Stress-Induced Performance Variations in 3D and Flexible Circuits

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Tengtao Li

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Sachin S. Sapatnekar, Advisor

August, 2019
Acknowledgements

First of all, I want to express my sincere gratitude to my incredible advisor, Prof. Sachin S. Sapatnekar for his support and guidance during my whole PhD study. It would not be possible to complete this thesis without his inspiration as well as his encouragement all along these years. As a wonderful mentor, he taught me how to be a qualified researcher, who needs not only abundant knowledge and solid understandings of his area, but also the ability to overcome challenging problems during research. More important, I have gained a lot of valuable experience during the association with him such as the rigorous attitude, critical thinking, thoroughness and attention to detail, which all helped shaping me to become a better researcher. It is indeed a privilege to work with one of the finest engineers and a highly regarded professor in the field of VLSI computer-aided design.

I want to thank my committee members at the University of Minnesota: Prof. Chris Kim, Prof. Antonia Zhai, and Prof. Ulya Karpuzcu, for their constructive feedback about my research. I am grateful to the National Science Foundation (NSF), for the resource and financial support towards my thesis projects. I want to express my gratitude to the ECE staff, especially Linda Jagerson and Carlos Soria, for their help and support. I express my thanks to Minnesota Supercomputing Institute for providing the resources and timely support to perform valuable large-scale simulations for my work.

I want to thank Dr. Vivek Mishra, Dr. Sravan Marella, Dr. Zhaoxin Liang for their help during the starting of my PhD study. Many thanks to Dr. Deepashree Sengupta, Meghna Mankalale, Dr. Farhana Sharmin Snigdha, Masoud Zabihii, Vidya Chhabria, Tonmoy Dhar, Susmita Dey Manasi, Kishor Kunal, Geraldo Pradipta and Mohammad Shohel for being wonderful labmates as well. I also would like to thank my friends, Dr. Bingzhe Li, Qiannan Li, Dr. Bin Li, Dr. Jiaxi Hu, Jianjun Yuan, Dingyi Liu,
Kaiyang Sun, Baogeng Ma, Chengyao Tan and Xiaonan Zhang for their companion and encouragement, which made my PhD life joyful and meaningful.

Finally, I want to thank my wonderful family. I want to thank my parents and parents-in-law for their persistent support and encouragement. Special thanks to my adorable wife, Dr. Qianqian Fan, for her inspiration, companion, encouragement and care.
Dedication

To my wonderful parents and dear wife.
Abstract

Mechanical stress is a significant source of variability in advanced VLSI technologies that impacts circuit performance. Unintentional mechanical stress induced either by the manufacturing process or during daily use affects transistor electrical parameters such as mobility and threshold voltage due to piezoresistivity and stress-induced band deformation, respectively. Consequently, the performance of circuits/systems is highly dependent on stress distributions. Thus, the modeling of stress effect is important and necessary in the performance analysis of circuits and systems to meet design specifications.

Mechanical stress is expressed in various ways in different integrated systems. In 3D ICs, which are implemented by stacking multiple ultra-thin chips (UTCs) in the vertical direction with through-silicon-vias (TSVs), thermomechanical stresses are induced due to the mismatches among various materials during the annealing process. For UTCs and organic thin-film transistors (OTFTs) that are widely adopted in the strongly emerging market for flexible electronics, such as flexible display and flexible system-in-foil (SiF), significant stress can be induced by deformations during normal operations. This thesis conducts stress-induced performance evaluations of three main application fields in 3D DRAMs, flexible displays, and flexible SiFs.

The first part of the thesis addresses TSV-based 3D-stacked DRAMs, which can significantly increase cell density and bandwidth while also providing lower power consumption than their 2D counterparts. However, 3D IC structures experience significant thermomechanical stress due to the differential rates of contraction of their constituent materials, which have different coefficients of thermal expansion. This induces stress impacts on circuit performance. We develop a procedure that performs a performance analysis of 3D DRAMs, capturing the impact of both layout-aware stress and layout-independent stress on parameters such as latency, leakage power, refresh power, area, and bus delay. The approach first proposes a semianalytical stress analysis method for the entire 3D DRAM structure, capturing the stress induced by TSVs, micro bumps, package bumps, and warpage. Next, this stress is translated to variations in device mobility and threshold voltage, after which analytical models for latency, leakage power,
and refresh power are derived. Finally, a complete analysis of performance variations is performed for various 3D DRAM layout configurations to assess the impact of layout-dependent stress. We explore the use of alternative flexible package substrate options to mitigate the performance impact of stress. Specifically, we explore the use of an alternative bendable package substrate made of polyimide to reduce warpage-induced stress and show that it reduces stress-induced variations, and improves the performance metrics for stacked 3D DRAMs.

The second part of the thesis addresses stress effects in flexible displays with OTFTs, which are widely used in flexible circuits such as flexible displays, sensor arrays, and radio frequency identification cards (RFIDs). These technologies offer features such as better flexibility, lower cost, and easy manufacturability using a low-temperature fabrication process. Due to their very nature, flexible displays experience significant mechanical strain/stress in the field due to the deformation caused during daily use. These deformations can impact device and circuit performance, potentially causing a loss in functionality. In this part of work, the effects of extrinsic strain due to two fundamental deformations modes, bending and twisting, are first modeled. Next, this strain is translated to variations in device mobility, after which analytical models for error analysis in the flexible display are derived based on the rendered image values in each pixel of the display. To rectify strain-induced errors, two error correction approaches for flexible displays are then proposed, based on voltage compensation and flexible clocking.

The third part of this thesis studies silicon-based UTCs that provide an excellent solution to build flexible SiFs for bio-sensing and bio-monitoring purposes. These UTCs use flexible chips that are thinned down to about 20µm, and utilize CMOS devices that deliver much higher performance than alternatives such as organic devices and thin-film transistors (TFTs), while being compatible with other SiF components such as flexible displays and flexible sensor arrays. Flexible SiFs experience significant mechanical stress in the field due to the deformation caused during normal use, which causes undesirable circuit performance shifts. We model the stress due to two types of packages schemes for UTCs with various chip dimensions, translating stress to shifts in device mobility and threshold voltage, and evaluate the system-level performance variations of two common SiF elements, an A/D converter and an SRAM.
## Contents

Acknowledgements i

Dedication iii

Abstract iv

List of Tables ix

List of Figures x

1 Introduction 1

1.1 Stress Effects in 3D ICs and Flexible Electronics . . . . . . . . . . . . . 3
1.1.1 Stress Effects in 3D DRAMs . . . . . . . . . . . . . . . . . . . . . 3
1.1.2 Stress Effects in Flexible Displays with Organic Transistors . . . 4
1.1.3 Stress Effects in Flexible SiFs with UTCs . . . . . . . . . . . . . 6

1.2 Thesis Organization . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2 Effects of Stress in Integrated Systems 10

2.1 Basic Principles of Stress . . . . . . . . . . . . . . . . . . . . . . . . . . 10
2.2 Finite Element Analysis (FEA) Approach . . . . . . . . . . . . . . . . . 11
2.3 Electrical Variations Due to Stress . . . . . . . . . . . . . . . . . . . . 12
  2.3.1 Stress Effect on CMOS Devices . . . . . . . . . . . . . . . . . . . . 12
  2.3.2 Stress Effect on OTFTs . . . . . . . . . . . . . . . . . . . . . . . . 13

3 Stress-Induced Performance Shifts in 3D DRAMs 16

3.1 Performance Evaluation of 3D DRAMs . . . . . . . . . . . . . . . . . . . 18
3.1.1 Memory Organization and Performance ........................................ 19
3.1.2 The Impact of Stress on 3D DRAM Performance ........................... 20
3.2 Stress Modeling of a 3D DRAM Stack ............................................. 21
3.3 Experimental Results in 3D DRAMs ............................................... 27
  3.3.1 Using FR-4 as the Package Substrate Material ............................... 28
  3.3.2 Using Polyimide as Package Substrate Material .............................. 32
3.4 Conclusion ....................................................................................... 39

4 Strain-Aware Performance Evaluation and Correction for OTFT-Based Flexible Displays
  4.1 Strain Modeling of a Flexible Display ............................................. 41
  4.2 Performance Evaluation of Flexible Displays .................................... 43
  4.3 Experimental Results in Flexible Displays ....................................... 47
  4.4 Correcting Strain-Induced Performance Loss .................................... 50
  4.5 Experimental Results with the Compensation Schemes ....................... 54
  4.6 Conclusion ....................................................................................... 56

5 Stress-Induced Performance Shifts in Flexible System-in-Foils Using Ultra-Thin Chips
  5.1 Stress Modeling of a Flexible System-in-Foil with Ultra-Thin Chips ........ 58
  5.2 Performance Evaluation of ADCs and SRAMs in SiFs .......................... 61
    5.2.1 Performance Evaluation of SAR ADCs ....................................... 62
    5.2.2 Performance Evaluation of SRAMs ............................................. 63
    5.2.3 The Impact of Stress on SRAM Performance ................................. 66
  5.3 Experimental Results in SiFs ......................................................... 66
    5.3.1 Stress vs. Chip Size and Corresponding Device Variations ............... 66
    5.3.2 Performance Variations in Flexible SAR ADCs ......................... 67
    5.3.3 Performance Variations in SRAMs ............................................ 70
    5.3.4 Compensating for Stress-Induced Variations ................................ 71
  5.4 Conclusion ....................................................................................... 72

6 Conclusion ...................................................................................... 73
References

Appendix A. Components of the Row Cycle Time in Memory Arrays 86
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Piezoresistivity Coefficients ($\times 10^{-12}\text{Pa}^{-1}$) in (100) Si</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>The Strain-Mobility Relationship in P3HT OTFT [1]</td>
<td>14</td>
</tr>
<tr>
<td>3.1</td>
<td>Dimensions of the TSVs, $\mu$-Bumps, and Package Bumps</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>Material Parameters in 3D DRAMs</td>
<td>22</td>
</tr>
<tr>
<td>3.3</td>
<td>Summary of Error Distributions between FEA and Volume Fraction Approximation in S1–S3 in 3D DRAMs</td>
<td>25</td>
</tr>
<tr>
<td>3.4</td>
<td>Summary of TSV Distributions in L1–L8</td>
<td>28</td>
</tr>
<tr>
<td>3.5</td>
<td>Row Cycle Time ($t_{RC}$), Leakage Power ($P_{\text{leak}}$), and Refresh Power ($P_{\text{ref}}$) for L1–L8 ($D_0 = 33.62\text{ns}$, $P_{\text{leak}}^{\text{nom}} = 50.66\text{mW}$, $P_{\text{ref}}^{\text{nom}} = 18.90\text{mW}$)</td>
<td>32</td>
</tr>
<tr>
<td>3.6</td>
<td>Area Overhead of TSV and KOZ for L1–L8 in 3D DRAMs</td>
<td>34</td>
</tr>
<tr>
<td>3.7</td>
<td>Bus Delay Variation for L1–L8 in 3D DRAMs</td>
<td>34</td>
</tr>
<tr>
<td>3.8</td>
<td>Row Cycle Time ($t_{RC}$), Leakage Power ($P_{\text{leak}}$), and Refresh Power ($P_{\text{ref}}$) for L1–L8 with PI substrate ($D_0 = 33.62\text{ns}$, $P_{\text{leak}}^{\text{nom}} = 50.66\text{mW}$, $P_{\text{ref}}^{\text{nom}} = 18.90\text{mW}$)</td>
<td>36</td>
</tr>
<tr>
<td>3.9</td>
<td>Area Overhead of TSV and KOZ for L1–L8 in 3D DRAMs with PI substrate</td>
<td>38</td>
</tr>
<tr>
<td>3.10</td>
<td>Bus Delay Variation for L1–L8 in 3D DRAMs with PI Substrate</td>
<td>38</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of SSIM Results in Flexible Displays with Stress</td>
<td>51</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary of Peak Device Parameter Variations</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>Stress-Induced Variations in SRAMs ($t_{\text{access}}^{\text{nom}} = 0.390\text{ns}$, $P_{\text{leak}}^{\text{nom}} = 0.639\text{mW}$)</td>
<td>70</td>
</tr>
<tr>
<td>5.3</td>
<td>Overdesign Parameters for SAR ADC and SRAM</td>
<td>71</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Structure of a 3D DRAM [2] and the layout of its rank 0 layer. \hspace{1cm} 3
1.2 Structure of a liquid crystal flexible display with OTFTs. \hspace{1cm} 5
1.3 Ultra-thin chip packaging schemes: (a) Middle Chip, between two flexible layers (b) Top Chip, atop a flexible substrate [3]. \hspace{1cm} 7
1.4 Block diagram of the flexible ECG system [4]. \hspace{1cm} 7
2.1 Structure of a bottom-gate top-contact OTFT. \hspace{1cm} 14
3.1 Organization of a 3D DRAM array. \hspace{1cm} 18
3.2 Stress maps showing the accuracy of the volume fraction approximation for TSVs, $\mu$-bumps, and package bumps. \hspace{1cm} 23
3.3 $\sigma_{11}$ component of $\sigma_{pkg-bump}$ along the vertical direction. \hspace{1cm} 26
3.4 Accuracy of our semianalytical model for a TSV cluster: the solid curve fits the blue sample points. The horizontal axis is the set of evaluated points (7 values of $r$ evaluated at 6 values of $w$). \hspace{1cm} 27
3.5 Contours of $\sigma_{11}$ in eight layouts corresponding to different TSV configurations. \hspace{1cm} 29
3.6 Variations in mobility and $V_t$ in L8. \hspace{1cm} 30
3.7 $t_{RC}$ variation contours in eight layouts corresponding to different TSV configurations. \hspace{1cm} 31
3.8 Subthreshold leakage current variations in eight layouts corresponding to different TSV configurations. \hspace{1cm} 33
3.9 Increase of wire length caused by the KOZ. \hspace{1cm} 33
3.10 3D DRAM structure using the PI substrate. \hspace{1cm} 34
3.11 Contours of $\sigma_{11}$ in the eight layouts with the PI substrate. \hspace{1cm} 35
3.12 $t_{RC}$ variation contours in the eight layouts using the PI substrate. \hspace{1cm} 35
3.13 Subthreshold leakage current variations in the eight layouts using the PI substrate. ................................................. 37
4.1 FEA simulation set up for two types of deformations in a flexible display. 41
4.2 FEA simulation results of Fig. 4.1 showing strain distributions in a flexible display. ............................................. 42
4.3 (a) An OTFT in a pixel and (b) its equivalent RC model [5]. ............... 44
4.4 $V_{\text{disp}}$ transients at the nominal and a shifted mobility value. ........... 45
4.5 Accuracy of the fitted function for $V_{\text{disp}}^\text{strain} (t_{\text{row}})$. ................. 46
4.6 Error histogram for the interpolation function for $V_{\text{disp}}^\text{strain}$. .......... 47
4.7 OTFT mobility variation results for two types of deformations. ............ 47
4.8 Results for displaying Lena after Barbara. .................................. 48
4.9 Results for displaying Cameraman after Girlface. ............................ 50
4.10 Schematic of the pentacene field-controllable strain sensor based on hybrid gate dielectrics (high-k Ta$_2$O$_5$+PVP) [6]. ......................... 52
4.11 An architecture showing $V_{\text{data}}$ and $t_{\text{row}}$ compensation in an OTFT array. 53
4.12 Results for displaying Lena with $V_{\text{data}}$ compensation in an OTFT array. 55
4.13 Results for displaying Lena with $t_{\text{row}}$ compensation in an OTFT array. 56
4.14 Relationship between SSIM and $t_{\text{row}}$. ........................................ 57
5.1 FEA simulation set up for the silicon-based UTC package. ................. 59
5.2 FEA simulation results of Fig. 5.1 showing stress distributions in flexible SiFs. ......................................................... 60
5.3 Peak stress values in silicon-based UTCs with different chip sizes. .......... 61
5.4 Architecture of an 8-bit SAR ADC. ............................................. 62
5.5 Organization of an SRAM array. ................................................. 64
5.6 Variations in mobility and $V_t$ in Top Chip packaging. ..................... 67
5.7 $V_{\text{out}}$ transients in the SAR ADC at the nominal and a shifted mobility value. ....................................................... 68
5.8 DNL of the SAR ADC of Fig. 5.4 under bending stress. .................... 69
Chapter 1

Introduction

Integrated circuits (ICs) serve as the core elements in almost all modern electronic products that influence our lives, from laptops, smartphones, to emerging 3D ICs, flexible electronics and bio-monitoring applications. Mechanical stress analysis plays a more and more important role in the performance evaluation and reliability analysis of a circuit/system especially in 3D ICs and flexible electronics. Since unintentional mechanical stress affects the electrical parameters of CMOS transistors and OTFTs such as mobility and threshold voltage due to piezoresistive effect and stress-induced band deformations, respectively. As a result, the performance of a circuit/system is highly dependent on stress distributions within the structure. In this thesis, we have analyzed the stress effects in three main application fields: 3D DRAMs with TSVs, flexible displays using OTFTs, and flexible SiFs in which UTCs are applied to implement A/D converters and on-chip SRAMs.

The sources of mechanical stresses can be divided into several categories depending on different applications. Stress can be induced either by the mismatches among different materials during the manufacturing of 3D ICs or by the operations during the daily usage of flexible displays and SiFs.

In 3D ICs, mechanical stress induced by the manufacturing may cause significant performance shifts in 3D ICs. 3D ICs can break the bottleneck of limited density and I/O number of conventional 2D structures with 3D stacking. Memory is considered to be an excellent platform that can leverage 3D stacking due to greatly increased cell density per unit footprint, large improvements over 2D structures in the latency and power
associated with communication, and low thermal overhead. 3D DRAMs can be built by stacking multiple DRAM layers in the vertical direction, with all layers are connected with TSVs that can transmit data, address, and power signals [2, 7, 8]. Each layer contains not only DRAM cells, but also addressing and other peripheral circuitry. Wide I/O 3D DRAMs achieve significant improvements in the memory bandwidth by using a large number of TSVs that traverse the 3D stack. However, despite the advantages of 3D DRAMs, significant stress can be induced into the structures in manufacturing. Since the manufacture of 3D DRAMs usually adopts high temperature, thermomechanical stress will be induced during the annealing process due to the mismatch between materials, such as the mismatch between TSV and surrounding Si chip and the mismatch between adjacent layers in the vertical direction. The stress affects the electronic parameters of CMOS devices, which in turn decides the performance of 3D DRAMs in latency, power and area.

Flexible electronics, unlike 3D ICs, may suffer significant mechanical stress not due to manufacturing process but due to deformations during daily operation. Flexible electronics adopt bendable, elastic, and lightweight materials, such as plastic, polymer, or even paper, for their substrate, and are being increasingly deployed in applications such as flexible displays [9], flexible sensor arrays [10], RFIDs [11], electronic papers, and SiFs [12]. Flexible electronics provide excellent compatibility with wearable, bio-sensing, and bio-monitoring systems, in which the system must be flexible to fit non-rigid surfaces such as the human skin, and are required to undergo various deformations during their use. However, significant mechanical stress can be induced by these deformations during the usage, which affects the electrical parameters of transistors depending on the technology adopted in different applications. Technologies that are currently being pursued to support the strong emerging market for flexible electronics [13] include organic electronics, TFTs on foil, and UTCs on foil. Thus, the stress induced by these deformations either affects the mobility of OTFTs, which is used in flexible displays in this thesis, or change parameters of CMOS devices in UTCs of flexible SiFs. As a result, stress-induced shifts in OTFTs can cause errors in flexible displays, while the variations of CMOS devices in UTCs in SiF applications can affect the accuracy of A/D converters and the performance of on-chip SRAMs.
Figure 1.1: Structure of a 3D DRAM [2] and the layout of its rank 0 layer.

