Stress-Induced Performance Shifts in 3D DRAMs

TENGTAO LI, University of Minnesota, USA
SACHIN S. SAPATNEKAR, University of Minnesota, USA

3D-stacked DRAMs can significantly increase cell density and bandwidth while also lowering power consumption. However, 3D structures experience significant thermomechanical stress due to the differential rate of contraction of the constituent materials, which have different coefficients of thermal expansion. This impacts circuit performance. This paper develops 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.

CCS Concepts: • **General and reference** → **Reliability**; **Performance**; • **Hardware** → **3D integrated circuits**; **Modeling and parameter extraction**;

Additional Key Words and Phrases: Stress, 3D DRAMs, Wide I/O, finite element analysis, performance analysis, package substrate

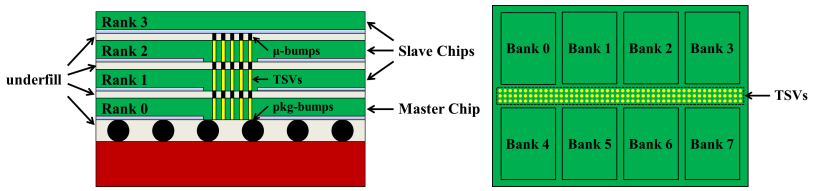
ACM Reference Format:
Tengtao Li and Sachin S. Sapatnekar. 2020. Stress-Induced Performance Shifts in 3D DRAMs. *ACM Trans. Des. Autom. Electron. Syst.* 1, 1, Article 1 (January 2020), 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

3D DRAMs are considered to be the first commercial product to bring 3D stacking and TSVs into the mainstream memory market [29]. Products related to this technology, used in application domains such as mobile phones [11], high-end servers [12] and graphics/computing units [9], stack multiple DRAM layers in the vertical direction, with all layers connected with through-silicon-vias (TSVs) that can transmit signals including data, address, and power signals [6, 11, 28]. Each layer contains not only DRAM cells, but also addressing and other peripheral circuitry. A primary benefit of 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 DRAM 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

Authors' addresses: Tengtao Li, University of Minnesota, 200 Union Street SE, Minneapolis, MN, 55455, USA, lixx2967@umn.edu; Sachin S. Sapatnekar, University of Minnesota, 200 Union Street SE, Minneapolis, MN, 55455, USA, sachin@umn.edu.

1084-4309/2020/1-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) 3D DRAM structure (b) Layout of Rank 0

Fig. 1. 3D DRAM structure [28] and the layout of the rank 0 layer.

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 [26]. Therefore, stacked 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 [28] (in some structures, multiple ranks may be placed in each layer [26]). 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 [28]. A typical configuration stacks all chips on a flip-chip package using back-to-face (B2F) bonding [4], 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.

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 [19]. 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 [25],
(b) μ-bump and package bump induced stress [15], 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, 30, 32], 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.

Traditionally, a rigid package 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 5.2.

In this work, we explore the use of polyimide (PI) 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 [5], radio frequency identification cards (RFIDs) [8], system-in-foil (SiF) [7], flexible sensor array [2] and DRAM array [31]. Compared to the traditional package substrate material, PI has smaller CTE [14], 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 5.2.

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 [25] 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 [15], 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.

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 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.

2 PERFORMANCE EVALUATION OF 3D DRAM

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 [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 [24], 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.

2.1 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

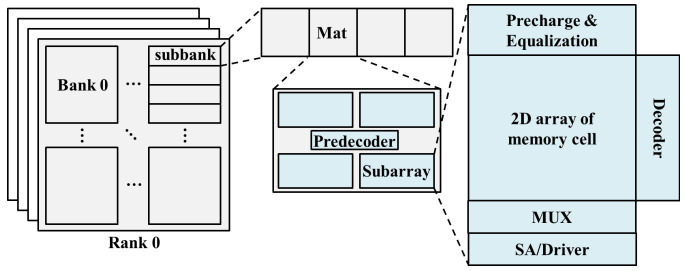


Fig. 2. Organization of a 3D DRAM array.

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 [23]:

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

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:

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

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:

$$P_{ref} = \frac{E_{\text{refresh}}}{T_{\text{refresh}}} \quad (3)$$

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 [23]. These quantities are independent of stress, but the refresh period is strongly affected by stress and influences P_{ref}.

2.2 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}.

3 STRESS MODELING OF A 3D DRAM STACK

3.1 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 1 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 and stress induced by package bumps. Intrinsic stress is also layout-independent.

By linear superposition, we can perform the tensor addition:

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

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.

3.2 Stress Analysis of a 3D DRAM Stack

Consider a 8Gb 3D DRAM with four stacked memory chips, similar to [28], as shown in Fig. 1. Each layer is thinned from the wafer thickness of $\sim 300\mu$m 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 1, where D, H, and P are the diameter, height, and pitch, respectively.

