Optimized Simulation Methodology of Warpage and Localized Stress Hotspot Prediction for Assembly Risk Assessment

Zhi Yang*
Groq Inc
Mountainview, CA, USA
zyang@groq.com

Krishna Mellachervu
Ansys Inc
San Jose, CA, USA
krishna.mellachervu@ansys.com

Igor Arsovski
Groq Inc
Mountainview, CA, USA
iarsovski@groq.com

Clint Harames
Groq Inc
Mountainview, CA, USA
charames@groq.com

Jim Miller
Groq Inc
Mountainview, CA, USA
jmiller@groq.com

***Abstract*— Generative Artificial Intelligence workloads, like Large Language Models, are growing in computational demand by 1000% every year, while Moore's Law scaling is only supplying 3% more transistors/mm^2 every year. To close the gap between these wildly diverging demand and supply exponentials, the industry not only needs better chip-to-chip interconnects, but also ways to integrate more silicon into a single package. This paper we will focus on advanced packaging modeling of the Groq Language Processing Unit (LPU^{TM}) inference engine, the highest performance Large Language Model Inference Engine to date. More specifically the paper will focus on the accurate warpage prediction, which has emerged as a pivotal challenge with profound implications for design reliability and manufacturability.**

Accurate warpage/stress modeling techniques are essential to identify and visualize localized thermomechanical stress caused by coefficient of thermal expansion (CTE) mismatch between dissimilar materials within the board. As a result, failure prone and/ or high-risk regions are promptly revealed at an early design stage and mitigated through design optimization. However, the computational costs needed for such high-fidelity 3D simulation methodology are extremely expensive, time-consuming and almost impractical in real life. To resolve such accuracy-computational cost dilemma, this study investigates different modeling techniques to recommend an optimal balance between efficient simulation and accuracy.

As the demand for higher performance of electronic devices with but lower power consumption requirement of electronic devices intensifies, smaller features, including finer line / space of copper traces, and higher aspect ratio vias between metal layers are becoming mainstream trends of today's board designs. However, this comes with even greater challenges of identifying the risk level for potential global warpage and localized fine features, such as traces and vias. Therefore, numerical finite element analysis (FEA) simulation plays an ever-increasing crucial role in the advanced packaging field, aiding in the reliability assessment and substrate / board risk mitigation during the assembly process.

Simplest lumped method for warpage modeling is based on the rule-of-mixture theory that showed outlier warpage prediction against experimental measurement data. Recognizing the limitations of lumped modeling, industry has now started to adopt trace mapping techniques, which considers the in-plane metal/ dielectric volume fractions and non-uniform distributions for warpage prediction. Due to the intrinsic nature of layer smeared effective properties in trace mapping approaches, no localized stress especially on traces/ vias can be extracted and visualized.

In this work, we introduce an innovative modeling technique called hybrid reinforcement methodology, where localized trace/via are modeled discretely as beam/shell elements embedded within a base material for warpage and stress prediction at both global and local scale. This hybrid proposed reinforcement modeling methodology demonstrated great alignment of absolute location for maximum warpage prediction with measurement data error within 4 mm accuracy. At the same time, it provides sufficient detailed stress information around traces and vias. This work deploys the test data validated innovative modeling methodology, which proactively assesses localized high risk regions during surface mount technology (SMT) process for package integration and identifies potential failure sites. Because of high modeling accuracy, this methodology has been applied to Groq next generation system pre-emptive derisk and optimization to further improve overall performance with lower testing cost.

***Keywords*— Warpage, localized via stress modeling, advanced packaging, finite element analysis, integrated circuits.**

I. INTRODUCTION

As the demand for higher performance but lower power consumption requirement of electronic devices intensifies, smaller features including finer line / space of copper traces, higher aspect ratio via between metal layers are becoming mainstream trends of today's board design. However, this comes with even greater challenges identifying the risk level for global warpage and localized fine features, such as traces and vias. Therefore, numerical FEA simulation plays an ever-increasing crucial role in the advanced packaging field, aiding in the reliability assessment and substrate / board risk mitigation during the assembly process.

