From CMP Surface Prediction to Defect Detection: An AI-Driven Virtual Metrology–TCAD Framework

Yeji Kim
Semiconductor R&D Center
Samsung Electronics
Hwaseong-si, Republic of Korea
yej2.kim@samsung.com
(ORCID: 0000-0003-4658-645X)

Min-Chul Park
Semiconductor R&D Center
Samsung Electronics
Hwaseong-si, Republic of Korea
m.c.park@samsung.com

Sangyeon Kim
Semiconductor R&D Center
Samsung Electronics
Hwaseong-si, Republic of Korea
syeon83.kim@samsung.com

Usuk Chae
Foundry Business
Samsung Electronics
Hwaseong-si, Republic of Korea
usuk.chae@samsung.com

Byungchul Shin
Foundry Business
Samsung Electronics
Hwaseong-si, Republic of Korea
bchul.shin@samsung.com

Segab Kwon
Semiconductor R&D Center
Samsung Electronics
Hwaseong-si, Republic of Korea
sg1987.kwon@samsung.com

SeongRyeol Kim
Semiconductor R&D Center
Samsung Electronics
Hwaseong-si, Republic of Korea
sr75.kim@samsung.com

Yoon-Suk Kim
Semiconductor R&D Center
Samsung Electronics
Hwaseong-si, Republic of Korea
ys1108.kim@samsung.com

Jae-Hyun Kang
Foundry Business
Samsung Electronics
Hwaseong-si, Republic of Korea
jh0717.kang@samsung.com

Young-Gu Kim
Foundry Business
Samsung Electronics
Hwaseong-si, Republic of Korea
yg09.kim@samsung.com

Joong-Won Jeon
Foundry Business
Samsung Electronics
Hwaseong-si, Republic of Korea
joongwon.jeon@samsung.com

Dae Sin Kim
Semiconductor R&D Center
Samsung Electronics
Hwaseong-si, Republic of Korea
daesin.kim@samsung.com

***Abstract*—Layout-dependent physico-chemical interactions in chemical mechanical polishing (CMP) processes often result in surface non-uniformities, which can lead to defects in subsequent manufacturing steps. We present an integrated approach that unifies large-scale data-driven virtual metrology with 3D TCAD simulations, leveraging a neural operator specialized in PDE-based modeling for high-resolution and high-speed predictions. As a result, the method enables scalable analysis of full-chip layouts containing around over two billion points within a four-hour window, thereby allowing early-stage detection of potential failure risks at the design level.**

Keywords—CMP-induced defect, Neural Operator, VM-TCAD

I. Introduction

As the stack heights in semiconductor processes decrease and process margins shrink, the surface non-uniformities introduced by chemical mechanical polishing (CMP) processes accumulate and lead to either an increase in the number of defects or an increase in the severity of defects in subsequent steps (Fig.1). CMP-induced defects induce malfunctions of entire chip, so that cause severe reduction in semiconductor production yield. [1,2]

CMP processes can be simulated by Complex partial differential equation (PDE)-based models and derive an estimate of wafer surface after CMP. However, layout-dependent factors and real-world variables which cannot be accounted for in the PDE limit the accuracy of such approaches. [3-5] To address this challenge, data-driven virtual metrology using large-scale measurement data has emerged, and neural operator (NO)-based architectures are have garnered attention for their capability to capture comprehensive physical phenomena. However, the prior research that combines these two approaches remains limitations, requiring calibration and too high computational costs to mitigate CMP-induced defects. [6,7]

This paper proposes an AI-driven metrology-TCAD framework for large-area CMP-induced defect analysis. The framework includes surface-morphology predictor from