Table 1. 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

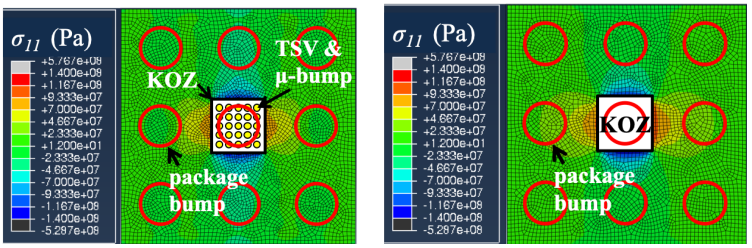
Table 2. Material Parameters

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

The entire 3D DRAM structure undergoes a thermal load of $\Delta T=-250°C$ as it is annealed from 275°C to 25°C (room temperature) to represent the annealing process. We consider the worst-case scenario here, since the operating temperature of 3D DRAMs 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 [16, 18]. The materials in the stack shrink differentially due to their differing CTEs, inducing thermal stress. All material parameters are summarized in Table 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 [21], 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



(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.

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:

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

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 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 and μ-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 VF 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 VF 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 and change the following parameters:

(a) The structure of the TSV arrangement: configuration S1 is a TSV stripe of 20×1 TSVs, as against the square configuration in Fig. 3.
(b) The number of TSVs: configuration S2 changes the different number of TSVs in the TSV cluster, as compared to Fig. 3.

(c) The density of TSVs: configuration S3 uses 5×5 TSVs, distributed with a larger pitch of 40μm. This leads to a lower TSV distribution density and a smaller volume fraction of the TSV region, according to (7).

The results and error distributions for these cases are as shown in Table 3 demonstrate that the volume fraction approximation is very accurate across a range of TSV numbers, distributions, and densities.

Table 3. Summary of Error Distributions between FEA and Volume Fraction Approximation in S1–S3

Structure	# TSVs	TSV pitch (μm)	Average error	Variance	Worst case error
S1	20×1	25	-0.25%	8.79E-05	-2.47%
S2	3×3	25	-0.43%	1.60E-04	-3.20%
S3	5×5	40	-0.41%	2.26E-04	-2.77%

(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.

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 [21]. 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, $\sigma_{warpage}$, and package bumps, $\sigma_{pkg-bump}$, by linear superposition:

$$\sigma_{LI} = \sigma_{warpage} + \sigma_{pkg-bump} \tag{9}$$

To compute the warpage-induced stress, we simulate the 3D stack with no TSVs, μ-bumps or package bumps and apply the thermal load of $\Delta T = -250°$C, and find the stress, $\sigma_{warpage}$ 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, $\sigma_{pkg-bump}$, we simulate another 3D structure with no TSVs or μ-bumps but containing one package bump in the underfill layer between the substrate and nethermost DRAM layer. The σ_{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 (9).

Fig. 4 shows the representative stress component σ_{11} of $\sigma_{pkg-bump}$, as a function of the the distance from the center of the package bump along vertical direction. The radius of the package bump is 50μm and the yellow region represents the region right above 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

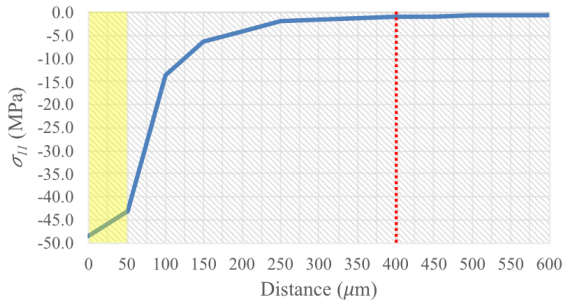


Fig. 4. σ_{11} component of $\sigma_{pkg-bump}$ along vertical direction.

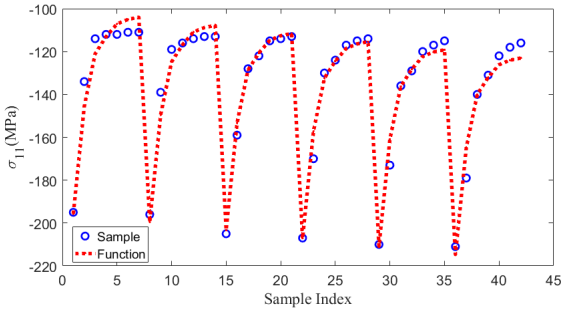


Fig. 5. The accuracy of our semianalytical model for a TSV cluster: the solid curve fits the blue sample points. The horizontal axis is the set of evaluated points (7 values of r evaluated at 6 values of w).

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:

$$\sigma_{pkg-bump} = \sum_{i=1}^{K} \sigma_{pkg-bump\_i} \tag{10}$$

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. Fig. 5 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 1.3×10^{-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:

$$\sigma_{total} = \sigma_{LI} + \sum_{i=1}^{N} \sigma_{TSV\_stripe\_i} + \sum_{i=1}^{M} \sigma_{TSV\_cluster\_i} \tag{11}$$

4 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 [20]. The relative change of mobility in the rotated coordinate system (x', y') is given by [20]:

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

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 $\pi/2$. The piezoresistivity coefficients are taken from [25].

