Stress-Induced Performance Shifts in Flexible System-in-Foils Using Ultra-Thin Chips

Tengtao Li and Sachin S. Sapatnekar
ECE Department, University of Minnesota, Minneapolis, MN 55455
e-mail: {lixx2967, sachin}@umn.edu

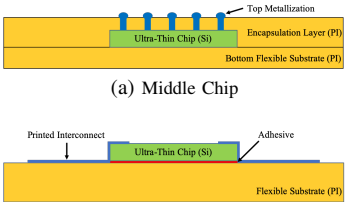
***Abstract*—Silicon-based ultra-thin chips (UTCs) are used to build flexible system-in-foils (SiFs) for bio-sensing and bio-monitoring, and utilize CMOS devices that deliver much higher performance than alternatives such as organic or thin-film transistors. Flexible SiFs experience significant mechanical stress in the field due to the deformation caused during daily use. These impact circuit performance, potentially causing a loss in functionality. This paper first models the stress due to two types of packages schemes for UTCs. Next, the stress is translated to shifts in mobility and threshold voltage of CMOS devices. Finally, the system-level performance variations of two common SiF elements, an A/D converter and an SRAM, are evaluated.**

I. Introduction

Flexible electronics adopt bendable, elastic, and lightweight materials, such as plastic, polymer, or even paper, as a substrate, and are being increasingly deployed in applications such as flexible displays [1], flexible sensor arrays [2], radio frequency identification cards (RFIDs) [3], electronic paper, and system-in-foil (SiF) [4]. Flexible electronics provides 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.

Technologies that are currently being pursued to support the emerging market for flexible electronics include organic electronics, thin-film-transistors (TFTs), and ultra-thin chips (UTCs). Despite their flexibility and low manufacturing cost, the performance of organic electronics and TFTs is limited by their carrier mobilities: typical mobility values range from 0.1–1 cm^2/Vs for amorphous-Si (a-Si) TFTs, with the best organic materials achieving mobility of 1–10 cm^2/Vs [5], as against $>$100 cm^2/Vs in single-crystalline silicon at room temperature. Thus, CMOS UTCs on foil have emerged as an excellent substrate for solutions that require high performance devices and dense interconnects. This technology is compatible with organic or TFT electronics components, e.g., 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 [6]. 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. 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:
Middle Chip [7], [8]: UTCs are placed between polyimide (PI) layers, with laser-drilled vias metallized by sputtering.
Top Chip [9], [10]: UTCs are placed over a PI substrate.
Under deformation, the Middle Chip scheme experiences lower stress than the Top Chip scheme, but the added process step makes this technology more expensive. For high-power applications, Middle Chip has worse heat removal paths, but typical SiF applications are low power, and the low thermal resistance of the ultra-thin substrate makes this a non-issue.



(a) Middle Chip

(b) Top Chip

Fig. 1: Ultra-thin chip packaging: (a) Middle Chip, between two flexible layers (b) Top Chip, atop a flexible substrate [8].

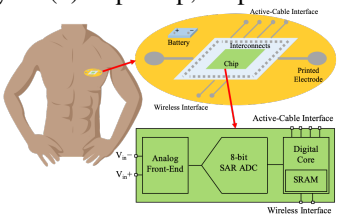


Fig. 2: Block diagram of the flexible ECG system [11].

Two critical components of any SiF system are the analog-to-digital converter (ADC), which converts a real-world analog signal to the digital domain for processing, and an on-chip memory. Consider the application area of wearable electrocardiogram (ECG) monitoring implemented in SiF using CMOS components [11], [12], shown in Fig. 2. 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 stores the digital signal in on-chip memory and transmits the data externally via wires or wirelessly. Similar structures are used in electroencephalogram (EEG), electromyogram (EMG), and temperature/blood pressure monitors.

Stress inside the chip affects device mobility and threshold voltage. For the ADC this induces linearity errors, and for the on-chip SRAM, the latency and leakage power are affected. Prior work has analyzed stress in digital systems can be applied to the digital core [13], 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 paper. This paper 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.