1.1 Stress Effects in 3D ICs and Flexible Electronics

In addition to process and environmental variations, which have been well studied, mechanical stress effects also contribute to performance variations in integrated circuits. Unintentional mechanical stresses in devices, which can be induced during the manufacturing process or by deformations in daily use, affect transistor electrical properties due to piezoresistivity and electronic band deformation [14]. Thus, the performance of 3D ICs and flexible electronics are affected by stress.

1.1.1 Stress Effects in 3D DRAMs

The structure of a 3D DRAM stack is illustrated in Fig. 1.1, in which each chip in the stack constitutes a rank, as in [2] (in some structures, multiple ranks may be placed in each layer [15]). One master chip, containing normal DRAM as well as control and datapath circuitry for every rank in the stack, is placed at the bottom, and several slave chips, each containing only normal DRAM and DRAM core test circuits are stacked above it [2]. A typical configuration stacks all chips on a flip-chip package using back-to-face (B2F) bonding [16], and the device layer appears near the bottom surface of each chip. The signals that are required to traverse multiple layers, such as data, address, and power, are transmitted through copper TSVs. A dielectric underfill layer is added between the DRAM layers which serves the purpose of isolation while also providing mechanical support, and typically constituted of SiO$_2$ or BCB. The TSVs in different 3D layers are connected using $\mu$-bumps, surrounded by underfill. Similarly, package bumps, which are also surrounded by underfill, are placed between the master chip and package substrate to enable the communication between memory and CPU.
An important consideration of the design of 3D DRAM structures is the need to address the stress induced by TSV fabrication and 3D stacking. The manufacturing process for a TSV requires a temperature of 275°C, while 3D stacking typically requires a temperature between 200°C to 400°C, depending on the bonding method and the types of materials that are used for the µ-bump [17]. When the structure cools down to room temperature by annealing, the mismatch in the coefficient of thermal expansion (CTE) of different materials may leave a residual stress in the structure [18]. DRAM performance is affected by this stress, which impacts transistors in the device layers of the DRAM chips. This extrinsic stress originates from:

(a) CTE mismatches between TSVs and the surrounding silicon [14],
(b) µ-bump and package bump induced stress [19], and
(c) warpage caused by the mismatch in the CTE of different layers, such as the DRAM layer and the underfill layer.

The stress tensor inside the 3D DRAM chip affects the band structure and crystal lattice in the channel of devices [20–22], causing shifts in device parameters, such as mobility and threshold voltage, and eventually translating to changes in memory performance parameters such as latency, leakage power, and refresh power.

1.1.2 Stress Effects in Flexible Displays with Organic Transistors

OTFTs offer an attractive technology for flexible circuits. OTFT circuits [23] have been demonstrated to implement ring oscillators, decoders, flip-flops, pulse-generators, voltage multipliers, charge amplifiers, etc. Several applications of OTFTs, including flexible displays [9], sensor arrays [10,24], RFIDs [11] and organic DRAM cells [25] have been reported. The speed of these systems lags that of cutting-edge CMOS, but they can provide the right level of performance in applications where flexibility is paramount. In flexible displays, the OTFT acts as a switch driving a capacitance in each pixel, which is charged to a voltage that renders an appropriate bias to the display cell [9]. For organic DRAM cells [25], the OTFT acts as a switch during read/write operations.

In this thesis, we consider the application area of flexible displays and focus on designing displays that, in addition to showing mechanical integrity under strain, are
also resilient to the performance variation caused by extrinsic strain. Specifically, we show how the electrical characteristics of the circuitry may be affected by strain (e.g., through shifts in device carrier mobilities due to piezoresistance), which may influence the functional correctness of the system.

A flexible display, shown in Fig. 1.2 [26], consists of multiple layers: a) a flexible substrate (composed of materials such as polyimide (PI), plastic, or even paper), which provides mechanical support for other layers; b) a matrix backplane comprising a layer of devices that control the display, driven by row/column drivers, as detailed at the top right of the figure; c) a functional display layer with an $n \times m$ liquid crystal matrix, each pixel of which is controlled by the voltage level of a separate storage capacitor; d) a front transparent plane that provides protection and isolation from air, water, and dust.

As compared to conventional CMOS technologies, OTFTs involve a simpler manufacturing process based on low-cost deposition of thin films of semiconducting organic materials on flexible substrates at relatively low temperature. The mechanical flexibility of OTFTs makes them compatible with flexible substrate for lightweight and foldable products [27]. However, these advantages come at a performance penalty: typical mobility values in OTFT devices range from 0.1–1 cm$^2$/Vs for organic devices, with the best organic materials achieving mobility of 1–10 cm$^2$/Vs [27,28].

Flexible displays are subject to changes in shape, which can be modeled using several modes: bending, pressing, and twisting [29–31]. The influence of strain on the mobility,
on-current and leakage current of an individual OTFT device is discussed in [1], but to our knowledge, there is no work that clearly evaluates the effect of strain on a typical OTFT application. As we will show, the strain induced by these deformations may have a significant influence on the quality of the image in a flexible display.

1.1.3 Stress Effects in Flexible SiFs with UTCs

While several low-cost technologies are being pursued to support the strong emerging market for flexible electronics [13], including organic electronics, TFTs on foil, and UTCs on foil, the performance of many of these devices (e.g., organic electronics and TFTs) is limited by their carrier mobilities, ranging from 0.1–10 cm$^2$/Vs. In contrast, the carrier mobilities in single-crystalline silicon used to build CMOS circuits are above 100 cm$^2$/Vs at room temperature. Thus, CMOS UTCs on foil have emerged as an excellent substrate for solutions where high performance devices and dense interconnects are required. This technology is also compatible with organic or TFT electronics components, such as flexible displays and flexible sensor arrays, in a hybrid SiF.

The technology for CMOS UTCs is well established, and includes various thinning techniques, such as back grinding and Chipfilm [26]. Back grinding is a widely used method for wafer/chip thinning. A small thickness about 20µm is achieved in two steps: coarse grinding, which removes the Si bulk quickly, followed by fine grinding to obtain a smooth surface. Chipfilm technology uses a buried cavity underneath the chip during fabrication, which defines the chip thickness [32]. With both techniques, the silicon chip thickness can go down to of 20µm or less. This flexible chip can then be packaged with a flexible substrate using schemes illustrated in Fig. 1.3, namely,

**Middle Chip** [3,33]: The UTC is embedded between two spin-on PI layers. Laser-drilled vias are metallized by sputtering.

**Top Chip** [34,35]: UTCs are placed face-up over a PI substrate.

Under deformation, the Middle Chip scheme experiences lower stress than the Top Chip scheme. However, the added process step makes this technology more expensive. For high-power applications, Middle Chip has worse heat removal paths to ambient, but typical SiF applications are low power, and the thermal resistance of the ultra-think substrate is low and does not raise temperature significantly.

Two critical components of any SiF system are an analog-to-digital converter (ADC),
which converts a real-world analog signal to the digital domain for processing, and an on-chip memory. For example, consider the application area of wearable electrocardiogram (ECG) monitoring implemented in SiF using CMOS components. The block diagram of the flexible ECG system [4, 36] is shown in Fig. 1.4. The SoC includes three key parts: an analog front-end (AFE), an 8-bit successive approximation register (SAR) ADC, and a digital core. The digital core temporarily stores the digital signal in the on-chip memory and finally transmits the data to the outside world via wires or a wireless module. This structure provides a prototype for the systems monitoring other electrical health signals as well, such as electroencephalogram (EEG), electromyogram (EMG), temperature, and blood pressure.

The stress tensor inside the chip affects the band structure and crystal lattice in the channel of devices [20–22], causing shifts in device parameters, such as mobility and
threshold voltage, and eventually affects the elements inside the system: specifically, the ADC where errors affect the accuracy, and the on-chip SRAM, whose latency and leakage power are affected. This part of the thesis develops an analysis and mitigation methodology for stress-induced performance variations in SRAMs and ADCs in a SiF under two commonly-used UTC technologies - Top Chip and Middle Chip. Prior work has analyzed stress in digital systems can be applied to the digital core [13, 37], but there is limited understanding today of stress-induced variations in non-digital blocks such as the ADC, and memory systems, and this is the focus of our thesis.

1.2 Thesis Organization

The thesis is organized as follows:

- Chapter 2 introduces the basic principles of mechanical stress and its effect on electrical parameters of CMOS and OTFT devices.

- Chapter 3 develops an unifying procedure that combines the impact of all sources of stress in the entire structure of a wide I/O 3D DRAM, and analyzing the impact of this stress on memory performance parameters. Compared to the expensive FEA method or other analytical methods in previous works, our semianalytical model provides a fast method for computing the stress in an entire wide I/O 3D DRAM by modeling the stress caused by TSV stripes and clusters accurately. We use this analysis technique to explore the impact of changes in the TSV layout on memory system performance in 3D DRAMs including latency, leakage power, refresh power, area, and wire length. Finally, we consider using polyimide to replace the traditional FR-4 as the package substrate material to reduce warpage-induced stress.

- Chapter 4 contributes an analysis the performance variations of flexible displays due to strain, and to rectify the errors induced due to this variation. The chapter is organized into three parts: (a) Strain analysis due to bending, twisting, and pressing. (b) Analysis of OTFT carrier mobility degradation analysis under strain, and its translation to the error in the displayed image. (c) Error correction using two approaches: voltage compensation and the use of a flexible clock.
• Chapter 5 performs an analysis of stress due to bending, for various chip sizes of a flexible SiF application, and then computes the stress-induced carrier mobility and threshold voltage shifts, finally, analyzes performance variations of the two key elements, the SAR ADC and the SRAM, inside the system, and provides overdesign methods to compensate for the effect of stress.

• Chapter 6 concludes the thesis.
Chapter 2

Effects of Stress in Integrated Systems

As introduced in Chapter 1, the CTE mismatch between different materials is the source of mechanical stress in 3D DRAMs, while significant stress can be induced in flexible displays and SiFs by the deformations during the use. This chapter introduces the analytical stress and electrical models used in this work. First, the fundamental equations of elasticity are presented. Then, the impact of mechanical stress on transistor electrical properties, such as mobility and threshold voltage, is captured. The fundamental models of piezoresistivity and deformation potential theory are presented to characterize the changes in mobility and threshold voltage, respectively, under stress. Subsequent chapters draw upon the basic equations from this chapter to obtain specific solutions for the relevant problems.

2.1 Basic Principles of Stress

In the theory of linear elasticity, the terms, displacement, strain and stress are frequently used. Displacement is defined as the actual change in dimensions of a deformed body due to applied forces.

Strain is a measure of deformation, representing the displacement between particles in an object relative to a reference length. The mechanical strain field can be represented
as the tensor:
\[
\epsilon = \epsilon_{ij} = \begin{pmatrix}
\epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\
\epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\
\epsilon_{31} & \epsilon_{32} & \epsilon_{33}
\end{pmatrix}
\] (2.1)
where the subscripts \( i, j \in \{1, 2, 3\} \) refer to the three coordinate axes, \( \epsilon_{ii} \) are normal strains, and \( \epsilon_{ij, i \neq j} \) are shear strains.

Stress physically corresponds to the reactionary internal forces per unit due to deformation of an object under external forces. The mechanical stress field can also be represented by a tensor:
\[
\sigma = \sigma_{ij} = \begin{pmatrix}
\sigma_{11} & \tau_{12} & \tau_{13} \\
\tau_{21} & \sigma_{22} & \tau_{23} \\
\tau_{31} & \tau_{32} & \sigma_{33}
\end{pmatrix}
\] (2.2)
where the subscripts \( i, j \in \{1, 2, 3\} \) refer to the three coordinate axes, \( \sigma_{ii} \) are normal stresses, and \( \tau_{ij} \) are shear stresses. Within the elastic limit, stress and strain are related by Hooke’s law.

### 2.2 Finite Element Analysis (FEA) Approach

The stress state of a system can be fully described by 15 unknowns: six stress components, six strain components, and three displacement components. Two main approaches can be applied to solve the stress state of a system: analytical solution approach and FEA approach. Closed-form solutions can be obtained for simple geometries using analytical models. However, for more general shapes and complex boundary conditions, the elasticity equations tend to be intractable and may require complicated mathematical analysis. In such scenarios, FEA method allows the elasticity equations to be solved using numerical methods. Considering the complex structure of 3D DRAMs and multiple stress sources, FEA methods are adopted in this thesis.

The FEA method solves the elastic problem in the following way. First, the structure is discretized into subdomains known as elements with special points called nodes. The elements usually take two-dimensional or three-dimensional polygonal shapes and the
nodes correspond to the corners of the polygon. Approximate solutions are developed for each element in terms of nodal values. Algebraic equations are then constructed among the nodal values, based on their physical connectivity and by applying continuity and prescribed boundary conditions. The algebraic equations are then solved to obtain the required values of the stress distributions. If the number of elements is sufficiently large, the solution can be considered to be accurate. In this thesis, we use the ABAQUS [38] finite element package for obtaining stress distributions in 3D DRAMs, flexible displays and flexible SiFs as shown in subsequent chapters.

2.3 Electrical Variations Due to Stress

2.3.1 Stress Effect on CMOS Devices

The cubic lattice structure of silicon crystal is typically defined in Miller notation, and the wafer orientation (typically, [001]) is normal to the plane of the wafer. Since transistors are oriented along [110] due to mobility considerations, we use a rotated coordinate system with the $x'$-axis along [110] and the $y'$-axis along $\overline{110}$. According to piezoresistivity theory, mobility can be expressed as a linear combination of the elements of stress tensor because the resistivity tensor which is related to mobility would vary with the stress tensor [21]. The relative change of mobility in the rotated coordinate system ($x'$, $y'$) is given by [21]:

$$\Delta \mu' = \left[ \pi'_{11} \sigma'_{x'x'} + \pi'_{12} \sigma'_{y'y'} + \pi_{12} \sigma_{zz} \right] \cos^2 \phi' + \left[ \pi'_{44} \tau_{x'y'} \right] \sin 2\phi'$$

$$+ \left[ \pi'_{11} \sigma'_{y'y'} + \pi'_{12} \sigma'_{x'x'} + \pi_{12} \sigma_{zz} \right] \sin^2 \phi'$$

(2.3)

where $\sigma'_{x'x'}$, $\sigma'_{y'y'}$, $\sigma_{zz}$ are normal stresses in the rotated coordinate system, $\tau_{x'y'}$ is the shear stress, $\pi'_{11}$, $\pi'_{12}$ and $\pi'_{44}$ are the piezoresistivity coefficients in the primed coordinate system, $\pi_{12}$ is the piezoresistivity coefficient in the original coordinate system, and $\phi'$ is the angle between the transistor channel and $x'$-axis, typically 0 or $\pi/2$. The piezoresistivity coefficients are taken from [14] and listed in Table 2.1.

Stress can also cause a shift in the transistor threshold voltage due to three effects: change in the silicon electron affinity, bandgap, and valence band density-of-states [39]. Mechanical strain in the transistor channel, given by the strain tensor $\epsilon$, could induce shifts and splits in the conduction band and balance band and therefore the threshold
Table 2.1: Piezoresistivity Coefficients \((\times 10^{-12}\text{Pa}^{-1})\) in (100) Si

<table>
<thead>
<tr>
<th></th>
<th>(\pi_{11})</th>
<th>(\pi_{12})</th>
<th>(\pi_{44})</th>
<th>(\pi'_{11})</th>
<th>(\pi'_{12})</th>
<th>(\pi'_{44})</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMOS</td>
<td>1022.0</td>
<td>-537.0</td>
<td>136.0</td>
<td>310.5</td>
<td>174.5</td>
<td>1559.0</td>
</tr>
<tr>
<td>PMOS</td>
<td>-66.0</td>
<td>11.0</td>
<td>-1381.0</td>
<td>-717.5</td>
<td>662.5</td>
<td>-77.0</td>
</tr>
</tbody>
</table>

Voltage is changed with strain tensor in Cartesian coordinate system. The stress and strain tensors can be related using Hooke’s law. The threshold voltage variations can be computed as [22]:

\[
q \Delta V_{tn} = m \Delta E_C - (m - 1) \Delta E_V
\]

\[
q \Delta V_{tp} = m \Delta E_V - (m - 1) \Delta E_C
\]

where \(\Delta V_{tn}\) and \(\Delta V_{tp}\) are the changes in NMOS and PMOS threshold voltages, respectively, \(q\) is the electron charge, and \(m\) is the body-effect coefficient and typically takes values ranging from 1.1 to 1.4. The term \(\Delta E_C\) is the minimum conduction band potential change over carrier band number \(i\), \(\Delta E_C^{(i)}\), while \(\Delta E_V\) denotes the maximum of the changes in valence band potentials between heavy-hole (hh) and light-hole (lh), which can be noted by \(\Delta E_V^{hh}\) and \(\Delta E_V^{lh}\). These are given by:

\[
\Delta E_C^{(i)}(\epsilon) = \Xi_d (\epsilon_{xx} + 2 \epsilon_{yy} + 3 \epsilon_{zz}) + \Xi_u \epsilon_{ii}, i \in \{x, y, z\}
\]

\[
\Delta E_V^{(hh, lh)}(\epsilon) = a (\epsilon_{xx} + \epsilon_{yy} + 2 \epsilon_{zz})
\]

\[
\pm \sqrt{\frac{b^2}{4} (\epsilon_{xx} + \epsilon_{yy} - 2 \epsilon_{zz})^2 + \frac{3 b^2}{4} (\epsilon_{xx} - \epsilon_{yy})^2 + d^2 \epsilon_{xy}^2}
\]

where \(\Xi_d\) and \(a\) are the hydrostatic deformation potential constants, which can induce shifts in the conduction band and valence band, respectively, while \(\Xi_u\), \(b\), and \(d\) are the shear deformation potential constants that affect the conduction and valence bands. The description of stress effect on CMOS devices is referred to [40].

### 2.3.2 Stress Effect on OTFTs

The structure of a single OTFT is illustrated in Fig. 2.1, and shows one bottom gate electrode buried in the gate dielectric material and another two electrodes, source and drain, on the top. The organic semiconductor material is placed between the top contacts and dielectric layer and the current in the channel is controlled by the voltage applied to the gate electrode [27, 28].
The OTFTs considered in this thesis are based on conjugated molecular systems with alternating single and double carbon-carbon bonds, and their molecular p-bonding orbitals are responsible for the ability of the material to transport charge [41,42]. Regioregular poly(3-hexylthiophene) (P3HT) is a type of alkyl-substituted polythiophenes, which have excellent solubility in a variety of organic solvents, and thin films are readily prepared by spin-coating, dip-coating, drop-coating, screen printing, or inkjet printing [28]. Stable doping of P3HT has been realized by adding strong molecular acceptors, such as 2,3,5,6-tetrafluoro7,7,8,8-tetracyanoquinodimethane (F_4TCNQ), which can undergo efficiently hole transfer for P3HT. Thus, the P3HT-based OTFT is a p-type device [43].

