CHIP/PACKAGE CO-ANALYSIS OF THERMAL-INDUCED STRESS FOR FAN-OUT WAFER LEVEL PACKAGING

Stephen Pan, Zhigang Feng, Norman Chang
ANSYS, Inc.
San Jose, CA, USA
stephen.pan, zhigang.feng, nchang@ansys.com

ABSTRACT
This paper presents an innovative co-analysis solution for thermal-induced stress of fan-out wafer level packaging (FOWLP). The reliability of FOWLP on either the fan-out package region or in the chip IDLs are of concern for loadings from large differential thermal expansions at Si chip and package/PCB interfaces and Tg (Glass Transition Temperature) effects in the dielectric materials. The dielectric layers in WLP and IDL of a chip are of weak ELK materials which are easy to develop crack at relatively low stress levels. The drastic property changes, for example, 30x change in Coefficient of Thermal Expansion (CTE) at Tg, aggravates the development of micro-flaws and the coalesce into long cracks which leads to eminent failure of the FOWLP. The approach in thermal-induced stress analysis includes the generation of thermal-aware chip power maps for multiple dies in FOWLP, the conversion of converged thermal profiles in FOWLP to thermal loadings for stress analysis, efficient model generation and analysis for Tg effect of non-linear material properties, and detailed sub-modeling of on-chip structures for thermal-induced stresses. Discussions of innovative thermal-induced stress modeling process and results extraction for a FOWLP test case are demonstrated.

Key words: CTM, CPS, FOWLP, SoC, RDL, IDL

INTRODUCTION
FOWLP (Fan-Out Wafer Level Packaging) is known for its low cost and high performance. Figure 1 is the cross-section view of a FOWLP with RDL fan-out layers which is thin in thickness. Like FCBGA, there could be multiple dies in FOWLP, either side-by-side or stacked. The dies could be flushed with the mold and exposed to air for better heat dissipation. The fan-out area could have Copper pillars up to the top to connect to a top package and form a Package on Package (PoP) configuration. Removing the need for a package substrate reduces the cost and makes the FOWLP a popular choice in low-power high-performance industry such as for Mobile devices.

The electrical and thermal performance of FOWLP was discussed in [1] that it has superior form factor, pin count, and thermal performance to FC-BGA and it paves the way to the solution from foundry today. The mechanical reliability of the dielectric layer of BOL which has similar structure as FOWLP was reviewed [2]. This shows the efforts of packaging house to provide a solution. This paper is from the perspective of EDA tool development to demonstrate what can be done for FOWLP in terms of thermal and stress analyses.

In the following sections, a FOWLP package with two side-by-side chips was analyzed for thermal and stress responses and checked for reliability related issues in CPS (Chip-Package-System) environment. One of the chip in this FOWLP was reviewed for thermal stresses in IDL (Inter-Dielectric Layer) which is known to be weak and has reliability concern at higher stresses. The focus of this paper is on the modeling and simulation capability developed for FOWLP and the SoC chips. The theory involved in the thermal-stress analyses are described. Then the important modeling methods in creating thermal and stress models are discussed with sample case results presented.

https://en.wikipedia.org/wiki/Embedded_Wafer_Level_Ball_Grid_Array
Figure 1. Structure of modern FOWLP

THEORY
Heat flow
The governing equations of heat condition are:

\[ \frac{\partial}{\partial x} \left( K_x \frac{\partial T}{\partial x} \right) + \frac{\partial}{\partial y} \left( K_y \frac{\partial T}{\partial y} \right) + \frac{\partial}{\partial z} \left( K_z \frac{\partial T}{\partial z} \right) = \rho c \frac{\partial T}{\partial t} \]

\( T = T(x,y,z,t) \) = Temperature Distribution
\( K_x \) = Thermal Conductivity in the x-direction
\( K_y \) = Thermal Conductivity in the y-direction
\( K_z \) = Thermal Conductivity in the z-direction
\( \rho \) = Mass Density
\( c \) = Heat Capacity
For isotropic steady state condition, the governing equation is the following.

