Full Chip Stress Model for Flash BEOL Crack Failure Risk Analysis

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

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

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

Hyoshin Ahn
CSE Team
Samsung Electronics
Hwaseong-si, Korea
hyoshin.ahn@samsung.com

Inkook Jang
CSE Team
Samsung Electronics
Hwaseong-si, Korea
inkook.jang@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

Satoru Yamada
Process TCAD Lab
Samsung Electronics
Yokohama, Japan
satoru.yamada@samsung.com

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

***Abstract*—Using a combination of domain decomposition, massive parallelization and dimensionality reduction, a full chip-size stress simulation flow was developed. By application of shell elements in the Finite Element Method (FEM) framework, the prediction of the stress distribution in a Flash memory die (area about 1 cm²) back end of line (BEOL) metallization layers with nanometer scale precision becomes possible within a half day. Model calibration for several Flash memory product generations allowed more than 90 percent accuracy of crack defect formation probability prediction.**

Keywords—FEM, layout, stress, full-chip, BEOL, Crack Risk, Crack Failure

I. Introduction

With advancements in the semiconductor technology and the tightening of design rules, the back end of line metallization density keeps increasing. It results in an increased rate of process failures due to a large mechanical stress accumulation induced by the thermal expansion coefficient mismatch between an insulation low-κ material and the line metal. It leads to the formation of specific delamination and cracking defects that are especially damaging for Flash memory devices generally. The accumulated stress level depends on both the chip-scale metallization density and local layout features, thus a full chip-scale stress analysis is needed to evaluate layouts for crack failure risks (Fig. 1).

The Finite Element Method (FEM) is typically used for mechanical stress analysis, but its performance is limited by the mesh element count. Since typical BEOL layout feature sizes are in order of tens of nanometers and FEM mesh element size should be small enough to resolve them, simulation of structures larger than 100 μm^2 is hardly feasible. At the same time, typical chip area is in a range of 10^8 μm^2 leading to 10^6 times domain scale gap making full-chip stress simulation seemingly impossible. Therefore, only specific local layout patterns are typically analyzed, limiting the simulation predictive power, as large scale layout effects cannot be taken into account. Also, the choice of these patterns is based on an empirical assumptions and thus simulation coverage cannot be guaranteed for a new layout.

To simplify the layout evaluation procedure for mechanical failure risks, a new simulation method that provides a full-chip BEOL layout stress prediction with nanometer scale precision is needed

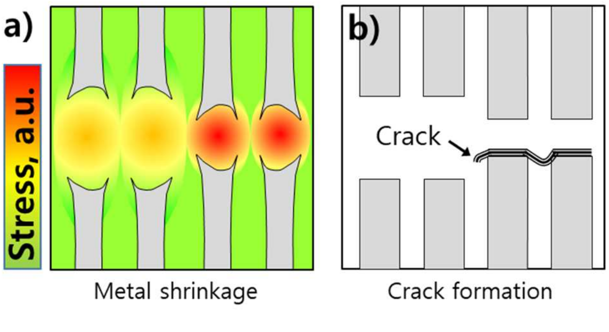

Fig. 1. (a) Sketch of metal shrinkage induced stress, (b) crack formation in the region of high stress.

II. Rapid Physics Stress Model

Since the stress in the BEOL metallization layers is mostly in the lateral directions, it is possible to reduce the problem dimensionality by replacing the 3D structure with a combination of shell elements [1,2] and an underlying bulk that represents the substrate (Fig. 2). Shell layers can have dense (but only 2D) mesh, keeping all nanometer scale features of the layout intact, while a much coarser mesh can be used for an underlying bulk structure, keeping total number of mesh elements relatively small and reducing simulation time. Typical full 3D TCAD simulation of BEOL stress using single 32 core server can cover area up to (5 µm)^2 and takes a few hours. It makes simulation of full-chip (area about 1 cm^2) completely unfeasible. Application of the shell and bulk element combination can increase an area of the chip processed by a single server to (100 µm)^2, making it possible to cover full chip area within single day (Fig. 3).