Table 2.2: The Strain-Mobility Relationship in P3HT OTFT [1]

<table>
<thead>
<tr>
<th>$\epsilon_{11}$ (%)</th>
<th>0.42</th>
<th>0.84</th>
<th>1.26</th>
<th>1.75</th>
<th>1.98</th>
<th>2.45</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta \mu/\mu$ (%)</td>
<td>2.80</td>
<td>4.20</td>
<td>5.60</td>
<td>7.20</td>
<td>7.90</td>
<td>9.50</td>
</tr>
</tbody>
</table>

The slope of the fitted linear relationship between the relative change in mobility and strain is 3.3. In other words, a tensile strain with the value of 1% will cause a 3.3% increase in P3HT OTFT, while a compressive strain with the value of −1% will cause 3.3% mobility degradation in OTFT. It can be shown that P3HT based OTFTs have a good mechanical flexibility, the sensitivity between mobility variation
and strain can reach a much higher value of 26 using alternative methods for building flexible electronics using amorphous silicon-based inorganic TFTs [45].
Chapter 3

Stress-Induced Performance Shifts in 3D DRAMs

3D DRAMs are considered to be the first commercial product to bring 3D stacking and TSVs into the mainstream memory market [46]. Products related to this technology, used in application domains such as mobile phones [7], high-end servers [47], graphics/computing units [48], low-cost hardware design with machine learning [49–55], and new storage devices [56–61], stack multiple DRAM layers in the vertical direction, with all layers connected with TSVs that can transmit signals including data, address, and power signals [2,7,8]. Each layer contains not only DRAM cells, but also addressing and other peripheral circuitry. A primary benefit of the implementation of 3D stacking and TSV is the significant improvement of DRAM cell density comparing to conventional DRAMs: since each layer is similar to a conventional DRAM, the packing density linearly increases with the number of stacked layers. Moreover, 3D DRAMs can improve the memory bandwidth with shorter latency paths that travel across 3D layers through the TSVs. Conventional DRAMs, such as the DDRx family, are pin-count-limited and must use long off-chip transmission lines to interconnect memory modules; in contrast, 3D DRAMs replace these off-chip lines with an on-chip bus within the 3D structure [15]. Therefore, stacked 3D DRAMs are excellent candidates for applications that show a demand for high bandwidth and low power memory, including mobile devices.

Traditionally, a rigid package substrate, using materials such as FR-4, has been
used to provide mechanical support and protection to a memory chip. However, as we will show, significant stress will be induced into the 3D DRAMs during the annealing process since the package substrate has the largest CTE and shrinks much faster than other layers. As a result, large warpage-induced stresses occur after the annealing, which are up to more than 40% of the total stress as shown in Section 3.3.2.

In this work, we explore the use of polyimide as an alternative to FR-4. PI is a bendable, elastic and lightweight material and is increasingly used as the package substrate material in various applications for flexible electronics, such as flexible displays [29], RFIDs [26], SiFs [11], flexible sensor array [10] and DRAM array [25]. Compared to the traditional package substrate material, PI has smaller CTE [62], which can effectively reduce the warpage-induced stresses and further reduce the impact on the performance of 3D DRAMs. Thus, we consider the impact of using PI as the package substrate instead of more rigid materials such as the traditional FR-4. We present detailed results for the PI substrate and a comparison with the rigid substrate in Section 3.3.2.

Pieces of the stress-induced performance variation analysis problem have attracted prior attention, but no work has addressed the complete problem of performance shifts in 3D-stacked memories incorporating all stress sources. The work in [14] discusses the stress caused by a single TSV rather than the total stress due to a large array of TSVs, of the type seen in 3D DRAMs. In [19], a method for obtaining the stress distribution in 3D ICs is proposed based on linear superposition of local-scale stress due to TSVs, μ-bumps, and package bumps. However, this approach still requires significant runtime for layouts with large numbers of TSVs, μ-bumps, and package bumps in 3D DRAMs. Both works have analyzed logic circuits, considering device-level or gate-level variations due to stress, rather than performance variations of a memory array. Additionally, there are some works to evaluate the performance of new storage devices [63–65] and systems [66,67]. For the storage reliability issues, new types of erase coding algorithms are exploited in [68,69].

The contributions of this chapter are in developing a unifying procedure that combines the impact of all sources of stress in the entire structure of a 3D DRAM, and analyzing the impact of this stress on memory performance parameters. Compared to the expensive FEA method or other analytical methods in previous works, our semianalytical model provides a fast method for computing the stress in an entire 3D DRAM
by modeling the stress caused by TSV stripes and clusters accurately. We use this analysis technique to explore the impact of changes in the TSV layout on memory system performance in 3D DRAMs.

### 3.1 Performance Evaluation of 3D DRAMs

Modern transistors use strained silicon, implemented by introducing *intrinsic* stress induced by materials that introduce lattice mismatches to enhance device mobilities, and hence the drive current and switching speed. We consider the effects of *extrinsic* stress caused by TSVs, µ-bumps, package bumps, and warpage.

Extrinsic stress on transistors perturbs the mobility and threshold voltage of MOS devices, with the magnitude of the perturbation being determined by the stress. These device parameter shifts are translated into variations in the performance of the 3D DRAM at the system level. Such an evaluation requires a system-level simulation, and we build upon the infrastructure of CACTI-3DD [70], an architecture-level integrated power, area, and timing modeling framework for 3D stacked DRAM main memory, to model the impact of stress-induced memory performance variations. Note that CACTI-3DD is built on top of CACTI 6.5 [71], and while it includes TSV models and 3D integration models to enable the evaluation of timing, power, and area for 3D DRAM, stress-induced variations are not modeled.

#### 3.1.1 Memory Organization and Performance

The 3D DRAM array model (Fig. 3.1) consists of multiple ranks with mutually exclusive access; each rank has several identical banks that can be accessed simultaneously. A
bank is divided into identical subbanks, each consisting of multiple mats. During a read/write access, all mats in a subbank are activated. There are four subarrays in a mat that share predecoding and decoding circuitry, and each subarray has DRAM cells with its own associated peripheral circuitry, such as precharge circuits, decoders, MUXes, and sense amplifiers.

**Timing:** The row cycle time is the time interval between two successive row accesses, and is limited by the time it takes to activate a wordline, sense the data, write back the data, and then precharge the bitlines. Thus row cycle time can be calculated as [72]:

\[
t_{RC} = t_{\text{row-dec-drv}} + t_{BL} + t_{\text{SA}} + t_{\text{writeback}} + t_{\text{WL-reset}}
\]

\[
+ \max(t_{\text{BL-pre}}, t_{\text{BL-mux-pre}}, t_{\text{SA-mux-pre}})
\]

(3.1)

Here, \(t_{\text{row-dec-drv}}\) is the delay of row decoding path including row predecoders, decoders and wordline drivers, \(t_{BL}\) and \(t_{SA}\) are the delay of bitline and sense amplifier, \(t_{\text{writeback}}\) is the time to write data back to DRAM cell after read operation, and \(t_{WL\text{-reset}}, t_{BL\text{-pre}}, t_{BL\text{-mux-pre}}, \text{and } t_{SA\text{-mux-pre}}\) are, respectively, the times to reset the wordline, and precharge the bitline, bitline MUX and sense amplifier MUX. These terms are described in the Appendix.

**Power:** The primary impact of leakage current in a DRAM is felt by the storage elements in the DRAM core. A 1T1C DRAM memory cell stores data in the capacitor and uses the access transistor to connect the cell to the bit lines. Leakage through the access transistor, when it is nominally off, impacts the retention time of the memory, and larger leakage necessitates more frequent refreshes, resulting in larger refresh power. The minimum refresh period, \(T_{\text{refresh}}\), is bounded by the retention time, \(T_{\text{retention}}\), of a DRAM array, which is given by:

\[
T_{\text{retention}} = \frac{C_{\text{cell}} \Delta V_{\text{cell}}}{I_{\text{leak}}}
\]

(3.2)

where \(\Delta V_{\text{cell}}\) is the worst-case capacitor voltage that leads to a read failure, and \(I_{\text{leak}}\) is the worst-case leakage in a DRAM cell.

The refresh power, \(P_{\text{ref}}\), of the 3D DRAM can be modeled as:

\[
P_{\text{ref}} = \frac{E_{\text{refresh}}}{T_{\text{refresh}}}
\]

(3.3)

where \(T_{\text{refresh}} = T_{\text{retention}}\) is the refresh period and \(E_{\text{refresh}}\) is the energy of a refresh operation. The contributors to \(E_{\text{refresh}}\) include the refresh predecoders, refresh decoder
drivers, and the refresh bitline, and correspond to charging/discharging capacitances, as detailed in [72]. These quantities are independent of stress, but the refresh period is strongly affected by stress and influences $P_{ref}$.

### 3.1.2 The Impact of Stress on 3D DRAM Performance

From the Appendix, it can be seen that the components of (3.1) correspond to a set of RC products, where the resistance is influenced by the device threshold voltage and mobility, which in turn are affected by extrinsic stress. For example, in computing gate delays, $R_{on} \propto 1/I_{on}$, and $I_{on}$ is directly affected by the variations of mobility and threshold voltage. The refresh power depends on the leakage current, $I_{leak}$, and is affected by the same transistor parameters. For current $I_x, x \in \{on, leak\}$, we model the perturbations as:

$$I_x^{stress} = I_x^{nom} + \frac{\partial I_x}{\partial V_t} \Delta V_t^{stress} + \frac{\partial I_x}{\partial \mu} \Delta \mu^{stress}$$  \hspace{1cm} (3.4)

where $I_x^{stress}$ is the current after incorporating the effect of extrinsic as well as intrinsic stress, $I_x^{nom}$ is the nominal current considering only intrinsic stress within the transistor, $\Delta V_t^{stress}$ and $\Delta \mu^{stress}$ are the stress-induced variations in threshold voltage and mobility, and $\partial I_x/\partial V_t$ and $\partial I_x/\partial \mu$ are the sensitivities corresponding to the variations in threshold voltage mobility, respectively.

We calibrate this linear model of $I_{on}$ and $I_{leak}$ for the range of mobility and threshold voltage shifts seen in our experiments. The leakage changes exponentially with the threshold voltage, but for the range of variation due to stress, we find that the above local linear approximation is sufficient. Under a 16nm PTM model, the maximum error of our perturbation model is 4.48% for $I_{leak}$ and 2.16% for $I_{on}$.

### 3.2 Stress Modeling of a 3D DRAM Stack

The equations that describe stress are linear, justifying the use of linear superposition to combine stress from various sources. The three extrinsic stress sources listed in Section 1.1.1 can be classified into:

- **Layout-dependent stress**, $\sigma_{LD}$, is induced by the stress sources related to layout, specifically stresses caused by the locations of the TSVs and $\mu$-bumps relative to
various blocks in the layout.

- **Layout-independent stress**, $\sigma_{LI}$, does not vary with the layout: here, this corresponds to warpage caused by the CTE mismatch between layers and stress induced by package bumps. Intrinsic stress is also layout-independent.

By linear superposition, we can perform the tensor addition:

$$\sigma_{total} = \sigma_{LD} + \sigma_{LI}$$ (3.5)

to compute the total stress, $\sigma_{total}$. We use this concept to conduct finite element analysis (FEA) simulations for core structures, use them to build semianalytical models for $\sigma_{LD}$ and $\sigma_{LI}$, and then apply these models to compute $\sigma_{total}$ for various TSV layouts. This method avoids expensive FEA simulations for stress on each layout.

Consider a 8Gb 3D DRAM with four stacked memory chips, similar to [2], as shown in Fig. 1.1. Each layer is thinned from the wafer thickness of $\sim 300\mu m$ thickness down to 50$\mu m$, and the chips are stacked in a B2F manner, with the device layer near the bottom surface of each DRAM layer. Based on the models within CACTI-3DD, the length, width, and height of the 3D DRAM stack are determined to be 4.5mm, 3.2mm, and 380$\mu m$, respectively.

TSVs are used to transmit data and power signals through the stack, and underfill layers and $\mu$-bumps are present between each memory chip layer. An underfill layer and a set of package bumps are added between the master chip and the package substrate. The dimensions of the TSV, $\mu$-bumps, and package bumps are listed in Table 3.1, where $D$, $H$, and $P$ are the diameter, height, and pitch, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$D$</th>
<th>$H$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSV</td>
<td>20$\mu m$</td>
<td>50$\mu m$</td>
<td>25$\mu m$</td>
</tr>
<tr>
<td>$\mu$-bump</td>
<td>20$\mu m$</td>
<td>10$\mu m$</td>
<td>25$\mu m$</td>
</tr>
<tr>
<td>Package bump</td>
<td>100$\mu m$</td>
<td>50$\mu m$</td>
<td>300$\mu m$</td>
</tr>
</tbody>
</table>

The entire 3D DRAM structure undergoes a thermal load of $\Delta T = -250^\circ C$ as it is annealed from 275$^\circ C$ to 25$^\circ C$ (room temperature) to represent the annealing process. We consider the worst-case scenario here, since the operating temperature of 3D DRAMs
Table 3.2: Material Parameters in 3D DRAMs

<table>
<thead>
<tr>
<th>Material</th>
<th>CTE (ppm/K)</th>
<th>Young’s Modulus (GPa)</th>
<th>Poisson Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>2.3</td>
<td>188</td>
<td>0.27</td>
</tr>
<tr>
<td>Cu</td>
<td>17</td>
<td>110</td>
<td>0.35</td>
</tr>
<tr>
<td>SiO₂</td>
<td>0.5</td>
<td>71</td>
<td>0.17</td>
</tr>
<tr>
<td>substrate</td>
<td>17.6</td>
<td>19.7</td>
<td>0.13</td>
</tr>
<tr>
<td>pkg-bump</td>
<td>22</td>
<td>44.4</td>
<td>0.35</td>
</tr>
<tr>
<td>µ-bump</td>
<td>20</td>
<td>26.2</td>
<td>0.35</td>
</tr>
<tr>
<td>HC_TSV</td>
<td>9.69</td>
<td>149</td>
<td>0.31</td>
</tr>
<tr>
<td>HC_µ-bump</td>
<td>10.3</td>
<td>48.5</td>
<td>0.26</td>
</tr>
</tbody>
</table>

will typically be higher than room temperature during their use. Note that although the TSVs and package assembly are conducted at different times, each has the same thermal load [73,74]. The materials in the stack shrink differentially due to their differing CTEs, inducing thermal stress. All material parameters are summarized in Table 3.2. We assume all materials are stress-free at the beginning of the annealing process and neglect the stress induced by wafer thinning. Backside grinding after CMOS fabrication is a widely used method for thinning wafers/chips down below 100µm for use in 3D stacking chips. The process of grinding induces defects on the backside of the wafer and causes stress that can bend an unsupported thin substrate. However, the maximum absolute value of the stress caused by wafer thinning is less than 10MPa [75], which is only about 3% compared to the total thermomechanical stress in this work and is therefore negligible.

In principle, it is possible to perform FEA to compute the resulting stress profile in the 3D structure. FEA proceeds by first meshing the structures into small polyhedral subdomains called elements, and then constructs a set of equations relating the stress at neighboring vertices of the polyhedra to each other, and enabling polynomial interpolation within the body of the element.

For sufficiently fine meshing, the FEA solution is accurate but can be computationally costly. For our problem, the TSV size is in tens of µm, implying that elements should be in the µm range. For a chip whose area of the chip is several mm, the number of elements becomes very large, and is computationally prohibitive for the problem.
of design planning, where multiple layout configurations must be explored. We introduce two simplifications that are effective in making the computation tractable while maintaining accuracy:

- Replacing a mass of TSVs in silicon by an equivalent material with the same volume fraction, and
- Building a semianalytical model, to be used with linear superposition, for stress analysis.

(1) Volume fraction: A rectangular region of dimension $W \times L$ containing $N$ TSVs, as shown in Fig. 3.2, can be replaced by a homogeneous cuboid. If $HC_{TSV}$ is the material (typically, silicon) of the cuboid that contains the TSV (where the TSV is typically made of copper), then the homogeneous approximation is a cuboid whose CTE is a weighted function of the CTEs of TSV and surrounding chip, where the weights correspond to the relative volume of each material. The volume fractions, $\alpha_{TSV}$ and $\alpha_{Si}$, are:

$$\alpha_{TSV} = N \cdot \left( \frac{\pi R^2 H}{W \cdot L \cdot H} \right) ; \quad \alpha_{Si} = 1 - \alpha_{TSV} \quad (3.6)$$

where $R$, is the radius of the TSV and $H$ is the height of the layer. The CTE of the homogeneous cuboid ($HC_{TSV}$) is then given by

$$CTE_{HC_{TSV}} = \alpha_{TSV} \cdot CTE_{TSV} + \alpha_{Si} \cdot CTE_{Si} \quad (3.7)$$

A similar method is also applied to $\mu$-bumps embedded in underfill to replace these nonhomogeneous regions by the equivalent homogeneous cuboids with an appropriate
CTE. Fig. 3.2 shows the results of FEA simulation for a cluster of $5 \times 5$ TSVs, as against the results when the TSVs and $\mu$-bumps are replaced by a volume fraction approximation.

To test the accuracy of volume fraction approximation, we tested 60 points in both (a) and (b). Then we compute the error caused by volume fraction in (b) by comparing the stress values sampled in (a). The updated error distribution has a mean of 0.30% and a variance of 1.84E-04. The worst-case error is $-3.10\%$, which demonstrates the accuracy of the volume fraction approximation.

To further verify the validity of the volume fraction approximation, we have determined the accuracy of the volume fraction approximation over different TSV distributions. We keep the dimension of the structure and the number of layers identical to the structure as shown in Fig. 3.2 and change the following parameters:

(a) The structure of the TSV arrangement: configuration S1 is a TSV stripe of $20 \times 1$ TSVs, as against the square configuration in Fig. 3.2.

(b) The number of TSVs: configuration S2 changes the different number of TSVs in the TSV cluster, as compared to Fig. 3.2.

(c) The density of TSVs: configuration S3 uses $5 \times 5$ TSVs, distributed with a larger pitch of 40\(\mu\)m. This leads to a lower TSV distribution density and a smaller volume fraction of the TSV region, according to (3.6).

The results and error distributions for these cases are as shown in Table 3.3 demonstrate that the volume fraction approximation is very accurate across a range of TSV numbers, distributions, and densities.

