Full Chip Stress Model for Defect Formation Risk Analysis in Multilayer Structures

2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) | 979-8-3315-1635-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/SISPAD62626.2024.10732998

Kyungmi Yeom
CSE Team
Samsung Electronics
Hwaseong-si, Korea
k.yeom@samsung.com

Geunsang Yoo
CSE Team
Samsung Electronics
Hwaseong-si, Korea
geunsang.yoo@samsung.com

Alexander Schmidt
CSE Team
Samsung Electronics
Hwaseong-si, Korea
alexander.shmidt@samsung.com

Anthony Payet
Process TCAD Lab
Samsung Electronics
Yokohama, Japan
a.payet@samsung.com

Joohyun Jeon
CSE Team
Samsung Electronics
Hwaseong-si, Korea
jooh.jeon@samsung.com

Seungmin Lee
CSE Team
Samsung Electronics
Hwaseong-si, Korea
sm101.lee@samsung.com

Yutaka Nishizawa
Process TCAD Lab
Samsung Electronics
Yokohama, Japan
y.nishizawa@samsung.com

Masaru Uchiyama
Process TCAD Lab
Samsung Electronics
Yokohama, Japan
m.uchiyama@samsung.com

Yasuyuki Kayama
Process TCAD Lab
Samsung Electronics
Yokohama, Japan
y.kayama @samsung.com

Chihak Ahn
TCAD Lab
Samsung Semiconductor Inc.
San Jose, United States
chihak.ahn@samsung.com

Woosung Choi
TCAD Lab
Samsung Semiconductor Inc.
San Jose, United States
woosung.c@samsung.com

Dae Sin Kim
CSE Team
Samsung Electronics
Hwaseong-si, Korea
daesin.kim@samsung.com

***Abstract*—To overcome the limitations of the previously developed stress simulation method for full-chip scale [1], which could only analyze a single layer of metallization due to its use of a shell element method, a simulation flow that can handle multiple layers was developed. Introducing stress simulation during the incremental formation of the back-end-of-line (BEOL) structure is crucial for predicting the risk of stress-induced defects not just on the surface, but throughout the entire 3D structure, including the chance of defects between metallization layers. To enhance the predictive capability for the larger-scale stress-induced defects, local stress averaging was utilized to balance simulation accuracy with coverage area. This methodology allowed for the expansion of the simulation domain beyond the chip level, thereby enabling the estimation of layout-induced deformations on a wafer scale**

Keywords— Chip Scale Simulation, Stress Failure, Defect Analysis, Risk Analysis, Multilayer Structure, Layout Analysis

I. Introduction

As semiconductor technology progresses and design rules become stricter, the density of back-end-of-line (BEOL) metallization continues to rise. The formation of the BEOL entails processing steps at elevated temperatures, inevitably leading to internal stresses arising from the mismatch in thermal expansion coefficients of the constituent materials. This can potentially lead to formation of multiple stress-induced defects (cracks and delamination sites both inside specific metallization layer and between the layers) that would reduce product yield, unless the design rules are developed that mitigate stress-induced defects.

To tackle this problem, a chip-scale stress simulation approach was developed [1], yet it is constrained by several significant limitations that must be addressed to enhance both coverage and accuracy. Primarily, metallization layer formation is an incremental process and therefore stress keeps changing before the final structure is formed. It may lead to formation of the defects that could not be predicted by "one shot" simulation considering only the last stage of the process. Secondly, stress of each layer may be affected by both preceding and subsequent metallization layers. Overcoming these limitations is a key to extend the methodology to be able to improve design rule of complex BEOL stacks.

Given the shift in semiconductor engineering from Front-End Process optimization – hampered by increasing complexity and development costs – to complex 3D integration, the necessity for more predictive and accurate simulation tools in this domain is undeniable.

II. Multilayer Full-Chip Scale Stress Simulation Methodology

To overcome limitation of huge size of simulation for a full-chip structure (considering that typical chip size is in the range of 100^2 mm and typical layout features can be as small as 10-100 nm, meaning that complete structure should have size of about 10^{12} nodes, we are applying layout dimension reduction methodology developed previously [2]. This methodology leverages dimensionality reduction (Fig. 1) to extend the manageable simulation size per computation node to several hundred μm and cut into optimum size to calculate the size of chip (Fig. 2). Absence of data exchange between individual tile simulations allows introduction of highly efficient parallelization, effectively running massively parallel execution w/o any overhead (Fig. 3). Individual tile size is the main optimization parameter for simulation TAT and it has strong non-linear dependence of tile size due to memory allocation

delays for large structures. At the same time, with larger tile size total number of tiles goes down. Fig. 3 shows results of tile size optimization: in spite of extremely fast simulations for small tiles, size below 100 µm*100 µm is not efficient.