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 [1]. 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 [30]:

$$q\Delta V_{tn} = m\Delta E_C - (m-1)\Delta E_V \tag{13}$$

$$q\Delta V_{tp} = m\Delta E_V - (m-1)\Delta E_C \tag{14}$$

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:

$$\begin{aligned}&\Delta E_C^{(i)}(\epsilon) = \Xi_d(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) + \Xi_u\epsilon_{ii}, i \in \{x, y, z\} \\ &\Delta E_V^{(hh,lh)}(\epsilon) = a(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) \\ &\pm\sqrt{\frac{b^2}{4}(\epsilon_{xx} + \epsilon_{yy} - 2\epsilon_{zz})^2 + \frac{3b^2}{4}(\epsilon_{xx} - \epsilon_{yy})^2 + d^2\epsilon_{xy}^2}\end{aligned} \tag{15}$$

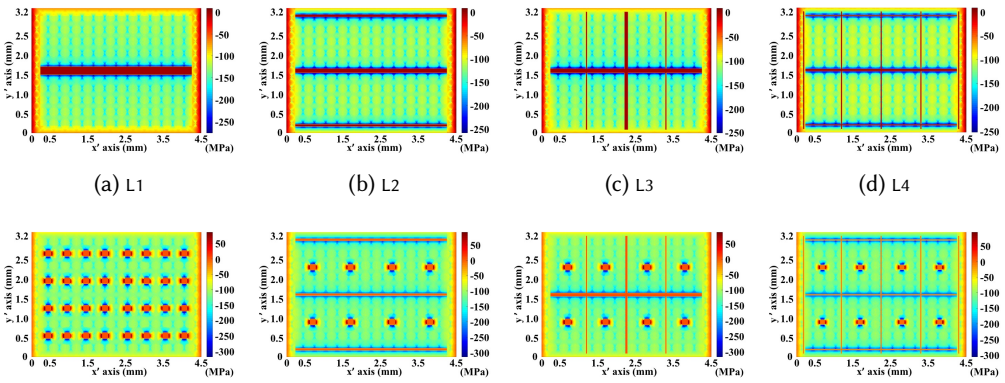
where Ξ_d and a are the hydrostatic deformation potential constants, which can induce shifts in the conduction band and valence band, respectively, while Ξ_u, b, and d are the shear deformation potential constants that affect the conduction and valence bands.

5 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 (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 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 6×6 arrangement of TSVs in each cluster; L8 uses a 5×5 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 4. 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



(a) L1 (b) L2 (c) L3 (d) L4

(e) L5 (f) L6 (g) L7 (h) L8

Fig. 6. Contours of σ_{11} in the eight layouts.

5.1 Using FR-4 as the package substrate material

The spatial distribution of TSVs in L1–L8 is apparent in Fig. 6, which shows the contours of σ_{11}, as a representative stress component, for each structure in the master chip placed at the bottom

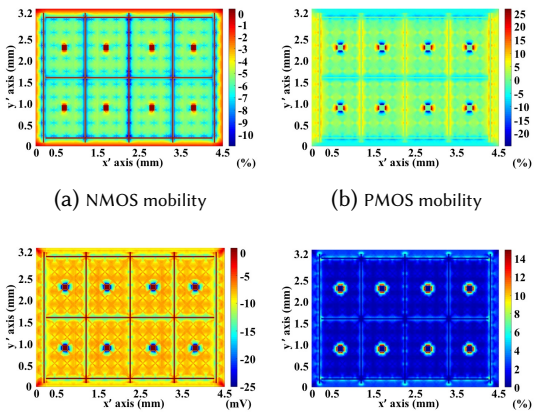
of the stack, which experiences the largest stress. Since the region of interest is outside the TSV clusters/stripes, for convenience the color code inside the TSV regions shows zero stress within. Each stress contour translates to a map of mobility and threshold voltage variations, and Fig. 7 shows the data corresponding to L8 in Fig. 6(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%.

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 −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 (1), but similar analyses can be performed for other timing metrics. The t_{RC} variation contours in L1–L8 are shown in Fig. 8 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.

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 5. 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 7.0%.

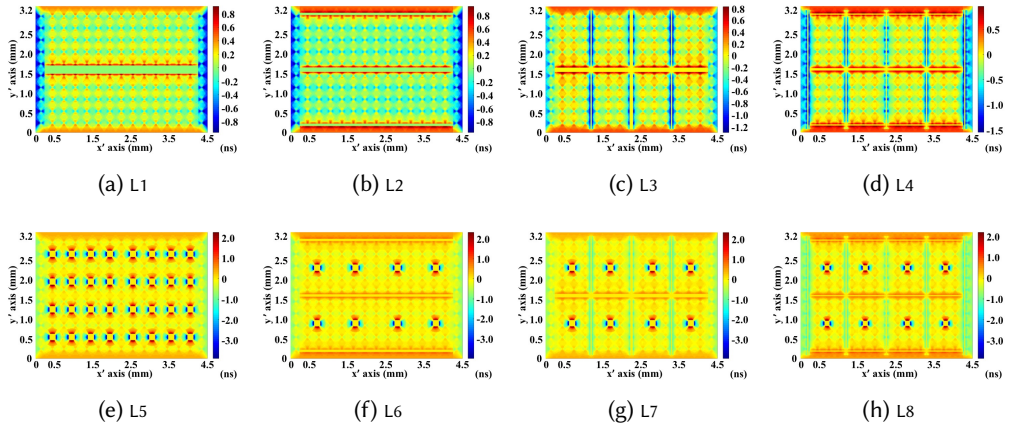
Power: Based on the shifts in V_t and mobility, the contours of I_{leak} are shown in Fig. 9. Transistors near TSV stripes suffer significant variations, with shifts of up to 50% seen in L1, with the widest TSV stripe. TSV clusters induce larger variations, of up to 88% in L5–L8.



