Stress-Aware Performance Evaluation of 3D-Stacked Wide I/O DRAMs

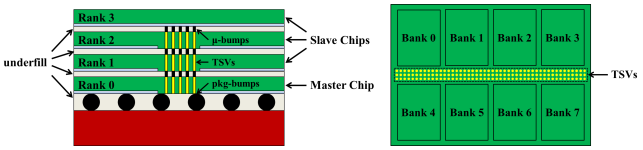
Tengtao Li and Sachin S. Sapatnekar
ECE Department, University of Minnesota, Minneapolis, MN 55455
e-mail: {lixx2967, sachin}@umn.edu

***Abstract*—3D-stacked wide I/O DRAM can significantly increase cell density and bandwidth while also lowering power consumption. However, 3D structures experience significant thermomechanical stress, which impacts circuit performance. This paper develops a procedure that performs a full performance analysis of 3D DRAMs, including latency, leakage power, refresh power, and area, while incorporating the effects of both layout-aware stress and layout-independent stress. The approach first proposes an analytic 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.**

I. Introduction

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 through-silicon-vias (TSVs) that can transmit data, address, and power signals [1]–[3]. 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. 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, wide I/O replaces these off-chip lines with an on-chip wide I/O bus within the 3D structure [4]. Therefore, wide I/O 3D DRAMs are excellent candidates for applications that show a demand for high bandwidth and low power memory, including mobile devices.

The structure of a 3D DRAM stack is illustrated in Fig. 1, in which each chip in the stack constitutes a rank, as in [1] (in some structures, multiple ranks may be placed in each layer [4]). 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 [1]. A typical configuration stacks all chips on a flip-chip package using back-to-face (B2F) bonding [5], 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 μ-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.



(a) 3D DRAM structure

(b) Layout of Rank 0

Fig. 1: 3D DRAM structure [1] and the layout of the rank 0 layer.

An important consideration the design of wide I/O 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 [6]. 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 [7]. 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 [8],
(b) μ-bump and package bump induced stress [9], 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 [10]–[12], 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.

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 [8] 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 [9], 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 wide I/O 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.

The contributions of this paper are in developing a 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.

II. Performance Evaluation of 3D DRAM

Modern transistors use strained silicon, implemented by introducing *intrinsic* stress induced by materials that introduce lattice

[0] This work was supported in part by the NSF under award CCF-1421606.

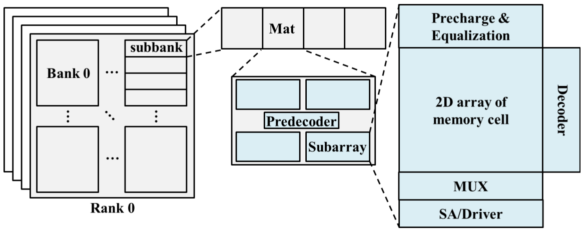


Fig. 2: Organization of a 3D DRAM array.

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 [13], 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 [14], 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.

A. Memory Organization and Peformance

The 3D DRAM array model (Fig. 2) 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 [15]:

$$\begin{aligned} t_{RC} = t_{\text{row-dec-drv}} + t_{\text{BL}} + t_{\text{SA}} + t_{\text{writeback}} + t_{\text{WL-reset}} \\ + \max(t_{\text{BL-pre}}, t_{\text{BL-mux-pre}}, t_{\text{SA-mux-pre}}) \end{aligned} \tag{1}$$

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 $t_{\text{WL-reset}}$, $t_{\text{BL-pre}}$, $t_{\text{BL-mux-pre}}$, and $t_{\text{SA-mux-pre}}$ 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_{refresh}, is bounded by the retention time, $T_{\text{retention}}$, of a DRAM array, which is given by:

$$T_{\text{retention}} = \frac{C_{cell} \Delta V_{cell}}{I_{leak}} \tag{2}$$

where ΔV_{cell} is the worst-case capacitor voltage that leads to a read failure, and I_{leak} is the worst-case leakage in a DRAM cell.