Table 3.3: Summary of Error Distributions between FEA and Volume Fraction Approximation in S1–S3 in 3D DRAMs

<table>
<thead>
<tr>
<th>Structure</th>
<th># TSVs</th>
<th>TSV pitch ((\mu)m)</th>
<th>Average error</th>
<th>Variance</th>
<th>Worst-case error</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>20 x 1</td>
<td>25</td>
<td>-0.25%</td>
<td>8.79E-05</td>
<td>-2.47%</td>
</tr>
<tr>
<td>S2</td>
<td>3 x 3</td>
<td>25</td>
<td>-0.43%</td>
<td>1.60E-04</td>
<td>-3.20%</td>
</tr>
<tr>
<td>S3</td>
<td>5 x 5</td>
<td>40</td>
<td>-0.41%</td>
<td>2.26E-04</td>
<td>-2.77%</td>
</tr>
</tbody>
</table>

(2) *Semianalytical Modeling and Superposition:* Our objective is to perform fast evaluation of a set of TSV layouts to determine the impact of stress-induced performance
shifts. According to (3.5), $\sigma_{LI}$ is independent of layout decisions, and therefore, we first generate a methodology to separate these stresses from the layout-dependent stresses, $\sigma_{LD}$. The stresses $\sigma_{LI}$ must be computed just once for a given die dimension and can be computed with FEA using a volume fraction simplification to curb the computation time. Layout-dependent effects are then computed using a semianalytical model and superposed through tensor addition to determine the total stress.

Layout-independent stress is caused by warpage and package bumps, which are largely independent of the layout of 3D DRAMs: package bump locations are determined by the choice of packaging, and warpage due to layout effects (as seen as the residual stress after backside grinding) has been shown to be negligible [75]. The major contributors to stress are (a) the CTE mismatch between different layers, which can cause significant warpage and induce stress into 3D DRAMs, and (b) CTE mismatch between package bumps and surrounding materials. Since stress analysis involves the solution of a linear partial differential equation, stress effects can be superposed. Thus, layout-independent stress can be described as the sum of the two stress sources, warpage, $\sigma_{\text{warpage}}$, and package bumps, $\sigma_{\text{pkg-bump}}$, by linear superposition:

$$\sigma_{LI} = \sigma_{\text{warpage}} + \sigma_{\text{pkg-bump}}$$  (3.8)

To compute the warpage-induced stress, we simulate the 3D stack with no TSVs, $\mu$-bumps or package bumps and apply the thermal load of $\Delta T = -250^\circ\text{C}$, and find the stress, $\sigma_{\text{warpage}}$ induced by the warpage due to CTE mismatch between different layers. Our interest is in computing stress in device layer, which means that the $z$ coordinate is a constant, and the layout-independent stress is a function only in term of the $x$ and $y$ coordinates.

To compute the stress caused by package bumps, $\sigma_{\text{pkg-bump}}$, we simulate another 3D structure with no TSVs or $\mu$-bumps but containing one package bump in the underfill layer between the substrate and nethermost DRAM layer. The $\sigma_{LI}$ caused by both warpage and the package bump is generated with FEA simulation. Then the stresses purely caused by the package bump are calculated using (3.8).

Fig. 3.3 shows the representative stress component $\sigma_{11}$ of $\sigma_{\text{pkg-bump}}$, as a function of the distance from the center of the package bump along vertical direction. The radius of the package bump is 50$\mu$m and the yellow region represents the region right above
Figure 3.3: $\sigma_{11}$ component of $\sigma_{pkg-bump}$ along the vertical direction.

the package bump. The maximum value of $\sigma_{pkg-bump}$ is -48.7MPa, and is reached at the center. A negative sign represents the compressive stress and the stress is positive (tensile) along the horizontal direction. The value of $\sigma_{pkg-bump}$ reduces quickly as we move away from the center and is effectively negligible beyond a certain distance, which we call the effective influence zone of the package bump. Based on this simulation, we set this effective influence zone for a package bump to a radius of 400µm, as shown by the red line in the figure. Since 3D DRAMs contain multiple package bumps, $\sigma_{pkg-bump}$ is the superposition of all $K$ package bumps in the effective influence zone:

$$
\sigma_{pkg-bump} = \sum_{i=1}^{K} \sigma_{pkg-bump,i}
$$

Our 4.5mm × 3.2mm die can accommodate 150 TSVs in a row and 120 TSVs in a column, and we consider TSVs laid out in rows, columns, and clusters of various sizes. For instance, for a TSV row, we consider five possible widths $w$ of 50µm to 250µm and

Figure 3.4: Accuracy of our semianalytical model for a TSV cluster: the solid curve fits the blue sample points. The horizontal axis is the set of evaluated points (7 values of $r$ evaluated at 6 values of $w$).
sample the stress at seven distances, $r$, from the edge of the row. Since stress typically reduces as $1/r$, the points are chosen appropriately spaced. Based on these 35 samples from FEA analysis using ABAQUS, we subtract out the layout-independent component, $\sigma_{LI}$, and build a semianalytical model of the form $\sigma_{LD} = k_1 + k_2r + k_3/r + k_4w$. A similar approach is taken for a TSV column and for a square TSV cluster, except that for a TSV cluster, we build separate models for $r$ above/below the cluster and to the left/right of the cluster. Note that like a single TSV, a TSV cluster would induce tensile $\sigma_{11}$ stress along the $x'$-axis and compressive $\sigma_{11}$ stress along the $y'$-axis. For TSVs and $\mu$-bumps distributed in row and column stripes, only compressive stress occurs in the area close to the long edges. Fig. 3.4 shows that the model provides excellent accuracy. The error distribution generated with 42 points shows an average error of 0.36% with a variance of $1.3 \times 10^{-3}$, and the worst-case error is 6.59%. Similar accuracies are obtained for TSV stripes (rows/columns).

The approach is generalizable to any layout and requires FEA-based precharacterizations of just three structures: rows, columns, and clusters. Repeated cheap evaluations of the semianalytical model can then be used the explore the space of TSV layouts, computing the stress for a layout with $N$ TSV stripes and $M$ TSV clusters as:

$$\sigma_{total} = \sigma_{LI} + \sum_{i=1}^{N} \sigma_{TSV\_stripe\_i} + \sum_{i=1}^{M} \sigma_{TSV\_cluster\_i}$$

(3.10)

### 3.3 Experimental Results in 3D DRAMs

We investigate a set of TSV layouts for an 8Gb 4-layer 3D DRAM array. The TSVs are arranged in some combination of (a) rows, where each row contains 150 TSVs, (b) columns, with 120 TSVs per column, and (c) clusters. The maximum number of TSVs in each row/column is decided by the size of 3D DRAM structure and the pitch of TSVs. Eight TSV layouts are described in Table 3.4. The rows may appear at the top, middle, or bottom, and the columns may appear in one of five equally spaced locations from left to right. The precise distribution of rows and columns is shown in parentheses. The TSV clusters appear in an array, with the number of rows and columns in parentheses. For example, for L1, all 1200 TSVs in L1 are distributed in the middle as a stripe containing 8 rows; L2 contains 3 TSV stripes, each of which has 2, 4, and 2 rows at the top, middle, and bottom, respectively; L3 and L4 contain both TSV rows and columns...
with the same total number of TSVs as L1 and L2; L5–L7 use a $6 \times 6$ arrangement of TSVs in each cluster; L8 uses a $5 \times 5$ arrangement of TSVs. The total number of TSVs is around 1200 in all cases, of which $2/3$ are used for data and $1/3$ for power distribution. Distributing the TSVs throughout the layout reduces data latency over a concentration of TSVs as in L1.

Table 3.4: Summary of TSV Distributions in L1–L8

<table>
<thead>
<tr>
<th>Layout</th>
<th>TSV rows</th>
<th>TSV columns</th>
<th>TSV clusters</th>
<th># TSVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>8 (0,8,0)</td>
<td>-</td>
<td>-</td>
<td>1200</td>
</tr>
<tr>
<td>L2</td>
<td>8 (2,4,2)</td>
<td>-</td>
<td>-</td>
<td>1200</td>
</tr>
<tr>
<td>L3</td>
<td>4 (0,4,0)</td>
<td>5 (0,1,3,1,0)</td>
<td>-</td>
<td>1200</td>
</tr>
<tr>
<td>L4</td>
<td>4 (1,2,1)</td>
<td>5 (1,1,1,1,1)</td>
<td>-</td>
<td>1200</td>
</tr>
<tr>
<td>L5</td>
<td>-</td>
<td>-</td>
<td>32 ($6 \times 6$)</td>
<td>1152</td>
</tr>
<tr>
<td>L6</td>
<td>6 (2,2,2)</td>
<td>-</td>
<td>8 ($6 \times 6$)</td>
<td>1188</td>
</tr>
<tr>
<td>L7</td>
<td>3 (0,3,0)</td>
<td>4 (0,1,2,1,0)</td>
<td>8 ($6 \times 6$)</td>
<td>1218</td>
</tr>
<tr>
<td>L8</td>
<td>3 (1,1,1)</td>
<td>5 (1,1,1,1,1)</td>
<td>8 ($5 \times 5$)</td>
<td>1250</td>
</tr>
</tbody>
</table>

Figure 3.5: Contours of $\sigma_{11}$ in eight layouts corresponding to different TSV configurations.
3.3.1 Using FR-4 as the Package Substrate Material

The spatial distribution of TSVs in L1–L8 is apparent in Fig. 3.5, which shows the contours of $\sigma_{11}$, as a representative stress component, for each structure in the master chip placed at the bottom of the stack, which experiences the largest stress. Since the region of interest is outside the TSV clusters/stripes, for convenience the color code inside the TSV regions shows zero stress within. Each stress contour translates to a map of mobility and threshold voltage variations, and Fig. 3.6 shows the data corresponding to L8 in Fig. 3.5(h). NMOS transistors near TSV stripes and clusters suffer a mobility degradation up to $-11\%$, while PMOS transistors lying over and under lateral TSV stripes and clusters suffer a mobility degradation up to $-24\%$. For PMOS transistors at the left and right edge of TSV columns and clusters, the mobility can increase by up to 26%.

![Graphs showing variations in mobility and $V_t$](image)

Figure 3.6: Variations in mobility and $V_t$ in L8.

For both NMOS and PMOS devices, the stress-induced shifts are negative for $\Delta E_C$ and positive for $\Delta E_V$. As a result, the bandgap is smaller so that the absolute values of threshold voltages for both NMOS and PMOS transistors decreases. The maximum variation occurs near TSV stripes and clusters, with threshold voltage variations for NMOS (PMOS) transistors of up to $-25mV$ ($15mV$). This leads to faster switching
speeds and larger leakage currents, i.e., latency is improved but leakage power and refresh power are aggravated.

**Timing:** The computed stress tensors translate to variations in transistor parameters. We now analyze the impact of stress on system timing for L1–L8. We focus on $t_{RC}$, defined in (3.1), but similar analyses can be performed for other timing metrics. The $t_{RC}$ variation contours in L1–L8 are shown in Fig. 3.7 for $\phi = \pi/2$, and it can be seen that $t_{RC}$ increases in the region above and below TSV rows and clusters, but decreases to the left and right of TSV columns and clusters (the latency variations would change signs if $\phi = 0$). Moreover, TSV clusters create larger $t_{RC}$ shifts than TSV rows or columns since they induce larger mobility variations, especially for PMOS transistors.

![Figure 3.7: $t_{RC}$ variation contours in eight layouts corresponding to different TSV configurations.](image)

The latency performance of a 3D DRAM is usually limited by the worst-case values of $t_{RC}$. The maximal and minimal $t_{RC}$ variations in L1–L8 are summarized in the columns 2–5 of Table 3.5. All percentage changes are with respect to $D_0$, the nominal $t_{RC}$ without the effect of stress for L1, and $\Delta D^+$ and $\Delta D^-$ are the best-case and worst-case shifts in $t_{RC}$, respectively. Structures with TSV clusters suffer more significant $\Delta D^-$ of up to 7.0%.

**Power:** Based on the shifts in $V_t$ and mobility, the contours of $I_{leak}$ are shown in Fig. 3.8. Transistors near TSV stripes suffer significant variations, with shifts of up to 50% seen in L1, with the widest TSV stripe. TSV clusters induce larger variations, of
Table 3.5: Row Cycle Time \( t_{RC} \), Leakage Power \( P_{\text{leak}} \), and Refresh Power \( P_{\text{ref}} \) for L1–L8

\( (D_0 = 33.62\text{ns}, P_{\text{leak}}^{\text{nom}} = 50.66\text{mW}, P_{\text{ref}}^{\text{nom}} = 18.90\text{mW}) \)

<table>
<thead>
<tr>
<th></th>
<th>Row Cycle Time ( t_{RC} )</th>
<th>Leakage ( P_{\text{leak}} )</th>
<th>Refresh ( P_{\text{ref}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \Delta D^+ ) (ns) ( \Delta D^+ ) (%) ( \Delta D^- ) (ns) ( \Delta D^- ) (%)</td>
<td>( \Delta P_{\text{leak}} ) (mW) ( \Delta P_{\text{leak}} ) (%)</td>
<td>( \Delta P_{\text{ref}} ) (mW) ( \Delta P_{\text{ref}} ) (%)</td>
</tr>
<tr>
<td>L1</td>
<td>-0.96 -2.9% 0.91 2.7%</td>
<td>12.35 24.4% 9.53 50.4%</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>-0.96 -2.9% 0.94 2.8%</td>
<td>12.12 23.9% 8.74 46.2%</td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td>-1.30 -3.9% 0.83 2.5%</td>
<td>11.52 22.7% 8.98 47.5%</td>
<td></td>
</tr>
<tr>
<td>L4</td>
<td>-1.54 -4.6% 0.95 2.8%</td>
<td>12.04 23.8% 8.85 46.8%</td>
<td></td>
</tr>
<tr>
<td>L5</td>
<td>-3.89 -11.6% 2.26 6.7%</td>
<td>13.06 25.8% 16.11 85.2%</td>
<td></td>
</tr>
<tr>
<td>L6</td>
<td>-3.98 -11.8% 2.32 6.9%</td>
<td>12.35 24.4% 16.38 86.7%</td>
<td></td>
</tr>
<tr>
<td>L7</td>
<td>-3.95 -11.7% 2.34 7.0%</td>
<td>11.89 23.5% 16.55 87.6%</td>
<td></td>
</tr>
<tr>
<td>L8</td>
<td>-3.87 -11.5% 2.20 6.5%</td>
<td>12.45 24.6% 15.68 83.0%</td>
<td></td>
</tr>
</tbody>
</table>

up to 88% in L5–L8.

The last four columns of Table 3.5 show the variations of leakage power, \( P_{\text{leak}} \), and refresh power, \( P_{\text{ref}} \), in L1–L8. All percentage changes are with reference to the nominal leakage power, \( P_{\text{leak}}^{\text{nom}} \), and the nominal refresh power, \( P_{\text{ref}}^{\text{nom}} \), for L1 in the absence of stress-induced leakage shifts. Across layouts, \( \Delta P_{\text{leak}} \) varies only slightly since it is dominated by layout-independent stress (layout-dependent stress is diluted when averaged over the chip). However, \( \Delta P_{\text{ref}} \) is bounded by the worst-case as it is constrained by the worst retention time, and is thus a serious problem, with TSV clusters (L5–L8) inducing larger \( \Delta P_{\text{ref}} \) than TSV stripes.

**Area:** Significant variations in timing and especially in refresh power are induced by the stress in memory chips, particularly near the TSVs. To avoid these, we maintain a keep-out-zone (KOZ) for a TSV array in which no transistor may be placed. We define the KOZ as a rectangular region within which \( \Delta P_{\text{ref}} \) larger than 30%, and measure the area overhead associated with the KOZ in Table 3.6. The figure of 30% was chosen to maintain a manageable area for the KOZ: the corresponding areas for a 25% threshold are much larger. Here, \( A_{\text{TSV}} \), \( A_{\text{KOZ}} \), and \( A_{\text{total}} \) are, respectively, the area overhead caused by TSVs, their KOZs, and the sum. The nominal area of each DRAM chip is 14.4mm\(^2\). The overhead lies between 13.2% and 26.0% and is largest for L8. Note that
Figure 3.8: Subthreshold leakage current variations in eight layouts corresponding to different TSV configurations.

L2, with three TSV stripes, has a higher area overhead than L3, with four TSV stripes since TSV stripes near the chip edge cause a larger $I_{\text{leak}}$ increase than those in the middle, as shown in Figs. 3.7(b) and (d), owing to the additional warpage stress which is more pronounced near the edge of the chip.

Table 3.6: Area Overhead of TSV and KOZ for L1–L8 in 3D DRAMs

<table>
<thead>
<tr>
<th>Layout</th>
<th>$A_{\text{TSV}}$ (mm$^2$)</th>
<th>$A_{\text{TSV}}$ (%)</th>
<th>$A_{\text{KOZ}}$ (mm$^2$)</th>
<th>$A_{\text{KOZ}}$ (%)</th>
<th>$A_{\text{total}}$ (mm$^2$)</th>
<th>$A_{\text{total}}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.75</td>
<td>5.2%</td>
<td>2.18</td>
<td>15.1%</td>
<td>2.93</td>
<td>20.4%</td>
</tr>
<tr>
<td>L2</td>
<td>0.75</td>
<td>5.2%</td>
<td>2.03</td>
<td>14.1%</td>
<td>2.78</td>
<td>19.3%</td>
</tr>
<tr>
<td>L3</td>
<td>0.75</td>
<td>5.2%</td>
<td>1.16</td>
<td>8.0%</td>
<td>1.91</td>
<td>13.2%</td>
</tr>
<tr>
<td>L4</td>
<td>0.75</td>
<td>5.2%</td>
<td>2.35</td>
<td>16.3%</td>
<td>3.10</td>
<td>21.5%</td>
</tr>
<tr>
<td>L5</td>
<td>0.72</td>
<td>5.0%</td>
<td>2.85</td>
<td>19.8%</td>
<td>3.57</td>
<td>24.8%</td>
</tr>
<tr>
<td>L6</td>
<td>0.74</td>
<td>5.2%</td>
<td>2.56</td>
<td>17.8%</td>
<td>3.30</td>
<td>22.9%</td>
</tr>
<tr>
<td>L7</td>
<td>0.76</td>
<td>5.3%</td>
<td>1.78</td>
<td>12.4%</td>
<td>2.54</td>
<td>17.7%</td>
</tr>
<tr>
<td>L8</td>
<td>0.78</td>
<td>5.4%</td>
<td>2.97</td>
<td>20.6%</td>
<td>3.75</td>
<td>26.0%</td>
</tr>
</tbody>
</table>

**Wire length and bus delay:** In 3D DRAMs, memory bus routing uses a spine-like structure [70] that is used to connect the logic control circuits to the banks and finally to each memory cell, as shown in Fig. 3.9. Due to the KOZ, the memory arrays must be
spaced away from the TSV array, as shown in the figure, to avoid the KOZ. As a result, the chip area increases and so does the wire length of global buses used in the memory. Fig. 3.9 shows the increase in wire length caused by KOZ for layout configuration L1. The increased wire length results in an increase in the bus delay. For each of the eight layouts, L1, ···, L8, Table 3.7 shows the percentage increase in the bus delay variation, $\Delta D_{bus}$, with respect to the nominal bus delay of 0.93ns. We see that L7 shows the largest $\Delta D_{bus}$ of 14.3%.

Table 3.7: Bus Delay Variation for L1–L8 in 3D DRAMs

<table>
<thead>
<tr>
<th>Layout</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
<th>L8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta D_{bus}$ (%)</td>
<td>1.5%</td>
<td>2.4%</td>
<td>2.5%</td>
<td>3.4%</td>
<td>9.6%</td>
<td>10.0%</td>
<td>14.3%</td>
<td>14.0%</td>
</tr>
</tbody>
</table>

Runtime: FEA is computational: an L1-like layout with 400 TSVs requires 4 hours of CPU time (Intel Xeon 5560 Nehalem, 2.80GHz); with 800 TSVs, it times out after a day. Our volume fraction method computes L1 (1200 TSVs) in 98s, and our semianalytical model only requires a few clock cycles (2 multiplies, 1 divide, 3 adds). Even for the L5 layout, which has the most TSV clusters, our model evaluates the entire chip using 64 multiplies, 32 divides, and 127 adds.