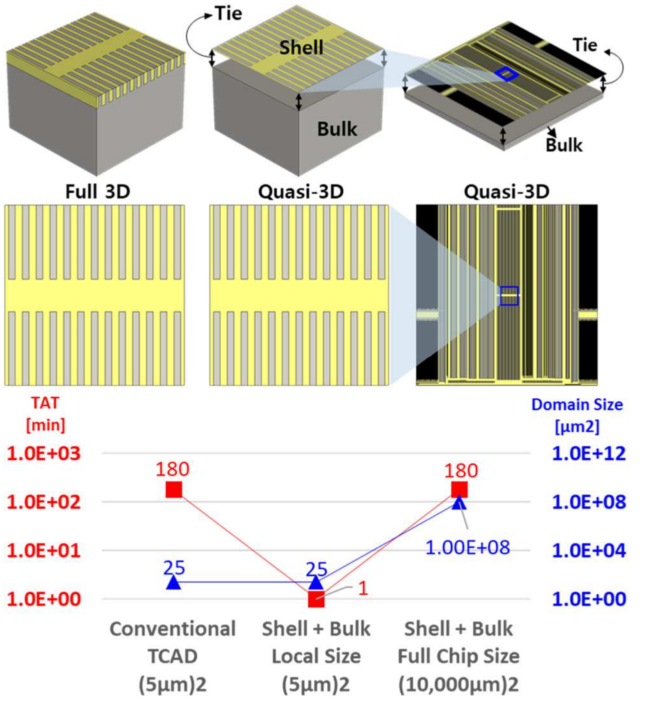

Fig. 2. The conversion of Full 3D to Shell + Bulk (Quasi-3D) element structures for large scale simulation. Simulation domain areas and typical simulation times are also compared.

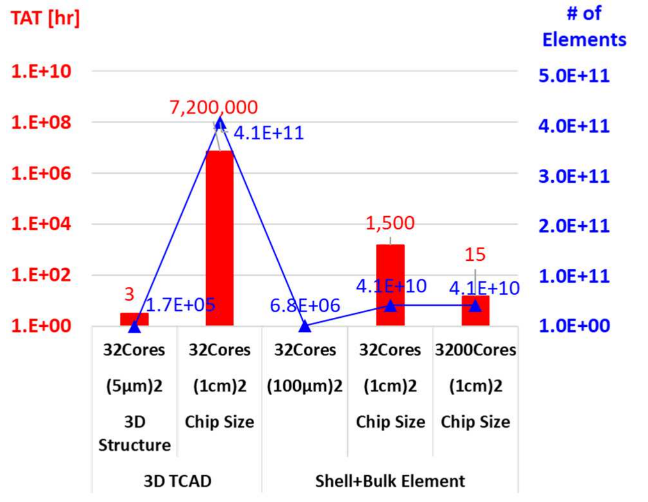

Fig. 3. The conversion of Full 3D to Shell + Bulk (Quasi-3D) element structures for large scale simulation. Simulation domain areas and typical simulation times are also compared.

An additional advantage of a large size simulation domain is that it is possible to get rid of artifacts generated by simulation cell boundaries. Since we are using fixed boundary conditions, only the stress values far enough from the cell boundaries should be taken into account. As shown in Fig. 4, for typical layouts we still can see some impact on the resulting stress for simulation cells below (30 µm)^2 and we always see some discrepancies in the results within ~10 µm from the cell boundaries.

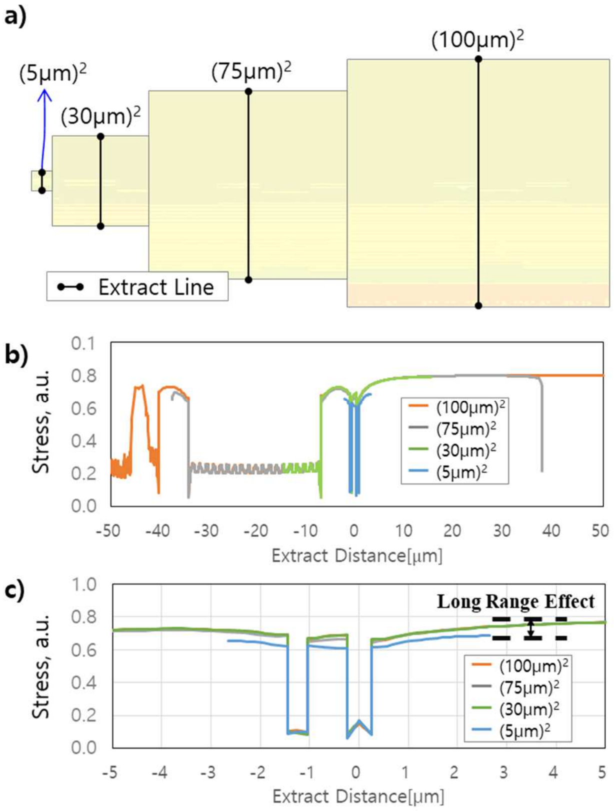

Fig. 4. (a) Simulation cells based on the same layout file and stress extraction lines (same position is used for all cases). Exact layout pattern details are not shown. (b) Extracted YY direction stress: simulation results are converging for large cell sizes, but a significant discrepancy within 10 µm from the cell boundaries is typically observed. (c) Close up image of long range stress effect for small tile size.