(a) NMOS mobility (b) PMOS mobility

(c) NMOS V_t (d) PMOS V_t

Fig. 7. Variations in mobility and V_t in L8.



(a) L1 (b) L2 (c) L3 (d) L4

(e) L5 (f) L6 (g) L7 (h) L8

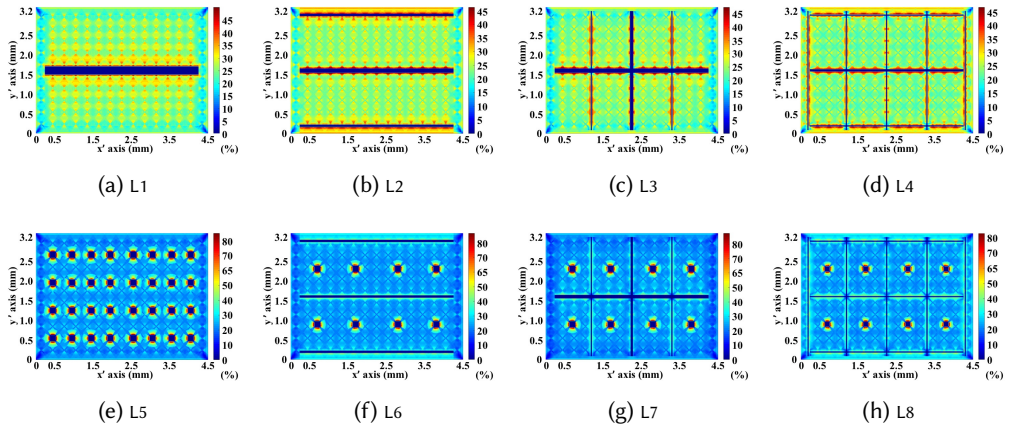
Fig. 8. t_{RC} variation contours in the eight layouts.

Table 5. 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	-0.96	-2.9%	0.91	2.7%	12.35	24.4%	9.53	50.4%
L2	-0.96	-2.9%	0.94	2.8%	12.12	23.9%	8.74	46.2%
L3	-1.30	-3.9%	0.83	2.5%	11.52	22.7%	8.98	47.5%
L4	-1.54	-4.6%	0.95	2.8%	12.04	23.8%	8.85	46.8%
L5	-3.89	-11.6%	2.26	6.7%	13.06	25.8%	16.11	85.2%
L6	-3.98	-11.8%	2.32	6.9%	12.35	24.4%	16.38	86.7%
L7	-3.95	-11.7%	2.34	7.0%	11.89	23.5%	16.55	87.6%
L8	-3.87	-11.5%	2.20	6.5%	12.45	24.6%	15.68	83.0%

The last four columns of Table 5 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 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_{TSV}, A_{KOZ}, and A_{total} are, respectively, the



(a) L1 (b) L2 (c) L3 (d) L4

(e) L5 (f) L6 (g) L7 (h) L8

Fig. 9. Subthreshold leakage current variations in the eight layouts.

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 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. 8(b) and (d), owing to the additional warpage stress which is more pronounced near the edge of the chip.

Table 6. 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%	2.18	15.1%	2.93	20.4%
L2	0.75	5.2%	2.03	14.1%	2.78	19.3%
L3	0.75	5.2%	1.16	8.0%	1.91	13.2%
L4	0.75	5.2%	2.35	16.3%	3.10	21.5%
L5	0.72	5.0%	2.85	19.8%	3.57	24.8%
L6	0.74	5.2%	2.56	17.8%	3.30	22.9%
L7	0.76	5.3%	1.78	12.4%	2.54	17.7%
L8	0.78	5.4%	2.97	20.6%	3.75	26.0%

Wire length and bus delay: In 3D DRAMs, memory bus routing uses a spine-like structure [13] that is used to connect the logic control circuits to the banks and finally to each memory cell, as shown in Fig. 10. Due to the KOZ, the memory arrays must be spaced away from the 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. 10 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, $\cdots$, L8, Table 7 shows the percentage increase in the bus delay variation, ΔD_{bus}, with respect to the nominal bus delay of 0.93ns. We see that L7 shows the largest ΔD_{bus} of 14.3%.

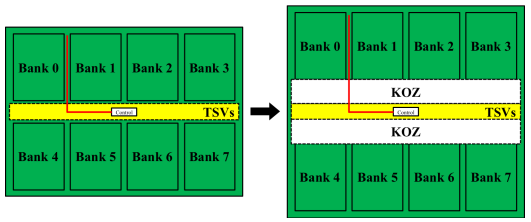


Fig. 10. The increase of wire length caused by the KOZ.