3.3.2 Using Polyimide as Package Substrate Material

As shown in Fig. 3.5, warpage-induced layout-dependent stress takes a large proportion of the total stress and the most significant $\sigma_{11}$ value can reach more than -110MPa, which is caused by the CTE mismatch between different layers and is more than 40% of the total stress in L1–L4. According to Table 3.2, it can be found that the CTE of the
package substrate layer (FR-4) is 17.6, which is much larger than those of underfill layer (SiO$_2$) and DRAM layer (Si). During the annealing process, substrate layer shrinks much faster than other layers resulting in a downward warpage and significant stress.

To reduce warpage-induced stress, we propose to substitute the bendable polyimide substrate material for the traditional package substrate, as shown in Fig. 3.10, to decrease the CTE mismatch between substrate and other layers. The CTE of the PI substrate can be as low as 3 ppm/K [62], which is close to that of DRAM layer (2.3 ppm/K) and underfill layer (0.5 ppm/K). As a result, the PI substrate shrinks much more slowly with temperature than the traditional FR-4 package substrate during the annealing process, which can reduce the warpage after the annealing. Furthermore, the underfill layer has the smallest CTE and can help to prevent the chip from deformation more strongly by competing against the weaker PI substrate.

![Figure 3.10: 3D DRAM structure using the PI substrate.](image)

We investigate the same set of TSV layouts as summarized in Table 3.4 to compare
the results between 3D DRAMs with the traditional package substrate and proposed PI substrate. Fig. 3.11 shows the contours of the representative stress component $\sigma_{11}$ for each layout. Comparing to the stress results for 3D DRAMs with the FR-4 package substrate as shown in Fig. 3.5, there is a shift of about 100MPa in each layout. For example, the range of $\sigma_{11}$ in layout L1 with the FR-4 package substrate, as shown in Fig. 3.5, is from $-250\text{MPa}$ to $0\text{MPa}$, where the negative sign represents compressive stress. However, $\sigma_{11}$ in the same layout with PI substrate ranges between $-150\text{MPa}$ and $0\text{MPa}$. In other words, the substitution of the substrate material can effectively reduce warpage and the corresponding stress. Based on our simulations, the most significant value of warpage-induced $\sigma_{11}$ in 3D DRAMs with the PI substrate is only about $-2\text{MPa}$, which is occurring in the center of the chip. All other layouts L2–L8, containing TSV stripes, TSV clusters or both, show the similar results. Moreover, TSV clusters induce more significant stress than TSV stripes, which can be seen by comparing contours of $\sigma_{11}$ in the first row and second row in Fig. 3.11.

![Stress Contours](image)

**Figure 3.12:** $t_{RC}$ variation contours in the eight layouts using the PI substrate.

The computed stress tensors are then translated into variations in electrical parameters with the approach as shown in Section 2.3.1. The method discussed in Section 3.1 is used to generate the stress-induced performance variation of 3D DRAMs with the PI substrate.

**Timing:** We analyse the effect of stress on row cycle time $t_{RC}$ for L1–L8 with the PI substrate. The contours of $t_{RC}$ variation in the eight layouts are shown in Fig. 3.12.
Table 3.8: Row Cycle Time ($t_{RC}$), Leakage Power ($P_{leak}$), and Refresh Power ($P_{ref}$) for L1–L8 with PI substrate

\[D_0 = 33.62\text{ns}, \ P_{nom}^{leak} = 50.66\text{mW}, \ P_{nom}^{ref} = 18.90\text{mW}\]

<table>
<thead>
<tr>
<th></th>
<th>Row Cycle Time $t_{RC}$</th>
<th>Leakage $P_{leak}$</th>
<th>Refresh $P_{ref}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta D^+$ (ns)</td>
<td>$\Delta D^+$ (%)</td>
<td>$\Delta D_{PI}^{FR}$ (ns)</td>
</tr>
<tr>
<td>L1</td>
<td>-0.58</td>
<td>-1.7%</td>
<td>0.88</td>
</tr>
<tr>
<td>L2</td>
<td>-0.56</td>
<td>-1.7%</td>
<td>0.79</td>
</tr>
<tr>
<td>L3</td>
<td>-1.31</td>
<td>-9.9%</td>
<td>0.82</td>
</tr>
<tr>
<td>L4</td>
<td>-1.27</td>
<td>-3.8%</td>
<td>0.69</td>
</tr>
<tr>
<td>L5</td>
<td>-3.67</td>
<td>-10.9%</td>
<td>2.16</td>
</tr>
<tr>
<td>L6</td>
<td>-3.74</td>
<td>-11.1%</td>
<td>2.23</td>
</tr>
<tr>
<td>L7</td>
<td>-3.76</td>
<td>-11.2%</td>
<td>2.24</td>
</tr>
<tr>
<td>L8</td>
<td>-3.66</td>
<td>-10.9%</td>
<td>2.09</td>
</tr>
</tbody>
</table>

From L1–L4 it can be found that the $t_{RC}$ variation caused by TSV stripes ranges between -1.31ns and 0.88ns. It is seen that TSV clusters can induce more significant $t_{RC}$ variation by comparing L5–L8, which contain TSV clusters, with L1–L4 with only TSV stripes. The maximum and minimum $t_{RC}$ variations are summarized in Table 3.8 in columns 2–7. All percentage changes are with respect to $D_0$, the nominal $t_{RC}$ without the effect of stress for L1, and $\Delta D^+$ and $\Delta D^-$ are the best-case and worst-case shifts in $t_{RC}$, respectively, and $\Delta D_{PI}^{FR}$ represents the change between structures with the FR-4 and PI substrates, with respect to $D_0$. The layouts with TSV clusters suffer more significant $t_{RC}$ variation because of the larger stress. The worst-case shift reaches 2.24ns (6.7%) as shown in L7. Moreover, $\Delta D_{PI}^{FR}$ results show that the PI substrate reduces the $t_{RC}$ variations in all the eight layouts by reducing warpage-induced stress. The worst-case $t_{RC}$ shifts is reduced by 0.1% to 0.8%, with respect to $D_0$.

**Power:** The contours of variations in the subthreshold leakage current, $I_{leak}$, for the eight layouts are shown in Fig. 3.13, incorporating the influence of stress-induced $V_t$ and mobility variations. The structures containing TSV clusters suffer larger $I_{leak}$ increase and the maximum $I_{leak}$ shift occurring in L7 is 68.3%. Moreover, 3D DRAMs with the PI substrate suffer much less $I_{leak}$ variation than those with FR-4 substrate, as seen by comparing Fig. 3.8 and Fig. 3.13.

The leakage power, $P_{leak}$, and refresh power, $P_{ref}$, are affected by $I_{leak}$, as discussed in Section 3.1, and columns 8–13 in Table 3.8 show the variations of $P_{leak}$ and $P_{ref}$. All percentage changes are with reference to the nominal leakage power, $P_{nom}^{leak}$, and
the nominal refresh power, $P_{\text{nom}}^{\text{ref}}$, and $\Delta^{\text{FR-4}}_{\text{PI}}$ denotes the variation between structures with the FR-4 substrate and PI substrate, with respect to $P_{\text{nom}}^{\text{leak}}$ and $P_{\text{nom}}^{\text{ref}}$, respectively. Typically, TSV clusters can induce more significant variation in both $P_{\text{leak}}$ and $P_{\text{ref}}$. By observing $\Delta^{\text{FR-4}}_{\text{PI}}$ results, it can be driven there is significant drop in both $P_{\text{leak}}$ and $P_{\text{ref}}$ in the structures with the PI substrate comparing to those with traditional FR-4 package substrate. With the substitution of PI for substrate material, there is an average variation of $-17.8\%$ in $P_{\text{leak}}$ and an average shift of $-19.3\%$ in $P_{\text{ref}}$, with respect to $P_{\text{nom}}^{\text{leak}}$ and $P_{\text{nom}}^{\text{ref}}$, respectively.

**Area:** The PI substrate can significantly reduce $I_{\text{leak}}$ and $P_{\text{ref}}$, and hence the area of the KOZ, which is used to avoid large increases in $P_{\text{ref}}$, can be reduced naturally. Using the threshold $\Delta P_{\text{ref}} > 30\%$ for deciding the KOZ, the results for L1–L8 are summarized in Table 3.9. Here, $\Delta^{\text{FR-4}}_{\text{PI}}$ denotes the difference in percentage between structures with the FR-4 substrate and PI substrates, with respect to the nominal DRAM area.

The TSV-induced area overhead remains the same since the chip layouts with the PI substrate are identical to those with the FR-4 substrate. However, the KOZ area $A_{\text{KOZ}}$ is greatly reduced in L1–L4, which contain only TSV stripes after applying PI substrate. Additionally, there is a significant reduction in $A_{\text{KOZ}}$ caused by TSV clusters as well, as shown in L5–L8. The average variation of $A_{\text{KOZ}}$ among L1–L8 is $-14.3\%$ and the maximum reduction is $-19.3\%$, occurring in L8.
Table 3.9: Area Overhead of TSV and KOZ for L1–L8 in 3D DRAMs with PI substrate

<table>
<thead>
<tr>
<th>Layout</th>
<th>$A_{TSV}$ (mm$^2$)</th>
<th>$A_{TSV}$ (%)</th>
<th>$A_{KOZ}$ (mm$^2$)</th>
<th>$A_{KOZ}$ (%)</th>
<th>$A_{FR-4}$ (mm$^2$)</th>
<th>$A_{total}$ (mm$^2$)</th>
<th>$A_{total}$ (%)</th>
<th>$A_{FR-4}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.75</td>
<td>5.2%</td>
<td>0.01</td>
<td>0.1%</td>
<td>-15.1%</td>
<td>0.76</td>
<td>5.3%</td>
<td>-15.1%</td>
</tr>
<tr>
<td>L2</td>
<td>0.75</td>
<td>5.2%</td>
<td>0.00</td>
<td>0.0%</td>
<td>-14.1%</td>
<td>0.75</td>
<td>5.2%</td>
<td>-14.1%</td>
</tr>
<tr>
<td>L3</td>
<td>0.75</td>
<td>5.2%</td>
<td>0.00</td>
<td>0.0%</td>
<td>-8.0%</td>
<td>0.75</td>
<td>5.2%</td>
<td>-8.0%</td>
</tr>
<tr>
<td>L4</td>
<td>0.75</td>
<td>5.2%</td>
<td>0.00</td>
<td>0.0%</td>
<td>-16.3%</td>
<td>0.75</td>
<td>5.2%</td>
<td>-16.3%</td>
</tr>
<tr>
<td>L5</td>
<td>0.72</td>
<td>5.0%</td>
<td>0.78</td>
<td>5.4%</td>
<td>-14.4%</td>
<td>1.50</td>
<td>10.4%</td>
<td>-14.4%</td>
</tr>
<tr>
<td>L6</td>
<td>0.74</td>
<td>5.1%</td>
<td>0.21</td>
<td>1.5%</td>
<td>-16.3%</td>
<td>0.95</td>
<td>6.6%</td>
<td>-16.3%</td>
</tr>
<tr>
<td>L7</td>
<td>0.76</td>
<td>5.3%</td>
<td>0.22</td>
<td>1.5%</td>
<td>-10.9%</td>
<td>0.98</td>
<td>6.8%</td>
<td>-10.9%</td>
</tr>
<tr>
<td>L8</td>
<td>0.78</td>
<td>5.4%</td>
<td>0.18</td>
<td>1.3%</td>
<td>-19.3%</td>
<td>0.96</td>
<td>6.7%</td>
<td>-19.3%</td>
</tr>
</tbody>
</table>

**Wire length and bus delay:** As the KOZ is reduced, the wire length and bus delay overhead decreases. Table 3.10 shows the bus delay variation in L1–L8 with the PI substrate and $\Delta_{FR-4}^{PI}$ is the difference between structures with the FR-4 and PI substrates. The figure shows a significant decrease in bus delay, especially in the layout with only TSV stripes, L1–L4. There is no bus delay overhead in L2–L4 after applying PI substrate.

Table 3.10: Bus Delay Variation for L1–L8 in 3D DRAMs with PI Substrate

<table>
<thead>
<tr>
<th>Layout</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
<th>L8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta D_{bus}$ (%)</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>9.0%</td>
<td>5.4%</td>
<td>5.4%</td>
<td>5.1%</td>
</tr>
<tr>
<td>$\Delta_{FR-4}^{PI}$ (%)</td>
<td>-1.0%</td>
<td>-2.4%</td>
<td>-2.5%</td>
<td>-3.4%</td>
<td>-0.6%</td>
<td>-4.6%</td>
<td>-9.0%</td>
<td>-9.0%</td>
</tr>
</tbody>
</table>

Heat dissipation is an important problem of 3D chips because of two reasons: (a) 3D stacking increases the dissipated heat per unit footprint, and (b) the conventional package substrate made of FR-4 has a much smaller thermal conductivity value than other layers in 3D chips. As reported in [76], the thermal conductivity of FR-4 is 0.29W/m·K while Si has a much larger thermal conductivity of 168W/m·K. PI has a similar thermal conductivity of 0.35W/m·K [76], which is marginally better than FR-4. Therefore, PI does not introduce new heat removal problems, and methods that are used in conventional FR-4 substrates can carry over to PI substrates. Furthermore, there are some works show that the thermal conductivity of PI can be enhanced up to
1.2 W/m·K by compositing with boron nitride (BN) [77, 78], which may even slightly ease the thermal bottleneck.

While all of these simulations indicate that for performance reasons, PI is significantly better than FR-4 as the substrate material for 3D-DRAMs, there are several reasons why FR-4 has been widely used. First, as a traditional rigid substrate, it is more robust. Despite the better flexibility, PI provides limited mechanical support for stacked chips. As a result, FR-4 has been widely used in many applications where rigid substrate support is very important. Second, the cost of PI is much higher (2–3× than that of FR-4) and it involves higher assembly complexity. Third, FR-4 has better performance in moisture absorption: as reported in [79], PI can absorb 3.5X moisture than FR-4 under the same environment.

3.4 Conclusion

We have presented an approach for fast semianalytical stress modeling with modest precharacterization costs, which enables the exploration of a variety of TSV layouts. As a general rule, clustered structures create substantially more stress than layouts with horizontal and vertical stripes. This results in a net area loss due to the cost of the larger KOZ, as well as larger penalties in delay, leakage power, and communication latency. Layouts that use a single strip in the middle of the chip show the lowest stress overhead. Finally, we show that a flexible PI substrate can effectively reduce warpage-induced stress and improve memory performance as compared to the traditional rigid substrate.
Chapter 4

Strain-Aware Performance Evaluation and Correction for OTFT-Based Flexible Displays

Flexible electronics are built using bendable, elastic, and lightweight materials. Flexible circuits are often implemented using organic semiconductor materials and permit significant strain when a deformation is induced in the structure. These technologies are amenable to low-cost roll-to-roll manufacturing techniques, can be conformally shaped, and can be used to construct large-area structures [5]. As a result, they are being increasingly deployed in flexible displays [9], flexible sensor arrays [10], RFIDs [11], electronic paper, and SiFs [26].

Typically, a flexible display is periodically refreshed at a rate that is in the order of μs (at these rates, OTFTs switching speeds are fast enough), with all \( n \) pixel rows activated one by one using a shift register. The OTFT is turned on by activating the row/scan line, and the storage capacitor is charged/discharged through the column/data line. The voltage of the storage capacitor determines the pixel value of each display pixel. For a grayscale display, the pixel value ranges from 0 to 255, while a color display uses three sub-pixels in each pixel, each with a value ranging from 0 to 255, representing RGB values of the pixel. This work focuses on grayscale images, with 8-bit pixel values, but the concepts can be extended easily to RGB images.
This chapter performs a strain analysis of a flexible display under two types of deformations, bend and twist+press. Next, the corresponding performance variations of flexible displays are evaluated. Finally, two compensation schemes are proposed to correct the errors induced by stress.

4.1 Strain Modeling of a Flexible Display

![Figure 4.1: FEA simulation set up for two types of deformations in a flexible display.](image)

We conduct finite element analysis (FEA) simulations for the flexible display under various types of deformations using ABAQUS, from which the strain and stress can be obtained and then translated to device variations. The dimensions of the flexible display used in our FEA simulation are $50\text{mm} \times 50\text{mm} \times 0.5\mu\text{m}$ (length $\times$ width $\times$ thickness). The display contains $512 \times 512$ pixels and the corresponding resolution is $260$ ppi. The material of the structure is polyimide (PI), which is widely used for the substrate of flexible displays, and its thickness is $500\mu\text{m}$. The Young’s modulus of PI is $2.5\text{GPa}$ and the Poisson’s ratio equals to $0.34$ [80]. In this FEA simulation structure, we omit the device layer, located near the top of the substrate, for two reasons: first, because the OTFT layer is very flexible by design, and second, because its thickness is much lower than the thickness of the substrate [80].

During the daily use of the flexible display, it may suffer various of deformations, such as bending, twisting, pressing/stretching. Manufactured systems are typically stress-tested to determine their response to such deformations [81,82]. We now show the boundary conditions (BCs) for FEA simulations for two typical deformations.
Bend: The FEA simulation is for bend deformation is set up as shown in Fig. 4.1(a). The plane in green is the top view of the flexible display which is made of PI. First, the BC $U_Z = 0$ is applied to the two red dotted lines, where $U_Z$ denotes the displacement along the $z$-axis. This BC is used to fix the two red dotted lines along the $z$-axis, but it allows the lines slide in the $x$-$y$ plane during the bend process. Then a rigid beam in the shape of cylinder is placed on the top of the display and is used to bend the display. The BC $U_Z$ along the negative $z$-axis is applied to the beam and then the display is bent by the beam. When $U_Z$ is large enough value, the bend radius is approximately equal to the radius of the cylindrical rigid beam.

Twist+press: Figure 4.1(b) illustrates the set-up for the FEA simulation for the twist+press deformation. First, the left $y$-$z$ surface of the display is fixed. Then a BC $U_{RX} = 90^\circ$ is applied to the right $y$-$z$ surface of the display, where $U_{RX}$ donates the rotational degree of freedom along the $x$-axis. As a result, the surface is rotated by $90^\circ$ along the red dotted line, which is located at the center of the display and is parallel to the $x$-axis. Meanwhile, another BC $U_X$ is applied to the same surface along the negative $x$-axis. With this BC, the display is pressed along the $x$-axis.

The results of FEA simulation showing the strain induced by these two deformations on the top surface, corresponding to the location of the device layer, are displayed in Fig. 4.2. A negative sign represents compressive strain, while a positive value donates tensile strain. A rigid beam with a radius of 5mm is used to bend the display in this simulation, similar to typical values used for bending tests [83]. In this case the strain...
on the top surface is negative (compressive); on the other hand, the bottom surface (not shown) experiences an equal and opposite positive (tensile) strain. Under normal bending conditions, strains change gradually in space. Similarly, Fig. 4.2(b) shows the $\epsilon_{11}$ result induced by the twist+press deformation. In this simulation, two BCs, $U_X = -5\text{mm}$ and $U_{RX} = 90^\circ$ are applied to the display.