\[ K \left( \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) = 0 \]

\[ K_x = K_y = K_z = K \]

Boundary conditions:

For known temperature at specific locations,

\[ T(x_0, y_0, z_0; t) = F(t) \]

For heat flux/power generation,

\[ Q = -K A \frac{\partial T}{\partial n} \]

\[ Q = \text{Heat Flux in the units of Watts} \]

\[ A = \text{Surface Area in the units of mm}^2 \]

\[ n = \text{Outer Normal of the Surface Area} \]

Typically used in IC power dissipation on the die surface.

For heat convection and radiation,

\[ q = -K \frac{\partial T}{\partial n} (h + h_r) (T - T_x) \]

\[ q = \text{Surface heat flux} \]

Use for heat loss to the ambient.

For the FOWLP study, steady state heat transfer was used in this study and average power map corresponding to a chip operating condition is assumed.

**Thermal-elasticity**

The stress (Figure 1A) is related to the strains by:

\[ \{\sigma\} = [D] \{\varepsilon^{el}\} \]

where:

\[ \{\sigma\} = \text{stress vector} = \begin{bmatrix} \sigma_x & \sigma_y & \sigma_z & \sigma_{xy} & \sigma_{xz} \end{bmatrix}^T \]

\[ [D] = \text{elasticity or elastic stiffness matrix or stress-strain matrix} \]

\[ \{\varepsilon^{el}\} = \{\varepsilon\} - \{\varepsilon^h\} = \text{elastic strain vector} \]

\[ \{\varepsilon\} = \text{total strain vector} = \begin{bmatrix} \varepsilon_x & \varepsilon_y & \varepsilon_z & \varepsilon_{xy} & \varepsilon_{xz} & \varepsilon_{yz} \end{bmatrix}^T \]

\[ \{\varepsilon^h\} = \text{thermal strain vector} \]

\[ \{\varepsilon^h\} = [D^T]^{-1} \begin{bmatrix} \alpha_x^{eh} & \alpha_y^{eh} & \alpha_z^{eh} & 0 & 0 \end{bmatrix}^T \]

Stress vector definition:

\[ \{\varepsilon\} = \{\varepsilon^h\} + [D]^{-1} \{\sigma\} \]

\[ \{\varepsilon^{eh}\} = [D^T]^{-1} \begin{bmatrix} 1/G_{xy} & 1/G_{xz} & 1/G_{yz} & 0 & 0 \end{bmatrix} \]

\[ [D]^{-1} = \begin{bmatrix}
  1/E_x & -v_{xy}/E_x & -v_{xz}/E_x & 0 & 0 & 0 \\
  -v_{yx}/E_y & 1/E_y & -v_{yz}/E_y & 0 & 0 & 0 \\
  -v_{zx}/E_z & -v_{zy}/E_z & 1/E_z & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 1/G_{xy} & 0 \\
  0 & 0 & 0 & 0 & 0 & 1/G_{xz} \\
  0 & 0 & 0 & 0 & 0 & 1/G_{yz}
\end{bmatrix} \]

where typical terms are:

\[ E_x = \text{Young’s modulus in the x direction} \]

\[ v_{xy} = \text{major Poisson’s ratio} \]

\[ v_{yx} = \text{minor Poisson’s ratio} \]

\[ G_{xy} = \text{shear modulus in the xy plane} \]

For isotropic materials (\( E_x = E_y = E_z \) and \( v_{xy} = v_{yx} = v_{xz} \)), and the \([D]^{-1}\) matrix is symmetric.