Therefore, if the simulation domain that can be covered by single server is large enough, it is possible to split the whole chip in large overlapping tiles and simulate them independently considering only central parts of each tile and discarding peripheral regions that are distorted due to potentially inconsistent boundary conditions. In this manner an efficient massive parallel execution of the simulation is possible since we do not have to share any data between the simulation instances thus removing all parallelization overhead (Fig. 5).

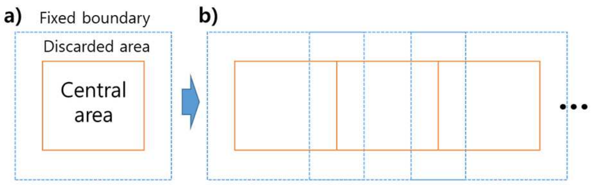

Fig. 5. (a) Single tile with fixed boundary and discarded peripheral area, (b) overlapping tiles covering full chip with their central areas.

A special layout data pre-processing tool has been developed for the preparation of corresponding shell elements and bulk mesh structures directly from a full chip BEOL layout OASIS format file. At first for a specific BEOL layer it performs a full chip layout tiling and cleaning, removing all redundant points from the layout polygons to ensure high quality boundary representation of metal lines. Afterwards, a shell element meshing is performed for each tile using Triangle engine [3] and uniform bulk element mesh is applied. Next, hybrid shell and bulk element stress simulations are performed with a highly efficient in-house stress simulation engine. Finally, a post-processing algorithm has been developed to extract the stress values at specific points of interest and perform analysis of potential layout weak spots. Thanks to the fact that after the completion of a relatively short tiling procedure all further steps are completely independent, special load management tool was implemented using DRMAA [4], that directly interacted with our computing farm scheduler software and ensured almost perfect workload balancing. Thereby, a massively parallel execution with about 3200 CPU cores allowed a full chip layout processing and data extraction within 15 hours.

For the BEOL delamination crack issue, the failure probability is highly correlated with the stress at the end of Metal Lines (ML) as shown in Fig. 6a. Therefore, an automated layout analysis and ML edge point stress extraction algorithm was added at the post-processing step. The results of reduced dimension shell element simulations are not exactly coinciding with full 3D FEM simulation, but are highly correlated (R^2=0.95, Fig. 6b) and therefore can be matched exactly if the mechanical properties of BEOL materials and underlying bulk substrate used in quasi-3D simulation are calibrated.

This approach allows for an accurate and efficient physics-based chip-level stress simulation, which we refer to as a "Rapid Physics Stress Model" (RPSM).

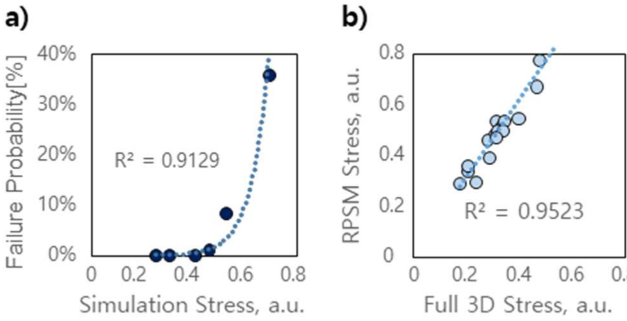

Fig. 6. Correlation of H/W failure rate with simulated stress (a) and RPSM correlation with 3D FEM simulation (b).

III. Stress Failure Risk Analysis Flow And Scoring

Since RPSM results can be obtained within a day, it is possible to introduce BEOL crack risk analysis in a Flash memory die layout sign-off flow to provide a stress-induced failure risk assessment after each layout revision step. To provide quantitative comparison between various layout versions and product generations, a crack risk scoring methodology is needed.