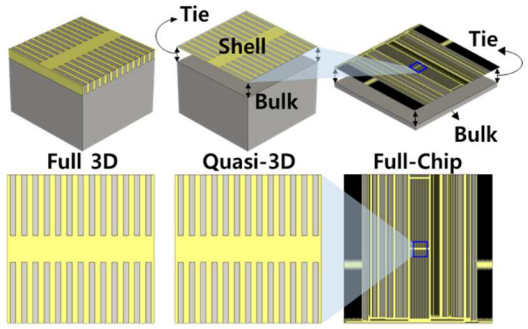

Fig. 1. Conversion of full 3D and Quasi-3D shell + bulk element structures and full chip simulation for a single metallization layer.

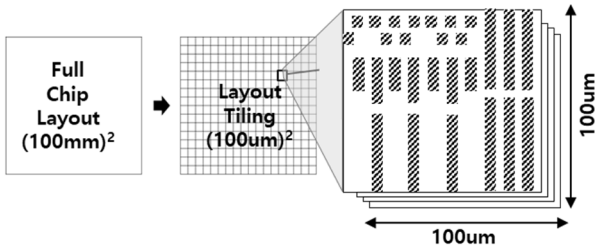

Fig. 2. Full chip simulation execution with independent large scale layout tiling.

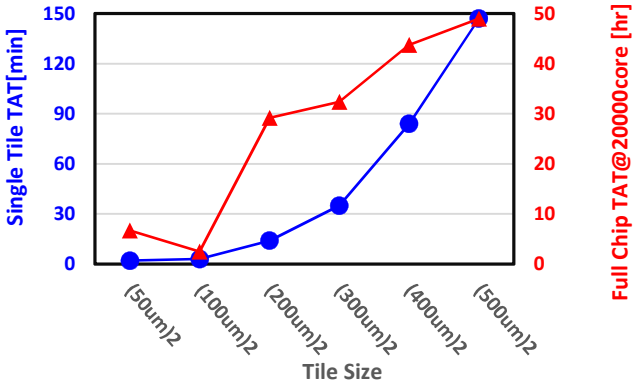

Fig. 3. Single tile simulation TAT & and full chip (100,000 µm*100,000 µm) simulation TAT. For VNAND typical BEOL layout tile size of 100 µm*100 µm is the most efficient.

The principal constraint of the current methodology was its restricted applicability for a single metallization layer. Consequently, a logical advancement should involve expanding the dimension reduction technique would be to extend the dimension reduction technique to cover not only the final processing step of a given layer but also every stage of the Damascene process, including the formation of the insulator layer, patterning, metal deposition, and the chemical-mechanical polishing (CMP) process. The process schematics are shown in Fig. 4 along with the resulting stress distribution extracted from the top of the structure. At each step, the structure is modeled as a combination of a bulk element (representing the substrate) and a stack of shell elements, each comprising up to two materials selected from Oxide, Metal, or Gas. To streamline the treatment of boundary conditions, a material with an extremely low bulk modulus is employed to represent Gas, ensuring continuity within the shell element. Implementation of multilayer structure treatment allowed not only more accurate prediction of stress distribution in the single metallization layer, but also gives an opportunity to predict lateral stress distribution in the structures with multiple metallization layers progressively.

An example of multilayer simulation flow in shown Fig. 5. The actual deposition process for each metallization layer includes four steps, as depicted in Fig. 4; therefore, the simulation in Fig. 5 comprises 12 intermediate steps. The stress distribution changes after each processing step. Consequently, the criterion for crack formation previously developed using the Weibull distribution [1] should be applied not only to the final stage but also to all intermediate stages of processing.

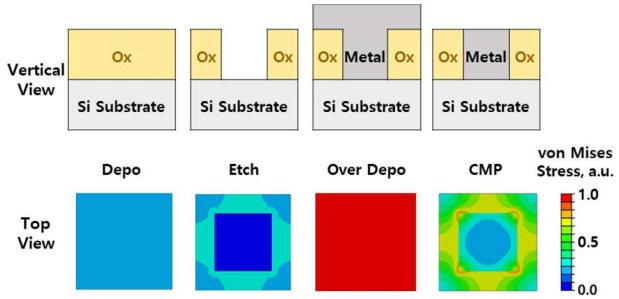

Fig. 4. Schematic of a progressive simulation for a simple structure: each step of a Single Damascene Process is followed and von Mises stress in a.u. is shown.

