A Comparative Simulation Study of 3D Through Silicon Stack Assembly Processes

Kamal Karimanal
Cielution LLC
kamal@cielution.com

Abstract

A memory stack on logic 3D IC stack was considered for comparative study of warpage response to two different process choices, namely, Die to Die (D2D) and Package to Die (P2D) assembly. Process and reliability modeling software CielMech, and Commercial Finite Element Analysis (FEA) software ANSYS Mechanical were utilized to simulate thermo-mechanical effects of sequential chip attach, underfilling and encapsulation process steps for the chosen flows. Warpage at room temperature as well as attach temperature after each attach step were compared. Results indicated that underfill, substrate, and mold compound thermal strains play important roles in warpage evolution. Significant differences in the final assembled state warpage was predicted and is attributable to path dependence of warpage evolution.

Introduction

FEA has been used in IC packaging for various purposes including package design [1], assembly process development [2], as well as reliability risk evaluation [3,4,5]. In general, three dimensional numerical analysis of new technologies yield valuable learning suited for physics based root cause investigation at a relatively low cost. Through Silicon Via (TSV) based 3D stacking is an emerging technology with various forks in process flow choices on which the industry is yet to standardize. Such technological alignment across the supply chain leads to economy of scale benefits and cost savings from standardized automation tools and processes. Until then, development of suitable process flow to build 3D stack involves steep and expensive learning curve involving the balancing of various competing objectives.

Various approaches such as Die to Die (D2D), Die to Wafer (D2W) and Wafer to Wafer (W2W) and Package to Die (P2D) are being considered [6] as the first assembly step in 3D IC packaging. Depending on the nature and dimensions of the chips to be stacked, certain approaches may be easier to rule out. For example, stacking of different sized dies will rule out W2W, due to wafer utilization reasons. From a business standpoint, P2D followed by additional chip attach may not suit a foundry centric stacking flow where the complete heterogeneous chip stack is supplied by the foundry to the assembly and test services vendor for substrate attach and packaging. On the other hand, P2D flow may aid sequential testing of each chip using traditional test probes designed for C4 bump. Above is just a couple of examples of the multiple factors influencing process flow choices.

Not to be ignored while making the above choice, are the thermo-mechanical implications on assembly yield and long term package reliability. These aspects are usually a component of the testing that's involved in test vehicle programs. However, a comprehensive exploration of design space, process and material choice knobs is out of the scope of test chip and test vehicle investigations, which also involves multiple other components of interest in the electrical, electrostatic and manufacturing domains. As a result, proactive FEA analysis is a necessary tool for evaluating feasibility and process/design parameter tuning prior to assembly.

In this work D2D and P2D stacking approaches are evaluated by comparison of chip/substrate warpage at various attach and post attach room temperature stages of assembly process flow. A notable competing flow is the W2D process, which is out of the scope of this work and will be addressed in a future work.

3D IC Stack Modeled

Schematic of the 3D IC stack modeled is shown in Figure 1.

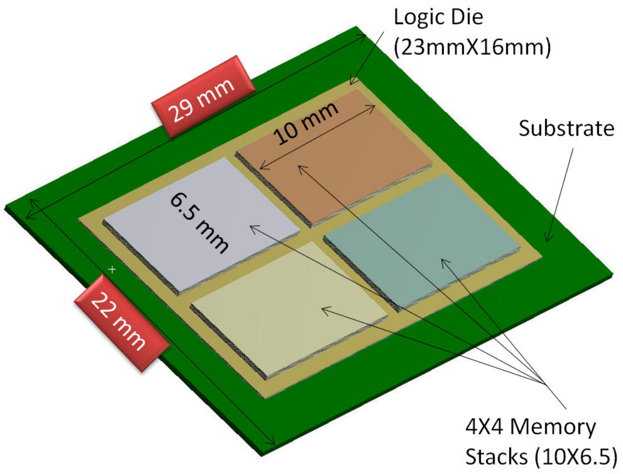


Figure 1: Schematic of the 4X4 Memory Stack on Logic 3D IC Stack.