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

$$P_{ref} = \frac{E_{\text{refresh}}}{T_{\text{refresh}}} \tag{3}$$

where $T_{\text{refresh}} = T_{\text{retention}}$ is the refresh period and E_{refresh} is the energy of a refresh operation. The contributors to E_{refresh} include the refresh predecoders, refresh decoder drivers, and the refresh bitline, and correspond to charging/discharging capacitances, as detailed in [15]. These quantities are independent of stress, but the refresh period is strongly affected by stress and influences P_{ref}.

B. The Impact of Stress on 3D DRAM Performance

From the Appendix, it can be seen that the components of (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:

$$I_x^{stress} = I_x^{nom} + \frac{\partial I_x}{\partial V_t} \Delta V_t^{stress} + \frac{\partial I_x}{\partial \mu} \Delta \mu^{stress} \tag{4}$$

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, 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.

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}.

III. Stress Modeling of a Wide I/O 3D DRAM stack

A. Basic Principles

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

$$\sigma = \sigma_{ij} = \begin{pmatrix} \sigma_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \sigma_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \sigma_{33} \end{pmatrix} \tag{5}$$

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

The 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 I can be classified into:

- *Layout-dependent stress*, σ_{LD}, is induced by the stress sources related to layout, specifically stresses caused by the locations of the TSVs and μ-bumps relative to various blocks in the layout.
- *Layout-independent stress*, σ_{LI}, does not vary with the layout: here, this corresponds to warpage caused by the CTE mismatch between layers. Intrinsic stress is also layout-independent.

By linear superposition, we can perform the tensor addition:

$$\sigma_{total} = \sigma_{LD} + \sigma_{LI} \tag{6}$$

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

B. Stress Analysis of a 3D DRAM Stack

Consider a 8Gb 3D DRAM with four stacked memory chips, similar to [1], as shown in Fig. 1. Each layer is thinned from the wafer thickness of ∼ 300μm thickness down to 50μm, 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 380μm, respectively.

TSVs are used to transmit data and power signals through the stack, and underfill layers and μ-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, μ-bumps, and package bumps are listed in Table I, where D, H, and P are the diameter, height, and pitch, respectively.

TABLE I: Dimensions of the TSVs, μ-bumps, and package bumps.

	D	H	P
TSV	20μm	50μm	25μm
μ-bump	20μm	10μm	25μm
Package bump	100μm	50μm	300μm

TABLE II: Material Parameters

Material	CTE (ppm/K)	Young's Modulus (GPa)	Poisson Ratio
Si	2.3	188	0.27
Cu	17	110	0.35
SiO$_2$	0.5	71	0.17
substrate	17.6	19.7	0.13
pkg-bump	22	44.4	0.35
μ-bump	20	26.2	0.35
HC_TSV	9.69	149	0.31
HC_μ-bump	10.3	48.5	0.26
HC_pkg-bump	2.38	68.7	0.19

The entire 3D DRAM structure undergoes a thermal load of ΔT=-250°C as it is annealed from 275°C to 25°C. The materials in the stack shrink differentially due to their differing CTEs, inducing thermal stress. All material parameters are summarized in Table II.

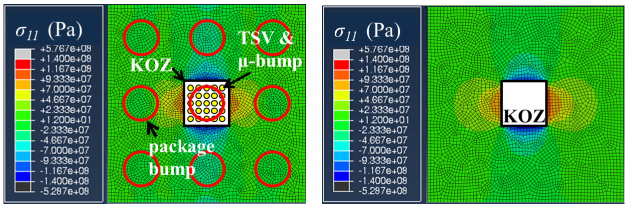
In principle, it is possible to perform FEA to compute the resulting stress profile in the 3D structure. FEA proceeds by first meshing the structures into small polyhedral subdomains called elements. and then constructs a set of equations relating the stress at neighboring vertices of the polyhedra to 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, 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, α_{TSV} and α_{Si}, are:

$$\alpha_{TSV} = N \cdot \left(\frac{\pi R^2 H}{W \cdot L \cdot H} \right) \quad ; \quad \alpha_{Si} = 1 - \alpha_{TSV} \tag{7}$$



(a) A 5 × 5 TSV cluster (b) Its volume fraction approximation

Fig. 3: Stress maps showing the accuracy of the volume fraction approximation for TSVs, μ-bumps, and package bumps.