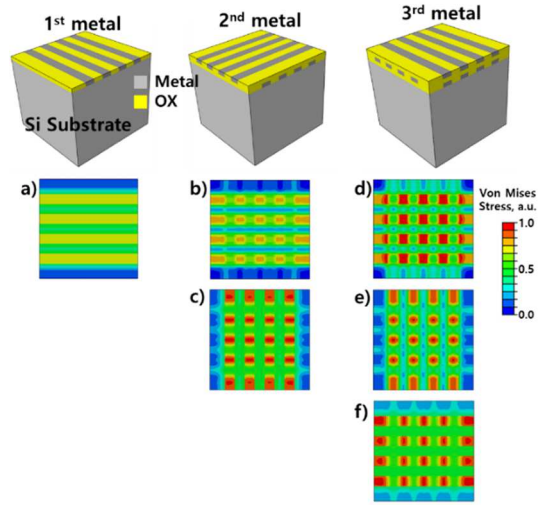

Fig. 5. Schematic of multilayer simulation and the effect of further layers. a) von Mises stress after first metal layer formation, b), c) von Mises stress after second metal layer formation, calculated in the first and the second layers, respectively, d), e), f) von Mises stress in the first, second and third layers, respectively, calculated after third metallization layer formation.

III. Extension to Global Structure Deformation

Layout-dependent stresses generated in a single chip would have effect at the wafer scale also. Typically, Stoney equation

with its multiple extensions are used as the solution to predict overall wafer warpage [3], but this approach has an inherent challenge to extract the anisotropic input stresses from the chip layout. It needs a full-chip stress simulation self-consistently, which would require huge computational resources even if dimension reduction and massive parallelization that we've successfully implemented before is applied, since in this case we need to exchange data between simulation nodes leading to unavoidable performance bottleneck.

Nevertheless, the methodology can be used in a hierarchical manner: first a single chip stress distribution is generated using tiling technique (Fig. 1-3.) and then resulting principal stresses XX, YY extracted and averaged over each single tile. Due to this averaging, a simulation mesh can be radically coarsened therefore it is possible to simulate stress at chip and wafer level at once, applying proper boundary conditions and extracting overall deformation depending on the layout. In this simulation, of course, large portion of information regarding local stress effects is lost, but since large scale deformations do not depend on local fine layout features, it is possible to improve simulation efficiency without significant degradation of the accuracy of results.

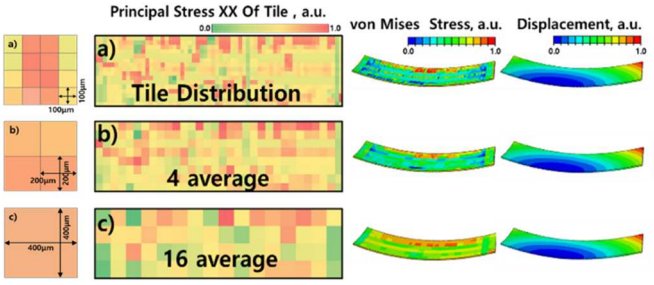

Fig. 6. Local stress averaging method: a) Principal stress distribution across tiles of 100 µm, b) averaging results for 4 tiles, and c) averaging results for 16 tiles. And Local stress averaging method demonstrates von Mises stress contour and 3D deformation as a result: even though there is difference in local stress contour, deformation level is nearly identical

User can control extent of the local averaging, Fig. 6 shows example of the local Stress XX averaging for tiles of 100 µm*100 µm, 200 µm*200 µm and 400 µm*400 µm. As shown in Fig. 6 even though local stress distribution strongly depends on averaging, overall chip deformation is much less sensitive and even if radical averaging is applied (Fig. 6c), results of chip-scale deformation are virtually identical to the results of simulation with less averaging (Fig. 6a, 6b). Local stress averaging reduces total tile number and generates similar contour deformation shown in Fig. 6. Thus, the method is suitable for assessing the probability of stress-induced defect formation in practical chip-scale applications and allows manageable TAT of simulation.