The memory stack consistes of 4 memory chips vertically stacked and there are 4 such stacks arranged as shown in Figure 1. The Memory stacks as well as the Logic die are 50 µm in thickenss. Details of the 4 memory chips making up each of the memory stacks is shown in Figure 2.

The properties of materials of the package are shown in Table 1.

Assembly Flows Considered

The D2D and P2D flows considered in this work are illustrated in Figure 3 and Figure 4. Even though different underfilling (such as No Flow (NUF), Molded Underfill (MUF) and Capillary Underfill(CUF)) are being considered we will not evaluate those forks in the process flow by considering CUF or NUF only. Since we are only modeling the distinct states before and after underfilling NUF and CUF are indistinguishable by our modeling approach.

In both the flows shown in Figure 3 and Figure 4, the memory stack appears as one unit, without details on the process flow that lead to its assembly. For this work it is assumed that that the memory stack went through a W2W assembly process [6] as shown in Figure 5. The effect of W2W flow was accounted for in the memory stack by assigning a reference temperature of 230C for the entire stack, where 230C is assumed as the W2W process temperature.

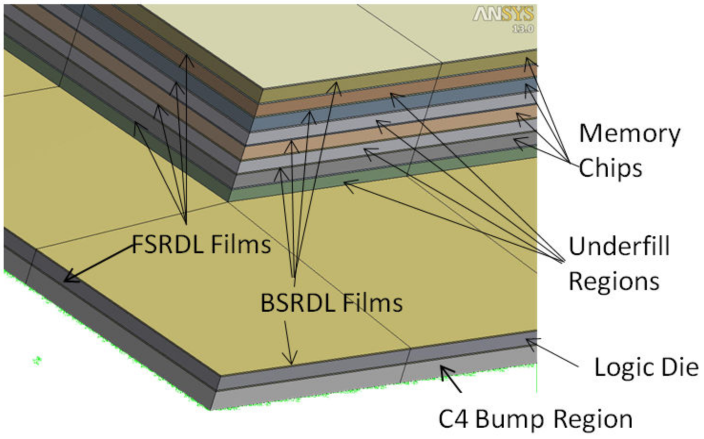


Figure 2: Close-Up View of Stack Details

Part	Material	Thickness (µm)	Properties		
			E (Gpa)	CTE (PPM/C)	ν
Memory and Logic Dies	Silicon	50	160	3	0.3
FSRDL (All chips)	TEOS Oxide	5	71.487	0.51	0.3
BSRDL (All chips)	Polyimide (PBO)	5	2	55	0.3
D2D Inter connections	25 µm dia Copper Micropillars	25	121	17.3	0.3
	Solder Cap(Sn 95.5, Ag 3.5)	10	50	20	0.3
	Underfill epoxy	35	8	30.06	0.28
Substrate to Logic die interconnection	100 µm dia C4 Bumps(Sn 95.5, Ag 3.5)	70	50	20	0.3
	Underfill epoxy	70	15	30.06	0.28
Encapsulation	Typical Mold Compound (MC)	700	26	15	0.3
Substrate	Hitachi MCL_E_700G	400	33	Planar: 8 Normal: 20	0.25

Table 1: Package Materials Used in the 3D IC Studied.

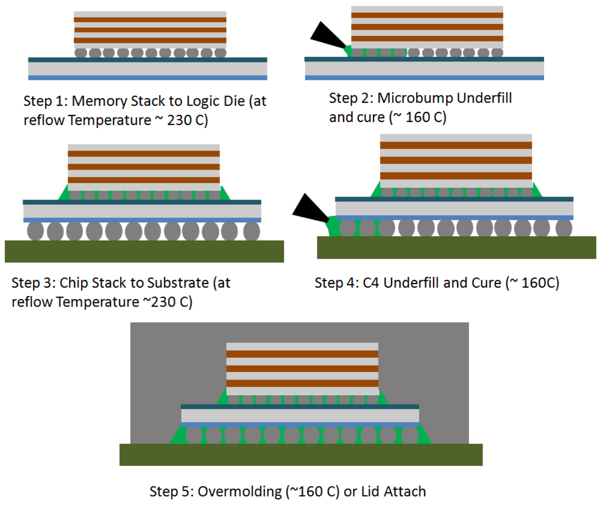

Step 1: Memory Stack to Logic Die (at reflow Temperature ~ 230 C)
Step 2: Microbump Underfill and cure (~ 160 C)
Step 3: Chip Stack to Substrate (at reflow Temperature ~230 C)
Step 4: C4 Underfill and Cure (~ 160C)
Step 5: Overmolding (~160 C) or Lid Attach