II. Stress Modeling of a UTC-based Flexible SiF

A. Stress Analysis of a Flexible SiF with Ultra-Thin Chip

Stress corresponds to the reactionary internal forces per unit area due to deformation of an object under external forces. The stress field can be represented as the tensor:

$$\sigma = \sigma_{ij} = \begin{pmatrix} \sigma_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \sigma_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \sigma_{33} \end{pmatrix} \quad (1)$$

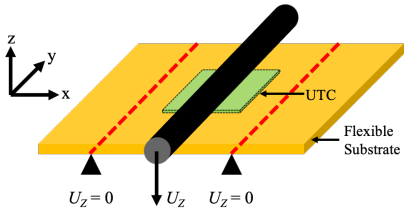


Fig. 3: FEA simulation set up for the UTC package.

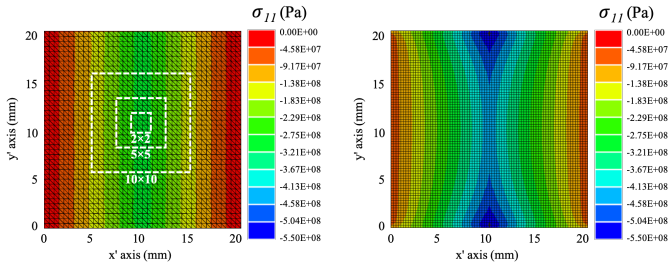
where $i, j \in \{1, 2, 3\}$ refer to the three coordinate axes. The terms σ_{ii} are normal stresses, while τ_{ij} are shear stresses.

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 [14]. 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 to translate these deformations to a bending test, and we analyze bending-induced stress that causes performance shifts at the device and ADC/SRAM block level.

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. The structure used in FEA simulations is as shown in Fig. 3 and the dimensions of the whole structure is 40mm $\times$ 40mm $\times$ 120μm (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 120μm for both packaging schemes. In Middle Chip packaging, the UTC is 20μm thick, and both the bottom substrate and the encapsulation PI layer are set to 50μm [7]. The Top Chip package uses a chip the same thickness and a substrate of 100μm, thus providing a fair stress comparison with the Middle Chip package by maintaining the same package thickness. Various chip sizes ranging from 20mm $\times$ 20mm to 2mm $\times$ 2mm 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 [15].

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 [14].

Fig. 4 shows the FEA simulation results for both Middle Chip and Top Chip packaging methods. The figures show the stress maps of σ_{11} near the top x-y surface of the UTC, where the devices are located. Fig. 4(a) shows the stress distribution of a 20mm $\times$ 20mm UTC: the stress reaches -218MPa 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. 4(b)), where σ_{11} goes up to -550MPa. Other stress



(a) Middle Chip (b) Top Chip

Fig. 4: FEA results of Fig. 3 showing stress distributions.

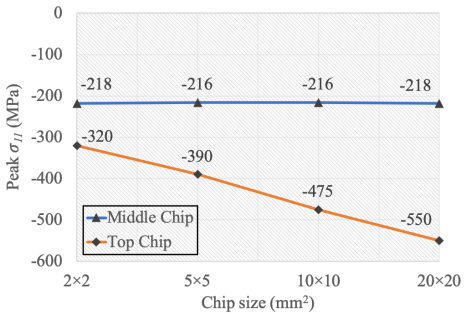


Fig. 5: Peak stress values with different chip sizes. The high stress values are attributed to large deformations for SiF.

tensors, such as σ_{22}, σ_{33} , and τ_{12} are also evaluated and used to analyze CMOS device performance in the SiF.

For a fixed chip thickness of 20μm, 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, 20mm $\times$ 20mm, 10mm $\times$ 10mm, 5mm $\times$ 5mm, and 2mm $\times$ 2mm chips show similar peak stress. The stress profile of a smaller chip is essentially the same as a cutout of that size from the stress profile of a 20mm $\times$ 20mm chip, as illustrated in Fig. 4(a), which shows the stress contours for various chip sizes. The most significant σ_{11} values range between -216MPa and -220MPa. 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). 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 σ_{11} can be reduced from -550MPa to -320MPa by reducing to chip size to 2mm $\times$ 2mm. Note that event with a smaller chip size, the peak stress value is still larger than that in a Middle Chip package.