IV. Combining Global Warpage and Local Stress-Induced Defect Risk Assessment

Based on the successful chip-scale layout-based stress analysis and extraction of the resulting chip warpage and deformation, we are striving to extend the analysis to local pattern dependent defect formation probability. While the chip warpage simulation inevitably demanded stress data averaging over large areas of the chip, local defect formation probability should mostly depend on local stress distribution. To overcome conundrum of the necessity to keep a few tens of nanometer-scale precision of the stress distribution with the consideration of mm-scale chip deformation of tens of micrometers, we are applying sub-modeling approach: local stress simulation is performed considering large scale stress distribution in the chip and corresponding deformation as the boundary condition.

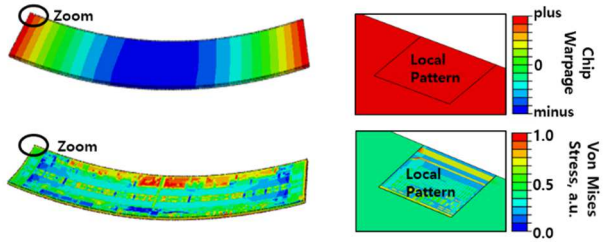

Fig. 7. Embedded local pattern added to the deformed chip shape for extraction of local effect of the warpage.

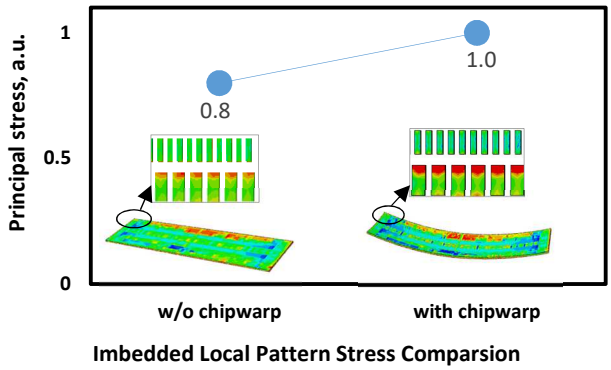

Fig. 8. Local stress contour for case when warpage is considered and not considered. Our methodology allows quantification of warpage effect on the local stress level and defect probability.

An example of the simulation of stress near the corner of the deformed chip is shown in Figs.7-8. At first the full-chip layer-by-layer stress simulation was applied and using averaging method principal stresses XX and YY were extracted and averaged. Resulting chip-scale warpage is applied as a boundary conditions for a local 3D stress simulation (Fig. 7). Existence of local stress induced by the whole chip layout can change local stress (Fig. 8), even though the extent of this effect may strongly depend on overall chip layout properties and materials used.

To quantify the probability of defect formation and connect it to the local stress value we've applied previously developed methodology that is using cumulative Weibull distribution given by:

$$F(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k},\ x \geq 0$$

where $F(x)$ is the defect formation probability, depending on stress x, λ is the characteristic critical stress and k defines how "brittle" is the material: i.e. how wide is the transition from non-defective stresses material to crack [1].

The limiting factor of the simulation is that the stress value and associated crack failure risk does not depend on the adhesive properties of the boundary material and therefore the approach

can be applied only if the same boundary metal is used. At the same time, local stress is generated by the metal line material properties and boundary metal has very small impact on overall stress level.

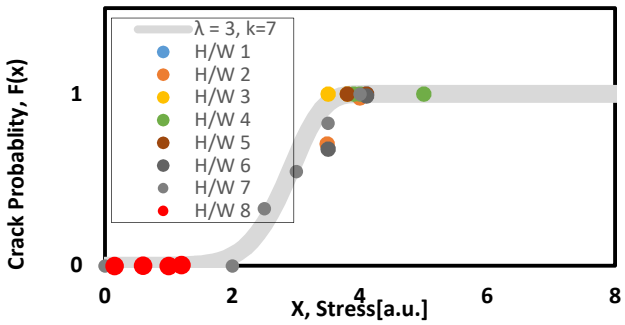

Fig. 9. Weibull Cumulative Distribution Curve. x – local stress, λ – critical stress, k – steepness of the transition region between "no crack" to "100% crack" regime. Over 1000 total cases were measured to extract probability to defect formation for each of 8 products. Only H/W with the same boundary material is summarized.