Figure 3: D2D Process Flow with Chip Attach Before Substrate Attach.

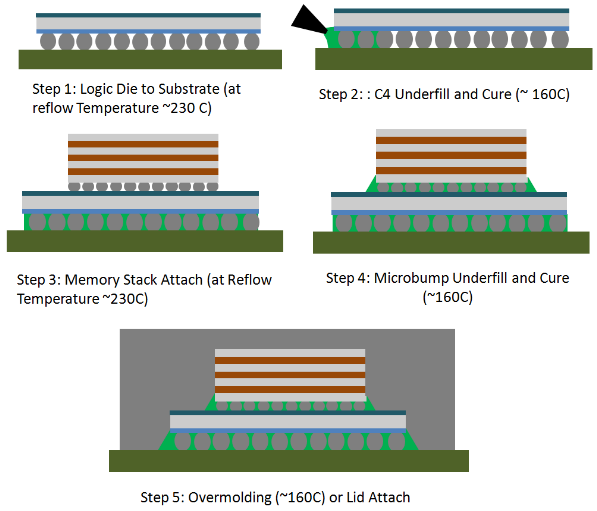

Step 1: Logic Die to Substrate (at reflow Temperature ~230 C)
Step 2: : C4 Underfill and Cure (~ 160C)
Step 3: Memory Stack Attach (at Reflow Temperature ~230C)
Step 4: Microbump Underfill and Cure (~160C)
Step 5: Overmolding (~160C) or Lid Attach

Figure 4: Substrate to Die Attach First Process Flow with Chip Attach Before Substrate Attach.

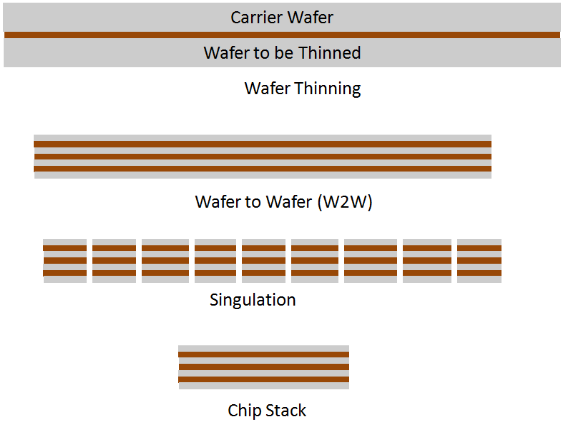


Figure 5: W2W Stacking

FEA Modeling Approach

CielMech from Cielution LLC was used to automate geometry creation, mesher setup, solver inputs and boundary conditions going into general purpose commercial FEA simulation software ANSYS Mechanical from ANSYS Inc. In addition to the standard procedure of geometry creation, meshing, materials and boundary conditions assignment, certain specific modeling techniques were used by CielMech to model sequential process steps. These techniques will be described here:

Warpage Due to Thin Films

Traditional full thickness chips used in monolithic flip chip packages have thin films of oxides, metallization and polyimide deposited on the front (active) side of the chip. Even though these films tended to warp the chip, including that effect was not always required in FEA models due to the fact that the CTE mismatch forces due to those thin films (~1 to 5 µm) were expected to be negligible compared to those from the relatively bulky substrate (~400 µm), chip (400 to 800 µm) and encapsulation (~500 µm).