First of all, due to enormous size of the data generated by the stress simulation, one has to choose appropriate points of interest. Since all cracks and delamination defects observed in experiments start near the interface between metal and low-κ dielectric, stress data extraction points are chosen based on the initial layout near the corners of metal masks. The number of resulting points is large, so the next step of the filtering process is to arrange results according to the stress values. Example of resulting stress histogram is shown in Fig. 7a. Since individual crack formation is generally a random process that depends on local variations of the lithography and metal deposition processes, the cumulative Weibull distribution was chosen as a basis of a failure risk scoring methodology. Risk score is calculated as an integral of the total stress histogram (number of points at ends metal lines having particular stress value) with the cumulative Weibull distribution function (1),

$$\mathrm{F}(\sigma,\lambda,k)=1-e^{(\sigma/\lambda)^{k}} \qquad (1)$$

where σ is stress value, λ is a critical stress for crack formation and k is a parameter that reflects probabilistic nature of delamination crack formation (Fig. 7b). Critical stress for crack formation may depend on specific material properties (for instance, barrier metal and low-κ dielectric compositions) and process conditions (e.g. BEOL metal deposition temperature and ramp-down rate, barrier metal thickness and roughness, etc). Critical stress values and other distribution parameters may be predicted from a crack formation modeling based on some fracture mechanics theory (see, for instance [5]), but in practice it is often more efficient to extract these values from the experimental crack failure rate data extracted for a fixed process conditions and special test layout.

The resulting RPSM crack risk score can be used to compare layout revisions (Fig. 7c). Moreover, by using Weibull distribution parameters calibrated for various BEOL process options, it is possible to evaluate whether a given layout can be manufactured within the target process cost with an acceptable low crack risk or if revision is mandatory.

$$RPSM\ score=\int Weibull(\sigma)\cdot Histogram(\sigma)\,d\sigma$$

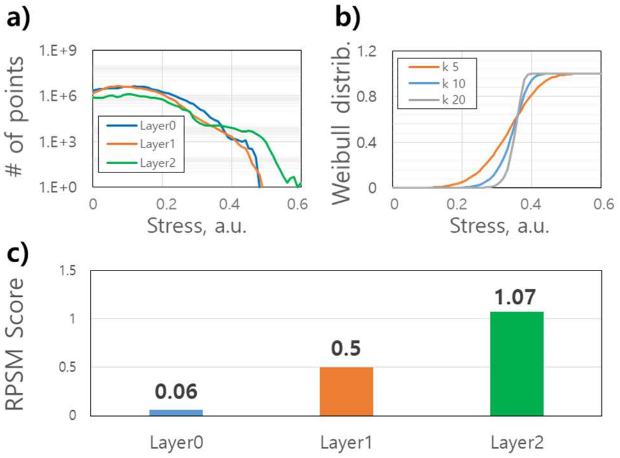

Fig. 7. RPSM stress failure score definition: (a) Stress histogram example, (b) Weibull distribution sensitivity for a k parameter value and c) comparison of the crack scoring for 3 BEOL layers of a Flash device. Layer 2 has the highest risk and so its design rule has to be revised for a yield enhancement.

Along with chip-scale risk analysis, it is often important to provide information about specific local layout patterns that lead to cracks during the manufacturing process. This information enables layout designers to improve design rules and avoid the generation of weak spots. To achieve this, we have developed an additional level of stress data filtering. For points with high local stress values, we extract local layouts (squares of 3x3μm centered on the high-stress points) from the original OASIS file. Subsequently, we categorize the local patterns based on their similarity in local layout. Since it is often the case that Flash BEOL layout has a huge number of similar elements, the total number of unique layout patterns with high stress is 2-3 orders of magnitude smaller than the total number of potential crack risk patterns that are extracted (Fig. 8). Due to the long-range effects of layout stress, it is quite common for identical local layout patterns to exhibit significantly different stress values depending on their position in the chip and its surrounding layout density. Therefore, when analyzing a specific weak pattern, it is important to consider not only the total number of occurrences on the die but also its stress distribution, average stress value, and the stress variation.

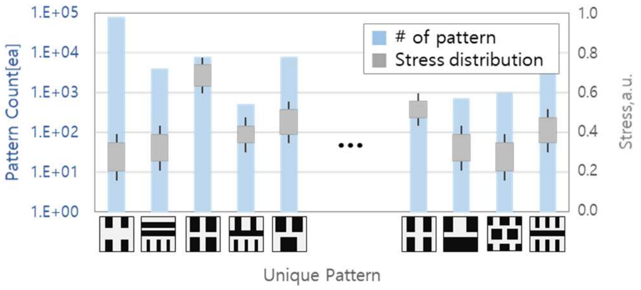

Fig. 8. Unique high stress pattern categories and corresponding total counts and stress distributions.

In some specific cases the patterns that generate large local stress in the simulation may not cause the formation of the crack due to non-ideal correspondence of the layout form the OASIS file with the real shapes generated during lithography. It may lead to overestimation of simulated stress, due to accumulation near the metal line corners that in reality cannot be manufactured exactly rectangular (Fig. 9.). Additional filtering step is added to ignore detected high stress points based on local metal like aspect ratios.