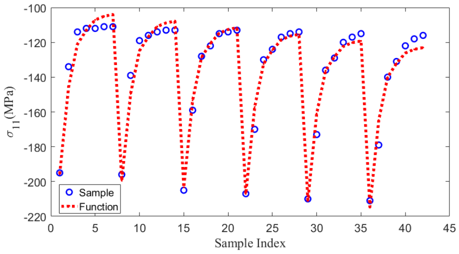


Fig. 4: The accuracy of our semianalytical model for a TSV cluster: the solid curve fits the blue sample points. The horizontal axis is the set of evaluated points (7 values of r evaluated at 6 values of w).

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

$$CTE_{HC\_TSV} = \alpha_{TSV} \cdot CTE_{TSV} + \alpha_{Si} \cdot CTE_{Si} \tag{8}$$

A similar method is also applied to μ-bumps and package bumps embedded in underfill to replace these nonhomogeneous regions by the equivalent homogeneous cuboids with an appropriate CTE. Fig. 3 shows the results of FEA simulation for a cluster of 5×5 TSVs, as against the results when the TSVs, μ-bumps, and package bumps are all replaced by a volume fraction approximation. The error of this approach is 0.02% on average, with a variance of 8.36E-04.

(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 (6), σ_{LI} is independent of layout decisions, and therefore, we first generate a methodology to separate these stresses from the layout-dependent stresses, σ_{LD}. The stresses σ_{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.

To compute the layout-independent stress, we simulate the 3D stack with no TSVs or μ-bumps and apply the thermal load of $\Delta T = -250$°C, and find the stress, σ_{LI} 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.

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 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, σ_{LI}, and build a semianalytical model of the form $\sigma_{LD} = k_1 + k_2 r + k_3/r + k_4 w$.

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 σ_{11} stress along the x'-axis and compressive σ_{11} stress along the y'-axis. For TSVs and μ-bumps distributed in row and column stripes, only compressive stress occurs in the area close to the long edges. For a TSV cluster, Fig. 4 shows that the model provides excellent accuracy (0.36% average error with 1.30E-03 variance). 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:

$$\sigma_{total} = \sigma_{LI} + \textstyle\sum_{i=1}^{N} \sigma_{TSV\_stripe\_i} + \sum_{i=1}^{M} \sigma_{TSV\_cluster\_i} \qquad (9)$$

IV. Electrical Variations Due to Stress

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 $[\bar{1}10]$. 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 with the stress tensor [11]. The relative change of mobility in the rotated coordinate system (x', y') is given by [11]:

$$\frac{\Delta\mu'}{\mu'} = [\pi'_{11}\sigma_{x'x'} + \pi'_{12}\sigma_{y'y'} + \pi_{12}\sigma_{zz}]\cos^2\phi' + \qquad (10)$$
$$[\pi'_{11}\sigma_{y'y'} + \pi'_{12}\sigma_{x'x'} + \pi_{12}\sigma_{zz}]\sin^2\phi' + [\pi'_{44}\tau_{x'y'}]\sin 2\phi'$$

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 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. The piezoresistivity coefficients are taken from [8].