The governing equation of thermal-elasticity are

\[ \text{Strain-displacement relationship:} \]
The equilibrium equations in terms of displacements for isotropic materials are

\[
\begin{bmatrix}
\frac{\partial}{\partial x} & 0 & 0 \\
0 & \frac{\partial}{\partial y} & 0 \\
0 & 0 & \frac{\partial}{\partial z}
\end{bmatrix}
\begin{bmatrix}
e_x \\
e_y \\
e_z
\end{bmatrix}
= \begin{bmatrix} u \\ v \\ w \end{bmatrix}
\]

FEM is used by reducing the equilibrium equations into matrix with displacement vector as unknowns. Once the displacements are solved, strain and stress can be calculated.

The material properties of E and \(\alpha\) could be temperature dependent as used in the FOWLP thermal-stress study. The boundary conditions in the thermal-stress problem are the displacement constraints at a local area to avoid rigid body motion, and temperature distribution of FOWLP on board. For warpage analysis, environment temperature change is used for differential thermal expansion on FOWLP.

MODELING OF FOWLP FOR ANALYSIS

ECAD-based FOWLP

modeling and task flow

The structure of the FOWLP with two chips is similar to that of a FC-BGA with molding compound, just that the package substrate layers are replaced with the RDL (ReDistribution Layer) layers which is much thinner in thickness. The total thickness of the RDL layers is around 60um vs. ~300um for typical FC-BGA. The mold material is placed around the dies as structural support to avoid excessive warpage due to the use of RDL layers which is thin in thickness in the fan-out zone. The configurations of the FOWLP are listed below.

- **FOWLP size**: 8x8mm
- **Chips**: 2x3x0.3mm with 450 Cu pillars
- **RDL thickness**: 0.06mm
- **Mold**: 7x6x0.3mm
- **2nd level joints**: 270 Cu pillars with solder paste

Figure 2 is the 3D geometry of the FOWLP as imported from ECAD in a chip thermal analyzer (CTA) [6] and the placement of the two chips with the molding flushed on die top so that the dies are exposed for better heat dissipation. The yellow pattern under the dies are the pins/pillars under the dies in contact with the RDL. Figure 3 is the zoom-in view of the traces, via, and pins. The target analysis model using FE (finite element) is to include the geometry details for most accurate solutions.

The thermal-stress task flow (Figure 4) for FOWLP starting with chip power map generation from a chip power calculation tool [6]. The chip power map is in the form of CTM to be described in the next section. The FOWLP-on-board geometry was then modeled using FE method with CTM as power input and still air at 20C ambient. Empirical based equivalent heat transfer coefficients on exposed surfaces are used in this study. For more complicate environment, such as multiple packages/boards and fans in a box and the moving air environment can only be simulated by CFD (Computational Fluid Dynamic) type of simulators [8]. The resulting thermal boundary conditions to FOWLP can be extracted from the CFD solutions and used in the conduction-based thermal analysis model in this study.

After the thermal analysis of FOWLP in a system, the analysis model can be converted for stress analysis using a structural analysis tool [7], to review the deformation and stress states on Cu traces/vias or dielectric which could be of weak ELK (Extreme Low K) or ULK (Ultra Low K) material.

**Figure 2.** Geometry configuration of the FOWLP in this study through ECAD import.

**Figure 3.** Zoom in view showing the details of trace, via, and pins/Cu pillars under the die
The CTM was rotated when applied to the chips in analysis. In this FOWLP as defined in ECAD, viewing from top, the left chip CTM was rotated 90 degree and the right chip CTM was rotated 270 degree before flipping for final placement on FOWLP. When using CTM in thermal analysis, instance power on tiles can be updated with the resulting temperature for the next power-thermal iteration. Typically, only 2 to 3 iterations are needed for the converged solution in CPS environment.