Table 7. Bus delay variation for L1–L8

Layout	L1	L2	L3	L4	L5	L6	L7	L8
ΔD_{bus} (%)	1.5%	2.4%	2.5%	3.4%	9.6%	10.0%	14.3%	14.0%

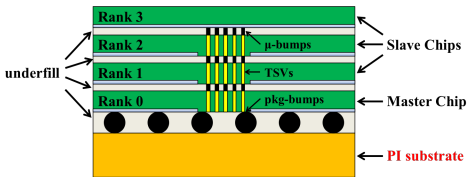


Fig. 11. 3D DRAM structure with the PI substrate.

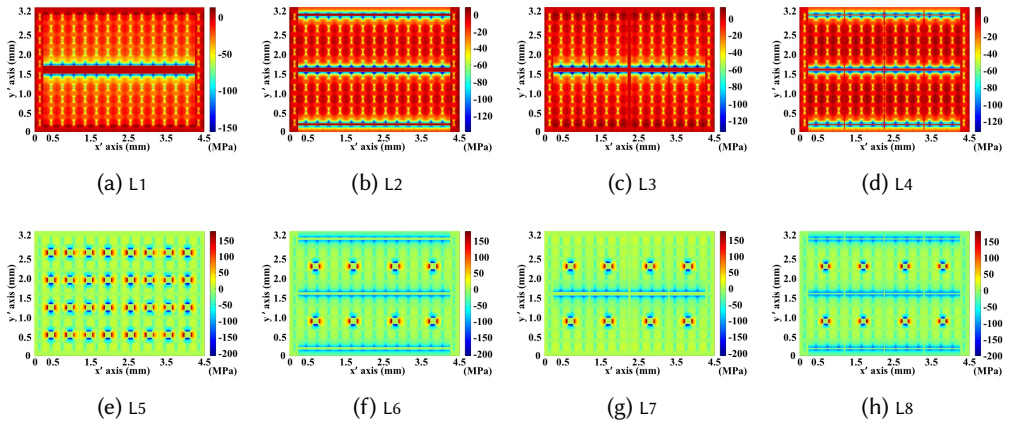
Runtime: FEA is computational: an L1-like layout with 400 TSVs requires 4 hours of CPU time (Intel Xeon 5560 Nehalem, 2.80GHz); with 800 TSVs, it times out after a day. Our volume fraction method computes L1 (1200 TSVs) in 98s, and our semianalytical model only requires a few clock cycles (2 multiplies, 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.

5.2 Using polyimide as package substrate material

As shown in Fig. 6, warpage-induced layout-dependent stress takes a large proportion of the total stress and the most significant σ_{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 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. 11, to decrease the CTE mismatch between substrate and other layers. The CTE of the PI substrate can be as low as 3 ppm/K [14], 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.

We investigate the same set of TSV layouts as summarized in Table 4 to compare the results between 3D DRAMs with the traditional package substrate and proposed PI substrate. Fig. 12 shows

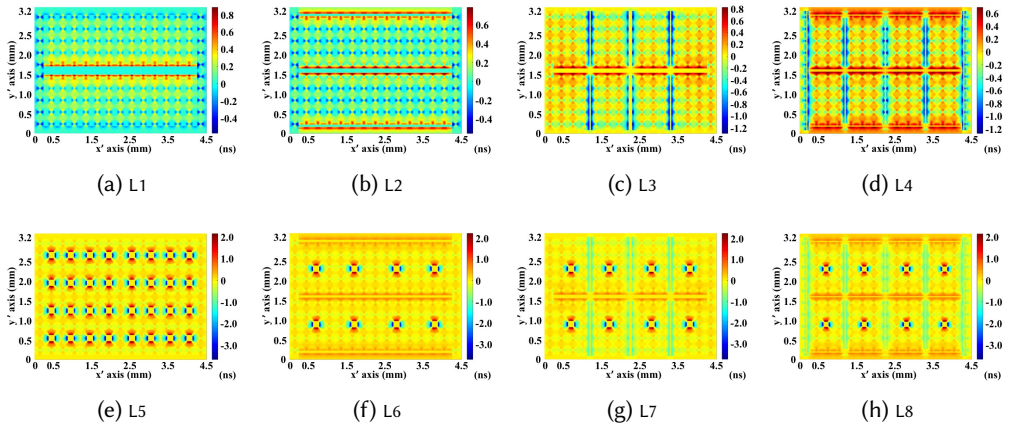


(a) L1 (b) L2 (c) L3 (d) L4

(e) L5 (f) L6 (g) L7 (h) L8

Fig. 12. Contours of σ_{11} in the eight layouts with the PI substrate.

the contours of the representative stress component σ_{11} for each layout. Comparing to the stress results for 3D DRAMs with the FR-4 package substrate as shown in Fig. 6, there is a shift of about 100MPa in each layout. For example, the range of σ_{11} in layout L1 with the FR-4 package substrate, as shown in Fig. 6, is from −250MPa to 0MPa, where the negative sign represents compressive stress. However, σ_{11} in the same layout with PI substrate ranges between −150MPa and 0MPa. 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 σ_{11} in 3D DRAMs with the PI substrate is only about −2MPa, 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 σ_{11} in the first row and second row in Fig. 12.



(a) L1 (b) L2 (c) L3 (d) L4

(e) L5 (f) L6 (g) L7 (h) L8

Fig. 13. t_{RC} variation contours in the eight layouts with the PI substrate.