III. Electrical Variations Due to Stress

The wafer orientation, defined by Miller indices (typically, [001]), is normal to the plane of the wafer, but transistors are oriented along [110]. We use a rotated coordinate system with the x'-axis along [110] and the y'-axis along [$\bar{1}$10]. 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

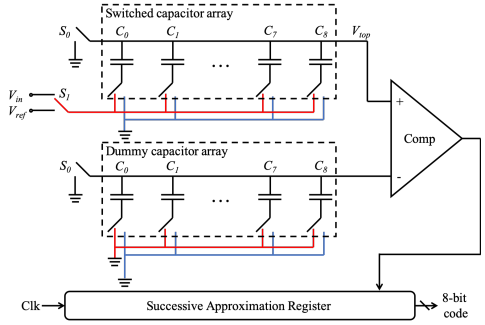


Fig. 6: Architecture of the 8-bit SAR ADC.

with the stress tensor. The relative change of mobility in the rotated coordinate system (x', y') is given by [16]:

$$\frac{\Delta\mu'}{\mu'} = [\pi'_{11}\sigma_{x'x'} + \pi'_{12}\sigma_{y'y'} + \pi_{12}\sigma_{zz}]\cos^2\phi' + [\pi'_{11}\sigma_{y'y'} + \pi'_{12}\sigma_{x'x'} + \pi_{12}\sigma_{zz}]\sin^2\phi' + [\pi'_{44}\tau_{x'y'}]\sin 2\phi' \quad (2)$$

where $\sigma_{x'x'}$, $\sigma_{y'y'}$, σ_{zz} are normal stresses in the rotated coordinate system, $\tau_{x'y'}$ is the shear stress, π'_{11}, π'_{12} and π'_{44} are the piezoresistivity coefficients [17] in the primed coordinate system, π_{12} is the piezoresistivity coefficient in the original coordinate system, and ϕ' is the angle between the transistor channel and x'-axis, typically 0 or π/2.

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. Mechanical strain in the transistor channel, given by the strain tensor ϵ, could induce shifts and splits in the conduction band and balance band and therefore the threshold 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 [18]:

$$q\Delta V_{tn} = m\Delta E_C - (m-1)\Delta E_V \quad (3)$$

$$q\Delta V_{tp} = m\Delta E_V - (m-1)\Delta E_C \quad (4)$$

where ΔV_{tn} and ΔV_{tp} are the changes in NMOS and PMOS threshold voltages, respectively, q is the electron charge, and m is the body-effect coefficient and takes values 1.1–1.4. The term ΔE_C is the minimum conduction band potential change over carrier band number i, $\Delta E_C^{(i)}$, while ΔE_V denotes the maximum of the changes in valence band potentials between heavy-hole (hh) and light-hole (lh), ΔE_V^{hh} and ΔE_V^{lh}.

IV. Performance of ADCs and SRAMs in SiFs

The performance variation of SiF systems, such as the flexible ECG monitoring system in Fig. 2, are strongly dependent on variations in the ADC and SRAM. According to Sec. III, 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 ADC and the on-chip SRAM to study the stress-induced performance variations.

A. Performance Evaluation of SAR ADCs

A standard 8-bit charge-redistribution SAR ADC implemented in a fully differential architecture is shown in Fig. 6 [19], [20]. 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 in 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 V_{in}. 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_{ref} and the other capacitors are switched to ground. The equivalent circuit is now a voltage divider between two equal capacitances, and V_{top} is:

$$V_{top} = -V_{in} + \frac{1}{2}V_{ref} \quad (5)$$

The comparator then performs the first comparison. If $V_{top} <$ 0V, then V_{in} is larger than a half of V_{ref} and MSB b_7 is 1 and the capacitor stays connected to V_{ref}. Otherwise, it is 0, and the largest capacitor is reconnected to ground. Then, the second largest capacitor C_1 is switched to V_{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:

$$V_{top}^{stress} = -V_{in} + (V_{ref}/2 + V_\Delta) \quad (6)$$

where V_{top}^{stress} is the voltage of the top plane with the effect of stress and V_Δ is the shift in charging process comparing to the stress-free situation. V_Δ is with a negative sign because of the stress-induced device degradation. Thus, it may affect the result of MSB when V_{in} is close to $1/2V_{ref}$ and the difference between is smaller than V_Δ. The result of MSB will become 0 from 1 and an error is induced. Similarly, stress can induce errors into the following bits as well during the distribution phase and thus will cause errors in an entire SAR ADC. Our analysis in Section V-B will determine the impact of such errors on standard ADC metrics such as INL and DNL.

B. 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 [21], an architecture-level integrated power, area, and timing modeling framework for SRAMs, to model the impact of stress-induced performance variations. As part of this work, we calibrate CACTI models for 16nm technology.

Small SRAM sizes in typical low power SiF applications such as the ECG implies that the memory array model 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 a 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 [22]:

$$t_{access} = \max(t_{\text{row-dec-path}}, t_{\text{bit-mux-dec-path}}, t_{\text{SA-dec-path}}) + t_{\text{request-network}} + t_{\text{reply-nework}} \quad (7)$$

Here, $t_{\text{request-network}}$ and $t_{\text{reply-network}}$ 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:

$$t_{\text{row-dec-path}} = t_{\text{predec}} + t_{\text{dec}} + t_{\text{driver}} + t_{\text{BL}} + t_{\text{SA}} \quad (8)$$

$$t_{\text{bit-mux-dec-path}} = t_{\text{mux}} + t_{\text{predec}} + t_{\text{dec}} + t_{\text{driver}} + t_{\text{SA}} \quad (9)$$

$$t_{\text{SA-dec-path}} = t_{\text{SA}} + t_{\text{mux}} + t_{\text{predec}} + t_{\text{dec}} + t_{\text{driver}} \quad (10)$$

where t_{predec}, t_{dec}, and t_{driver} are delay of wordline/bitline decoding path including predecoders/decoders/drivers; t_{BL}, t_{SA} ,and t_{mux} are the delay of bitline, sense amplifier, and the MUX gate. These terms are detailed in [22], and depend on I_{on}.

Power: The stress-induced shifts in device leakage current in turn affect the SRAM leakage power, modeled as [22]

$$P_{leak} = P_{\text{leak-request-network}} + P_{\text{leak-reply-nework}} + P_{\text{leak-predec}} + P_{\text{leak-dec}} + P_{\text{leak-driver}} + P_{\text{leak-SA}} + P_{\text{leak-mem-cell}}$$

Here, $P_{\text{leak-request-network}}$ and $P_{\text{leak-reply-nework}}$ 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_{leak}. For example, the leakage power of driver, an inverter, can be calculated as

$$P_{\text{leak-inv}} = 0.5(W_{\text{pmos}} I_{\text{leak-pmos}} + W_{\text{nmos}} I_{\text{leak-nmos}}) V_{\text{DD}} \quad (11)$$

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

$$P_{\text{leak-mem-cell}} = V_{\text{DD}} I_{\text{mem-cell}} \quad (12)$$

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_{leak}, which in turn highly depends on stress-induced shifts in threshold voltage.

C. The Impact of Stress on SRAM Performance

The components of (7) 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 leakage power depends on the leakage current, I_{leak}, and is affected by the same transistor parameters. Although I_{leak} and I_{on} are very nonlinearly dependent on V_t and μ, over the relatively small range of perturbations to these parameters associated with stress, a simpler sensitivity-based model based on a Taylor series expansion may be used. For current $I_x, x \in \{on, leak\}$,

$$I_x^{stress} = I_x^{nom} + \frac{\partial I_x}{\partial V_t}\Delta V_t^{stress} + \frac{\partial I_x}{\partial \mu}\Delta \mu^{stress} \quad (13)$$

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, ΔV_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. The perturbations in I_{on}^{stress} can be reasonably captured by linear model where the sensitivities are constant. However, since I_{leak}^{stress} is a more nonlinear function, we find that an accurate fit requires a piecewise linear model for the perturbation.