Unlike the bend deformation, which causes compressive strain in the upper layer and tensile strain in the lower layer (not shown), in these simulations, the twist+press deformation induces compressive strain only, all through the display structure (if the press deformation is lower, tensile strain can also be seen). Reversing the BC direction to $U_X = +5\text{mm}$ along the positive $x$-axis results in tensile strain.

Based on the contours in Fig. 4.2, this twist+press deformation can induce much more significant strain in the display than the bend deformation described here; as a result, it will cause more significant influence on the OTFT devices and thus on the entire display.

### 4.2 Performance Evaluation of Flexible Displays

Each pixel of the display contains a capacitor charged through an OTFT switch, as shown in Fig. 1.2. The whole display frame is refreshed periodically with a specific system-defined refresh frequency. The frame refresh period $t_f$ is the time interval in which the refresh operation for the whole display should be completed, and is the reciprocal of the refresh frequency [5]. Since each refresh operation activates the $n_{\text{row}}$ rows in the display in sequence, the row refresh period, $t_{\text{row}}$, is given by:

$$t_{\text{row}} = t_f/n_{\text{row}}$$

(4.1)

All pixels in the row are refreshed simultaneously.

A circuit representation of each pixel is shown in Fig. 4.3. Here, $V_{\text{data}}$ and $V_{\text{scan}}$ represent the voltage level of data and scan lines, respectively. The p-type OTFT switch, controlled by $V_{\text{scan}}$, allows access to the storage capacitor, $C_{st}$, whose charge reflects the pixel value. In each step, $C_{st}$ stores the pixel value at the current moment. This can be altered when the OTFT is turned ON through $V_{\text{scan}}$, when the voltage level of
$C_{st}$ is changed to new pixel value by charging/discharging the capacitor such that:

$$t_{\text{charge/discharge}} \leq t_{row} \quad (4.2)$$

where $t_{\text{charge/discharge}}$ is the time required to charge/discharge $C_{st}$ to $V_{data}$ successfully. If (4.2) is not satisfied, then the process of charging/discharging $C_{st}$ may be incomplete, and errors may be induced into the pixel values. In such a display, with 512 rows, at a refresh frequency of 120Hz, from (4.1), the refresh period $t_{row} = 16.3 \mu s$ represents the time to complete charge/discharge for every pixel.

As discussed in Sections 2.3.2 and 4.1, the deformations in a flexible display induce strain. In OTFTs, compressive strain causes mobility degradation due to the piezoresistive effect. Since the charge/discharge current through the OTFT is proportional to the device mobility, strain-induced mobility degradation will reduce the charge/discharge current, necessitating a longer time for $C_{st}$ to reach accurate pixel value. The range of $V_{DD} = 1$V is distributed across 256 pixel values, so that each grayscale level corresponds to 3.9mV. Assuming an allowable noise margin of half a level, charging is complete when the pixel value is $\pm 1.95$mV of the final value.

We perform HSPICE simulations to determine $V_{\text{disp}}$, the voltage across $C_{st}$ that is provided to the liquid crystal display pixel, for the circuit in Fig. 4.3(a). The OTFT device model is provided by the Organic Process Design Kit (OPDK) [84], a design kit that was specifically developed for OTFTs. We use a corner-based model for the transistor to capture the worst-case impact of process variations. Fig. 4.4 shows the transient voltage across $C_{st}$ as it charges from a pixel value of 0 (0V) to 255 (1V). The transient at the nominal (strain-free) mobility value, denoted by $V_{\text{disp}}$, is shown in the
upper curve, and has a charging period of 4.166 µs after $V_{data}$ switches at 2 µs. The same operation under a 50% mobility degradation is shown by the lower plot in the $V'_{disp}$ curve, and is seen to require a significantly longer period of 8.288 µs.

Note that if the voltage in the upper curve is examined before 4.166 µs, it will provide an approximation to the final value. In some cases, this approximation may be adequate and may be invisible to the user, but in other cases, it may corrupt the image significantly. The image is inherently resilient to such errors, and an error of about 5% in the pixel value may be acceptable in many cases. Even so, we will show that strain causes visible problems in rendered images.

The $V_{disp}$ charging transient is a function of mobility and the applied voltage, $V_{data}$, and the point of interest on this transient is its value at time $t_{row}$, when the charging process ends. At a fixed value of $t_{row}$, we build a linear model that captures the impact of strain on the final voltage value, over various values of $V_{data}$:

$$V_{disp}^{strain}(t_{row}) = V_{disp}^0 + k_VV_{data} + k_\mu\mu^{strain}$$

(4.3)

where $V_{disp}^{strain}$ is the voltage level of $V_{disp}$ after incorporating strain-induced mobility variation for a transition from 0 to $V_{data}$, $V_{disp}^0$, $k_V$, and $k_\mu$ are fitting constants, and $\mu^{strain}$ is the mobility under strain.

These models are built at a set of fixed values of $t_{row}$, and linear interpolation is used to evaluate the model at any intermediate values of $t_{row}$. Each such model is based
on HSPICE simulations using the OPDK. We generate data at 35 points (7 values of $V_{data} \times 5$ values of $\mu$) at each $t_{row}$ to obtain the function. This is a one-time exercise for a given technology and the cost of characterization is not large.

Combining the fitted linear model with the ability to interpolate at any value of $t_{row}$, we represent our display voltage model as:

$$V_{disp}^{strain} = f(V_{data}, \mu^{strain}, t_{row}) \quad (4.4)$$

Although the above model has been characterized for a transition from 0 to $V_{data}$, it may be used for any transient where the gap between the initial and final values of capacitor voltage is $V_{data}$. Intuitively, this is because the transition time depends on the RC time constant of the transition and the gap between the initial and final voltage levels. To see this more concretely, we use the RC model in Fig. 4.3, where $R$ is the equivalent OTFT resistance. Then

$$V_{disp}(t) = V_0 + (V_{data} - V_0)(1 - e^{-\frac{t}{RC_{st}}}) \quad (4.5)$$

The time required to charge $V_{disp}$ to a voltage $V$ is

$$t = RC_{st} \cdot \ln \frac{V_{data} - V_0}{V_{data} - V} \quad (4.6)$$

For the same charging range $V_{data} - V_0$, the time required to get within $\Delta V = V_{data} - V$ of the final value is the same. As our result is always determined by the $\Delta V$ value (e.g., 1.95mV for half a pixel of accuracy), our model can be used for any initial $V_0$ and any final $V_{data}$. Discharge transients are handled using similar principles.
Figure 4.6: Error histogram for the interpolation function for $V_{\text{strain}}^{\text{disp}}$.

(a) Bend

(b) Twist+press

Figure 4.7: OTFT mobility variation results for two types of deformations.

Fig. 4.5 shows the accuracy of the fitted linear model in Eq. (4.3) at the 35 sample points, corresponding to 7 values of $\mu$ and 5 values of $V_{\text{data}}$. The red dotted line shows the value evaluated from the linear function for $V_{\text{strain}}^{\text{disp}}$ at $t_{\text{row}} = 7.8\mu$s while the blue circles represent the 35 samples. To evaluate the quality of interpolation, we evaluate the interpolation at another 80 points, which differ from the points used to build the model. The error distribution of the error in percentage between the interpolation model in Eq. (4.4) and the samples is as shown in Fig. 4.6. As shown in the figure, the interpolation has a high accuracy while the mean and variance of the interpolation model is $-0.34\%$ and $2.69 \times 10^{-4}$, respectively.

4.3 Experimental Results in Flexible Displays

The mobility variation results for the two types of deformations in Section 4.1 are shown in Fig. 4.7. Since the bend deformation and twist+press deformation both induce
compressive strain with negative sign, the devices will suffer mobility degradation in both deformations. As shown in Fig. 4.7(a), the bend deformation can cause a mobility degradation up to $-14\%$, and the maximum degradation occurs at the center. The twist+press deformation will induce larger mobility degradation because it will induce more significant compressive strain. The maximum mobility degradation occurs near the corner and even the average mobility degradation can reach $-36\%$.

We apply the analysis from the previous sections to present a series of results that show the effect of the extrinsic strain caused by deformations. We consider the deformations in Section 4.1 for this evaluation. Note that other types of deformations may also be evaluated, but the intent behind using these specific cases is to mimic the behavior of the display after typical stress tests such as this. A similar approach may be followed to evaluate the degradation in the quality of a displayed image for any other stress test.

We use the SSIM [85,86], a widely-used metric that measures the structural similarity
between two images, to compare rendered images with the ideal. The SSIM metric ranges from $-1$ to $1$, and is close to $1$ when the images are nearly identical. The baseline image in our evaluations is the original image to be rendered.

The image Lena is chosen as the input that provides the inputs $V_{data}$ to all pixels of the display. The initial values of all pixels are based on the image Barbara, as shown in Fig. 4.8(a), and this corresponds to viewing the display switch from one image (Barbara) to another (Lena). These initial values are the initial voltages $V_{disp}$ across the storage capacitor $C_{st}$ in each pixel. The clock period $t_{row}$ is set to $7.8\mu s$. This value is chosen to ensure that the worst-case error is below $5\%$ in the absence of stress: Fig. 4.8(b) shows the transition from Barbara to Lena in the strain-free case. Here, the $5\%$ error margin provides excellent rendering at the scale shown.

Next, the same evaluation is repeated when the display undergoes the bend and twist+press deformations (described in Section 4.1), respectively. These results are shown in Fig. 4.8(c) and (d), respectively. Under strain, the charging/discharging process does not complete fully in the allocated time, $t_{row}$, and ghosts of the previous picture Barbara can still be distinguished in the result. The SSIM values for these three images are $0.9823$ for the strain-free image (b), and $0.6695$ and $0.6314$ for the images in (c) and (d), respectively.

Another example is shown in Fig. 4.9. The flexible display is refreshed from picture Girlface as shown in (a), to render the picture Cameraman. The stress-free case is shown in (b), while (c) and (d) show the result for bend and twist+press deformations. The difference between the pictures is very apparent visually, and through the SSIM metrics: $0.9665$ for (b), $0.7184$ for (c), and $0.6910$ for (d). The errors caused by strain can be more significant for displaying videos with frequent refreshes or in a display with more rows.

More SSIM results are summarized in Table 4.1. We show the results for $15$ pictures such as Baboon as shown in the first column with the influence under the strain from the two kinds of deformations. It is assumed that each picture is refreshed from Black, Lena, Barbara and Fingerprint respectively as shown in the first row. It can be found that strain in both deformations can cause significant errors. Larger errors will be caused by more significant strain induced in twist+press deformation. Furthermore, the SSIM result is affected by input pictures. An initial black image will barely affect
the structure of the next picture to be displayed, and is less disruptive; however the other three pictures will interfere with the structure of the new image in different ways, depending on the similarity between the images.

4.4 Correcting Strain-Induced Performance Loss

As discussed in Section 4.1–4.2, the strain induced in the use of the flexible display can cause the mobility degradation of OTFT devices. This mobility degradation can cause errors in the pixel values among the pixels of the flexible display. In this section, we propose two methods to compensate for the errors caused by the extrinsic strain.

Our approach is based on placing compact, flexible strain sensors in another layer added between the flexible substrate and matrix backplane to sense the presence of an extrinsic strain. Each sensor covers multiple rows and columns of the display array.
Table 4.1: Summary of SSIM Results in Flexible Displays with Stress

<table>
<thead>
<tr>
<th>Image</th>
<th>SSIM (Bend)</th>
<th>SSIM (T.+p.)</th>
<th>SSIM (Bend)</th>
<th>SSIM (T.+p.)</th>
<th>SSIM (Bend)</th>
<th>SSIM (T.+p.)</th>
<th>SSIM (Bend)</th>
<th>SSIM (T.+p.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baboon</td>
<td>0.9878</td>
<td>0.9832</td>
<td>0.8491</td>
<td>0.8137</td>
<td>0.8359</td>
<td>0.8017</td>
<td>0.7358</td>
<td>0.6904</td>
</tr>
<tr>
<td>Boat</td>
<td>0.9608</td>
<td>0.9486</td>
<td>0.7990</td>
<td>0.7622</td>
<td>0.7685</td>
<td>0.7324</td>
<td>0.6392</td>
<td>0.5911</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.9738</td>
<td>0.9624</td>
<td>0.8804</td>
<td>0.8494</td>
<td>0.8644</td>
<td>0.8312</td>
<td>0.8189</td>
<td>0.7824</td>
</tr>
<tr>
<td>Cameraman</td>
<td>0.8225</td>
<td>0.8151</td>
<td>0.6692</td>
<td>0.6341</td>
<td>0.6162</td>
<td>0.5806</td>
<td>0.4217</td>
<td>0.3897</td>
</tr>
<tr>
<td>Clown</td>
<td>0.6638</td>
<td>0.6234</td>
<td>0.7873</td>
<td>0.7485</td>
<td>0.7376</td>
<td>0.6964</td>
<td>0.7020</td>
<td>0.6637</td>
</tr>
<tr>
<td>Couple</td>
<td>0.9714</td>
<td>0.9619</td>
<td>0.7749</td>
<td>0.7337</td>
<td>0.7883</td>
<td>0.7490</td>
<td>0.6787</td>
<td>0.6340</td>
</tr>
<tr>
<td>Crowd</td>
<td>0.9571</td>
<td>0.9358</td>
<td>0.7984</td>
<td>0.7714</td>
<td>0.7638</td>
<td>0.7322</td>
<td>0.7009</td>
<td>0.6715</td>
</tr>
<tr>
<td>Feather</td>
<td>0.9490</td>
<td>0.9327</td>
<td>0.3983</td>
<td>0.3448</td>
<td>0.3899</td>
<td>0.3319</td>
<td>0.3538</td>
<td>0.3272</td>
</tr>
<tr>
<td>Girlface</td>
<td>0.8049</td>
<td>0.7884</td>
<td>0.7199</td>
<td>0.6814</td>
<td>0.6430</td>
<td>0.6085</td>
<td>0.4820</td>
<td>0.4395</td>
</tr>
<tr>
<td>Goldhill</td>
<td>0.9768</td>
<td>0.9650</td>
<td>0.8312</td>
<td>0.7955</td>
<td>0.7748</td>
<td>0.7364</td>
<td>0.7126</td>
<td>0.6718</td>
</tr>
<tr>
<td>Pirate</td>
<td>0.9730</td>
<td>0.9615</td>
<td>0.8084</td>
<td>0.7722</td>
<td>0.7564</td>
<td>0.7202</td>
<td>0.6493</td>
<td>0.6084</td>
</tr>
<tr>
<td>Plane</td>
<td>0.9953</td>
<td>0.9915</td>
<td>0.8027</td>
<td>0.7813</td>
<td>0.7483</td>
<td>0.7247</td>
<td>0.5376</td>
<td>0.5093</td>
</tr>
<tr>
<td>Tank</td>
<td>0.9887</td>
<td>0.9861</td>
<td>0.7258</td>
<td>0.6766</td>
<td>0.7345</td>
<td>0.6930</td>
<td>0.5588</td>
<td>0.4982</td>
</tr>
<tr>
<td>Truck</td>
<td>0.9772</td>
<td>0.9702</td>
<td>0.7867</td>
<td>0.7490</td>
<td>0.7394</td>
<td>0.6991</td>
<td>0.5692</td>
<td>0.5202</td>
</tr>
<tr>
<td>Zelda</td>
<td>0.9200</td>
<td>0.8980</td>
<td>0.7508</td>
<td>0.7102</td>
<td>0.6850</td>
<td>0.6466</td>
<td>0.5899</td>
<td>0.5510</td>
</tr>
</tbody>
</table>

Because the thickness of the sensor layer is small (100nm) and it is in close proximity to the *matrix backplane* layer and is effective in sensing the strain there.

Data from these sensors is fed back to a controller that applies compensating excitations to maintain image quality. We use the pentacene-semiconductor-based Wheatstone bridge strain sensor whose structure is shown in Fig. 4.10 [6]. The sensor is fabricated on the flexible substrate. A hybrid film of Ta$_2$O$_5$ and PVP is used as the gate dielectric layer and a thin pentacene film acts as the sensing area. A bottom-buried gate electrode controls the carrier channel in the pentacene layer while the other four top contacts are used as the input (the lateral pair) and output (the vertical pair) signals of the strain sensor. The two input contacts of the sensor bridge are used to apply the bias voltage.
Figure 4.10: Schematic of the pentacene field-controllable strain sensor based on hybrid gate dielectrics (high-k Ta$_2$O$_5$+PVP) [6].

while the output contacts read the output signal, which varies with the change in resistivity induced by the mechanical deformation, and can be translated into the amount of stress/strain and mobility variation.

From Eq. (4.4), $V_{disp}$ depends on $V_{data}$, $t_{row}$, and $\mu$. Under strain, the voltage waveform may be shifted from the nominal $V_{disp}$ to $V'_{disp}$ even as $V_{data}$ and $t_{row}$ are unchanged, i.e.,

$$V'_{disp} = f(V_{data}, \mu', t_{row})$$

(4.7)

To make the image immune to strain, we can return $V'_{disp}$ to $V_{disp}$ either by compensating $V_{data}$, by changing the input pixel values, or by using a longer $t_{row}$, adjusted using a voltage-controlled (VCO).

**V$_{data}$ compensation**: Instead of using the original input pixel value $V_{data}$, a compensating input $V'_{data}$ is generated with Eq. (4.7) to make

$$f(V'_{data}, \mu', t_{row}) = V_{disp}$$

(4.8)

Comparing to Eq. (4.7), a compensating $V'_{data}$ is applied instead of $V_{data}$ to enable the output to reach the target pixel value $V_{disp}$ at time $t_{row}$ under the changed mobility, $\mu'$.

**$t_{row}$ compensation**: The charge/discharge time is extended from the nominal value, $t_{row}$, to a larger value, $t'_{row}$. Since the mobility degradation reduces the OTFT current, it slows down the charge/discharge process: this compensation allows more to complete the transient.

$$f(V_{data}, \mu', t'_{row}) = V_{disp}$$

(4.9)
The value of $t'_{row}$ is controlled by clock signal (CLK) used for the column driver, and a voltage-controlled oscillator (VCO) [87], which is placed on a separate (nonflexible) board with other control circuits such as row/column drivers, allows its period to be changed.

The architecture for $V_{data}$ and $t_{row}$ compensation is shown in Fig. 4.11. First, a sensor layer with $10 \times 10$ sensors is added between flexible substrate and matrix backplane. The chosen $10 \times 10$ granularity reflects the earlier observation that under normal deformations, the spatial change in strains is gradual. Because the output of the sensor is an analog signal, an analog-to-digital (A/D) converter is added to each strain sensor. The 4-bit feedback-type A/D converters are used here, each of which consists of a 4-bit counter, a 4-bit D/A converter and a comparator; this can also be placed on the board with the drivers and does not need to be flexible. The compensating outputs are provided by a look-up table (LUT), which is implemented with a ROM. The digital data from the strain sensor addresses the LUT, and for each $V_{data}$ value, and compensating output is generated, depending on the compensation method that is used.