Table 8. Row Cycle Time (t_{RC}), Leakage Power (P_{leak}), and Refresh Power (P_{ref}) for L1–L8 with PI substrate (D_0 = 33.62ns, P_{leak}^{nom} = 50.66mW, P_{ref}^{nom} = 18.90mW)

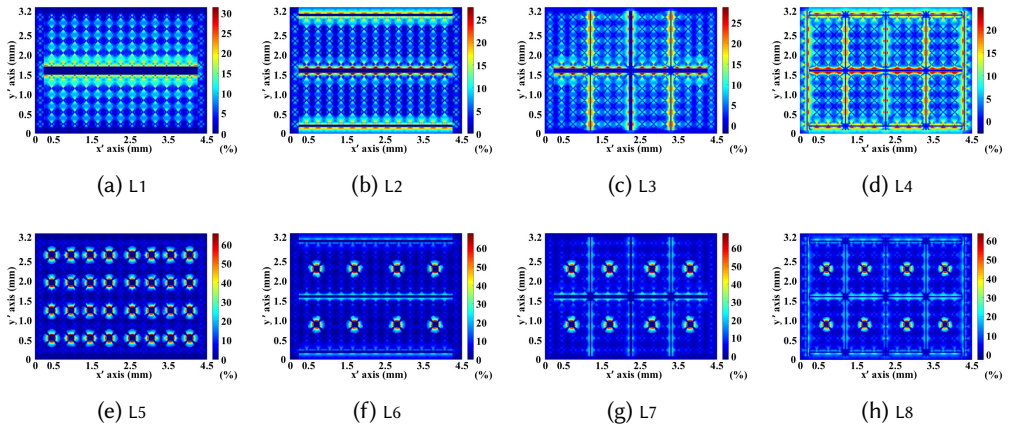
	Row Cycle Time t_{RC}						Leakage P_{leak}			Refresh P_{ref}		
	ΔD^+ (ns)	ΔD^+ (%)	Δ_{PI}^{FR-4} (%)	ΔD^- (ns)	ΔD^- (%)	Δ_{PI}^{FR-4} (%)	ΔP_{leak} (mW)	ΔP_{leak} (%)	Δ_{PI}^{FR-4} (%)	ΔP_{ref} (mW)	ΔP_{ref} (%)	Δ_{PI}^{FR-4} (%)
L1	-0.58	-1.7%	1.2%	0.88	2.6%	-0.1%	3.10	6.1%	-18.2%	5.94	31.4%	-19.0%
L2	-0.56	-1.7%	1.2%	0.79	2.3%	-0.4%	3.10	6.1%	-17.8%	5.23	27.7%	-18.6%
L3	-1.31	-3.9%	0.0%	0.82	2.4%	-0.1%	2.66	5.3%	-17.5%	5.45	28.8%	-18.7%
L4	-1.27	-3.8%	0.8%	0.69	2.1%	-0.8%	3.03	6.0%	-17.8%	4.72	25.0%	-21.9%
L5	-3.67	-10.9%	0.7%	2.16	6.4%	-0.3%	3.83	7.6%	-18.2%	12.42	65.7%	-19.5%
L6	-3.74	-11.1%	0.7%	2.23	6.6%	-0.3%	3.33	6.6%	-17.8%	12.85	68.0%	-18.7%
L7	-3.76	-11.2%	0.6%	2.24	6.7%	-0.3%	3.10	6.1%	-17.3%	12.92	68.3%	-19.2%
L8	-3.66	-10.9%	0.6%	2.09	6.2%	-0.3%	3.49	6.9%	-17.7%	12.06	63.8%	-19.2%

The computed stress tensors are then translated into variations in electrical parameters with the approach as shown in Section 4. The method discussed in Section 2 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. 13. 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 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 ΔD^+ and ΔD^- are the best-case and worst-case shifts in t_{RC}, respectively, and Δ_{PI}^{FR-4} 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, Δ_{PI}^{FR-4} 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. 14, 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. 9 and Fig. 14.

The leakage power, P_{leak}, and refresh power, P_{ref} are affected by I_{leak}, as discussed in Section 2, and columns 8–13 in Table 8 show the variations of P_{leak} and P_{ref}. All percentage changes are with reference to the nominal leakage power, P_{leak}^{nom}, and the nominal refresh power, P_{ref}^{nom}, and Δ_{PI}^{FR-4} denotes the variation between structures with the FR-4 substrate and PI substrate, with respect to P_{leak}^{nom} and P_{ref}^{nom}, respectively. Typically, TSV clusters can induce more significant variation in both P_{leak} and P_{ref}. By observing Δ_{PI}^{FR-4} results, it can be driven there is significant drop in both P_{leak} and P_{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_{leak} and an average shift of -19.3% in P_{ref}, with respect to P_{leak}^{nom} and P_{ref}^{nom}, respectively.



(a) L1 (b) L2 (c) L3 (d) L4

(e) L5 (f) L6 (g) L7 (h) L8

Fig. 14. Subthreshold leakage current variations in the eight layouts with the PI substrate.