Warpage modeling addresses the challenges posed by the thermomechanical stresses arising from various sources, such as coefficient of thermal expansion (CTE) mismatches, the curing shrinkage of underfill materials, and temperature fluctuations during manufacturing and device operation. Accurate warpage predictions expose issues like solder joint failures, delamination, and electrical interconnect reliability problems, which may lead to catastrophic device malfunction and premature failure. Advanced packaging techniques, especially for 2.5D and 3D stacking, introduce additional complexities to warpage prediction due to the diverse materials, non-uniform heat dissipation, and intricate interconnect architectures. Hence, high fidelity warpage modeling becomes imperative not only for ensuring reliable device performance but also for reducing the design cycle time and minimizing costly design iterations.

However, achieving high accuracy in warpage modeling often comes at the expense of extensive computational resources and time. Finite Element Analysis (FEA) simulations, which offer detailed insight into the thermomechanical behavior of packages, require significant computational effort for large and complex structures.

The mainstream warpage prediction methodology is largely based on homogenized material properties obtained from rule-of-mixture theory using material volume fractions. It's mathematically straightforward and cost-effective to predict global level thermomechanical behavior for simple board designs. Thus, disregarding the heterogeneous nature induced by detailed non-uniformly located traces and other complex features present within the design. This limitation can lead to inaccuracies in warpage predictions, compromising product reliability.

Another approach with improved accuracy for warpage prediction is trace mapping. This approach discretizes board in-plane properties based on black and white design bitmap images to account for localized metal/ dielectric materials fraction. With sufficient discretization mesh, it generates very accurate global and local warpage behavior. One drawback is that it overlooks the out of plane layer to layer interactions, for example via impact on warpage constraints and localized stress.

Therefore, a more comprehensive warpage prediction methodology to fill that missing piece of puzzle is proposed as reinforcement modeling methodology. Reinforcement methodology explicitly models the vertical interconnections between layers such as buried via, micro-via, plated through hole (PTH), via in pad (VIP) structures as structural elements. As a result, via stress localization can also be sufficiently captured and visualized along with accurate global and local warpage behavior. Obviously higher fidelity modeling leads to very extensive computational resource consumption.

This study explores the optimal modeling methodology balancing accuracy and computational time to accelerate warpage predictions while maintaining a desirable precision level. Along with accurate warpage prediction, such optimal modeling methodology also reveals localized stress level for fine pitch trace/ via to obtain comprehensive understanding of risk evaluation.

II. Groq LPU^{TM} System PCIe Board

Figure 1 shows the bare fab top and bottom picture of Groq LPU^{TM} system PCIe board, shiny yellow regions indicate copper pad on the board, whereas dark gray areas represent dielectric materials. In total, this board is composed of 16 conductor layers, which are responsible for system power delivery, signal propagation and ground purposes. In Table 1, we list a few technical details regarding the Groq LPU^{TM} PCIe board, such as physical dimensions of the entire board and package shadow area and specific conductor metal density. As expected, the designer fully considers upper- and lower-layers metal balancing, which is key for warpage control, current carrying capacity optimization, yield improvement and so on.

