The Increasing Role of Polymers in Advanced Packaging – From Stress Buffer Layers to Wafer Level Underfills and Beyond

Andy Miller*, Kenneth J. Rebibis, Fabrice D. D. Duval, Teng Wang, John Slabbekoorn, Joeri De Vos, and Eric Beyne

IMEC's 3D Industrial Affiliation Program, Imec vzw, Kapeldreef 75, Leuven, 3001, Belgium
*Andy.Miller@imec.be

With the continued expansion of computing capability and reduction in technology node for advanced logic or memory devices the challenges encountered due to packaging also increase. Driven by Power, Performance and Area requirements of sub 14nm node devices and their associated system architectures, packaging solutions must also enable optimum performance whilst maintaining an affordable approach, especially within the mobile applications market. Due to this, the role of polymers within the integrated system of multi die packaging becomes ever more important. Polymers have become a vital enabler from both a technical and cost reduction viewpoint, where they have a role to play in various position within the integration scheme. Stress buffers layers are common, but we now move into the era of commercialized wafer level underfills, the use of silicone style materials for Chip Scale Packaging (CSP) to enable more flexible interconnects, Temporary Bonding Materials (TBM) compatible with alternative polymer materials, even pushing to the point of TBM compatibility with overmold materials. Even with all this research the thermal aspects of performance computing means we must also seek improved Thermal Interface Materials (TIMs) in addition to all of the afore mentioned materials.

Keywords: Underfills, Fan Out Wafer Level Packaging (FOWLP), Microbumps, Overmold, Thermal Interface Materials

1. Introduction

Within semiconductor processing polymers have traditionally been associated with certain specific steps enabling a low cost approach to passivation or reliability improvements. Materials such as Polyimides, BCB and even spin on Low k dielectrics have long been utilized to enable certain critical applications. With the emergence of DRAM stacks, device stacking on interposers and eventually combinations of Logic, DRAM and sensors to satisfy the Internet of Things (IoT) explosion require further focus on the use of polymers. To satisfy the increased electrical connection speed of these new system architectures we have been required to combine the more traditional transistor and local interconnect scaling with additional scaling in the 3rd dimension.

This 3rd dimensional scaling has seen the development of high aspect ratio Through Silicon Vias (TSVs), the associated die thinning (to 50 μm or less of Silicon) and the reduction of traditional bump interconnects down to pitches of 40μm currently and below 10 μm in the future. The ability to scale the μBump pitch is one of the key driving factors for improved polymer approaches, especially as we go below 10μm pitch. When considered with the differing

Received December 2, 2016
Accepted December 23, 2016
bonding techniques, the material properties of Wafer Level Underfills (WLUFs) become even more specific from the perspective of process capability, but also from the perspective of compatibility with other materials required for the process.

2. Integration options and complexity

The 3D landscape is an increasingly complex area, combining sensors, logic devices, stacked memory and interposers. How these different devices are combined is highly application specific. As a result the co-existence of 2.5D and 3D processing is a fundamental requirement from the perspective of a system designer. What this means in real terms is that the integration scheme from the perspective of packaging must be specific to the application. Something which works for High Performance Applications systems will not be appropriate for the mobile market due to higher cost requirements, use of TSVs and combinations of die with interposers. The corollary of this is die aimed at the mobile market. These die must be commodities, made in huge numbers and thus must be extremely low cost to be adopted by a mobile device manufacturer. Figure 1 describes a typical construct of a High Performance module utilizing Si Interposer, DRAM cube, Logic device and associated convertors and Input / Output chips.

![Fig. 1. Conceptual diagram of a potential High Performance module.](image)

Figure 2 shows how this concept has been brought to reality in the past few months by various chip design houses, in this case NVidia but before them AMD etc.

This kind of device illustrates the combination of 2.5D (interposer style stacking) alongside 3D (DRAM stacking through the High Bandwidth Memory module) stacking bringing the concepts to reality. If contrasted with the latest mobile market options the dual targets of low cost and lower performance can clearly be seen.

![Fig. 2. NVidia Tesla P100 depicting central processing chip surrounded by 4 High Bandwidth Memory modules on a passive Si interposer.](image)

With this approach of Package on Package (PoP) as illustrated in Fig. 3 there is no need for expensive TSV technology or controlled heat dissipation. However the high speed interconnects required are unsupported in this format, thus reducing the performance of the device package. To enable the extension of this technique towards higher performance capability the introduction of PoP Fan Out Wafer Level Packaging as shown in Fig. 4 thrust the world of overmolding materials into the limelight once again.

![Fig. 3. Mobile device architecture utilizing Package on Package technology.](image)

![Fig. 4. PoP FOWLP: Die embedding logic die in package substrate or wafer-level fan-out processing on reconstructed wafers.](image)
With this approach industry has been able to pack additional functionality onto a low cost package construct. This can be further extended by embedding 2 FOWLP packages in a reconstructed wafer onto the processor die as shown in Fig. 5.

Unfortunately this technique cannot currently provide the high density interconnect connections required by Wide I/O memory due to the classic challenges encountered by FOWLP of reconstructed wafer warpage affecting resolution capability of RDL etc. Ultimately this connection requirement could be enabled by the cartoon outlined in Fig. 6 – Flip Chip FOWLP.