Table 9. Area Overhead of TSV and KOZ for L1–L8 in 3D DRAMs with PI substrate

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

Area: The PI substrate can significantly reduce I_{leak} and P_{ref}, and hence the area of the KOZ, which is used to avoid large increases in P_{ref}, can be reduced naturally. Using the threshold $\Delta P_{ref} > 30\%$ for deciding the KOZ, the results for L1–L8 are summarized in Table 9. Here, Δ_{PI}^{FR-4} 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_{KOZ} is greatly reduced in L1–L4, which contain only TSV stripes after applying PI substrate. Additionally, there is a significant reduction in A_{KOZ} caused by TSV clusters as well, as shown in L5–L8. The average variation of A_{KOZ} among L1–L8 is -14.3% and the maximum reduction is -19.3%, occurring in L8.

Wire length and bus delay: As the KOZ is reduced, the wire length and bus delay overhead decreases. Table 10 shows the bus delay variation in L1–L8 with the PI substrate and Δ_{PI}^{FR-4} 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 10. Bus delay variation for L1–L8 with PI substrate

Layout	L1	L2	L3	L4	L5	L6	L7	L8
ΔD_{bus} (%)	0.4%	0.0%	0.0%	0.0%	9.0%	5.4%	5.4%	5.1%
Δ_{PI}^{FR-4} (%)	-1.0%	-2.4%	-2.5%	-3.4%	-0.6%	-4.6%	-9.0%	-9.0%

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 [10], 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/mâŃĚK [10], 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 1.2W/m·K by compositing with boron nitride (BN) [22, 27], 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 [3], PI can absorb 3.5X moisture than FR-4 under the same environment.

6 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.

APPENDIX

The components of the row cycle time, t_{RC}, are detailed below [23]:
(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:

$$t_{\text{BL}} = \begin{cases} \sqrt{2t_{step}\frac{V_{DD}-V_{tn}}{m}} & \text{if } t_{step} \le 0.5\left(\frac{V_{DD}-V_{tn}}{m}\right) \\ t_{step} + \frac{V_{DD}-V_{tn}}{2m} & \text{if } t_{step} > 0.5\left(\frac{V_{DD}-V_{tn}}{m}\right) \end{cases} \tag{16}$$

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] Herring, C. and Vogt, E. 1956. Transport and deformation-potential theory for many-valley semiconductors with anisotropic scattering. *Physical Review* 101, 3 (1956), 944–961.
[2] Wang, C., Hwang, D., Yu, Z., Takei, K., Park, J., Chen, T., Ma, B., and Javey, A. 2013. User-interactive electronic skin for instantaneous pressure visualization. *Nature Materials* 12, 10 (2013), 899–904.
[3] Coombs, C. F. and Holden, H. T. 2008. *Printed Circuits Handbook.* McGraw-Hill, New York, NY.
[4] Tan, C. S., Chen, K.-N., and Koester, S. J. 2012. *3D Integration for VLSI Systems.* Pan Stanford, Singapore.
[5] Jin, D.-U., Lee, J.-S., Kim, T.-W., An, S.-G., Straykhilev, D., Pyo, Y.-S., Kim, H.-S., Lee, D.-B., Mo, Y.-G., Kim, H.-D., and Chung, H.-K. 2009. World-largest (6.5") flexible full color top emission AMOLED display on plastic film and its bending properties. In *Proceedings of SID Symposium Digest of Technical Papers.* 983–985.
[6] Lee, D. U., Kim, K. W., Kim, K. W., Lee, K. S., Byeon, S. J., Kim, J. H., Cho, J. H., Lee, J., and Chun, J. H. 2015. A 1.2 V 8 Gb 8-channel 128 GB/s high-bandwidth memory (HBM) stacked DRAM with effective I/O test circuits. *IEEE Journal of Solid-State Circuits* 50, 1 (2015), 191–203.
[7] Cantatore, E. 2013. *Applications of Organic and Printed Electronics.* Springer, New York, NY.
[8] Burghartz, J. 2010. *Ultra-Thin Chip Technology and Applications.* Springer, New York, NY.
[9] Macri, J. 2015. AMD's next generation GPU and high bandwidth memory architecture: FURY. In *Proceedings of IEEE Hot Chips Symposium.* 1–26.
[10] Lienhard IV, J. H. and Lienhard V, J. H. 2011. *A Heat Transfer Textbook.* Dover Publications, Inc, Mineola, NY.
[11] Kim, J.-K., Oh, C. S., Lee, H., Lee, D., Hwang, H. R., Hwang, S., Na, B., Moon, J., Kim, J.-G., Park, H., Ryu, J.-W., Park, K., Kang, S. K., Kim, S.-Y., Kim, H., Bang, J.-M., Cho, H., Jang, M., Han, C., Lee, J.-B., Choi, J. S., and Jun, Y.-H. 2012. 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 Journal of Solid-State Circuits* 47, 1 (2012), 107–116.
[12] Pawlowski, J. T. 2011. Hybrid memory cube (HMC). In *Proceedings of IEEE Hot Chips Symposium.* 1–24.
[13] Chen, K., Li, S., Muralimanohar, N., Ahn, J. H., Brockman, J. B., and Jouppi, N. P. 2012. CACTI-3DD: architecture-level modeling for 3D die-stacked DRAM main memory. In *Proceedings of IEEE Design, Automation and Test in Europe.* 33–38.
[14] Hasegawa, M. and Horii, S. 2007. Low-CTE polyimides derived from 2,3,6,7-naphthalenetetracarboxylic dianhydride. *Polymer Journal* 39, 6 (2007), 610–621.
[15] Jung, M., Pan, D. Z., and Lim, S. K. 2013. Chip/package mechanical stress impact on 3-D IC reliability and mobility variations. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 32, 11 (2013), 1694–1707.
[16] Jung, M., Mitra, J., Pan, D. Z., and Lim, S. K. 2014. TSV stress-aware full-chip mechanical reliability analysis and optimization for 3D IC. *Commun. ACM* 57, 1 (2014), 107–115.
[17] Garrou, P., Bower, C., and Ramm, P. 2011. *Handbook of 3D Integration, Volume 1: Technology and Applications of 3D Integrated Circuits.* Wiley, Weinheim, Germany.
[18] Montméat, P., Enot, T., De Oliveira, G. L., and Fournel, F. 2018. Polymer bonding temperature impact on bonded stack morphology and adherence energy. *Microsystem Technologies* 24, 1 (2018), 793–799.
[19] Zou, Q., Zhang, T., Kursun, E., and Xie, Y. 2013. Thermomechanical stress-aware management for 3D IC designs. In *Proceedings of IEEE Design, Automation and Test in Europe.* 1255–1258.
[20] Jaeger, R. C., Suhling, J. C., Ramani, R., Bradley, A. T., and Xu, J. 2000. CMOS stress sensors on [100] silicon. *IEEE Journal of Solid-State Circuits* 35, 1 (2000), 85–95.
[21] Teixeira, R. C., De Munck, K., De Moor, P., Baert, K., Swinnen, B., Van Hoof, C., and Knüttel, A. 2008. Stress analysis on ultra thin ground wafers. *Journal Integrated Circuits and Systems* 3, 2 (2008), 83–89.