Die thinning, an essential step in 3D IC fabrication will result in significantly thinner chips (50 µm for active chip integration and 100 µm for interposer based integration). Additionally, 3D ICs may require redistribution layer (RDL) on the back side of the chip to re-route signals from the pads on the back side of parent chip to the pads on the front side of the daughter chip (in front to back attach approach). These Redistribution layers can be made of oxide or polyimide (PI) dielectric materials, which are also on the order of 1 to 5 µm in thickness, which may not necessarily be negligible effect in warping the chip out of spec limits for assembly. Additionally, shrinking warpage spec limits due to thinner micro-bump interconnects make it very important to accurately model all contributors to chip warpage.

In this work, every chip was lined with a front and back side film, whose 'effective' material properties and stress free temperatures were allowed to be independent of each other. The emphasis on 'effective' is due to the fact these films are composites of Back End of Line (BEOL), metallization, and RDL layers. An approach to determine such effective properties and reference temperatures is outlined in[9]. In this work, all those films will be considered as unpatterned blanket films. Hence native material properties and assumed process temperatures were be used in the simulations.

Effective bump array model

Typical logic dies used in high end flipchip components typically have thousands of bumps in the C4 interconnection array. These bumps are made of viscoplastic lead free solder, which are encapsulated in viscoelastic underfill material. Most engineering applications for monolithic flipchip ICs tended to lump this region using effective linear properties [9] for the global package level models. Some works in which the balls were modeled in detail at the global level either addressed scenarios with only hundreds of bumps/balls or certain other special treatment using beam elements [10]. In addition to the rigor of modeling above mentioned nonlinearities, the high mesh count (about 100,000 node points for 10,000 bumps) due to the bump count makes detailed modeling of the bumps at the global model level computationally prohibitive.

3D IC package modeling requires many such bump array interfaces to be modeled. This emphasizes the need for effective lumped region characterization approaches for 3D ICs. However, the justifications for effective linear region treatment of bump-underfill region (~100 µm thick) in monolithic packages with thick dies (400 to 800 µm) may or may not hold for 50 µm thin ICs with ~35-50 µm bump-underfill region thickness in the case of 3D IC stacks.

In this work we have used a characterization approach for each distinct interconnect interface using strip model consisting of only one row of bump array spanning the diagonal length of the interface regions. Actual viscoelastic properties for underfill and viscoplastic material properties for solder regions were used in the detailed strip model. This model was compared with a strip model of same dimensions with lumped interface to describe the bump/bump-underfill region. The effective linear properties using rule of mixtures [9] were imposed on the elements forming the lumped region. Model results showed that time dependence of warpage due to viscoelasticity and viscoplasticity were not noticeable and the steady state warpage of the detailed strip and the lumped strip matched. Based on this justification, rule of mixtures was used to estimate the bump array region properties for the substrate-logic die, and the logic die to memory stack interfaces. Table shows references to nonlinear properties used in detailed model and the effective properties used in the lumped strip model as well as the actual 3D process simulation.

Detailed Model			Lumped model Properties
Copper Pillar	25 µm dia, 25 µm tall	E=120GPA, CTE=17e-6, ν=0.3	E=2.4E10 , ν=0.3 prior to underfilling E=2.9E10,n=0.3, CTE=25.6E-6 after underfill cure
Solder Cap	10 µm tall		
Underfill		Viscoleastic model from Park and Feger [11]	
Solder Bump	100 µm, dia, 70 µm tall	Viscoplastic Model From Wang et al [12]	E=9.5E9 , ν=0.3 prior to underfilling E=1.5E10,n=0.3, CTE=26.6E-6 after underfill cure
Underfill		Viscoleastic model from Park & Feger[11]	