We calibrate this model of I_{on} and I_{leak} using SPICE under the PTM model 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} (using a two-segment PWL model) and 0.5% for I_{on}. Based on these SPICE-calibrated models, we then establish the technology file in CACTI for the 16nm technology node, and capture the stress-induced system-level variations by modifying I_{on} and I_{leak} according to the stress values obtained from our analysis.

V. Evaluation of the Analysis Technique

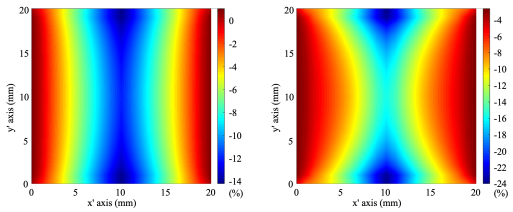
A. Stress Vs. Chip Size and Corresponding Device Variations

Fig. 7 shows the shifts in mobility and threshold voltage (V_t) of both NMOS and PMOS devices corresponding to the stress map in Fig. 4(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}.

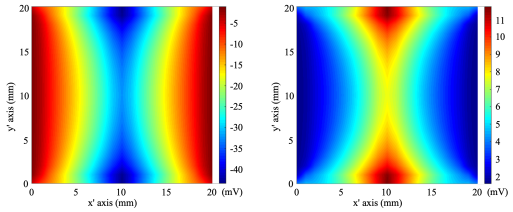
TABLE I: Summary of Peak Device Parameter Variations

	Middle Chip	Top Chip			
D_{chip}(mm^2)	2×2, 5×5, 10×10, 20×20	2×2	5×5	10×10	20×20
$\Delta\mu_N$(%)	5.6%	−8.3%	−10.1%	−12.3%	−14.2%
$\Delta\mu_P$(%)	10.3%	−14.0%	−17.1%	−20.8%	−24.0%
ΔV_{t_N}(mV)	17.9	−25.3	−30.8	−37.5	−43.4
ΔV_{t_P}(mV)	5.0	6.8	8.3	10.1	11.7

The peak stress-induced variations for various chip sizes and packaging schemes are shown in Table I. For the Middle



(a) NMOS mobility (b) PMOS mobility



(c) NMOS V_t (d) PMOS V_t

Fig. 7: Variations in mobility and V_t in Top Chip packaging.

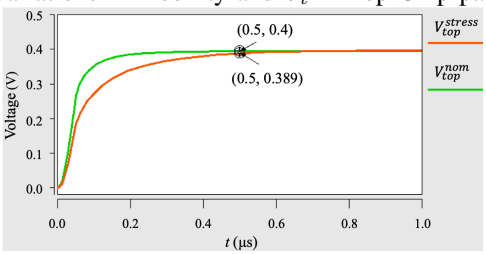


Fig. 8: V_{top} transients at the nominal and a shifted μ value.

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. 4 (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. Since the deformation can appear in any part of the chip, it is possible for the worst-case stress to appear anywhere in the chip. To design the circuit to ensure that it works correctly under all deformations, we consider the worst-case stress that induces the largest performance perturbation due to bending stress in the SiF structure.

B. Performance Variations in Flexible SAR ADCs

As shown in Sec IV, 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. 8 shows the charging transients of the MSB in the SAR ADC with SPICE simulations. Here, V_{top}^{nom} donates the voltage shift of top common plate of the switched capacitor array when the system is free of stress, while V_{top}^{stress} is the value under the influence of stress. In this experiment, the mobility variation of the switch device is set to -24% and the threshold voltage shift is 11mV, which correspond to the PMOS variations in Top Chip package as shown in Fig. 7 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

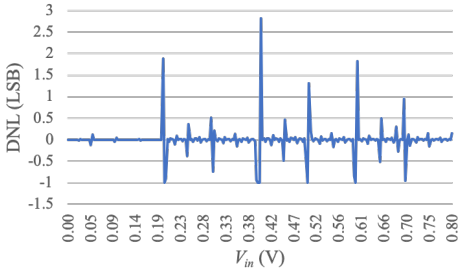


Fig. 9: DNL of the SAR ADC under bending stress.