[22] Diaham, S., Saysouk, F., Locatelli, M.-L., Belkerk, B., Scudeller, Y., Chiriac, R., Toche, F., and Salles, V. 2015. Thermal conductivity of polyimide/boron nitride nanocomposite films. *Journal of Applied Polymer Science* 132, 34 (2015), 42461–1–42461–9.

[23] Thoziyoor, S., Muralimanohar, N., Ahn, J. H., and Jouppi, N. P. 2008. *CACTI 5.1.* Technical Report HPL-2008-20. HP Labs, Palo Alto, CA.

[24] Thoziyoor, S., Muralimanohar N., Ahn, J. H., Wilton, S., and Jouppi, N. 2009. CACTI tools. http://www.hpl.hp.com/research/cacti/.

[25] Marella, S. K. and Sapatnekar, S. S. 2015. A holistic analysis of circuit performance variations in 3D-ICs with thermal and TSV-induced stress considerations. *IEEE Transactions on VLSI Systems* 23, 7 (2015), 1308–1321.

[26] Zhang, T., Xu, C., Chen, K., Sun, G., and Xie, Y. 2014. 3D-SWIFT: a high-performance 3D-stacked wide IO DRAM. In *Proceedings of ACM Great Lakes Symposium on VLSI.* 51–56.

[27] Li, T.-L. and Hsu, S. L.-C. 2010. Enhanced thermal conductivity of polyimide films via a hybrid of micro-and nano-sized boron nitride. *The Journal of Physical Chemistry B* 114, 20 (2010), 6825–6829.

[28] Kang, U., Chung, H.-J., Heo, S., Park, D.-H., Lee, H., Kim, J. H., Ahn, S.-H., Cha, S.-H., Ahn, J., Kwon, D., Lee, J.-W., Joo, H.-S., Kim, W.-S., Jang, D. H., Kim, N. S., Choi, J.-H., Chung, T.-G., Yoo, J.-H., Choi, J. S., Kim, C., and Jun, Y. H. 2010. 8 Gb 3-D DDR3 DRAM using through-silicon-via technology. *IEEE Journal of Solid-State Circuits* 45, 1 (2010), 111–119.

[29] Kim, W., Kim, D.-H., Hong, H. I., Milor, L., and Lim, S. K. 2014. Impact of die partitioning on reliability and yield of 3D DRAM. In *Proceedings of IEEE Interconnect Technology Conference/Advanced Metallization Conference.* 389–392.

[30] Zhang, W. and Fossum, J. G. 2005. On the threshold voltage of strained-Si–$Si_{1-x}Ge_x$ mosfets. *IEEE Transactions on Electron Devices* 52, 2 (2005), 263–268.

[31] Zhang, W., Ha, M., Braga, D., Renn, M. J., Frisbie, C. D., and Kim, C. H. 2011. A 1V printed organic DRAM cell based on ion-gel gated transistors with a sub-10nW-per-cell refresh power. In *Proceedings of IEEE Solid-State Circuits Conference Digest of Technical Papers.* 326–328.

[32] Sun, Y., Thompson, S. E., and Nishida, T. 2007. Physics of strain effects in semiconductors and metal-oxide-semiconductor field-effect transistors. *Journal of Applied Physics* 101, 10 (2007), 104503–1–104503–22.