*, 47(3):560–568, Mar. 2000.

[39] J. N. Burghartz, D. C. Edelstein, and M. Soyuer. Rf circuit design aspects of spiral inductors on silicon. *IEEE J. Solid State Circuits*, 33(12):2028–2034, Dec. 1998.

[40] J. R. Long and M. A. Copeland. The modeling characterization, and design of monolithic inductors for silicon rf ics. *IEEE J. Solid State Circuits*, 32(3):357–369, Mar. 1997.

[41] S. M. Yim, Y. Chen, and K. K. O. The effects of ground shield on spiral inductors fabricated in a silicon bipolar technology. *IEEE J. Solid State Circuits*, 37(2):237–244, Feb. 2002.

[42] C. P. Yue and S. Wong. On-chip spiral inductors with patterned ground shields for si-based rf ics. *IEEE J. Solid State Circuits*, 33(5):743–752, May. 1998.

[43] W. Y. Yin, S. J. Pan, and L. W. Li. Comparative characteristics of on-chip single and double-level square inductors. *IEEE Trans. on Magnetics*, 39(3):1778–1783, May. 2003.

[44] H. J. De Los Santos. On the ultimate limits of ic inductors - an rf mems perspective. *IEEE Elecronic Components and Technology Conference, San Diego, CA*, May. 2002.

[45] J. W. Lin, C. C. Chen, and Y. T. Cheng. A robust high q micromachined rf inductor for rfic applications. *IEEE Trans. on Electron Devices*, 52(7):1489–1496, Jul. 2005.

[46] Y. J. Kim and M. G. Allen. Surface micromachined solenoid inductors for high frequency applications. *IEEE Trans. on Components, Packing, and Manufacturing Technology, Part C*, 21(1):26–33, Jan. 2002.

[47] W. B. Kuhn, X. He, and M. Mojarradi. Modeling spiral inductors in sos process. *IEEE Trans. on Electron Devices*, 51(5):677–683, May. 2004.

[48] J. G. Fiorenza. *Design and Fabrication of an RF Power LDMOSFET on SOI.* PhD dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering, 2002. Jim Fiorenza's Thesis: MIT 2002.

[49] Hyperphysics Online, at http://hyperphysics.phy-astr.gsu.edu/hbase/tables/magprop.html.

[50] M. F. Iskander. *Electromagnetic Fields and Waves.* Prentice–Hall, Englewood Cliffs, NJ, 1992.

[51] C. A. Chang, S. P. Tseng, J. Y. Chuang, S. S. Jiang, and J. A. Yeh. Characterization of spiral inductors with patterned floating structures. *IEEE Trans. on Microwave Theory and Techniques*, 52(5):1375–1381, May. 2004.

[52] C. A. Balanis. *Antenna Theory.* John Wiley and Sons, New York, NY, second edition, 1997.

[53] From Hypertextbok, at http://hypertextbook.com/facts/2005/YasminSinclair.shtml.

[54] K. L. Mittal. *Contact Angle, Wettability, and Adhesion*, volume Second. VSP BV, Utrecht, The Netherlands, 2002.

[55] L. L. Schramm. *Surfactants: Fundamentals and Applications in the Petroleum Industry.* Cambridge University Press, Cambridge, UK, 2000.

[56] R. G. Laughlin. *The Aqueous Phase Behavior of Sufactants.* Academic Press, London, UK, 1996.

[57] D. H. Smith. *Surfactant-Based Mobility Control: Progress in Miscible-Flood Enhanced Oil Recovery.* American Chemical Society, Washington, DC, 1988.

[58] M. R. Porter. *Handbook of Surfactants.* Chapman and Hall, Glasgow, UK, 1994.

[59] B. B. Mikic and W. M. Rohsenow. Thermal contact resistance. *MIT Heat Transfer Lab, Report No 4521-41*, Sept 1969.

[60] M. M. Yovanovich, Y. S. Muzychka, and J. R. Culham. Spreading resistance of isofulx rectangles and strips on compound flux channels. *Journal of Thermophysics and Heat Transfer*, 13(4):495–500, Oct-Dec 1999.

[61] G. Ellison. Maximum thermal spreading resistance for rectangular source and plates with nonunity aspect ratios. *IEEE Trans. Components and Packing Technologies*, 26(2):439–454, Jun 2003.

[62] M. M. Yovanovich. Thermal resistances of circular source on finite circular cylinder with side and end cooling. *J. of Electronic Packaging*, 125(4):169–177, Oct-Dec 1999.

[63] Y. L. Chow and M. M. Yovanovich. The shape factor of the capacitance of a conductor. *J. of Applied Physics*, 53(12):8470–8475, Dec 1982.

[64] Y. S. Muzychka, M. R. Sridhar, and M. M. Yovanovich. Thermal spreading resistance in multilayered contatacts: Applications in thermal contact resistance. *J. of Thermophysics and Heat Transfer*, 13(4):489–494, Oct-Dec 1999.

[65] S. Lee, S. Song, V. Au, and K. P. Moran. Constriction / spreading resistance model for electronics packaging. *ASME/JSMP Thermal Engineering Conference*, 4:199–206, 1995.