For SoC or gate-level chip design, instances are composed of transistors with different logic functions and power calculations are done for the instances. There are more than 8 million instances in an area of 6 mm² for each of the two chips in this FOWLP. Each instance is a heating object, the thermal modeling efforts will be overwhelming in current FEM technology. The current approach is to lump them into smaller number of heat sources. A tile-based mapping from instances was performed and the resulting power map based on 10x10um tiles with totally about 120 thousand tiles or heat sources in the resulting power map for thermal analysis. Figure 6 is the contour display of the power map at an uniform temperature condition. The total power in this power map is about 1.5W. The transistor or instance power is typically temperature dependent due to leakage current component. The total power in a device includes the dynamic switching power in Figure 5 and the leakage power. From the chip power integrity tool, a collection of the power map calculated at several temperature points was generated in Chip-Thermal-Model (CTM) which is a temperature-dependent power map library for the chip. The generation of CTM was from either vectorless or event-driven power scenarios. It is the power on the chip at a specific operating mode averaged over a period of time. The on/off of the CTM operating modes can be assembled into transient scenarios for analysis [5].

CPS thermal modeling
Analysis model using FEM was generated in chip thermal analyzer with the CTM power applied to the chips of FOWLP. It ran through thermal/power iterations until the temperature profile and power map are consistent, i.e., at a converged state. Figure 7 is the converged thermal profile on top of FOWLP. In this FOWLP study, the materials used were defaults from the chip thermal analyzer for typical BGA. However, this will not affect our purpose to demonstrate the thermal-induced stress flow. The same FE model is used in the subsequent thermal stress analysis. In this FOWLP case, the FE model has 1.8M nodes, 3.3M elements, and the model details can be reviewed in the structural analysis software [7] (Figure 8).

Figure 9 shows the FE details of the traces and vias in RDL of FOWLP. The details of geometry in analysis model is critical to thermal response accuracy, especially sensitive for smaller packages [4] like this one. If smeared properties are used on RDL layers, the Tmax will be underestimated by 14% and Theta_JA of this FOWLP will be reduced from 37C/W to 31C/W. The observation is that many heat flow path through narrow vias from layer to layer in detailed model. But in the smeared model, the layers are shorted by the material of equivalent conductivity in a much larger region near each via. Heat is easier to spread out through ‘shorted’ conductive materials, though less conductive than copper for via.

On the other hand, stress analyses of traces and vias are not possible in a model using simplified and smeared properties.
Smeared model is only good in estimating overall warpage, die stress, and solder joint reliability, not for prediction of failures in RDL of FOWLP.

One of the issues in detailed trace outline modeling by FE is the existence of degassing holes on the PG nets in RDL with large metal pieces. While degassing holes is necessary in the formation of the RDL in FOWLP, it makes the mesh generation much more difficult (Figure 10). Comparison of results of the thermal models with and without degassing holes using FE model showed negligible differences in either temperature distribution and maximum temperature on chips. This is probably due to the very low heat flux on the PG layers so that the existence of the degassing holes has minimal impact on the overall heat dissipation path. For stress analysis, warpage results were compared and seeing negligible impacts from the degassing holes, possibly due to the small contribution of the thin RDL layers to the stiffness of the whole FOWLP structure. A function in the chip thermal analyzer was created to identify and remove the degassing holes automatically for simulation purposes.

The number of degrees of freedom in this FOWLP model is about 1.8M and the structural tool [7] took about 14 min. and 7.3GB peak memory to solve with a single core in a Linux server with 256GB of memory.