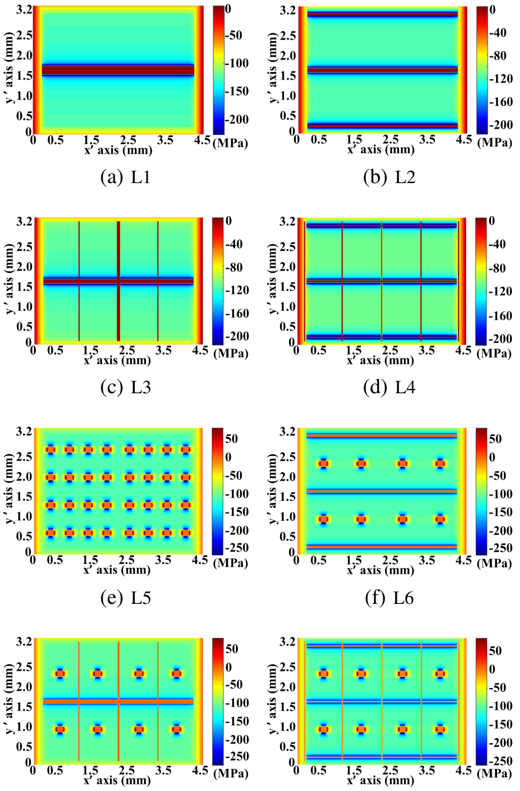
Stress can also cause a shift in the transistor threshold voltage due to three effects: change in the silicon electron affinity, bandgap, and valence band density-of-states [16]. Mechanical strain in the transistor channel, given by the strain tensor ϵ, could induce shifts and splits in the conduction 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 [12]:

$$q\Delta V_{tn} = m\Delta E_C - (m-1)\Delta E_V \qquad (11)$$
$$q\Delta V_{tp} = m\Delta E_V - (m-1)\Delta E_C \qquad (12)$$

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), which can be noted by ΔE_V^{hh} and ΔE_V^{lh}. These are given by:

$$\begin{aligned}
&\Delta E_C^{(i)}(\epsilon) = \Xi_d(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) + \Xi_u\epsilon_{ii}, i \in \{x, y, z\}\\
&\Delta E_V^{(hh,lh)}(\epsilon) = a(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz})\\
&\pm\sqrt{\frac{b^2}{4}(\epsilon_{xx} + \epsilon_{yy} - 2\epsilon_{zz})^2 + \frac{3b^2}{4}(\epsilon_{xx} - \epsilon_{yy})^2 + d^2\epsilon_{xy}^2}
\end{aligned} \qquad (13)$$

where Ξ_d and a are the hydrostatic deformation potential constants, which can induce shifts in the conduction band and valence band, respectively, while Ξ_u, b, and d are the shear deformation potential constants that affect the conduction and valence bands.



(a) L1 (b) L2

(c) L3 (d) L4

(e) L5 (f) L6

(g) L7 (h) L8

Fig. 5: Contours of σ_{11} in the eight layouts.

V. Experimental Results

We investigate a set of TSV layouts for an 8Gb 4-layer 3D DRAM array. The TSVs are arranged in some combination of *rows*, where each row contains 150 TSVs; *columns*, with 120 TSVs per column; and *clusters*. Eight TSV layouts are described in Table III. 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. While L8 uses a 5×5 arrangement of TSVs in each cluster, L5–L7 use a 6×6 arrangement. 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 III: Summary of TSV Distributions in L1–L8

Layout	TSV rows	TSV columns	TSV clusters	# TSVs
L1	8 (0,8,0)	-	-	1200
L2	8 (2,4,2)	-	-	1200
L3	4 (0,4,0)	5 (0,1,3,1,0)	-	1200
L4	4 (1,2,1)	5 (1,1,1,1,1)	-	1200
L5	-	-	32 (6 × 6)	1152
L6	6 (2,2,2)	-	8 (6 × 6)	1188
L7	3 (0,3,0)	4 (0,1,2,1,0)	8 (6 × 6)	1218
L8	3 (1,1,1)	5 (1,1,1,1,1)	8 (5 × 5)	1250

The spatial distribution of TSVs in L1–L8 is apparent in Fig. 5, which shows the contours of σ_{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. 6 shows the data corresponding to L8 in Fig. 5(h). NMOS transistors near TSV stripes and clusters suffer a mobility degradation up to -10%, while PMOS transistors lying over and under lateral TSV stripes and clusters suffer a mobility degradation up to -23%. For PMOS transistors at the left and right edge of TSV columns and clusters, the mobility can increase by up to 25%.