Table 2: Actual Material Properties and Corresponding Lumped Equivalents

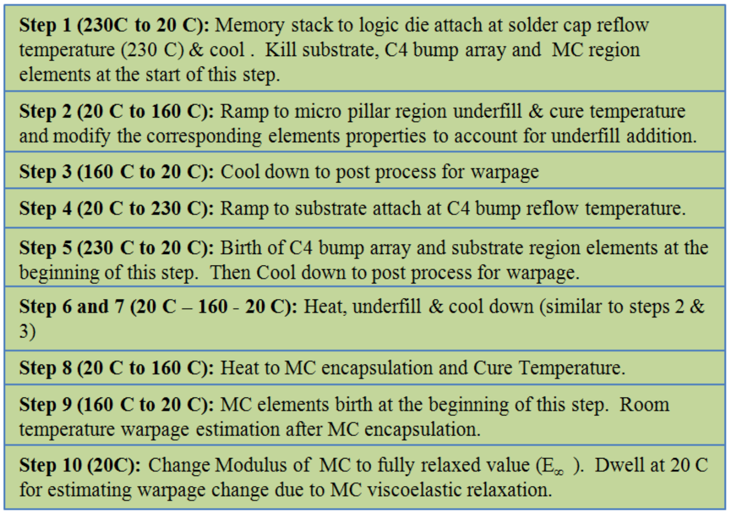
An additional modeling nuance related to bump array lumping to be mentioned is the technique used for capturing the underfill addition in the lumped region. The non existence of underfill prior to capillary epoxy filling can be captured geometrically with the help of element birth and death in the detailed model. However, the geometry of the lumped region is identical before and after underfilling. So this was modeled by enforcing a change in lumped region properties (modulus, CTE, poisson's ratio, and reference temperature).

Element Birth and Death

Element bird and death (EB&D) as it is known in ANSYS Mechanical software is an often used technique for modeling melting, solidification, bonding, thinning and other materials processing steps by which effect of addition or removal of materials at various temperatures are simulated. Details of this approach are well documented [13].

After creating the geometry and mesh for all parts of the final 3D IC package to be modeled, element death (ekill command) was used to artificially reduce stiffness of all elements belonging to the parts that are non-existent in the first step of the fabrication to negligible levels. Then through a combination of setting all element temperature to the process effective temperature and sequentially activating (ealive command) specific elements of the parts that are added by the process steps, the fabrication process is simulated. At each of the material addition step, the stress free temperature of the materials being introduced into the model are set to the temperature at which they are introduced. In the case of materials that appear pre-warped at the assembly step (for example due to thin film deposition), the stress free temperature may also be set to a different effective temperature [8]. This stress free temperature is used in conjunction with the Coefficient of Thermal Expansion (CTE) input to simulate thermal strain in material. Such individual material addition and stress free temperature assignment eliminates the need to assume a single effective stress free temperature for the entire assembly.

Table 3 and Table 4 show the sequential modeling steps involved in simulating D2D and P2D process steps of the package chosen.



Step 1 (230C to 20 C): Memory stack to logic die attach at solder cap reflow temperature (230 C) & cool . Kill substrate, C4 bump array and MC region elements at the start of this step.
Step 2 (20 C to 160 C): Ramp to micro pillar region underfill & cure temperature and modify the corresponding elements properties to account for underfill addition.
Step 3 (160 C to 20 C): Cool down to post process for warpage
Step 4 (20 C to 230 C): Ramp to substrate attach at C4 bump reflow temperature.
Step 5 (230 C to 20 C): Birth of C4 bump array and substrate region elements at the beginning of this step. Then Cool down to post process for warpage.
Step 6 and 7 (20 C – 160 - 20 C): Heat, underfill & cool down (similar to steps 2 & 3)
Step 8 (20 C to 160 C): Heat to MC encapsulation and Cure Temperature.
Step 9 (160 C to 20 C): MC elements birth at the beginning of this step. Room temperature warpage estimation after MC encapsulation.
Step 10 (20C): Change Modulus of MC to fully relaxed value (E_∞). Dwell at 20 C for estimating warpage change due to MC viscoelastic relaxation.