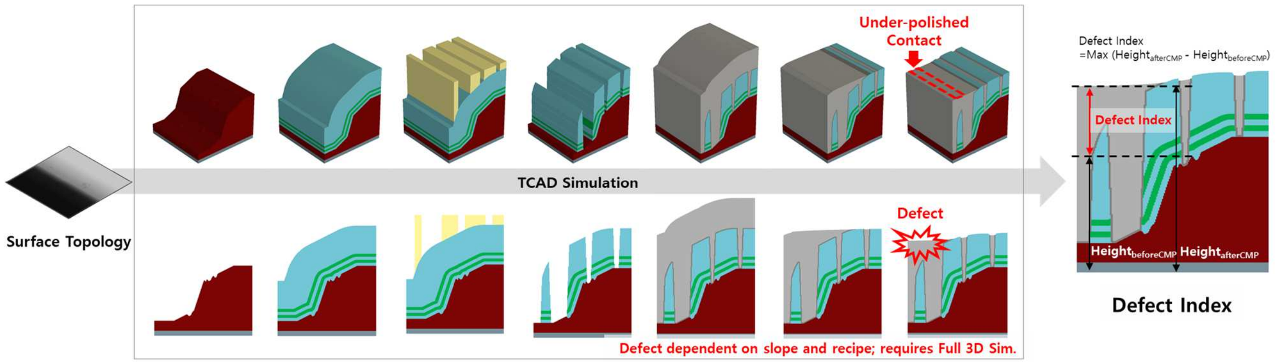

Fig. 1. Flow of full 3D Simulation for CMP process and definition of CMP-induced defect index.

layout inputs via virtual metrology and provides a high-speed TCAD surrogate model, applying several light-weighting computation technologies and data division skills.

II. Methods

This paper proposes an AI-driven metrology-TCAD framework for large-area CMP-induced defect analysis. The framework includes surface-morphology We proposes an integrated framework that combines Virtual Metrology (VM) with a real-time TCAD model to tackle CMP-induced surface non-uniformities and subsequent defect prediction, as illustrated in Fig. 2.

First, we adopt a DeepONet [8] as our VM model, featuring a NO architecture—shown in Fig. 3(a)—designed to capture the underlying PDE dynamics for surface topology during CMP processes. A key element of this predictive capability is the Response Corrector module, which learns to compensate for noise levels unique to each field-of-view (FoV). The module plays an essential role in the process of integrating and utilizing multi-FoV data, as different measurement data have different reference points or offsets due to measurement methods, measurement equipment, and other noise-related factors. By estimating offset-changes in the output for each FoV index based on training data, this module substantially enhances the model's overall accuracy. Furthermore, because the model accepts output coordinates as inputs, it supports high-resolution smooth surface predictions at arbitrary points while maintaining high accuracy as shown in Fig. 3(b).

Second, we employ the 3D TCAD simulation that takes the predicted surface morphology and the subsequent process mask layout as inputs to model the following process steps and detect potential defects. From these simulations, we derive a defect index to quantify the severity of CMP-induced failures. Figure 1 illustrates this simulation steps, showing how both surface topology and subsequent layout patterns are resulted to the final structure and defined defect index. To mitigate the high computational cost associated with traditional PDE-based TCAD solvers, we leverage a NO model that accelerates the simulation while preserving accuracy. The NO model trains TCAD simulation outputs, pairs of defect index and layout images which matched to the defect index value, thus surrogates the 3D TCAD simulation. Prior to the execution of our framework, it is necessary to execute the TCAD simulation for CMP processes and obtain pairs of layout images and CMP-induced defect.

Prediction-acceleration techniques were applied to defect prediction phase for high-speed large-area CMP defect analysis. There are two main bottlenecks to shorten simulation TAT: pre-processing for large layout and large data prediction by AI models. Layout rasterization with high-resolution, matched to output target resolution, is considered as a primary pre-processing step for layout. We applied segmented rasterization and prediction from rasterized large image so that finish full-chip CMP defect analysis in several days.