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 [20].

According to Eq. (5), after the MSB charging process, there should be an increase of $V_{ref}/2 = 0.4$V at the top plate. From Fig. 8 it can be found the V_{top}^{nom} can reach the designed voltage level within the time limit. However, at $t = 0.5\mu$s, V_{top}^{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 also induce errors in other bits. The cumulative performance variation induced by stress, over all bits of the 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 [23]:

$$DNL(i) = \frac{H(i) - H_{ideal}(i)}{H_{ideal}(i)} \tag{14}$$

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

$$INL(i) = \sum_{j=1}^{i} DNL(j) \tag{15}$$

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 -1LSB. Figure 9 shows the DNL errors in unit of LSB with V_{in} varying from 0 to V_{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_{in} between 0.389V and 0.4V are wrongly set to 128, the width of code 128 becomes longer than 1LSB. This in turn conducts two missing codes (126 and 127) with DNL equalling to -1LSB. 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 -2.9LSB, which occurs at $V_{ref}/2$. The errors above are induced by stress comparing to an ideal ADC and can be eliminated by the compensation schemes in Sec V-D.

C. Performance Variations in SRAMs

As discussed in Sec. IV-B, stress-induced shifts in μ and V_t can affect I_{on} and I_{leak}, 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 [11].

I_{on} and I_{leak}: Table II 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

TABLE II: Stress-Induced Variations in SRAMs (t_{access}^{nom} = 0.390ns, P_{leak}^{nom} = 0.639mW)

D_{chip}(mm^2)	Middle Chip	Top Chip			
	2×2, 5×5, 10×10, 20×20	2×2	5×5	10×10	20×20
ΔI_{on_N}(%)	−2.2%	−3.3%	−4.0%	−4.9%	−5.6%
ΔI_{on_P}(%)	−6.2%	−8.5%	−10.3%	−12.6%	−14.5%
ΔI_{leak_N}(%)	72.4%	81.9%	116.0%	145.4%	171.2%
ΔI_{leak_P}(%)	5.8%	6.2%	9.0%	11.3%	13.3%
t_{access}(ns)	0.399	0.407	0.408	0.411	0.414
Δt_{access}(%)	2.3%	4.3%	4.5%	5.2%	6.1%
P_{leak}(mW)	0.898	0.927	1.053	1.157	1.250
ΔP_{leak}(%)	40.5%	45.0%	64.7%	81.0%	95.4%

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_{access}, regardless of D_{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 $P_{leak}^{nom} = 0.639$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.

D. 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 III: Overdesign Parameters for SAR ADC and SRAM

	D_{chip}(mm^2)	Middle Chip	Top Chip			
		2×2, 5×5, 10×10, 20×20	2×2	5×5	10×10	20×20
ADC	Δt_{ADC}(%)	29.6%	40.3%	49.2%	59.8%	69.0%
	ΔW_{ADC}(%)	6.3%	8.5%	11.0%	14.2%	17.9%
SRAM	ΔV_{DD_SRAM}(%)	1.3%	2.3%	2.5%	2.8%	3.3%
	ΔW_{SRAM}(%)	2.6%	4.7%	5.0%	5.8%	6.7%

Table III summarizes the overdesigned parameters for the ADC and SRAM required to keep the system within specifications. Here, Δt_{ADC} and ΔW_{ADC} are the shifts in charging time and CMOS device widths for the ADC, while ΔV_{DD_SRAM} and ΔW_{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 effects.

VI. 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 a SAR ADC and increases in latency and leakage power for on-chip SRAM: these shifts can be compensated using safety margins.

References

[1] J.-h. Lee, *et al.*, "World's largest (15-inch) XGA AMLCD panel using IGZO oxide TFT," in *Proc. SID Symposium*, pp. 625–628, 2008.