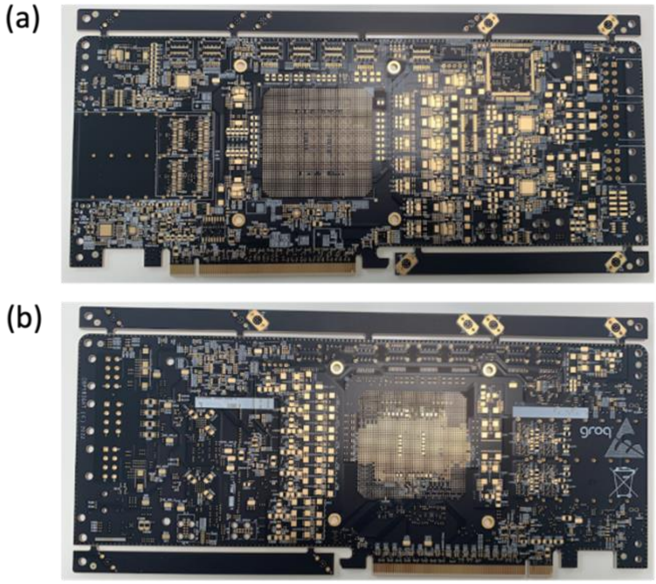


Fig 1: Bare Fab Picture (a) for Top View (b) for Bottom View.

Table 1. Groq PCIe Board Design High Level Summaries

Groq LPU PCIe Card	**Details**
Number of Metal Layers	16
Body Size	267 x 111.15 mm
Package Shadow Area Size	72.5 x 72.5 mm
Package Size	52.5 x 52.5 mm
M1 Cu%	52.5
M2 Cu%	90.8
M15 Cu%	91

Groq LPU PCIe Card	Details
M16 Cu%	42

III. Warpage Simulation Modeling Methodology

Rule-of-mixture Methodology (In-plane focus)

To effectively represent the multi-layer stackup in a board design, it is important to derive the effective properties at each layer based on volume fraction of copper and dielectric material. A common approach is to apply the rule-of-mixture (ROM) theory to effectively calculate Young's Modulus E, CTE and Poisson's ratio. The ROM is formulated through the assumption of uniform strain within the in-plane orientation. Based on force balancing on the entire board, the resultant expressions for the effective E, CTE, and Poisson's ratio can be obtained from the derived mathematical equations. The graphical representation of the laminate's structural stackup is illustrated as shown in Figure2.

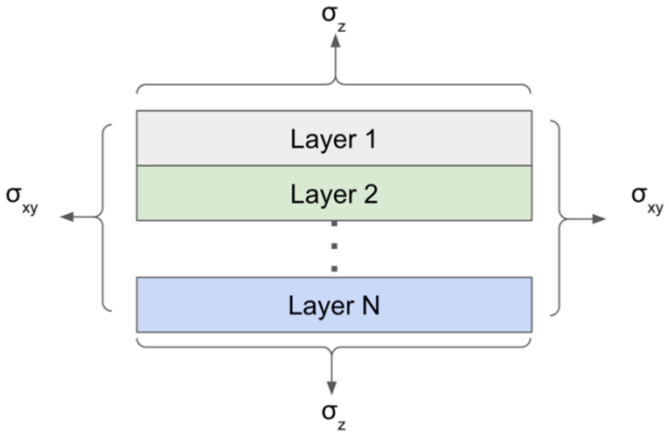


Fig 2: Laminated Structure Stackup Schematic

Since out of plane warpage is caused by in plane direction CTE mismatch between adjacent bonded materials, here in plane orientation effective properties are derived. Applying force balance on laminate structure:

In plane orientation:

$$\sigma_{eff} * A = \sum_{i=1}^{N} \sigma_i \ * A_i \quad (1)$$

Assuming entire laminate experience same strain:

$$E_{eff} * \varepsilon * A = \sum_{i=1}^{N} E_i \ * \varepsilon * A_i \quad (2)$$

$$E_{eff} = \sum_{i=1}^{N} E_i \ * C_i \quad (3)$$

Where C_i is volume fraction of ith material

Similarly derive for in plane effective Poisson's ratio (ν_{eff}) and CTE (α_{eff})

$$\nu_{eff} = \sum_{i=1}^{N} \nu_i \ * C_i \quad (4)$$

$$\alpha_{eff} = \sum_{i=1}^{N} \frac{E_i * C_i * \alpha_i}{E_i \ * C_i} \quad (5)$$

Once the effective material properties are extracted, the interactions between stackup layers are assumed as bonded, non-slip conditions. Under different temperature conditions, each effective layer expands/ shrinks at different rates based on their properties. Therefore, warpage prediction can be performed.