To address potential limitations of purely data-driven approaches, we further perform root-cause analysis, by Integrated Gradients (IG) analysis on layout, one method of

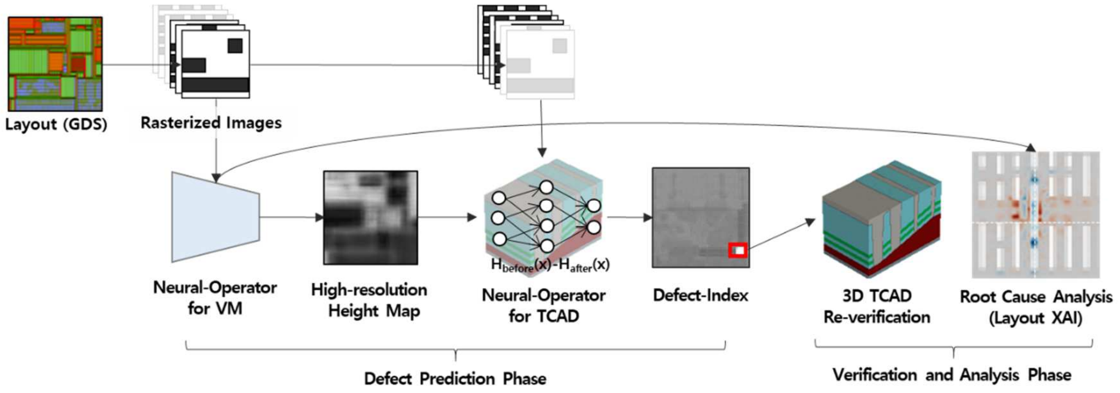

Fig. 2. Two-step AI-Simulation hybrid framework: (1) defect prediction phase, with virtual-metrology (VM) model and real-time process TCAD (RTT) surrogate model, and (2) verification and analysis phase.

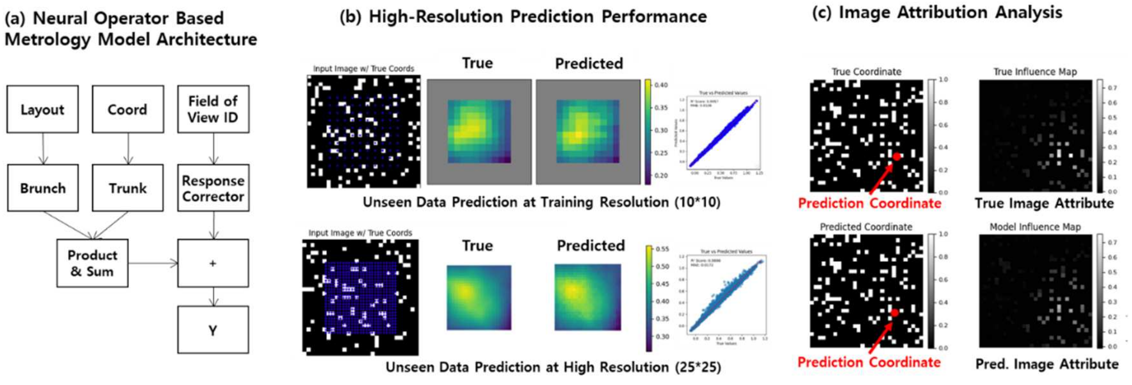

Fig. 3. Method and results for virtual metrology model: (a) the architecture of virtual metrology model with input of layout image, coordination, and identification number (ID) for field of view, (b) high-resolution predictions validated against a synthetic layout, showing minimal performance degradation from training, and (c) its image attribution by integrated gradients analysis.

eXplainable AI (XAI)—depicted in Fig. 3(c)—which reveals how layout pattern combinations and surrounding features influence the predicted surface morphology.[9,10] The analysis offers a substantial information for defect-risky points, even experienced engineers or traditional analyses limit broad and complex layout pattern combination issues.

III. Results

The proposed framework was implemented to 3nm logic full-chip analysis. We first acquired layout data encompassing both active and dummy patterns across all layers pertinent to the CMP process. Layout pattern data was furnished as a three-dimensional image array with dimension corresponding to (number of horizontal pixels, number of vertical pixels, number of used layers), extracted from a specified window-size of the layout. All framework parameters, including window-size, image-size, layer classification, and layer operation, were optimized according to the chip type. We established the optimal window-size at 10um based on L2-regularization-based feature selection utilizing numeric features such as pattern density. Additionally, the image-size was determined to be 64 pixels through grid search optimization.