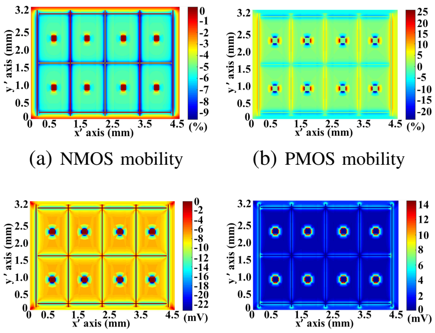
For both NMOS and PMOS devices, the stress-induced shifts are negative for ΔE_C and positive for ΔE_V. As a result, the bandgap is smaller so that the absolute values of threshold voltages for both NMOS and PMOS transistors decreases. The maximum variation occurs near TSV stripes and clusters, with threshold voltage variations for NMOS (PMOS) transistors of up to -23mV (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 paramaters. We now analyze the impact of stress on system timing for L1–L8. We focus on t_{RC}, defined in (1), but similar analyses can be performed for other timing metrics. The t_{RC} variation contours in L1–L8 are shown in Fig. 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.

TABLE IV: Row Cycle Time (t_{RC}), Leakage Power (P_{leak}), and Refresh Power (P_{ref}) for L1–L8
(D_0 = 33.62ns, P_{leak}^{nom} = 50.66mW, P_{ref}^{nom} = 18.90mW)

	Row Cycle Time t_{RC}				Leakage P_{leak}		Refresh P_{ref}	
	ΔD^+ (ns)	ΔD^+ (%)	ΔD^- (ns)	ΔD^- (%)	ΔP_{leak} (mW)	ΔP_{leak} (%)	ΔP_{ref} (mW)	ΔP_{ref} (%)
L1	-0.67	-2.0%	0.69	2.1%	12.22	24.1%	8.16	43.2%
L2	-0.67	-2.0%	0.84	2.5%	11.69	23.1%	7.89	41.7%
L3	-0.99	-2.9%	0.64	1.9%	11.00	21.7%	7.61	40.3%
L4	-1.36	-4.0%	0.86	2.6%	11.50	22.7%	7.85	41.5%
L5	-3.65	-10.9%	2.11	6.3%	12.44	24.6%	15.02	79.4%
L6	-3.64	-10.8%	2.10	6.2%	11.73	23.1%	14.96	79.2%
L7	-3.61	-10.7%	2.14	6.4%	11.37	22.4%	15.27	80.8%
L8	-3.64	-10.8%	2.05	6.1%	11.96	23.6%	14.57	77.1%

The latency performance of 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 IV. All percentage changes are with respect to D_0, the nominal t_{RC} without the effect of stress for L1, and ΔD^+ and ΔD^- are the best-case and worst-case shifts in t_{RC}, respectively. Structures with TSV clusters suffer more significant ΔD^- of up to 6.4%.



(a) NMOS mobility

(b) PMOS mobility

(c) NMOS V_t

(d) PMOS V_t

Fig. 6: Variations in mobility and V_t in L8.

Power: Based on the shifts in V_t and mobility, the contours of I_{leak} are shown in Fig. 8. Transistors near TSV stripes suffer significant variations, with shifts of up to 32% seen in L1, with the widest TSV stripe. TSV clusters induce larger variations, of up to 60% in L5–L8.

The last four columns of Table IV show the variations of leakage power, P_{leak}, and refresh power, P_{ref}, in L1–L8. All percentage changes are with reference to the nominal leakage power, P_{leak}^{nom} and the nominal refresh power, P_{ref}^{nom}, for L1 in the absence of stress-induced leakage shifts. Across layouts, ΔP_{leak} varies only slightly since it is dominated by layout-independent stress (layout-dependent stress is diluted when averaged over the chip). However, ΔP_{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 ΔP_{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 ΔP_{ref} larger than 30%, and measure the area overhead associated with the KOZ in Table V. 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_{TSV}, A_{KOZ}, and A_{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 10.8% and 43.9% and is largest for L5. Note that 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_{leak} increase than those in the middle, as shown in Figs. 7(b) and (d), owing to the additional warpage stress which is more pronounced near the edge of the chip.