Figure 11 is the temperature profile on the top metal traces of RDL which is connected to die by Cu pillar contacts (small squares with equivalent areas to the contacts) at the center of circular pads. The heat flux vector display in Figure 12 shows the critical path in heat dissipation near the die. Figure 13 shows the heat flux vector on Cu pads which connects the dies to the RDL. It seems that the right die in Figure 7 has higher heat flux on the Cu contacts and the temperature is a little lower than the other one. This was unexpected as the design is in general symmetric. A review of the model found that the materials filling the spaces under the two dies were not the same that leads to minor asymmetry for the thermal and stress results as discussed below. In FOWLP production, the filling material should be molding compound. Figure 14 is the heat flux vector plot of Cu pillars under RDL into the thermal board underneath. Most of the heat dissipations are under the die footprints through board, i.e., about 95% of all the power (~3 W) from the chips.
Hence, the temperature loading for thermal-stress analysis can be either from the chip-package thermal analysis run or regenerated in the structural analysis software in the coupled thermal-stress analysis. When exporting to thermal model from the thermal analysis, converged power maps were on the chips already, ready to rerun thermal analysis in the structural tool. When exporting as a structural model from chip-package thermal analysis tool, the existing thermal profile was automatically applied to all the FE nodes as thermal loads, ready to run thermal-stress analysis.

For thermal-stress analysis, the element type needs to be changed from thermal, e.g., SOLID70, to structural e.g., SOLID185. The material properties should be reviewed to include mechanical properties, e.g., E, v, and α (CTE). The typical structural boundary conditions are the nodal constraints to avoid rigid body motion, i.e., large translation and rotation of the FOWLP+board structure. A script was generated to hold the top center of the molding between the dies against the rigid body motions.

Linear elastic properties are used in the thermal-stress analysis, assuming that the FOWLP on board is stress free at room temperature of 20°C. The stress due to differential thermal expansion will build up in all components of FOWLP+board when the dies were powered with CTM at about 3W total.

Figure 15 is the warpage or Uz (Z displacement) contour of FOWLP due to power on. Again, the asymmetry of Uz was due to different α and E in the underfill material surrounding the Cu contacts directly under the dies. There were about 2.4x difference in α and 2.35x difference in E. Figure 16 is the Seqv (equivalent stress) on top metal traces. The Seqv is mostly at ~200MPa level and it could be ok given ultimate stress of copper wire at around 323MPa. For dielectric materials in RDL, we could check on S1, the first principal stresses as the dielectric material could be brittle and fragile at high tensile stress state. Figure 17 is the S1 (tensile) contour of the dielectric material.

CPS thermal-stress modeling
The Tmax on chip are almost the same when solved in chip-package thermal analysis (Figure 7) and in the structural analysis tool [7] which is also a thermal solver (Figure 8), only a negligible 0.022% difference observed in this case.
The FE model for thermal stress has about 5.5M degrees of freedoms (displacements on nodes) in solution. The linear model took about 10 min from matrix formulation to solution, using the same Linux server.

**Warpage of FOWLP at reflow**

When mounting FOWLP on board at elevated temperature (>200C) to melt the solder material in reflow oven, the FOWLP must remain flat so that all the connections by solder materials are in good condition. Assuming that the package is perfectly flat at room temperature (20C), when temperature rises in reflow oven, the FOWLP will deform due to differential thermal expansion of molding, chips, RDL metal, and dielectric materials. One issue for temperature changes over a large range like reflow is the existence of glass transition temperature (Tg) (Figure 18) in organic materials like molding and dielectric in RDL. For example, the thermal expansion coefficient α of the dielectric material was smaller (12ppm/C) below Tg of 100C and jumped to 30x above Tg (Figure 19). This analysis of FOWLP was using assumed material properties and only Tg in dielectric material is included and there were no changes in Young’s modulus at Tg.

For this nonlinear material property, multiple load steps were setup to monitor the development of displacements and stresses as temperature rises from 20C to 200C with Tg = 100C for dielectric in RDL.

Figure 20 is the warped FOWLP at 200C with contours of displacement in Z direction showing significant uplift at corners (3.48mm). The Copper pillars under the RDL moves 2.94mm vertically (Figure 21) which will fail the coplanarity check at reflow as the solder paste cannot fill the big gap at the corner joints. Further review of the materials used are needed by selecting proper CTE and E. Figure 22 shows that the glass transition effect shown at the end of step2 which corresponding to Tg=100C and the significant rise of Uz from 100C to 200C due to high α in the material.
For chip designers, the high thermal resistance of FinFET is difficult to model and simulate using field solution like FEM as the details of the FinFET geometry, the fin/finger configurations and materials are all properties of semiconductor foundries. Many foundries are now providing thermal resistance of instances or delta T of the FinFET devices based on measurements and/or simulations [9].