[2] C. Wang, *et al.*, "User-interactive electronic skin for instantaneous pressure visualization," *Nature Materials*, vol. 12, no. 10, pp. 899–904, 2013.

[3] E. Cantatore, *Applications of organic and printed electronics: a technology-enabled revolution*. New York, NY: Springer, 2013.

[4] C. Harendt, *et al.*, "Hybrid systems in foil (HySiF) exploiting ultra-thin flexible chips," *Solid-State Electronics*, vol. 113, pp. 101–108, 2015.

[5] H. Klauk, "Organic thin-film transistors," *Chemical Society Reviews*, vol. 39, no. 7, pp. 2643–2666, 2010.

[6] J. Burghartz, *Ultra-thin chip technology and applications*. New York, NY: Springer, 2010.

[7] W. Christiaens, *et al.*, "UTCP: a novel polyimide-based ultra-thin chip packaging technology," *IEEE Transactions on Components and Packaging Technologies*, vol. 33, no. 4, pp. 754–760, 2010.

[8] S. Gupta, *et al.*, "Ultra-thin chips for high-performance flexible electronics," *NPJ Flexible Electronics*, vol. 2, no. 8, pp. 1–17, 2018.

[9] R. S. Dahiya and S. Gennaro, "Bendable ultra-thin chips on flexible foils," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 4030–4037, 2013.

[10] G. Ezhilarasu, *et al.*, "A flexible, heterogeneously integrated wireless powered system for bio-implantable applications using fan-out wafer-level packaging," in *Proc. IEEE International Electron Devices Meeting*, pp. 29.7.1–29.7.4, 2018.

[11] G. Yang, *et al.*, "Bio-patch design and implementation based on a low-power system-on-chip and paper-based inkjet printing technology," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1043–1050, 2012.

[12] S. Izumi, *et al.*, "A wearable healthcare system with a 13.7 μA noise tolerant ECG processor," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, no. 5, pp. 733–742, 2015.

[13] A. Vilouras, *et al.*, "Modeling of CMOS devices and circuits on flexible ultrathin chips," *IEEE Transactions on Electron Devices*, vol. 64, no. 5, pp. 2038–2046, 2017.

[14] H. Heidari, *et al.*, "Bending induced electrical response variations in ultra-thin flexible chips and device modeling," *Applied Physics Reviews*, vol. 4, no. 031101, pp. 1–20, 2017.

[15] S. Logothetidis, *Handbook of flexible organic electronics: materials, manufacturing and applications*. Elsevier, Cambridge, UK, 2015.

[16] R. C. Jaeger, *et al.*, "CMOS stress sensors on (100) silicon," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 1, pp. 85–95, 2000.

[17] S. K. Marella and S. S. Sapatnekar, "A holistic analysis of circuit performance variations in 3D-ICs with thermal and TSV-induced stress considerations," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 23, no. 7, pp. 1308–1321, 2015.

[18] W. Zhang and J. G. Fossum, "On the threshold voltage of strained-Si−$Si_{1-x}Ge_x$ MOSFETs," *IEEE Transctions on Electron Devices*, vol. 52, no. 2, pp. 263–268, 2005.

[19] G. Yang, *et al.*, "A multi-parameter bio-electric ASIC sensor with integrated 2-wire data transmission protocol for wearable healthcare system," in *Proc. ACM the Conference on Design, Automation and Test in Europe*, pp. 443–448, 2012.

[20] N. Verma and A. P. Chandrakasan, "An ultra low energy 12-bit rate-resolution scalable SAR ADC for wireless sensor nodes," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 6, pp. 1196–1205, 2007.

[21] "CACTI tools." http://www.hpl.hp.com/research/cacti/.

[22] S. Thoziyoor, *et al.*, "CACTI 5.1," *HPL-2008-20, HP Labs, Palo Alto, CA*, 2008.

[23] M. Flores, *et al.*, "INL and DNL estimation based on noise for ADC test," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 5, pp. 1391–1395, 2004.