Trace Mapping Methodology

Trace mapping technology represents a transformative shift from the ROM approach, addressing the limitations by enabling true representation of detailed board traces within the simulation environment. By mapping the intricacies of the board's trace layout onto the corresponding material properties calculation, such as E and CTE, it allows for a far more faithful simulation of the real-world behavior of the board. This innovation is particularly significant as it inherently captures the anisotropic properties induced by the trace distribution, playing a pivotal role in evaluating warpage patterns and interconnect reliability. Trace mapping applies a black/ white bitmap discretization of actual design layout and automatically calculates effective properties of each discretization mesh. Such discretization facilitates the calculation of effective material properties by meticulously considering the volume fractions of copper and dielectric materials within each discrete mesh block. The higher mesh density is equivalent to more realistically copper/ dielectric material distribution, therefore local and global warpage behavior are well captured. Because the solver engine smears discretized mesh properties, the additional computational cost overhead is minimized.

Mapping takes place in two stages. As illustrated below, during the first stage, a representation of the layout is built upon a rectangular grid using the data from a specified ECAD layout design file. The cell size of the grid is governed by the smallest features in the layout that must be resolved. This size can be controlled by the user and should be specified based on the resolution required. A metal fraction value is assigned to each cell depending on the contribution of metal to that cell. The metal fraction value ranges from 0 to 1, where the 0 value represents a pure dielectric material and 1 a pure metal material. The conduction paths that connect the metal traces between the different layers, that is, the vias, can be specified as either hollow or solid (default). During the second stage, the metal fraction values are mapped from the source grid to the target mesh. Once the mesh is created, Mechanical then generates the mapped metal fractions. The sequence of this construction is illustrated below.

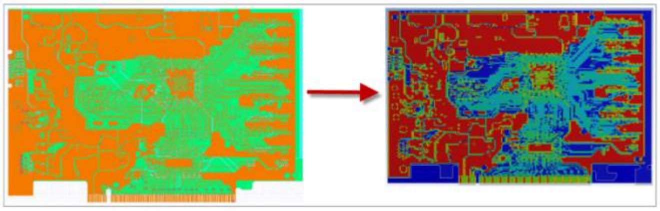


Fig 3: Actual Groq LPU PCIe Layout Design File Metal Distribution and Trace Mapping at Fine Discretization Mesh Density

Under varying temperature conditions, distinct layers within the stackup undergo expansion at differing rates due to their specific material properties. This asymmetrical expansion leads to out of plane warpage behavior of the board.

Reinforcement Trace Mapping Methodology

The reinforcement trace modeling technology developed by Ansys introduces a paradigm shift by enabling the accurate import and mapping of detailed traces onto the simulation domain. Unlike previously mentioned trace mapping methods that merely approximate the metal patterns, this technology offers a sophisticated approach that considers both the in-plane and cross-plane interactions of traces and vias. By leveraging the detailed geometrical and material information of traces and vias, it offers new simulation capabilities to visualize localized vertical structure stress distributions.

Trace reinforcement workflow allows modeling of traces and vias as “reinforcement elements”. Ansys Mechanical provides reinforcement specification for line bodies (discrete reinforcing) and surface bodies (smeared reinforcing). Each line body specified as reinforcement basically represents a reinforcing fiber arbitrarily oriented in space. Each surface body specified as reinforcement basically represents a reinforcing layer. This reinforcing layer can be either a homogeneous reinforcing layer (membrane) or reinforcing layer with evenly spaced fibers.

It uses a mesh independent method for creating reinforcing elements. The procedure uses MESH200 elements to represent the reinforcing member locations inside the generated reinforcing geometry(traces). When the solution is initiated, the application temporarily defines the reinforcement locations using MESH200 elements along with the base elements. During the solution process, the application internally creates the element REINF265 for surface bodies based on the intersection of corresponding MESH200 and base elements. Vias represented by line bodies, are meshed using BEAM 188 elements and EEMBED command is used to constrain all the beams with the base elements using REINF265 elements. Ansys Sherlock- Ansys Mechanical provides an automated way of generating this workflow within the workbench platform.