**$V_{data}$ compensation:** The inputs of the LUT for $V_{data}$ compensation method are the input pixel value for each column and the digital strain value from the nearest strain sensor. The output is the compensating input pixel value $V'_{data}$. Finally, by applying the compensating input pixel value $V'_{data}$ to the column driver, the accurate output pixel values are expected to show on the flexible display. The LUT takes 12 inputs – the 8-bit $V_{data}$ value and the 4-bit encoding of the strain sensor output from the A/D converter.
and outputs the compensated 8-bit $V'_{data}$ value. Thus, the size of the LUT is $4K \times 8$, i.e., 32Kb.

$t_{row}$ compensation: The input of the LUT for $t_{row}$ compensation approach is the digital strain value from the strain sensor and A/D converter. Since this compensation for $t_{row}$ is applied row-wise, it must be safe for each cell within the row, i.e., the compensating $t'_{row}$ must satisfy the worst-case charge/discharge time for any pixel in the row. This implies that we must consider the maximum strain in the row. Accordingly, the input strain value to the LUT is the maximum value among the strain sensors in the row, and the output $V'_{\text{tune}}$ is sent to the VCO to change its frequency.

As before, the LUT uses 12 inputs, corresponding to 8 $V_{data}$ lines 4 sensor data lines. An 8-bit output is provided: because the range of the $t_{row}$ is from 0 to $30\mu s$, this corresponds to a resolution of about $0.1\mu s$ in $t_{row}$. Thus, the capacity requirement for the LUT is the same as for $V_{data}$ compensation, and equals to 32Kb ($4K$ entries $\times 8$ bit $V'_{\text{tune}}$). The digital value of $V_{\text{tune}}$ is converted to analog form and sent to an 8-bit D/A converter that converts the digital value to an analog input signal for the VCO.

Assuming that both compensation modes are supported, the total overhead of the compensation approach is marked by the yellow blocks in Fig. 4.11. This includes the $10 \times 10$ sensor array, $10 \times 10$ A/D converters, a 32Kb LUT, a VCO and a 8-bit D/A converter. As stated above, the flexible sensors are placed in the sensor layer while other circuits are implemented with stable (inflexible) CMOS technology.

4.5 Experimental Results with the Compensation Schemes

As discussed in Section 4.4, two methods can be used to compensate the error induced by strain: altering $V_{data}$ and changing $t_{row}$ adaptively. The results of $V_{data}$ compensation are as shown in Fig. 4.12, where the display is refreshed from Barbara to Lena, as in Fig. 4.8(b). When the compensating input $V'_{data}$ is generated using the $V_{data}$ compensation method as discussed in Section 4.4, the corresponding value of $V'_{data}$ in each pixel would produce the image in (a) if the display were not stressed. The ghosts in this image (which is never rendered in this way, because the display is actually under stress) create a watermark that is the inverse of the case when $V_{data}$ is used on the stressed display. This effectively cancels out the ghosts in the bent picture, and render
the ghost-free image in (b). The SSIM result of (b), compared to the perfect Lena image is 0.9546, with the $10 \times 10$ spatial sensor resolution being the main accuracy limiter. With $512 \times 512$ sensors the scheme can fully compensate the errors and reach a SSIM result equals to 1.

![New input $V_{\text{data}}$](image1)

![Result with compensation](image2)

Figure 4.12: Results for displaying Lena with $V_{\text{data}}$ compensation in an OTFT array.

A second method to correct the strain-induced error is through $t_{\text{row}}$ compensation, which stretches the clock to lengthen the row charge/discharge time. Fig. 4.13 shows several results under the bend-induced strain as the display changes from Barbara to Lena, as before. The four images to SSIM values of 0.65, 0.70, 0.80, and 0.99. The corresponding values of $t_{\text{row}}$ are $3.7 \mu s$, $12.3 \mu s$, $21.4 \mu s$, and $29.1 \mu s$, respectively, in (a), (b), (c), and (d). It can be seen that when $t_{\text{row}}$ is large enough, the result is very close to the input. As compared to $V_{\text{data}}$ compensation, this method requires a $\sim 4 \times$ larger value of $t_{\text{row}}$ for comparable quality. As seen in Fig. 4.4, the amount of time required to reach high levels of accuracy can increase greatly because the $V_{\text{disp}}$ curve has a very low slope late in its transient. This increase in $t_{\text{row}}$ may be acceptable in some applications and not others.

For a more fine-grained range of $t_{\text{row}}$, Fig. 4.14 shows the relationship between SSIM results and $t_{\text{row}}$, where the blue line represents the curve with the bend-induced strain while the orange line is the relationship under the deformation of twist+press, both from Section 4.1. As expected, a better quality result can be obtained with longer charge/discharge time, at the cost of slower rendering.

As compared to $V_{\text{data}}$ compensation, $t_{\text{row}}$ compensation significantly reduces the
number of LUT lookups. The former requires $512 \times 512$ lookups, one for each pixel. In $t_{\text{row}}$ compensation, each row of the strain sensor array covers multiple rows of the display array, and the largest strain in any row of the sensor array determines the $t_{\text{row}}$ value for all display rows that it covers. Thus, this scheme requires only 10 lookups to the LUT. However, to reach high quality, $t_{\text{row}}$ compensation needs a longer refresh time $t'_{\text{row}}$, unlike $V_{\text{data}}$ compensation, which keeps $t_{\text{row}}$ unchanged.

### 4.6 Conclusion

We have presented a performance analysis of flexible displays while incorporating the impact of realistic extrinsic strain. It is seen that deformations can significantly affect the quality of an image on the OTFT display. Based on data from an array of flexible strain sensors, two compensation methods are proposed: $V_{\text{data}}$ compensation, which
Figure 4.14: Relationship between SSIM and $t_{row}$.

adjusts the data value sent to each pixel, and $t_{row}$ compensation, which alters the time allowed for charging/discharging each pixel. $V_{data}$ compensation can be used in scenarios in which a fixed $t_{row}$ is preferred, while $t_{row}$ compensation can be used for applications where fewer ROM accesses are preferred, the system can tolerate a change in $t_{row}$, and slightly lower quality results are acceptable if the acceptable increase in $t_{row}$ is limited.
UTCs are adopted in the flexible SiF applications where high performance is requested while provides excellent flexibility by thinning the Si chip down to 20µm. In this chapter, the stress analysis is first performed with two packaging schemes, Middle Chip and Top Chip. Next, the performance variations of the two core elements, an SAR ADC and an SRAM, in a typical flexible ECG SiF application are analyzed. Particularly, the bend-induced stress could cause errors in SAR ADCs and increase latency and power of SRAMs. Thus, overdesign parameters are introduced to compensate stress-induced variations in both elements.

5.1 Stress Modeling of a Flexible System-in-Foil with Ultra-Thin Chips

The bending test is a widely used method to test the reliability of flexible electronics. The performance variations of device and circuits are captured by bending the system with a radius ranging from tens of millimeters to 2mm [88,89]. For example, the wearable ECG monitoring SiF is designed to fit the curved surfaces of the human skin and must be
capable of sustaining such deformations during the daily usage. It is common practice to translate these deformations to a bending test, and we analyze bending-induced stress and corresponding performance shifts at the device level and the ADC/SRAM block level in this work. Note in contrast to the stress modeling of flexible displays as discussed in Section 4.1, in which the OTFT device layer is ignored due to its small thickness, the interest in this chapter is to analyze the stress distributions of UTCs in SiF applications. Here, UTCs, which is made of Si and contains CMOS transistors, can be packaged either between two PI layers or atop a flexible PI substrate.

We conduct finite element analysis (FEA) simulations using ABAQUS for the flexible system with UTC packages for the two SiF schemes shown in Fig. 1.3. The structure used in FEA simulations is as shown in Fig. 5.1 and the dimensions of the whole structure is $40\text{mm} \times 40\text{mm} \times 120\mu\text{m}$ (length $\times$ width $\times$ thickness). The plane in yellow is the top view of the flexible substrate made of PI, and the UTC Si chip is shown in green. Note that in Middle Chip packaging scheme, the chip is buried between two PI layers, while it is placed on the top of the flexible substrate in Top Chip package. The total thickness of the flexible system is set to $120\mu\text{m}$ for both packaging schemes. In Middle Chip packaging, the UTC is $20\mu\text{m}$ thick, and both the bottom substrate and the encapsulation PI layer are set to $50\mu\text{m}$ [33]. The Top Chip package uses a chip the same thickness and a substrate of $100\mu\text{m}$, thus providing a fair stress comparison with the Middle Chip package by maintaining the same package thickness. Various chip sizes ranging from $20\text{mm} \times 20\text{mm}$ to $2\text{mm} \times 2\text{mm}$ are simulated in this work to determine the relationship between chip size and stress distribution in both schemes. The Young’s modulus of Si is 188GPa, the Poisson’s ratio is 0.27, while the Young’s modulus of PI
is 2.5GPa and the Poisson’s ratio is 0.34 [80].

The boundary conditions (BCs) used in the FEA simulations are summarized here. First, the BC $U_Z = 0$ is applied to the two red dotted lines, where $U_Z$ denotes the displacement along the $z$-axis. This BC is used to fix the two red dotted lines along the $z$-axis, but it allows the lines to slide in the $x$-$y$ plane during the bend process. A rigid beam in the shape of cylinder is then placed on the top of the structure and is used to bend the SiF. The BC $U_Z$ along the negative $z$-axis is applied to the beam and the structure is bent by it. The value of $U_Z$ is set so that the bend radius equals to the radius of the cylindrical rigid beam, which is 5mm in this work [88, 90].

Fig. 5.2 shows the FEA simulation results for both Middle Chip and Top Chip packaging methods. The figures show the stress maps of $\sigma_{11}$ near the top x-y surface of the UTC, where the devices are located. Fig. 5.2(a) shows the stress distribution of a 20mm $\times$ 20mm UTC: the stress reaches $-218$MPa at the center, where the negative sign represents compressive stress. For the same chip dimension and bending conditions, a larger stress is induced by the Top Chip packaging scheme (Fig. 5.2(b)), where $\sigma_{11}$ goes up to $-550$MPa. Other stress tensors, such as $\sigma_{22}$, $\sigma_{33}$, and $\tau_{12}$ are also evaluated and used to analyze the performance of CMOS devices in the flexible system.

For a fixed chip thickness of 20$\mu$m, the relationship between the chip size and stress is studied for various chip sizes from 20mm $\times$ 20mm to 2mm $\times$ 2mm, based on FEA simulations.

**Middle Chip:** Simulations show that the bending-induced stress is independent of the
chip size. As illustrated in Fig. 5.3, different chip sizes – 20mm × 20mm, 10mm × 10mm, 5mm × 5mm, and 2mm × 2mm – show similar peak stress values. In fact, the stress profile of a smaller chip is essentially the same as a cutout of that size from the stress profile of a 20mm × 20mm chip. This is illustrated in Fig. 5.2(a), where the stress contours for various chip sizes can be seen. The ranges of the most significant $\sigma_{11}$ values range between $-216\text{MPa}$ and $-220\text{MPa}$. Interestingly, while the peak stress is the same for all chip sizes, the average stress goes down with the chip size as regions with lower stress are included in the chip.

**Top Chip:** Similar FEA simulations show that the peak stress here is very dependent on the chip sizes (Fig. 5.3). However, the peak stress is related to the chip size and the less significant stress is induced by the smaller chip in Top Chip packaging. The peak value of $\sigma_{11}$ can be reduced from $-550\text{MPa}$ to $-320\text{MPa}$ by reducing to chip size to 2mm × 2mm. Note that even with a smaller chip size, the peak stress value is still larger than that in a Middle Chip package.

### 5.2 Performance Evaluation of ADCs and SRAMs in SiFs

The performance variation of SiF systems, such as the flexible ECG monitoring system in Fig. 1.4, are strongly dependent on variations in the ADC and SRAM. According to Sec. 2.3.1, the bending-induced stress would cause the degradation in device mobility for both NMOS and PMOS, and the absolute values of threshold voltages are lowered by the stress in both types of devices. The shifts in these device parameters then affect the performance of the circuits in the system. In this work, we choose the 8-bit SAR
Figure 5.4: Architecture of an 8-bit SAR ADC.

ADC and the on-chip SRAM to study the stress-induced performance variations.

5.2.1 Performance Evaluation of SAR ADCs

A standard 8-bit charge-redistribution SAR ADC implemented in a fully differential architecture is shown in Fig. 5.4 [91, 92]. It consists of an array of binary weighted capacitors plus one additional capacitor of weight corresponding to the least significant bit (LSB), switches which connect the plates to certain voltages, and a comparator. The capacitive binary search array is composed of 256 digitally controlled unit capacitors, with a unit capacitance of 124fF each, resulting in a total capacitance of 31.7pF. The upper common plate of the switched capacitor array is connected to one terminal of the comparator. In order to cancel the charge injection errors induced by CMOS switches and achieve a high linearity, an identical dummy capacitor array is used to connect the other terminal of the comparator. The system ADC operates at a sample rate of 100KS/s with an external clock of 1MHz and a reference voltage of 0.8V.

A conversion is accomplished by a sequence of three stages. First, at the sample phase, the top plate is connected to ground and the bottom plates to the input voltage $V_{in}$. This results in a stored charge on the top plate which is proportional to the input voltage. Next, at the hold phase, the top grounding switch is then opened, and the bottom plates are connected to ground. Since the charge on the top plate is conserved, its potential $V_{top}$ goes to $-V_{in}$. Third, the redistribution phase begins by testing the
value of the most significant bit (MSB). The largest capacitor \( C_0 \) is switched to the reference voltage \( V_{\text{ref}} \) and the other capacitors are switched to ground. The equivalent circuit is now actually a voltage divider between two equal capacitances. As a result, the voltage \( V_{\text{top}} \) is increased by half of the reference voltage:

\[
V_{\text{top}} = -V_{\text{in}} + \frac{1}{2} V_{\text{ref}} \tag{5.1}
\]

The comparator then performs the first comparison. If \( V_{\text{top}} < 0 \), then \( V_{\text{in}} \) is larger than a half of \( V_{\text{ref}} \) and MSB \( b_7 \) is 1 and the capacitor stays connected to \( V_{\text{ref}} \). Otherwise, it is 0, and the largest capacitor is reconnected to ground. Then, the second largest capacitor \( C_1 \) is switched to \( V_{\text{ref}} \), the comparator determines the next bit, and so on until the LSB is decided after 8 comparisons for the 8-bit ADC.

Stress can affect the accuracy of SAR ADCs. The capacitor array is charged sequentially from MSB to LSB with switches implemented by CMOS transistors. And the device parameters, including mobility and threshold voltage, decide the current in device channel and charging speed. Since the switch transistors are affected by stress, the charging process may not reach the designed voltage level. Thus, errors will be induced in SAR ADCs. For example, assume the system is bent and the stress is induced when the test of MSB starts. Based on the stress result, the mobility increases and the threshold voltage increase in NMOS devices. The voltage of the top plane is:

\[
V_{\text{top}}^{\text{stress}} = -V_{\text{in}} + \left( V_{\text{ref}}/2 + V_{\Delta} \right) \tag{5.2}
\]

where \( V_{\text{top}}^{\text{stress}} \) is the voltage of the top plane with the effect of stress and \( V_{\Delta} \) is the shift in charging process comparing to the stress-free situation. \( V_{\Delta} \) is with a negative sign because of the stress-induced device degradation. Thus, it may affect the result of MSB when \( V_{\text{in}} \) is close to \( 1/2V_{\text{ref}} \) and the difference between is smaller than \( V_{\Delta} \). The result of MSB will become 0 from 1 and an error is induced.

### 5.2.2 Performance Evaluation of SRAMs

Extrinsic stress on transistors of an SRAM in a SiF perturbs the mobility and threshold voltage of MOS devices, with the magnitude of the perturbation being determined by the stress. These device parameter shifts are translated into variations in the performance of the SRAM. Such an evaluation requires a system-level simulation, and we build upon the
infrastructure of CACTI [71], an architecture-level integrated power, area, and timing modeling framework for SRAMs, to model the impact of stress-induced performance variations.

Small SRAM sizes in typical low power SiF applications such as the ECG implies that the array model (Fig. 5.5) is substantially simpler than CACTI, where the memory array has multiple identical banks/subbanks/mats/subarrays that can be concurrently accessed. Here, a single memory array is adequate, and the CACTI performance models have been adapted for our simpler array with a row decoder, bitline MUX decoder, and sense amplifier.

**Timing:** The access time is the time interval between an access request to an SRAM, and the access being completed by returning the requested data. The access time is limited by the delay on request network for address, reply network for data, and the maximum of the delays of row decoder path, bitline MUX decoder path, and sense amplifier path as these circuits operate in parallel. Thus we have [72]:

$$
t_{access} = \max(t_{row-dec-path}, t_{bit-mux-dec-path}, t_{SA-dec-path}) + t_{request-network} + t_{reply-network}
$$

(5.3)

Here, $t_{request-network}$ and $t_{request-network}$ are calculated as the product of unit wire length delay and the wire length and are independent to device parameters. The three path
delays are calculated as

$$t_{\text{row-dec-path}} = t_{\text{predec}} + t_{\text{dec}} + t_{\text{driver}} + t_{\text{BL}} + t_{SA}$$ (5.4)

$$t_{\text{bit-mux-dec-path}} = t_{\text{mux}} + t_{\text{predec}} + t_{\text{dec}} + t_{\text{driver}} + t_{SA}$$ (5.5)

$$t_{SA-\text{dec-path}} = t_{SA} + t_{\text{mux}} + t_{\text{predec}} + t_{\text{dec}} + t_{\text{driver}}$$ (5.6)

where $$t_{\text{predec}}$$, $$t_{\text{dec}}$$, and $$t_{\text{driver}}$$ are delay of wordline/bitline decoding path including predecoders/decoders/drivers; $$t_{\text{BL}}$$, $$t_{SA}$$, and $$t_{\text{mux}}$$ are the delay of bitline, sense amplifier, and the MUX gate. These terms are detailed in [72], and are dependent on the device current $$I_{\text{on}}$$.