As can be noted, this approach places significant effort on materials interactions. RDL layers, high density connection enabling WLUFs, overmolds enabling this reconstructed wafers, secondary layer overmolds for embedding memory die and even the possibility of Chip Scale Packaging (CSP) enabling materials such as silicones. With this in mind we will first discuss the evolution of WLUF.

3. The Evolution of underfills and bonding techniques

Traditionally underfill materials have been applied via small volume dispense, relying on capillary action to fully cover the surface between two die. With the continued scaling of microbumps the use of capillary underfill ceases to be a viable option. Figure 7 illustrates the relationship between Chip I/O bump area and microbump pitch.

With this reduction in microbump pitch comes in reduction in microbump height and thus a reduction in the gap which must be filled by the underfill material.

| Technology Node (nm) | Microbump Pitch (µm) | Resulting UF Gap (µm) | Underfill Type for die stacking |
|----------------------|----------------------|-----------------------|-------------------------------|
| 65                   | >50                  | 30                    | Capillary (applied post bonding of dies) |
| 32                   | 40-50                | 20                    | No Flow (pre-applied on die)     |
| 14                   | 10-20                | 10-15                 | Laminated Film – (pre-applied on wafer) |
| 7                    | 1-10                 | <10                   | Spin on – (pre-applied on wafer) |

Once pre-applied underfills become a requirement, the introduction of Thermal Compression Bonding (TCB) (as opposed to reflow bonding) also becomes a necessity to enable both alignment requirements but also to ensure no underfill material is trapped between the bumps to be connected. Additionally, as the pitch reduces the distribution of filler particle becomes ever more challenging to ensure homogeneous filler content supporting the stacked interfaces as shown in Fig. 8.
Fig. 8. XSEM micrograph of a 20µm pitch stacked microbump with uniformly distributed filler particles.

To enable full industry adoption of this stacking technique a complete bond cycle of <5 seconds is being requested by the primary device manufacturers. As such 2 seconds for Die pick and alignment is acceptable, leaving 2 - 3 seconds for complete placement, thermal ramp and force application. This requirement means the underfill material must snap cure within 2 secs when we use Die to Wafer (D2W) bonding. Further considerations must also be taken into account when performing multi die stacking. Figure 9 illustrates a multi die stack with underfill between each stack.

Fig. 9. SEM micrograph of multi die stack where each stacked die is 50µm thick.

In this scenario we must consider whether the same underfill material may be used for each layer, or different underfills pending the temperature seen by each underfill at their distinct bonding interface.

The next challenge facing stacking is the impact of the continued pitch reduction of microbumps to below 10µm pitch. In this case the D2W bonding equipment is at the edge of stacking alignment capability. Thus wafer level bonders, with much improved alignment capability enter the equation. In this scenario two wafers are bonded directly together, enabling extremely fast stacking of die. The impact for the underfill material is significant however, with the ability of the bond chamber to ramp to the desired temperature now taking several minutes rather than seconds. Thus the material properties of underfills used in wafer to wafer bonding must be fundamentally different in terms of curing and outgassing than those used in D2W bonding.

4. Chip package interaction and chip scale packaging

Chip Package Interaction (CPI) is a vital area of research for the chip making industry. Effectively all chips are packaged in one way or another and understanding how the chip interacts with the different Printed Circuit Boards (PCB) is critical to the reliability of the devices at point of use. A key requirement is to reduce the Coefficient of Thermal Expansion (CTE) mismatch between the Si chip and the PCB. Again, as interconnect pitch scaling continues, the impact of this mismatch increases. As such we must generate methods to reduce the CTE mismatch, an area polymers can excel in. In particular the high performance computing market is a key driver for CPI understanding. Figures 10 and 11 depict failure analysis XSEM micrographs of the impact of packaging without using stress reduction techniques in the device build up.

The build up of the metal layers leading to the final connection pad is a key area allowing polymers such as stress buffer layers to reduce the lateral forces incumbent upon the BEOL. The use of a thick RDL layer in combination with an appropriate stress buffer polymer allows the Cu pillar to withstand a much larger shearing force than a simple Cu pillar built upon a thin RDL, as illustrated in Fig. 12.

A natural extension of this approach is to then design the RDL in terms of being a “springlike” structure, further enabling stress dissipation. This structure, termed within IMEC as a “Flexi-Bump” is currently under development. This system involves the use of multiple polymer layers and RDLs being formed into a spring, thus minimising the impact of lateral shearing force still further.
Fig. 10. Low magnification SEM micrograph of stress induced Back End Of Line (BEOL) cracking in the vicinity of a Cu pillar.

Fig. 11. High magnification SEM micrograph of stress induced Back End Of Line (BEOL) cracking in the vicinity of a Cu pillar showing underfill filler entrapment.

Fig. 12. Thick RDL and polymer buffer layer enabling stress dissipation of lateral shearing force.

Fig. 13. Cross sectional cartoon of potential Flexi-Bump style construct.