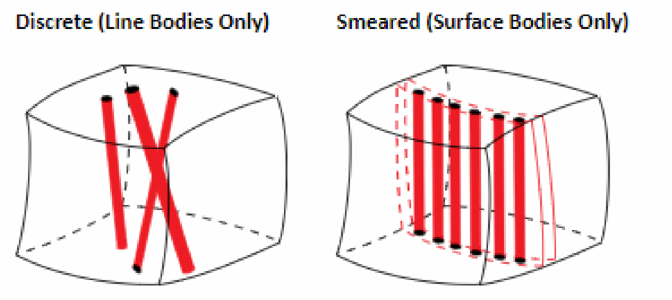
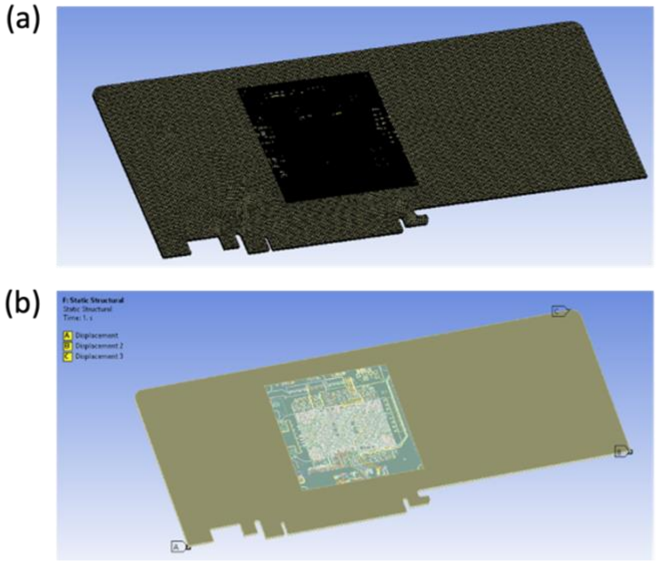


Fig 4: Reinforcement Modeling Schematic of PCB Design and Explicit Via Modeling as Beam

Proposed Modeling Approach

Based on above studies, a combined trace mapping and reinforcement methodology is finally proposed to build the optimized warpage prediction model for Groq LPU PCIe board. Reinforcement methodology is used in the package shadow area to capture more accurate local stress and warpages whereas trace mapping method was used elsewhere in the board. A meshed board model is shown in the Figure5. Each board layer is modeled explicitly along with non-uniform distributed traces and vias structures. Due to the high-fidelity requirement under package shadow area, the localized mesh is much finer to match high resolution of trace/ via embedding through reinforcement method. Figure5 (b) shows embedded trace/ via geometry from the actual design file.



F: Static Structural
Static Structural
Time: 1. s

A Displacement
B Displacement 2
C Displacement 3

Fig 5: (a) Combined Approach with Localized Reinforcement Methodology (Package Shadow Region), (b) Trace Mapping Methodology (everywhere else) for Groq PCIe Board Warpage/ Stress Risk Assessment

The directional deformation in out-of-plane(Z) direction is shown below in Figure6. It clearly shows under the package shadow area, the board is in convex shape. This contour indicates potential SMT risk during the package assembly process. Since electronic packages are not perfectly flat during SMT process, the mismatched contour direction between package and board will further increase the assembly failure likelihood. Excessive stress on vias/ traces during and after SMT process are posing great reliability risk to the system.