**Power:** The stress-induced shifts in device leakage current in turn affect the leakage power of SRAMs which is modeled as [72]

$$P_{\text{leak}} = P_{\text{leak-request-network}} + P_{\text{leak-reply-network}} + P_{\text{leak-predec}}$$

$$+ P_{\text{leak-dec}} + P_{\text{leak-driver}} + P_{\text{leak-SA}} + P_{\text{leak-mem-cell}}$$ (5.7)

Here, $$P_{\text{leak-request-network}}$$ and $$P_{\text{leak-reply-network}}$$ are the leakage power of request network and reply network independent to device parameters. $$P_{\text{leak-predec}}$$, $$P_{\text{leak-dec}}$$, $$P_{\text{leak-driver}}$$, $$P_{\text{leak-SA}}$$, and $$P_{\text{leak-mem-cell}}$$ are, respectively, the leakage power of predecoders, decoders, drivers, sense amplifiers, and memory cells. The terms related to predecoders, decoders, and drivers, are composed of basic logic gates and modeled as a function of leakage current, $$I_{\text{leak}}$$. For example, the leakage power of driver, an inverter, can be calculated as

$$P_{\text{leak-inv}} = 0.5(W_{\text{pmos}} I_{\text{leak-pmos}} + W_{\text{nmos}} I_{\text{leak-nmos}}) V_{DD}$$ (5.8)

where $$I_{\text{leak-pmos}}$$ and $$I_{\text{leak-nmos}}$$ are the subthreshold current per unit width for PMOS and NMOS, $$W_{\text{pmos}}$$ and $$W_{\text{nmos}}$$ donate the PMOS and NMOS widths. The leakage power of a single SRAM cell is:

$$P_{\text{leak-mem-cell}} = V_{DD} I_{\text{mem-cell}}$$ (5.9)

where $$I_{\text{mem-cell}}$$ is the sum of the leakage currents in the standby devices in a memory cell. Thus, the leakage power of SRAM is directly affected by $$I_{\text{leak}}$$, which in turn highly depends on stress-induced shifts in threshold voltage.
5.2.3 The Impact of Stress on SRAM Performance

The components of (5.3) correspond to a set of RC products, where the resistance is influenced by the device threshold voltage and mobility, which in turn are affected by extrinsic stress. For example, in computing gate delays, \( R_{on} \propto \frac{1}{I_{on}} \), and \( I_{on} \) is directly affected by the variations of mobility and threshold voltage. The leakage power depends on the leakage current, \( I_{leak} \), and is affected by the same transistor parameters. For current \( I_x, x \in \{on, leak\} \),

\[
I_x^{stress} = I_x^{nom} + \frac{\partial I_x}{\partial V_t} \Delta V_t^{stress} + \frac{\partial I_x}{\partial \mu} \Delta \mu^{stress}
\]

(5.10)

where \( I_x^{stress} \) is the current after incorporating the effect of extrinsic as well as intrinsic stress, \( I_x^{nom} \) is the nominal current considering only intrinsic stress within the transistor, \( \Delta V_t^{stress} \) and \( \Delta \mu^{stress} \) are the stress-induced variations in threshold voltage and mobility, and \( \partial I_x/\partial V_t \) and \( \partial I_x/\partial \mu \) are the sensitivities corresponding to the variations in threshold voltage mobility, respectively.

We calibrate this linear model of \( I_{on} \) and \( I_{leak} \) for the range of mobility and threshold voltage shifts seen in our experiments. The leakage changes exponentially with the threshold voltage, but for the range of variation due to stress, we find that the above local linear approximation is sufficient. Under a 16nm PTM model, the maximum error of our perturbation model is 4.8% for \( I_{leak} \) and 0.5% for \( I_{on} \).

5.3 Experimental Results in SiFs

5.3.1 Stress vs. Chip Size and Corresponding Device Variations

Fig. 5.6 shows the shifts in mobility and threshold voltage \( (V_t) \) of both NMOS and PMOS devices corresponding to the stress map in Fig. 5.2(b). Both NMOS and PMOS devices suffer mobility degradation, and the peak shifts reach \(-14\%\) and \(-24\%\), respectively. This leads to a reduction in \( I_{on} \), which may cause errors in SAR ADCs and increase the access time of SRAMs. The stress also induces a reduction in the absolute value of \( V_t \) in both NMOS and PMOS, which in turn leads to increased SRAM \( I_{leak} \) and \( P_{leak} \).

A comparison of the peak stress-induced device parameter variations among different chip sizes and packaging schemes is shown in Table 5.1. For the Middle Chip packaging,
the peak shifts remain the same regardless of the chip size since the peak stress values are identical, but for Top Chip, the peak variations become larger as the chip size increase, is consistent with the peak stress trends. Moreover, from Fig. 5.2 (the trends are similar for all chip sizes), it can be concluded that with the same chip size, more significant stress-induced device-level variations are observed in Top Chip packages comparing to Middle Chip packages.

5.3.2 Performance Variations in Flexible SAR ADCs

As shown in Sec 5.2, the stress-induced mobility degradation can result in incomplete charging of the binary capacitor array during the successive approximation steps and thus induce errors into SAR ADCs. Fig. 5.7 shows the charging transients of the MSB in the SAR ADC. Here, $V_{nom top}$ donates the voltage shift of top common plate of the switched capacitor array when the system is free of stress, while $V_{stress top}$ is the value under the influence of stress. In this experiment, the mobility variation of the switch
Table 5.1: Summary of Peak Device Parameter Variations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Middle Chip</th>
<th>Top Chip</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{chip} (\text{mm}^2)$</td>
<td>$2 \times 2, 5 \times 5, 10 \times 10, 20 \times 20$</td>
<td>$2 \times 2, 5 \times 5, 10 \times 10, 20 \times 20$</td>
</tr>
<tr>
<td>$\Delta \mu_N (%)$</td>
<td>5.6%</td>
<td>$-8.3%$</td>
</tr>
<tr>
<td>$\Delta \mu_P (%)$</td>
<td>10.3%</td>
<td>$-14.0%$</td>
</tr>
<tr>
<td>$\Delta V_{t,N} (\text{mV})$</td>
<td>17.9</td>
<td>$-25.3$</td>
</tr>
<tr>
<td>$\Delta V_{t,P} (\text{mV})$</td>
<td>5.0</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Figure 5.7: $V_{out}$ transients in the SAR ADC at the nominal and a shifted mobility value.

The device is set to $-24\%$ and the threshold voltage shift is 11mV, which corresponds to the PMOS variations in Top Chip package since PMOS is used as the switch transistor for charging and NMOS is for discharging. The clock frequency is 1MHz and thus the value for each bit should be determined within 1µs, including the time of charging/discharging the capacitor array, the delay of the comparator and the logic circuit. The comparator uses half the clock cycle for the signal to settle and starts the comparison at 0.5µs, before which the charging/discharging should be completed [93].

According to Eq. (5.1), after the MSB charging process, there should be an increase of $V_{ref}/2 = 0.4\text{V}$ at the top plate. From Fig. 5.7 it can be found the $V_{top}^{\text{nom}}$ can reach the designed voltage level within the time limit. However, at $t = 0.5\mu\text{s}$, $V_{top}^{\text{stress}}$ can only reach 0.389V and there is a gap of 11mV. In other words, for any analog input value of $V_{in}$ between 0.389V and 0.4V, the MSB result should be 0 since $V_{in}$ is smaller than 0.4V, but with the effect of stress, the threshold of 0.389V is sufficient for the MSB to go to 1, causing an MSB error. Similar situations can induce errors in other bits as well.
The performance of an ADC is captured by metrics such as differential nonlinearity (DNL), integral nonlinearity (INL), and missing codes. The DNL measures the difference in code width from the ideal width of one LSB level and can be calculated as [94]:

\[
DNL(i) = \frac{H(i) - H_{\text{ideal}}(i)}{H_{\text{ideal}}(i)}
\]  

(5.11)

where \(H(i)\) is the width of code \(i\), \(H_{\text{ideal}}(i)\) represents ideal width of code \(i\), which equals to 1LSB. The DNL errors accumulate and cause the INL error, also measured in LSBs, and formulated as:

\[
INL(i) = \sum_{j=1}^{i} DNL(j)
\]  

(5.12)

Missing codes are the ones that are missing from the transfer characteristics of an ADC. As a result, the missing code gives a DNL error of \(-1\)LSB. Figure 5.8 shows the DNL errors in unit of LSB with \(V_{\text{in}}\) varying from 0 to \(V_{\text{ref}}\). From the figure, it can be found that the maximum DNL is 2.8LSB which occurs at 0.4V. This is caused by the uncompleted charging for MSB. As a result, the result code \(V_{\text{in}}\) between 0.389V and 0.4V are wrongly set to 128, the width of width becomes longer than 1LSB. This in turn conducts two missing codes (126 and 127) with DNL equalling to \(-1\)LSB. The stress-induced errors in the following bits cause some large DNL errors and missing codes as well. In total, stress can cause 5 missing codes and the worst INL is \(-2.9\)LSB,
which occurs at $V_{ref}/2$. The errors above are induced by stress comparing to an ideal ADC and they can be eliminated by the compensation schemes proposed in Sec 5.3.4.

### 5.3.3 Performance Variations in SRAMs

As discussed in Sec. 5.2.2, the stress-induced shifts in mobility and $V_t$ can affect $I_{on}$ and $I_{leak}$ of devices, and impact memory by increasing access time and the leakage power. A 14kb SRAM is simulated with CACTI to capture the stress effect on SRAMs [4].

**Table 5.2: Stress-Induced Variations in SRAMs**

<table>
<thead>
<tr>
<th>$D_{chip}$ (mm$^2$)</th>
<th>Middle Chip</th>
<th>Top Chip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2x2, 5x5, 10x10, 20x20</td>
<td>2x2, 5x5, 10x10, 20x20</td>
</tr>
<tr>
<td>$\Delta I_{on,N}$ (%)</td>
<td>-2.2%</td>
<td>-3.3%</td>
</tr>
<tr>
<td>$\Delta I_{on,P}$ (%)</td>
<td>-6.2%</td>
<td>-8.5%</td>
</tr>
<tr>
<td>$\Delta I_{leak,N}$ (%)</td>
<td>72.4%</td>
<td>81.9%</td>
</tr>
<tr>
<td>$\Delta I_{leak,P}$ (%)</td>
<td>5.8%</td>
<td>6.2%</td>
</tr>
<tr>
<td>$t_{access}$ (ns)</td>
<td>0.399</td>
<td>0.407</td>
</tr>
<tr>
<td>$\Delta t_{access}$ (%)</td>
<td>2.3%</td>
<td>4.3%</td>
</tr>
<tr>
<td>$P_{leak}$ (mW)</td>
<td>0.898</td>
<td>0.927</td>
</tr>
<tr>
<td>$\Delta P_{leak}$ (%)</td>
<td>40.5%</td>
<td>45.0%</td>
</tr>
</tbody>
</table>

$I_{on}$ and $I_{leak}$: Table 5.2 summarizes the variations in $I_{on}$, $I_{leak}$, access time $t_{access}$, and leakage power $P_{leak}$ under the influence of stress for various chip sizes, $D_{chip}$. Since delay is limited by the worst case value of $I_{on}$, we examine the largest current shift, $\Delta I_{on,(N,P)}$. For leakage power we examine the average shift in CMOS leakage current, $\Delta I_{leak,(N,P)}$, over the entire SRAM region when it is placed at the upper central region of the chip, which corresponds to the worst-case stress for both the Middle Chip and Top Chip packages. It is seen that for Top Chip, $I_{on}$ decreases while $I_{leak}$ increases with chip size for both NMOS and PMOS, but for Middle Chip, both $I_{on}$ and $I_{leak}$ are invariant with chip size since the stress distribution in the SRAM region is identical in various chip sizes.

Access time: Variations in the access time and leakage power can be attributed to shifts in $I_{on}$ and $I_{leak}$, respectively. From the nominal access time, $t_{access}^{nom} = 0.390$ns,
the Middle Chip package induces an increase of 2.3% in $t_{\text{access}}$, regardless of $D_{\text{chip}}$, but Top Chip sees an increase of 4.3%–6.1%, with a larger shift for a larger chip size.

**Leakage power:** The leakage power is significantly increased by stress due to the shift in $V_t$. From the nominal value $P_{\text{nom}}^{\text{leak}} = 0.639\, \text{mW}$, Middle Chip packages see an increase of 40.5%, and Top Chip between 45.0%–95.4%. Larger chips see a larger shift for Top Chip but an equal shift for Middle Chip since the average stress is invariant.

### 5.3.4 Compensating for Stress-Induced Variations

To ensure that the SRAM and ADC in the SiF work correctly under stress-induced deformation, we add margins to account for stress and overdesign these circuits. Specifically, stress-induced errors in the SAR ADC can be eliminated by increasing the charging time for the capacitor array, or increasing the device size to enhance the speed. The compensation for the shifts in access time for SRAM, can be achieved by increasing the device size and supply voltage.

<table>
<thead>
<tr>
<th>$D_{\text{chip}}$(mm$^2$)</th>
<th>Middle Chip</th>
<th>Top Chip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2×2, 5×5, 10×10, 20×20</td>
<td>2×2, 5×5, 10×10, 20×20</td>
</tr>
<tr>
<td><strong>ADC</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta t_{\text{ADC}}$(%)</td>
<td>29.6%</td>
<td>40.3%</td>
</tr>
<tr>
<td>$\Delta W_{\text{ADC}}$(%)</td>
<td>6.3%</td>
<td>8.5%</td>
</tr>
<tr>
<td><strong>SRAM</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta V_{\text{DD,SRAM}}$(%)</td>
<td>1.3%</td>
<td>2.3%</td>
</tr>
<tr>
<td>$\Delta W_{\text{SRAM}}$(%)</td>
<td>2.6%</td>
<td>4.7%</td>
</tr>
</tbody>
</table>

Table 5.3 summarizes the overdesigned parameters for the ADC and SRAM required to keep the system within specifications. Here, $\Delta t_{\text{ADC}}$ and $\Delta W_{\text{ADC}}$ are the shifts in charging time and CMOS device widths for the ADC, while $\Delta V_{\text{DD,SRAM}}$ and $\Delta W_{\text{SRAM}}$ are the supply voltage and device size shifts for SRAM. The largest overdesign parameters are observed in the 20mm × 20mm Top Chip package. To eliminate the error in SAR ADC, the charging time is increased by 69.0%, or device sizes are increased by 17.9%. For the SRAM, the supply voltage can be increased by 3.3%, or the device size increased by 6.7%, to compensate stress-induced delay variation.
5.4 Conclusion

We have presented an analysis of stress-induced performance variations for key non-
digital components – the ADC and SRAM – of flexible SiF applications, under two
types of UTC packaging schemes and various chip sizes. Top Chip packages create
substantially more peak stress than Middle Chip ones; Top Chip shows a larger average
stress with size, while the opposite is true of Middle Chip. Stress-induced variations
result in errors in an SAR ADC and increases in latency and leakage power for on-chip
SRAM: these shifts can be compensated using safety margins.
Chapter 6

Conclusion

This thesis has developed approaches to analyze the effects of mechanical stress on the performance of the circuits/systems, and appropriate compensation methods are proposed to counter these effects. In 3D DRAMs, the fabrication process induces significant thermomechanical stress due to CTE mismatch between different materials, while in flexible displays and flexible SiFs, deformations due to daily operations will also induce significant stress. The stress then affects the electrical parameters such as mobility and threshold voltage of both CMOS devices in 3D DRAMs and UTCs in flexible electronics, as well as OTFTs in flexible displays, which in turn can cause serious performance degradation and/or logic errors in different applications.

For 3D DRAMs, we have proposed a fast semi-analytical approach that can calculate the stress within layout configurations containing TSV stripes and clusters by considering both layout-dependent stress and layout-independent stress. Based on our simulation results, TSV clusters create substantially more significant stress than TSV stripes, resulting in worse performance in latency, power and a larger KOZ area for avoiding excessive leakage current. Finally, we show that using PI as the package substrate material instead of traditional FR-4, enables the reduction of warpage-induced stress and the improvement of the 3D DRAM performance.

For flexible OTFT-base displays, we analyze the extrinsic stress induced by realistic bending operations. In particular, two typical deformations, bend and twist+press, are simulated in this thesis. As a general rule, the twist+press deformation induces more stress than the bend deformation and thus causes larger mobility degradation of the
OTFT devices that are used as switches in each pixel. As a result, both deformations can cause errors in the flexible display because the process of charging/discharging each pixel may not be completed in time. Finally, we propose two compensation schemes to rectify the stress-induced errors: $V_{data}$ compensation adjusts the data values sent to each pixel, and $t_{row}$ compensation tunes the time to charge/discharge each pixel. Both schemes are based on the stress values obtained from the on-chip stress sensors. $V_{data}$ compensation can be used in the applications where the high accuracy and a fixed $t_{row}$ are preferred, while $t_{row}$ compensation is a better choice where fewer ROM accesses are preferred.

For flexible SiFs with UTCs, the stress distributions of two different package schemes, Middle Chip and Top Chip, are first analyzed. The Top Chip scheme shows a larger average stress as the chip size increases, while the opposite is true of Middle Chip. Top Chip packages induce more stress than Middle Chip packages and thus cause larger variations in electrical parameters of CMOS devices within the SiF. Next, we present an analysis of stress-induced performance variations for key non-digital CMOS-based components, the ADC and SRAM, in flexible SiF applications. Stress-induced variations result in errors in an SAR ADC and increases in latency and leakage power for on-chip SRAM. Finally, we propose overdesign approaches for both ADC and SRAM to compensate stress-induced shifts taking the use of safety margins.
References


Appendix A

Components of the Row Cycle Time in Memory Arrays

The components of the row cycle time, $t_{RC}$, are detailed below [72]:

1. The term $t_{row-dec-driv}$ relates to predecoders, decoders, and drivers, composed of basic logic gates. The delay of a gate is:

   $t_d = \tau_0 \sqrt{(\ln V_s)^2 + 2\alpha \beta (1 - V_s)}$, where $\tau_0 = R_{on} C_{load}$ is the intrinsic delay for a load, $C_{load}$, $R_{on}$ is the output resistance (low-gain region), $V_s$ is the switching voltage, $\alpha = \tau_t/\tau_0$, $\tau_t$ is the input transition time, and $\beta = 1/(g_m R_{on})$, where $g_m$ is the transistor transconductance (high-gain region). Rise/fall delays are computed separately.

2. The bitline delay is given by:

   $t_{BL} = \begin{cases} \sqrt{2 t_{step} V_{DD} - V_{tn}} m, & \text{if } t_{step} \leq 0.5 \left( \frac{V_{DD} - V_{tn}}{m} \right) \\ t_{step} + \frac{V_{DD} - V_{tn}}{2m}, & \text{if } t_{step} > 0.5 \left( \frac{V_{DD} - V_{tn}}{m} \right) \end{cases}$

   (A.1)

   where $V_{tn}$ is the threshold voltage of the NMOS in the wordline decoding circuit, $m$ is the slope of wordline signal, and $t_{step} = 2.3 \frac{V_{DD}}{I_{on}} \frac{C_{cell}}{C_{cell} + C_{bl}}$, where $C_{bl}$ is the bitline capacitance, $C_{cell}$ is the DRAM cell capacitance, and $I_{on}$ is the access transistor drive current.

3. The sense amplifier delay is $t_{SA} = \frac{C_{bl}}{g_{mn} + g_{mp}} \ln \left( \frac{V_{DD}}{\Delta V} \right)$ where $\Delta V$ is the differential input voltage of the sense amplifier, $g_{mn}$ ($g_{mp}$) are the transconductance of the NMOS (PMOS) in the sense amplifier.
(4) The time required to write data back into the DRAM cell, $t_{\text{writeback}}$, is the product of the resistance of the access transistor ($V_{DD}/I_{on}$).

(5) The component $t_{\text{WL-reset}}$ is the product of the resistance of the final wordline driver, an inverter, and the wordline capacitance. Similarly, $t_{\text{BL-mux-pre}}$ and $t_{\text{SA-mux-pre}}$ are the delays of the MUX gate, which consists of NAND gates and inverters, modeled as in (1). Delays $t_{\text{writeback}}$, $t_{\text{WL-reset}}$, $t_{\text{BL-mux-pre}}$, and $t_{\text{SA-mux-pre}}$ are modeled as functions of $I_{on}$. 