Temperature rises on wires embedded in IDL in modern chip designs are expected to be higher than before as the wires are narrower in width and thinner in thickness that make the Joule heating effect more severe, in addition to the higher temperature impact from FinFET. As the dielectric material surrounding the wires are of low thermal conductivity, e.g., typically 1 W/m°C as compared to ~130 of silicon, even with small Joule heating power, the delta T on wire still will be significant, e.g., 3.3°C from 0.002mW on wire segment of 22nm width. While some foundries provided empirical formulas, this delta T on wire can be simulated accurately using FE sub-modeling technique.

The influence of the Joule heating on a wire dies out over distance quickly. There is no need to consider the thermal coupling for wires more than 5 um away from the previous study (Figure 23) [3].

The thermal coupling from FEOL and neighboring wires will add to the delta T from Joule heating of the wire and push the final wire temperature higher (Figure 24). Note that the final wire temperature could be calculated as the sum of the CPS temperature near the location on chip and the delta T for wire/device as described above.
CHIP sub-modeling for thermal and stress

While the full chip modeling with detailed wire layout is difficult by FEM, a smaller piece of the chip like the geometry in Figure 24 can be modeled by FEM in an automated way. Figure 25 shows the FE model of wires in a 10x10um region of a chip in the FOWLP. In Figure 26, the wires are embedded in low-k dielectric material (purple) and with silicon substrate (red) at bottom. From the previous study of delta T decay behavior, the range extending beyond the target wire of interest does not need to be much greater than 5um if the purpose of the analysis is to determine the thermal influence of the target wire to the neighboring wires. This FE model (Figure 26) has 368K nodes and 702K elements, not a very large model. For either thermal or structure analysis, the modeling range could still be extended and seeing good performance in analysis.

This model was used to test the capability in thermal stress on wires. The analysis was for free thermal expansion with environmental temperature rise of 120C, e.g., from CPS T profile. The model was hold fixed at top center to avoid rigid body motion. Figure 27 is the resulting deformed shape and displacement contours showing more thermal expansion in the interconnection layers than that in the silicon substrate. Figure 28 is the von Mises equivalent stress on the wires. A zoom-in view (Figure 29) identifies that the higher stress is due to shearing/bending on the vias of the bottom layers due to differential thermal expansion. The maximum stress of 441 MPa seems at the same level of the yielding or tensile strength of some Copper alloy. The boundary condition of this stress sub-model which allows free expansion is not real as there are still adjacent structures to support it. The better solution is using cut-boundary in the CPS or warpage model to extract the boundary displacements surrounding the submodel in FE tools in future studies.
For thermal sub-modeling, the wire Joule heating power can be allocated to all the wires in the selected region along with delta T or power on devices (Figure 30) based on chip power calculations in chip power integrity tools [6], the realistic thermal responses on wires can be retrieved (Figure 31). As mentioned in the previous section, the Joule and self heating effects with full BEOL/FEOL coupling are possible using the sub-modeling capability in chip-package thermal analysis tool and the structural analysis tools.

SUMMARY
The methods of thermal-induced stress analysis of FOWLP was discussed using a sample case from EDA tool perspectives. Automation and ease of use is the top consideration in the flow development. Detailed outline modeling of FOWLP is emphasized as that would be the key to accurate thermal and stress results. Recent development of chip sub-modeling can help in verifying the Joule heating on wires, the self-heating in FinFET structures, and therefore improve the reliability of chip designs.

ACKNOWLEDGEMENTS
Thanks to T. Park, S. Pytel for tool integration discussion.

REFERENCES