Figure 13 shows a potential X-Sectional build up of such a Flexi-Bump. This type of bump was built and then subjected to non-destructive shear testing to enable understanding of the viscoelastic deformation. Figure 14 shows the Cu pillar bump subjected to greater than 4 µm of lateral movement without damage to the bump or RDL.

Fig. 14. Non-destructive shear testing result of IMEC’s Flexi-Bump construct.

When considering the implications of flexible bumping with respect to packaging, the concept clearly extends itself to Chip Scale Packaging (CSP). In this case we can attach the chip directly to the PCB without the need for a package substrate. This approach requires greater flexibility of the bumps or pillars to take into account the even greater CTE mismatching which is a direct consequence of the package substrate removal. In this case rather than the use of Cu pillars,
the flexible connection via RDL layers embedded within stress absorbent materials are connected to solder balls as shown in Fig. 15.

Fig 15. Cartoon image of flexible RDL with solder ball attached.

When designed appropriately this type of flexible RDL and solder ball attachment allows direct placement of the Si device onto the PCB as shown in Fig. 16.

Fig 16. Picture of 5 imec chips directly placed onto PCB via flip chip attachment without the use of package substrates.

5. Fan out wafer level packaging (FOWLP)

Fan Out WLP is a concept that has been utilized by the industry for several years. Initially pioneered by Infineon with their eWLB (embedded Wafer Level Ball Grid Array) process, the approach has seen a proliferation of competitors in the last 2 – 3 years all aiming to reduce the cost whilst increasing the performance of their devices. Additionally the number of FOWLP packages appearing in mobile devices is also beginning to increase and is predicted to do so in an increasing fashion in the coming years (as illustrated in Fig. 17 from Yole Development [2]).

Fig. 17. Evolution of number of FOWLP packages in a mobile phone.

This new, low cost approach, has brought with it many challenges to the traditional semiconductor device manufacturing process line.

- Warpage of reconstructed wafers
- Reliability constraints
- Processing temperature constraints
- Failure analysis.

The warpage of the reconstructed wafers has been the primary focus for the process technology development teams over the recent years. As shown in Fig. 18 [3] the final thickness of the reconstructed wafer has a significant impact on the warpage, which in turn impacts the overall form factor of the package. As a result new methods must be investigated to enable thinner devices within the FOWLP approach.
Fig. 18. Impact of wafer thickness on warpage of a reconstructed wafer.

This warpage has significant impact on equipment chucking capabilities plus resolution limits for lithographic patterning. As such reducing the overall warpage of the reconstructed wafers is fundamentally required to extend FOWLP into the fine pitch RDL domain. The idea of utilizing flip chip technology to extend FOWLP into the mid performance regime is now something being researched as part of IMEC’s 3D program. Figure 19 (a similar but slightly different construct to that shown in Fig. 6) illustrates a potential package construct which enables the extension of FOWLP into the Wide IO connection domain, but requires extremely good control of warpage for the reasons outlined.

As can be imagined, this approach places many challenges onto the materials involved.

6. Thermal Interface Materials

The final piece of the jigsaw in terms of materials for advanced packaging is the Thermal Interface Materials (TIMs). These materials are designed to dissipate heat from the working devices as efficiently as possible. Figure 20 shows an example of where these materials fit into a package.

Fig. 19. Cartoon of potential high density interconnect Flip Chip FOWLP construct under investigation within IMEC’s 3D Program.

Fig. 20. Cartoon depicting location of TIM in classic package construct.

Here we see the 1st TIM placed between the chip and the heat spreader, with the 2nd TIM placed between the heat spreader and the heat sink. For high performance device cooling the impact of the TIMs actually becomes the dominant factor (see Fig. 21).

Fig. 21. Impact of TIMs on Thermal Resistance vs other thermally dissipating elements of the package.
As can be seen, this means we need significant research into both bulk and surface contact properties of TIMs to ensure cooling capabilities consistent with high performance devices.

7. Summary

With continued scaling of devices and the increased performance expectations on single digit node chips, packaging requirements also must increase in complexity. Polymers of varying types and classes can offer several low cost and flexible solutions to critical issues encountered within advanced packaging. Within this paper we have discussed several key areas of research enabling advanced packages to keep pace with front end scaling. Many different integration approaches can be envisioned, covering several different types of polymers such as wafer level underfills, low temperature curing materials, overmolds and bonding materials and even thermal interface materials are all required to successfully package the different types of devices needed for the new applications related to advanced packaging and even the Internet of Things (IoT). IMEC’s 3D System Integration Program is investigating many of these approaches and materials with the help of our equipment and material supplier partners.

References

1. G. Beyer, “Approaches, advantages and disadvantages to underfills”, Assembly and Packaging Challenges, IMAPS 2016, Arizona, March 2016
2. R. Beica, “3D packaging at the Forefront of the Semiconductor Industry”, 3D TSV Summit, Grenoble, 2016.
3. M. Prashant, S. W. Yoon, Y. Lin, and P. C. Marimuthu, Cost Effective 300mm Large Scale eWLB (embedded wafer level BGA) Technology, 13th EPTC 7th – 9th December 2011.