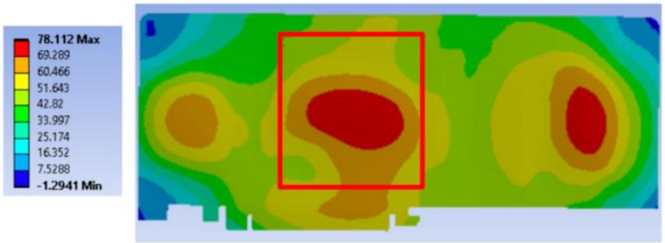


Fig 6: Groq PCIe Board Out-of-plane Directional Deformation (point of interest is package shadow area as indicated in red dotted box)

IV. Warpage Measurement Result And Model-to-hardware Correlation Analysis

To validate the accuracy of above-mentioned modeling methodologies, three Groq PCIe boards are used for warpage measurement by 3 dimensional DIC technique. Digital Image Correlation (DIC) is a non-contact methodology distinguished by its capacity for high-resolution and comprehensive optical metrology. This method facilitates the precise quantification of both in-plane and out-of-plane displacements or strains through the correlation of successive images captured during the object's deformed and undeformed states. Compared to traditional warpage measurement, DIC takes advantage of advances in optical imaging and computational analysis to provide a holistic understanding of warpage phenomena. DIC relies on the principles of image analysis and pattern recognition to track and quantify the displacements and strains across the surface of a specimen subjected to external loading or environmental conditions. The technique involves applying a random speckle pattern to the surface of the specimen, which serves as a unique identifier for each point on the surface. Images of the specimen are captured before and after deformation, and sophisticated algorithms are employed to match the speckle pattern and determine the displacements and strains. Furthermore, the non-contact nature of DIC offers several advantages, including the preservation of the specimen's integrity, the ability to measure delicate or fragile materials, and the capacity to monitor dynamic deformations in real-time.

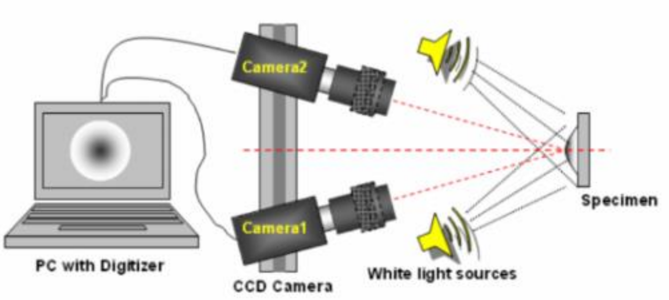


Fig 7: 3D DIC Warpage Measurement Setup Schematic

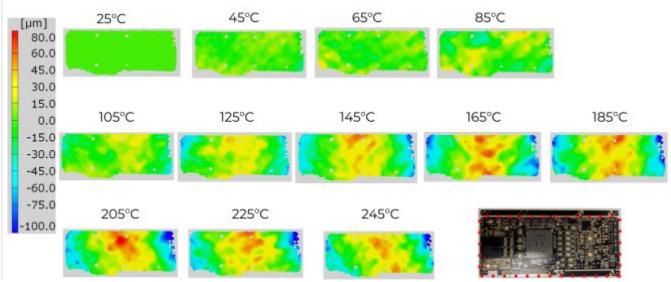


Fig 8: DIC Warpage Measurement Result on Out-of-plane Warpage at Multiple Temperatures (relative warpage extracted wrt room temperature)

In Figure 8, relative warpage with respect to room temperature is plotted. In this way, we eliminate the impact of initial warpage of board at room temperature and only focuses the relative warpage delta deviating from room temperature. Figure 8 presents the out-of-plane warpage behavior of Groq PCIe board, the package shadow region clearly shows higher position compared to the rest of the board at both room temperature and high temperature condition.. Normally for board level warpage measurement, warpage direction doesn’t flip under a wide range of temperature. Similar trend is observed in this test for all 3 boards. The potential reason may be attributed to the epoxy curing phenomenon during the board manufacturing process. Figure 9 shows model-to-hardware correlation considering part to part variation, overall numerical simulation shows a very similar trend against measurement data with error band around +/- 20 um overall and +/-5 um under package shadow area. Both trace mapping and reinforcement with embedded trace/ via representation successfully indicate the absolute minimum point on the board. This marks the highest risk region for future SMT processes. The main indicator from simulation is to predict accurately where the highest warpage is. From test data to model correlation, the worst warpage location precision error band is at 2 to 4 mm level, which again validates the accuracy level of such warpage modeling methodology. Because hybrid approach explicitly modeled via constraints and properties, it tends to have a stronger integration behavior, leading to higher stiffness and lower warpage.