Table 3: Simulation Process Sequence for D2D Flow

Step 1 (230C to 20 C): Substrate to logic die attach at C4 bump reflow temperature (230 C) & cool . Kill memory stack, copper pillar array and MC region elements at the start of step.
Step 2 (20 C to 160 C): Ramp to C4 bump region underfill & cure temperature and modify the corresponding elements properties to account for underfill addition.
Step 3 (160 C to 20 C): Cool down to post process for warpage
Step 4 (20 C to 230 C): Ramp to memory stack attach at solder cap reflow temperature. Post process for attach stage warpage
Step 5 (230 C to 20 C): Birth of lumped pillar array region and substrate regions elements Cool down to post process for warpage
Step 6 and 7 (20 C – 160 - 20 C): Heat, underfill & cool down (similar to steps 2 & 3)
Step 8 (20 C to 160 C): Heat to MC encapsulation and cure temperature.
Step 9 (160 C to 20 C): MC elements birth at 160 C. Cool down to room temperature for warpage estimation after MC encapsulation.
Step 10 (20C): Change Modulus of MC to fully relaxed value (E_∞). Dwell at 20 C for estimating warpage change due to MC viscoelastic relaxation.

Table 4: Simulation Process Sequence for P2D Flow Steps 1, & 5 which Differ from Chip first flow are highlighted in different color)

Finally, due to the geometric as well as loading condition symmetry, quarter symmetry model was used in the simulation studies.

Results and Discussion

Logic chip and substrate warpage over the diagonal span of the attached region was reported from simulation runs. Warpage is defined as the maximum normal (Z) direction displacement of the probed point relative to the un-deformend state at the beginning of the processing. The convention used in this work considers the substrate bottom as the lowest Z co-ordinate location. Relative to the substrate, the logic die and the memory stack are stacked in the positive Z direction. Negative and positive signs of the warpage reported also follow the same convention.

The parts being assembled at the beginning of step 1 (at 230 C) were un-deformed based on an assumption that the stress free temperature of all front and back side films were also 230 C. The stress free temperature depends completely on the actual film processing temperatures and intrinsic stresses in the films at the time of deposition [8]. The process modeling methodology used by CielMech is capable of accounting for intrinsic stresses through specification of film stress free temperatures adding additional process steps prior to initial reflow attach step.

Figure 6 shows that for the D2D attachment process, the room temperature warpage after logic die to memory attachment was estimated to be positive. This is due to the higher effective CTE of the memory stack due to the high CTE underfill layers which are sandwiched between the silicon memory stack layers. An important contributing factor to this warpage directionality is the cumulative effect of all the film stresses at room temperature. The FSRDL oxide shrinks less than the silicon and the BSRDL polyimide shrinks more than the silicon after cool down from reflow due to their respective CTEs. This supports stack curvature in the same direction. So, one of the warpage mitigation knobs is to ensure that the thermal strains of the films on either side of the chip tend to fight each other to maintain a more or less flat profile after process steps. However, supply chain and business reasons may not always present such a mitigation knob. For example, if the front side is processed by one supplier and the back side integration is performed by another, both companies need to have tools in place to ensure matching CTE and process to realize similar intrinsic stresses in front and back side films. More importantly, the front side and back side will require different metallization due to inherent differences in routing needs between the two sides.

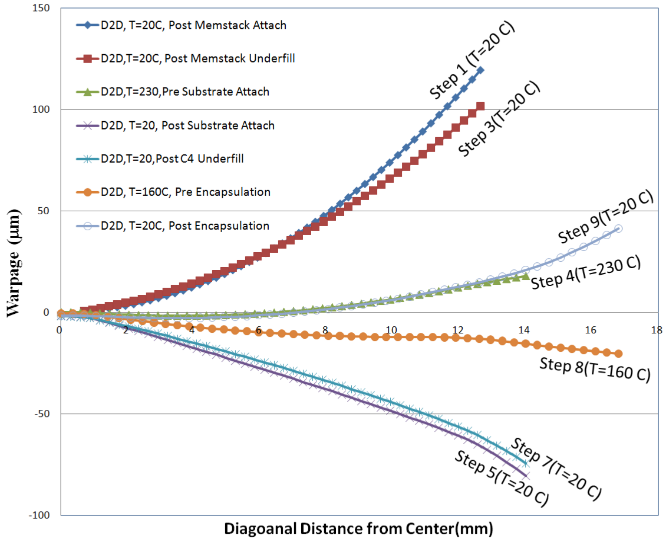


Figure 6: D2D Process Assembly Warpage at Various Process Stages.