Our virtual metrology model, trained on millions of layout patterns and corresponding optical height measurements, attained a test accuracy with a coefficient of determination (R^2) score of 0.76. The 3nm logic full-chip with a 1.2 cm^2 area, yielded 2.5 billion prediction points with discretization at 2 nm intervals. By leveraging 400 CPU cores in parallel, we successfully completed full-chip defect detection in 4 hours. To accelerate prediction, we employed gradual-segmentation techniques with per-side-2mm- and per-side-400um-size tiles on the full-chip layout.

Prior to executing the model, it is imperative to validate our TCAD simulation for the CMP process and defect-index that have been defined followed by layer thickness. A total of 200 sample points, including 100 CMP defects and 100 defect-free samples, were obtained and simulated. These samples were acquired through bright-field (BF) microscopy imaging and optical height measurement. The veracity of the sample points is challenging to ascertain, necessitating the collection of diverse, random true site for validation. The simulated defect index values of the sample points are represented by red and blue dashed lines in Fig.4(a). It was observed that the defect and defect-free samples exhibited distinct defect-index distributions with respect to 1. Therefore, it was concluded that the utilization of a simulated defect-index facilitates defect judgement.

The TCAD-surrogate model demonstrated a test R^2 score of 0.75, trained on a thousand of pairs of layout-pattern and corresponding simulated defect-index. The predicted defect-index values of the sample points are displayed in Fig.4(a) using a color-filled bar graphs with solid line. It was determined that the distributions of the simulated and predicted values exhibit a significant overlap in both under-1 and over-1 ranges. Critically, the models generalize to other chips sharing identical design rules and process technologies when trained with defect and defect-free samples.

The 1.2 cm^2-size chip is detected for 200nm-resolution CMP defect risk with 400 parallel CPU cores in 4 hours, consistent with the virtual metrology model. In contrast, a conventional numerical PDE solver require a year for the same resolution analysis, thereby demonstrating a 2,283-times speedup. (Fig.4(b)) Each 100 um^2-size single simulation requires a range of 8 to 20 hours with 1 CPU core, although each defect index of a single point is finished within several seconds. The TCAD-surrogate model facilitates real-time prediction of CMP defect-index even for over 1.2 cm^2 large-size chips. Predicted height and defect-index map of the full-chip map shown in Fig.5(a). CMP-defect-risky points, of which defect-index is over 1, are matched to the points with sharply-slopped height change and it is convincing with layout patterns and surrounding topology. (Fig.5(b))

3D TCAD re-verification was performed on flagged defect

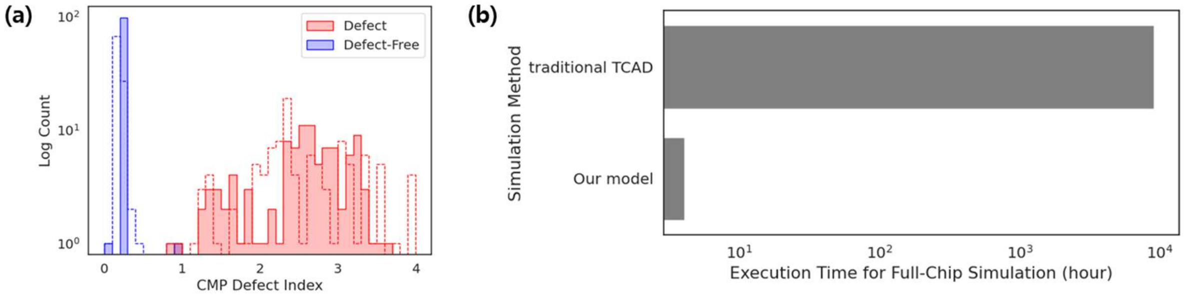

Fig. 4. Three result of each step in the framework: (a) plot of simulated (colorless bar with dashed line) and predicted (color-filled bar with solid line) defect-index for defect-verified samples, (b) plot of execution time for 3nm logic full-chip CMP simulation.