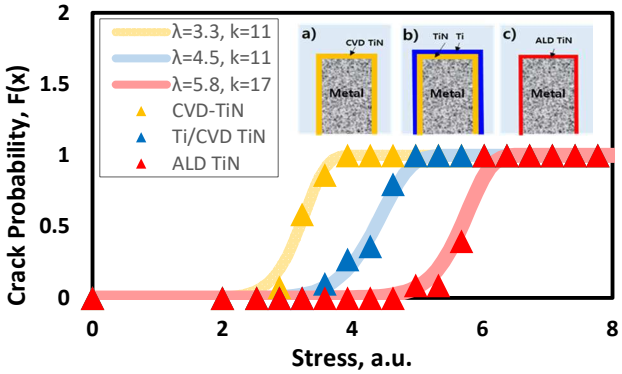

Fig. 10. Adhesion material model sketch (simulation does not consider boundary material explicitly, only as the shift of Weibull defect formation probability curve) : a) CVD TiN b) CVD TiN & Ti c) ALD TiN. Three kinds of boundary material dependent failure analysis with experimental data (totally 140 data points for 3 processes were correlated with respective cumulative Weibull distributions).

Therefore, it is possible to separate stress simulation and crack formation probability simulation and extend our approach to treat boundary metal effect. To do so, we first analyze the experimental data for the same metal patterns with different boundary metals (Fig. 10). All patterns would generate the same stress level since the layout and the volume of the metal line is the same. At the same time, depending on the boundary material and deposition process type, crack risk would be quite different. ALD TiN is known to have the best adhesive properties among tested boundary materials due to high uniformity of ALD process and higher quality of chemical bonding. At the same time, this process is the slowest and therefore the most expensive among all options. CVD-TiN deposition is much faster, but adhesive properties are quite poor, leading to defect formation even with low local stress levels. Combination of Ti/CVD-TiN is the middle ground between pure CVD and ALD processes.

At low stress levels, defects do not occur regardless of the number of stress points. As stress increases, there is a transition zone where the probability of crack formation becomes significant. At high stress levels, once a defect has formed, the risk of further local failure remains unchanged, as the device is already defective. The summary of 8 different products H/W data on crack formation probability depending on local stress is summarized in Fig. 9. Once we applied cumulative Weibull distribution analysis to failure rate measurement for all three boundary material types, we've founds a distinctive trend (Fig. 10): critical stress value is increasing (lowest value is for CVD-TiN, followed by Ti/CVD-TiN combination, and finally ALD TiN has the highest critical stress). Therefore, we can analyze BEOL layouts manufactured with different processes and define design rules depending on the boundary materials used. It allows optimization of layout and process cost at the early stages of new product design, leading to reduced number of testing lots and overall product MTO schedule improvement.

V. Conclusions

A physics-based, bottom-up methodology for assessing the risk of stress-induced defect formation in multilayer BEOL metallization structures was developed. Combination of multiple shell element simulation with incremental structure modification allows extraction of dynamically changing stress during BEOL process.

By applying local averaging to the resultant stresses, we can incorporate long-range stress-induced deformations into our analysis. This enables us to assess and consider chip and wafer-scale defects through efficient, reduced-dimension chip-scale stress calculations. Furthermore, using cumulative Weibull distribution analysis allows us to predict the probability of local crack formation based on local stress levels. Additionally, through analysis of experimental data, we can determine critical stress values that depend on the adhesive properties of boundary materials.

Finally, our methodology can be extended beyond the chip-scale level by utilizing the tiling approach initially applied in full-chip scale simulations, which can be hierarchically applied to multiple chips on a wafer. In that case stress averaging can be applied at chip scale and then the results can be used to perform a wafer-scale deformation simulation, enabling wafer-level warpage analysis based on chip-scale layout.

References

[1] K. Yeom, G. Yoo, A. Payet, A. Schmidt, H. Ahn, I. Jang & D. S. Kim, (2023, September). Full Chip Stress Model for Flash BEOL Crack Failure Risk Analysis. In 2023 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) pp. 29-32 (2023).

[2] K. Y. Sze, Three-dimensional continuum finite element models for plate/shell analysis. Progress in Structural Engineering and Materials, 44, 400-407 (2002).

[3] M. Li, J. Wu, J. He, N. Liu, T. Han, G. Zhang & T. Yu, An extended Stoney's formula including nonlinear deformation for large size wafer of multilayers with arbitrary thicknesses. Scripta Materialia, 186, 29-32 (2020) .

[4] G. G. Stoney, The Tension of Metallic Films Deposited by Electrolysis, Proceeding of the Royal Society of London, Serial A, Vol.82, pp. 172-177 (1909)