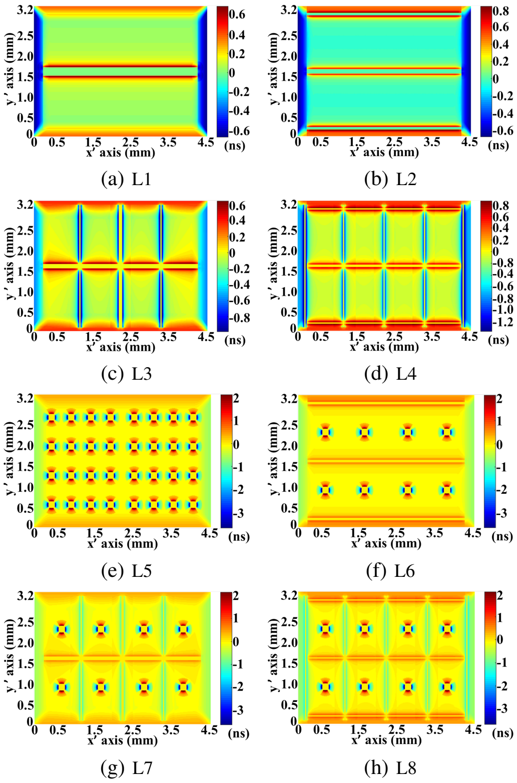
TABLE V: Area Overhead of TSV and KOZ for L1–L8

Layout	A_{TSV} (mm^2)	A_{TSV} (%)	A_{KOZ} (mm^2)	A_{KOZ} (%)	A_{total} (mm^2)	A_{total} (%)
L1	0.75	5.2%	0.80	5.6%	1.55	10.8%
L2	0.75	5.2%	1.20	8.3%	1.95	13.5%
L3	0.75	5.2%	1.14	7.9%	1.89	13.1%
L4	0.75	5.2%	2.16	15.0%	2.91	20.2%
L5	0.72	5.0%	5.60	38.9%	6.32	43.9%
L6	0.74	5.2%	2.51	17.4%	3.25	22.6%
L7	0.76	5.3%	1.70	11.8%	2.46	17.1%
L8	0.78	5.4%	2.55	17.7%	3.33	23.1%

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.

VI. 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 and leakage power. Layouts that use a single strip in the middle of the chip show the lowest stress overhead. These could be worse for communication latencies, but improved stress profiles compensate for this loss.



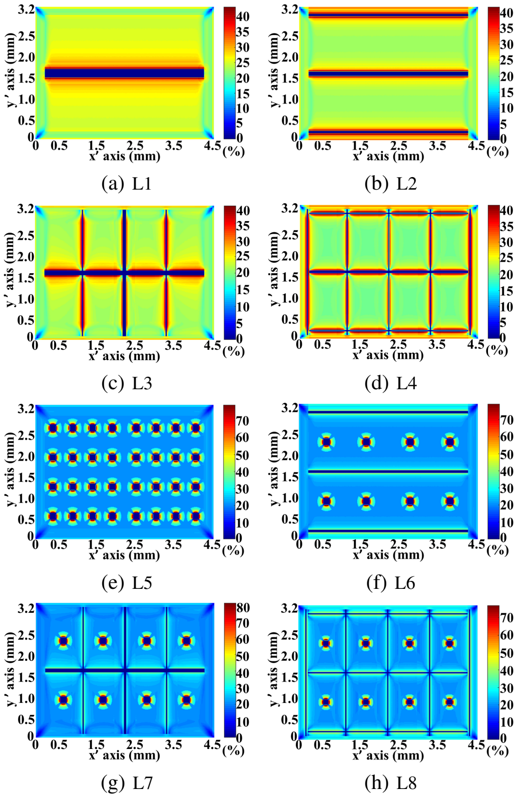
(a) L1 (b) L2

(c) L3 (d) L4

(e) L5 (f) L6

(g) L7 (h) L8

Fig. 7: t_{RC} variation contours in the eight layouts.