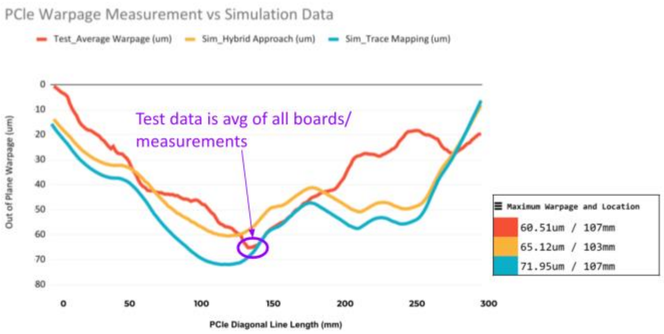


Fig 9**:** Groq PCIe Board Diagonal Line Out-of-plane Warpage Model-to-hardware Correlation.

Table 2: Different FEA Modeling Methodologies Comparison

Methodology	**Mesh count**	**Computational Time**	**Modeling Effort**
Trace Mapping	1x	1x	Low
Hybrid (Trace Mapping & Reinforcement)	1.6x	2.1x	Moderate
Full Reinforcement	High		

In addition to accurate global warpage prediction from both simulation methodologies, reinforcement with embedded element approach also provides localized stress visualization as a bonus. Such detailed stress check greatly facilitates design optimization. Figure 10 below shows highest localized stress on both traces and vias, as clearly indicated in the simulation result, such high-risk areas need to be optimized to ensure the maximum von mises stress level is below critical threshold.

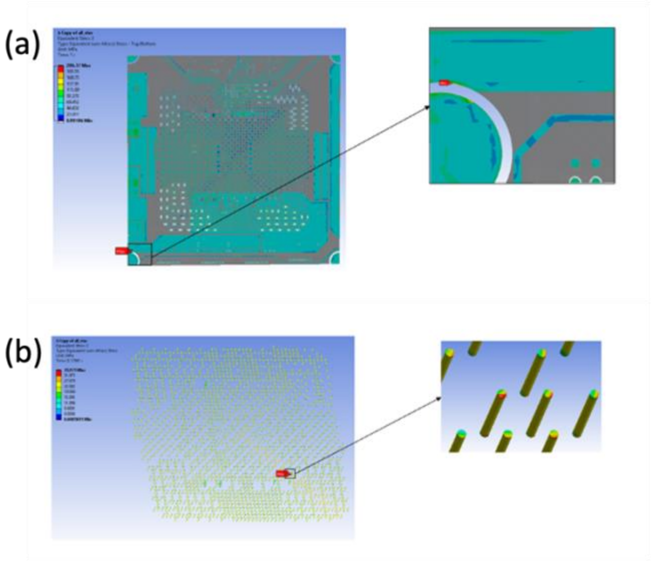


Fig 10: (a) Package Shadow Area Localized Stress with Zoomed in Trace Stress; (b) Package Shadow Area Localized Stress with Zoomed in Via Stress

Table 2 listed the quantitative comparison of three different modeling approaches for warpage prediction. Trace mapping method is relatively easy to implement for accurate warpage location prediction, with no further information on stress level on traces/ vias. Hybrid approach (localized reinforcement + everywhere else trace mapping) not only predicts accurate warpage location, but also indicates high stress regions on traces and vias. The cost is relative high mesh count in local regions and more than doubled computational time. Full reinforcement offers the most accurate details representation for both warpage location and stress levels, however due the extreme high effort required for meshing and solving, the data is not shown here. To provide guidance on tradeoffs and most appropriate method based on point of interest, in Table 3, we list best application use case for each modeling methodology.