From an assembly feasibility standpoint, the mitigating or matching curvature of warpages at attach temperature is most crucial. Due to the film, substrate and the silicon stress free temperatures being 230 C, the only contributor to warpage at this temperature is the underfill with its stress free temperature at the cure temperature of 160 C, and its CTE mismatch with silicon. As a result the warpage at chipsStack to Substrate attach temperature is relatively less at 18 µm. This can be further reduced by curing at higher temperature.

Subsequent cool down shifts the warpage to negative direction (-80 µm) due to the overpowering effect of the 400 µm thick substrate shrinkage. Finally, counter effect of encapsulation reduces the total package warpage after step 9 to +42 µm.

Figure 7 shows warpage evolution for the P2D assembly process. Since the process starts with substrate attach to 50 µm thin silicon, the overpowering effect of the substrate causes a 250 µm negative warpage. This warpage is progressively reduced due to the counteracting effect of the underfill and MC thermal strains in the subsequent steps. Again the attach temperature warpage is relatively low (-26 µm) due to the stress free temperature of all parts of the assembly other than C4 underfill being the same as attach temperature.

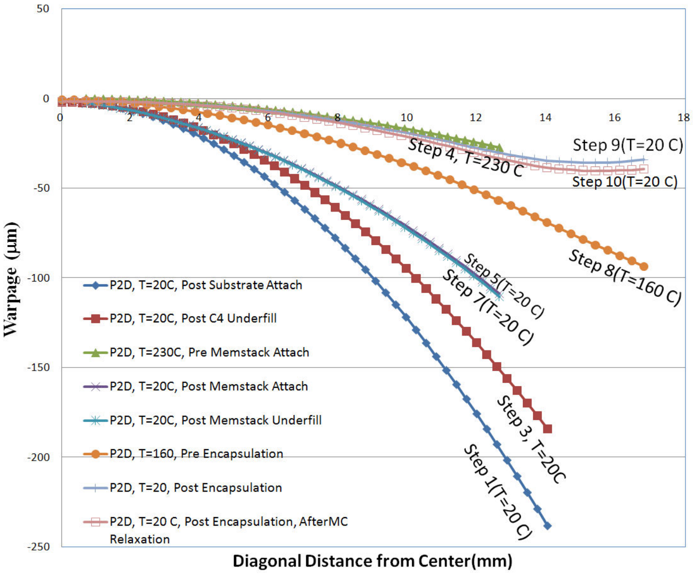


Figure 7: P2D Process Assembly Warpage at Various Process Stages.

Comparison of D2D and package first process flow shows that final warped state is completely different even though the final package structures are identical. This is due to the fact that some crucial warpage influencing process steps such as MC encapsulation happened at an already warped state which were different in the two process flows. For example the warpage at the point of MC encapsulation (step 8) was -20 µm for D2D flow compared to -94 µm for package first flow. Eventhough the cool down after MC encapsulation step resulted in similar amplitude change in warpage of about 110 µm, the difference in encapsulation stage curvature resulted in a different final warped state. Such path dependence will not be captured if geometric non-linearity effects are not included in the FEA simulation. This was accounted in this work through the use of 'nlgeom' command in ANSYS (also known as large deformation effects).

It should be noted that MC, underfill and substrate materials are viscoelastic in nature. In order to understand the potential magnitude of warpage change over time due to viscoelastic relaxation, Viscoelasticity should be included along with MC shrinkage during cure. From a simple and accurate methodology development standpoint, it is desirable to evaluate the relative sensitivity of warpage to these nonlinear effects. For example, ignoring underfill viscoelasticity in global models may have been a justifiable simplification while estimating package warpage in traditional thicker dies due to its low volume compared to that of the chip and the substrate. Whether such simplification is still valid in the context of thinned dies need to be re-evaluated through sensitivity studies. Such studies will result in a body of knowledge which will be useful for establishing simpler modeling practices which don't compromise accuracy.