(a) L1 (b) L2

(c) L3 (d) L4

(e) L5 (f) L6

(g) L7 (h) L8

Fig. 8: Subthreshold leakage current variations in the eight layouts.

Appendix

The components of the row cycle time, t_{RC}, are detailed below [15]: (1) The term $t_{\text{row-dec-drv}}$ 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, $\alpha = \tau_t/\tau_0$, τ_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:

$$t_{\text{BL}} = \begin{cases} \sqrt{2t_{step}\frac{V_{DD}-V_{tn}}{m}} & \text{if } t_{step} \le 0.5\left(\frac{V_{DD}-V_{tn}}{m}\right) \\ t_{step}+\frac{V_{DD}-V_{tn}}{2m} & \text{if } t_{step} > 0.5\left(\frac{V_{DD}-V_{tn}}{m}\right) \end{cases} \qquad (14)$$

where V_{tn} is the threshold voltage of the NMOS in the wordline decoding circuit, m is the slope of wordline signal, and $t_{step} = 2.3\frac{V_{DD}}{I_{on}}\frac{C_{cell}C_{bl}}{C_{cell}+C_{bl}}$, 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 $t_{\text{SA}} = \frac{C_{bl}}{g_{mn}+g_{mp}}\ln\left(\frac{V_{DD}}{\Delta V}\right)$ where ΔV is the differential input voltage of 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}.

References

[1] U. Kang, *et al.*, "8 Gb 3-D DDR3 DRAM using through-silicon-via technology," *IEEE J Solid-St Circ*, vol. 45, no. 1, pp. 111–119, 2010.

[2] J.-S. Kim, *et al.*, "A 1.2 V 12.8 GB/s 2 Gb mobile wide-I/O DRAM with 4×128 I/Os using TSV based stacking," *IEEE J Solid-St Circ*, vol. 47, no. 1, pp. 107–116, 2012.

[3] D. U. Lee, *et al.*, "A 1.2 V 8 Gb 8-channel 128 GB/s high-bandwidth memory (HBM) stacked DRAM with effective I/O test circuits," *IEEE J Solid-St Circ*, vol. 50, no. 1, pp. 191–203, 2015.

[4] T. Zhang, *et al.*, "3D-SWIFT: A high-performance 3D-stacked wide IO DRAM," in *Proc. GLSVLSI*, pp. 51–56, 2014.

[5] S. J. Koester, *3D integration for VLSI systems*. Pan Stanford, Singapore, 2011.

[6] P. Garrou, *et al.*, *Handbook of 3D integration*. Wiley, Weinheim, Germany, 2011.

[7] Q. Zou, *et al.*, "Thermomechanical stress-aware management for 3D IC designs," in *Proc. DATE*, pp. 1255–1258, 2013.

[8] 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 T VLSI Syst*, vol. 23, no. 7, pp. 1308–1321, 2015.

[9] M. Jung, *et al.*, "Chip/package mechanical stress impact on 3-D IC reliability and mobility variations," *IEEE T Comput Aid D*, vol. 32, no. 11, pp. 1694–1707, 2013.

[10] Y. Sun, *et al.*, "Physics of strain effects in semiconductors and metal-oxide-semiconductor field-effect transistors," *J Appl Phys*, vol. 101, no. 10, 2007.

[11] R. C. Jaeger, *et al.*, "CMOS stress sensors on [100] silicon," *IEEE J Solid St Circ*, vol. 35, no. 1, pp. 85–95, 2000.

[12] W. Zhang and J. G. Fossum, "On the threshold voltage of strained-Si$-$Si$_{1-x}$Ge$_x$ MOSFETs," *IEEE T Electron Dev*, vol. 52, no. 2, pp. 263–268, 2005.

[13] K. Chen, *et al.*, "CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory," in *Proc. DATE*, pp. 33–38, 2012.

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

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

[16] C. Herring and E. Vogt, "Transport and deformation-potential theory for many-valley semiconductors with anisotropic scattering," *Phys Rev*, vol. 101, no. 3, p. 944, 1956.