Table 3: FEA Modeling Methodologies Tradeoff Comparison and General Recommendation

Nature	**Rule of Mixture**	**Trace Mapping**	**Reinforcement**	**Hybrid Approach**
	Lumped representation	***Localized in plane representation***	***Global in and out of plane representation***	***Localized in and out of plane representation***
Accuracy	Lowest	Medium	Highest	High
Computational cost	Lowest	Medium	Highest	High
Main application	Quick turnaround global warpage check	Accurate global warpage check	Accurate global warpage and stress check	Accurate global warpage and localized stress check

V. Conclusion

In this work, the focus areas are absolute warpage location prediction and localized stress risk assessment of bare board. To achieve this goal with reasonable computation cost, three approaches are extensively investigated for comparison.

Rule-of-mixture approach offers a straightforward and easy to apply methodology to quickly obtain global warpage behavior with moderate accuracy. Trace mapping approach provides more accurate global and local warpage prediction with reasonable computational resource requirements. However, the drawback is lack of stress visibility. On the other side, fully reinforcement modeling approach with embedded trace/via representation approach offers the most sophisticated and close to true design warpage prediction advantages, at the same time, it's possible to visualize high stress regions between layer-to-layer interconnections through vias. It comes with the most extensive computational cost. Overall, we recommend a hybrid approach with reinforcement methodology only at critical regions with trace mapping method for everywhere else. This way it not only enables FEA modeling prediction sufficient accuracy level (especially highest warpage locations and localized stress identification) at critical locations, but also reduces computational resource overhead compared to full reinforcement approach. Other things to be noted when performing board level warpage predictions are: 1. Part to part variation due to manufacturing processes; 2. Repeatability test during temperature cycling; 3. Stress free temperature identification.

Our next step is to deploy a machine learning based algorithm to bridge the gap between test and simulation results. Accurate warpage modeling is not only crucial for assembly process risk mitigation but also lays the foundation for next step board level reliability risk assessment. To further improve the correlation between simulation and test, a physics-based machine learning approach can be used to capture the residual physics existing in test compared to ideal conditions used in simulation. This will not only help provide additional insights to warpage control and prediction, but also reduce prototype testing design of experiments (DOEs) and cost.

References

[1] C. -Y. Lien, Y. -C. Chuang, Y. Yao, E. Charn and E. Chen, "Block-Based Finite Element Modeling, Simulation, and Optimization of the Warpage of Embedded Trace Substrate," *2018 IEEE 20th Electronics Packaging Technology Conference (EPTC)*, Singapore, 2018, pp. 802-806, doi: 10.1109/EPTC.2018.8654342.

[2] V. -L. Pham, H. Wang, J. Xu, J. Wang, S. Park and C. Singh, "A Study of Substrate Models and Its Effect On Package Warpage Prediction," *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, Las Vegas, NV, USA, 2019, pp. 1130-1139, doi: 10.1109/ECTC.2019.00175.

[3] Ansys® MAPDL,24.1,Help Manual, 14.1 Reinforcing Workflow 14.1. Reinforcing Workflow (ansys.com)

[4] Ansys® MAPDL,24.1,Help Manual, 14.2 Direct-Embedding Workflow14.2. Direct-Embedding Workflow (ansys.com)

[5] Z.Yang, Zhi, K.Rivera, J.Patel, E.Tremble, D.B. Stone, K.Choi, and E.Blackshear. . "4-2-4 Laminate Hotspot Identification and Joule Heating Effect Assessment via Thermoelectrical Simulation. " *2019 IMAPSource Proceedings*,2019 (1): 120–26. doi.org/10.4071/2380-4505-2019.1.000120.

[6] C Cai, Z Yang, Y Li, P Jain, T Kang, K Mellachervu, "Electromigration Risk Assessment and Circuit Optimization using Innovative Multiphysics Modeling", *Journal of Microelectronics and Electronic Packaging* ,20 (1), 9-16, doi.org/10.4071/imaps.1882247