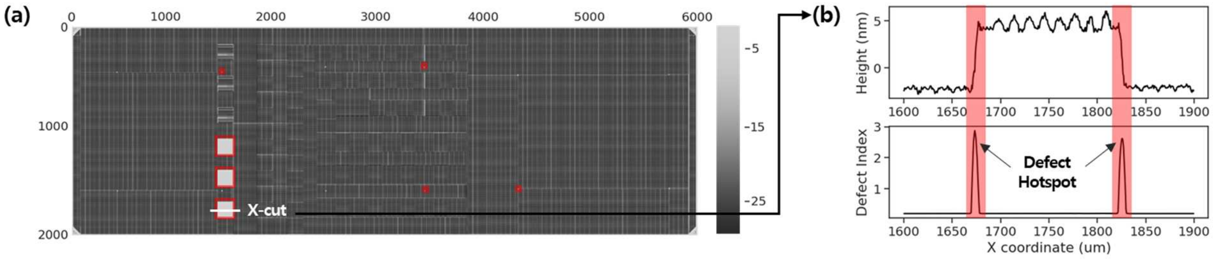

Fig. 5. Final output of the framework: (a) predicted height (gray) and defect-index (red) map for 3nm logic full-chip, and (b) predicted height (top) and defect index (bottom) graph of defect hotspot area from the map.

hotspots to both validate results and provide process engineers with defect-structure visualizations. It complements the proposed methodology that does not guarantee perfect accuracy despite of the substantial acceleration of TCAD simulation. The VM model and the surrogate model integrate an image attribution module, thereby enabling root-cause layout-pattern analysis of defects.

IV. Conclusion

We present an integrated powerful framework that unifies VM model with a real-time TCAD-surrogate model to tackle CMP-induced defects. For PDE-based CMP modeling, we adopt an NO architecture for the VM model and TCAD simulation, demonstrating robust generalization and high-resolution capabilities, further enhanced explainability through XAI. By combining the predicted surface topology with the accelerated TCAD surrogate flow, we achieve over 2,000× speedup compared to conventional PDE solvers, so that complete near real-time defect analysis. Finally, critical locations are re-verified via detailed TCAD, and image attribution provides layout-level DR insights to reduce CMP defects. The framework is valid for any other chip layout designed by identical rules and process technologies with trained layout. Future research would improve model accuracies and generalization, and expand our framework usage to other layers or processes.

References

[1] B. Suryadevara, ed. Advances in chemical mechanical planarization (CMP). Woodhead Publishing, 2016.

[2] L. Zhang, S. Raghavan, and M. Weling, “Minimization of chemical-mechanical planarization (CMP) defects and post-CMP cleaning,” Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena 17.5 (1999): 2248-2255.

[3] W.-T. Tseng, “Approaches to defect characterization, mitigation and reduction,” Advances in Chemical Mechanical Planarization (CMP). Woodhead Publishing, 2022. 467-503.

[4] H.-M. Yu, C.-C. Lin, M.-H. Hsu, Y.-T. Chen, K.-W. Chen, T. Luoh, “CMP process optimization engineering by machine learning,” IEEE Transactions on Semiconductor Manufacturing 34.3, 2021, pp.280-285.

[5] H. Bao, L. Chen. “A CNN-based CMP planarization model considering LDE effect,” IEEE Transactions on Components, Packaging and Manufacturing Technology 10.4, 2020, pp.723-729.

[6] M. C. Park, et al. “Deep pattern solution for interpreting systematic defects,” DTCO and Computational Patterning IV. Vol. 13425. SPIE, 2025.

[7] M. C. Park, et al. “Realistic and Scalable TCAD for Yield-Aware Full-Chip DTCO,” 2025 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2025

[8] L Lu, P. Jin, G. E. Karniadikis. “Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators,” Nature machine intelligence 3.3, 2021, pp.218-229.

[9] M. Sundararajan, T. Ankur, Y. Qiqi, “Axiomatic attribution for deep networks,” International conference on machine learning. PMLR, 2017.

[10] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G. Z. Yang, “XAI—Explainable artificial intelligence,” Science robotics 4.37 (2019): eaay7120.