In this work a first cut estimate of the change in package warpage due to MC relaxation was made by reducing the room temperature MC modulus by 10% after the final room temperature stage and solving for one more load step at constant temperature. This is to simulate the change in warpage due to viscoelastic relaxation of the MC during the long dwell period after packaging and before attachment to PCB. The choice of 10% modulus decrease was based on sub Tg relaxation data reported for MC and underfill materials in literature [14 and 11]. Results showed that the warpage increased by about 10% to -35.5 µm to -40 µm due to the decrease in counteracting moment due to reduced MC stiffness.

Conclusions

Step by step evolution of warpage of a 3D IC was studied using FEA simulation. Results showed that the assembly temperature warpage for second reflow was relatively low compared to room temperature warpage. However, even the 20 µm range warpage at reflow temperature may require some external force to avoid opens during micropillar attachment. Since the micropillar gaps (25 µm to 50 µm) are much smaller than traditional C4 bump heights (~100 µm), it is even more important to balance film as well as other stresses to ensure assembly with minimal external force. Results also showed that counteracting strains in MC and substrate tend to reduce final state warpage. Since the newer substrate materials with significantly lower CTE (6 top 8 PPM) are being introduced, it is important to appropriately match the MC dimensions as well as material to appropriately balance these counteracting forces. Comparison of the D2D and P2D process warpages also indicated a strong path dependency in the final warpage of the package.

References

1. Zhai, C. J., et al. "Reliability Modeling of Lidded Flip Chip Packages", Proceedings of the 57th IEEE Electronic Components & Technology Conference, 2007.
2. Besser, P. and Zhai, C. J., "Modeling and Measurement of Stress and Strain Evolution in Cu Interconnects", Seventh International Workshop on Stress-Induced Phenomena in Metallization, June 2004
3. Darveaux, Robert, "Effect of simulation methodology on solder joint crack growth correlation", Proceedings of the 50th IEEE Electronic Components & Technology Conference, 2000.
4. Wang et al, "Packaging effects on reliability of Cu/low-k interconnects", IEEE Transactions on Device and Materials Reliability, Vol 3, Issue 4, Dec 2003
5. Syed, Ahmer, R., "Accumulated Creep Strain and Energy Density Based Thermal Fatigue Life Prediction Models for SnAgCu Solder Joints", Proceedings of the 54th IEEE Electronic Components & Technology Conference, 2004.
6. Patti, R., "Chapter 4: Homogenous 3D Integration" in "Three Dimensional System Integration. IC process Stacking and Design" Edited by Papanikolaou, et. al, Springer, 2011
7. Zhang, G. Q. et al, "Mechanics of Microelectronics", 2006, Springer.
8. Zhai, C. J., et al., "Process-Oriented Stress Modeling and Stress Evolution During Cu/Low-K BEOL Processing", MRS Proceedings, Volume 812, 2004
9. Park, Seungbae, et al., "Predictive Model for Optimized Design Parameters in Flip-Chip Packages and

Assemblies", Ieee Transactions on Components and Packaging Technologies, vol. 30, No. 2, June 2007

10. Perkins, Andrew E., "Investigation And Prediction Of Solder Joint Reliability For Ceramic Area Array Packages Under Thermal Cycling, Power Cycling, And Vibration Environments", Ph. Dissertation, Georgia Institute of Technology, May 2007.
11. Park, Soojae and Feger, Claudius, “Underfill Fracture Toughness as a Function of Cooling Rate”, Proceedings of the 58th IEEE Electronic Components & Technology Conference, 2008.
12. Wang, G. Z. et al., "Applying Anand Model to Represent the Viscoplastic Deformation Behavior of Solder Alloys", ASME Journal of Electronic Packaging, 2001, Vol. 123, pp 247-253.
13. “ANSYS 14.0 Help System”, ANSYS Incorporated, 2011.
14. Xingjia Huang; Lee, S.W.R., "Stress relaxation in plastic molding compounds," Electronic Materials and Packaging, 2002. Proceedings of the 4th International Symposium on , vol., no., pp.37,42, 4-6 Dec. 2002