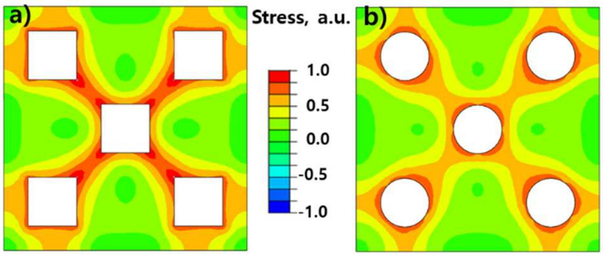

Fig. 9. Example of lithography-induced real device structure discrepancy with simulation assumption that may lead to false alarm of crack formation risk detection. (a) Traget lithgraphy shape defined in layout file and the stress calculated for that shape, (b) real shape that is observed after lithography step. Stress near corners is significantly lower in case (b).

IV. Conclusion

A multi-scale Rapid Physics Stress Model was developed for a full-chip BEOL layout stress calculation. The implementation of a hybrid shell and bulk element method, as well as highly efficient layout parsing and processing algorithms, coupled with a high performance in-house stress simulation tool, allowed a full chip stress distribution extraction within 15 hours. The accuracy of in-plane stress simulation was maintained at a level >95% when compared to local stress simulation results. Furthermore, large size of the simulation cell that can be used in RPSM mitigates boundary-related numerical artifacts.

The flowchart of the RPSM methodology for full-chip layout analysis is shown in Fig. 10. When combined with the crack risk scoring method based on the Weibull distribution, it allows for a quick comparison of the average defect risk across different layers within Flash BEOL and across various product generations. The methodology was applied to assess the advanced Flash memory product BEOL layout for predicting the risks of crack formation (correlation with crack-induced process failure H/W data is >90%). The methodology not only provides an averaged chip crack risk estimation but also offers stress statistics data on individual layout patterns that generate significant local stress. This allows for the improvement of design rules to reduce the crack-induced failure risk.

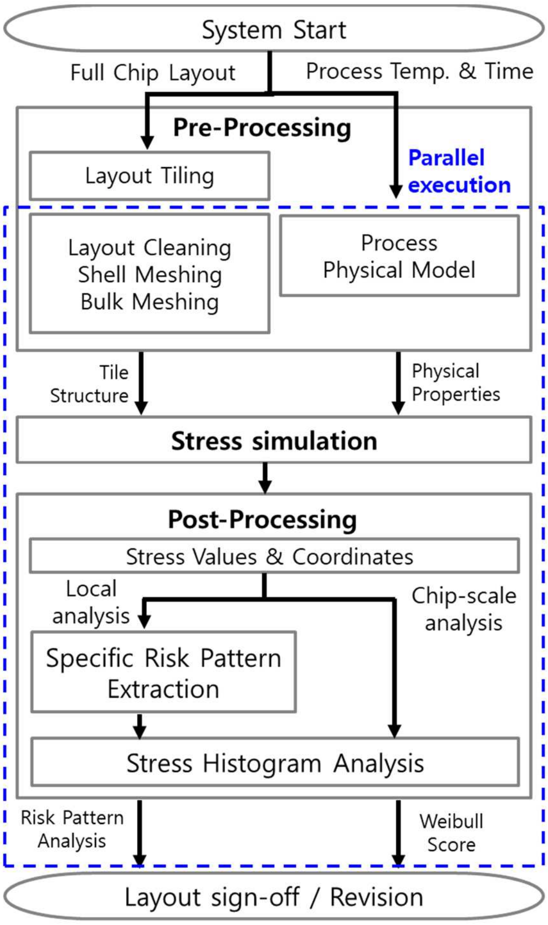

Fig. 10. Full flow of RPSM-based layout stress risk assessment and sign-off process. The major part of the flow is executed in parallel allowing for a rapid processing after each major layout revision.

References

[1] T.J. Hughes, The finite element method: linear static and dynamic finite element analysis. Courier Corporation, 2012.

[2] K.Y. Sze, "Three-dimensional continuum finite element models for plate/shell analysis," Progress in Structural Engineering and Materials, vol. 4(4), pp. 400-407, October 2002.

[3] J.R. Shewchuk, "Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator," In Workshop on applied computational geometry. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996 pp. 203-222.

[4] P. Troger, H. Rajic, A. Haas, and P. Domagalski, "Standardization of an API for distributed resource management systems," in Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid'07), pp. 619-626, May 2007.

[5] J.W. Hutchinson, and Z. Suo, "Mixed mode cracking in layered materials," Advances in applied mechanics, vol. 29, 1991, pp.63-191.