Wafer Level Packaging (WLP): Fan-in, Fan-out and Three-Dimensional Integration

Xuejun Fan
Department of Mechanical Engineering
Lamar University
PO Box, 10028, Beaumont, Texas, USA
xuejun.fan@lamar.edu

Abstract

In this paper, the state-of-the-art results of research and development in wafer-level packaging (WLP) is reviewed. The paper starts from the introduction of several fan-in wafer-level packaging technologies. The focus is given on the fan-in WLP reliability performance as related to the structural differences. New failure mechanisms that appear in build-up stack layers are discussed. Next, emerging fan-out wafer level packaging technologies are introduced. Several key challenges in fan-out WLP technologies are examined. Finally, Three-dimensional (3-D) integration of through-silicon-via (TSV) technology and wafer-level bonding technology with WLP, especially in MEMS and image sensor applications, is discussed.

1. Introduction

Demand for wafer level packaging (WLP) is not only driven by the need to shrink package size and height, simplify the supply chain and provide a lower overall cost by using the infrastructure of a batch process, but also for performance reasons [1]. Wafer-level packaging offers the benefits of flip chip in that the chip is face-down to the printed circuit board (PCB), enabling the shortest electrical path. Increased speed and reduced parasitics are the positive results of this method. Cost reduction is another driving force behind wafer-level packaging. Devices are gang packaged, in that the entire wafer is packaged all at once. The cost for packaging the devices on a given wafer does not change with the number of dies per wafer, as all processes are additive and subtractive steps performed with mask steps.

In general, there are two categories of WLP technologies: 'fan-in' and 'fan-out' wafer level packages (WLPs) [2]. Conventional (fan-in) WLPs are formed on the dies while they are still on the uncut wafer. The final packaged device is the same size as the die itself. Singulation of the device occurs after the device is fully packaged. Thus, fan-in WLPs are a unique form of packages and have the distinction of being truly die-sized. WLPs with fan-in designs are typically for low input/output (I/O) count and smaller die sizes.

In the last few years, 'fan-out' WLPs have emerged. Fan-out WLP starts with the reconstitution or reconfiguration of individual dies to an artificial molded wafer. The molded reconstituted wafer forms a new base to apply a batch process that features build-up and metallization constructions, as in the conventional fan-in WLP back-end processes, to form the final packages. Figure 1 shows the comparison of a fan-in versus a fan-out WLP.

Historically, the use of WLPs enabled Motorola to introduce its RAZR phone, which became the thinnest mobile phone available at the time when it was introduced [3]. With the increased price for gold wires, some companies are also considering to adopt WLPs as a low-cost alternative compared to wire-bond packages, especially for higher pin count devices. Recently, WLPs have also been widely used in image sensor applications. Through-silicon-via (TSV) technology has been incorporated with WLP solutions for packaging image sensors. Many developments are underway to realize three-dimensional (3-D) integration with WLP technologies [3].

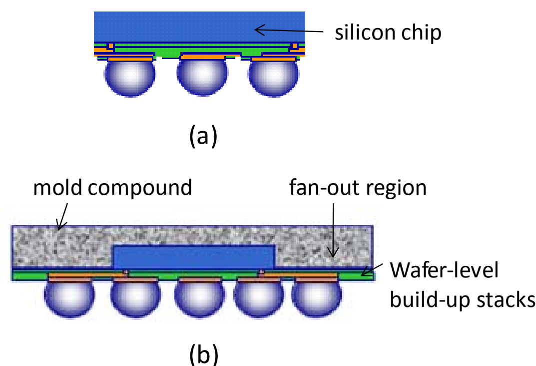

Figure 1. A fan-in WLP versus a fan-out WLP: (a); fan-in WLP; (b) fan-out WLP

In this paper, wafer level packaging technologies including fan-in, fan-out WLPs, and 3-D integration are reviewed. A variety of fan-in WLP technologies, such as ball on nitride (or ball on I/O), ball on polymer, and copper post WLPs, are described. The solder ball reliability and thin film/redistribution copper trace failures under various loading conditions are discussed. Key processes and challenges of fan-out WLP technologies are examined. The developments of 3-D integration in WLP technology are reviewed.

2. Fan-in WLPs

2.1 Overview

Conventional (fan-in) WLPs are formed on the dies while they are still on the uncut wafer. The process can be thought of as an extension of front-end manufacturing in that it involves the entire wafer, but is more similar to bumping processes for flip chip packages. The final packaged device is the same size as the die itself. Thus, fan-in WLPs are a unique form of packages and have the

978-1-4244-7027-310/$26.00 ©2010 IEEE

distinction of being truly die-sized, not "chip scale." For WLPs, underfill is typically not applied. Since there exists an intrinsic difference in the coefficient of thermal expansion (CTE) between silicon chip (~2.6ppm/°C) and PCB (~17ppm/°C), solder ball reliability under thermal cycling loading is apparently limited by die-size [4-7]. The larger the die size is, the greater thermal stresses are developed at the outmost solder balls due to the effect of distance from neutral point (DNP). To improve solder ball reliability performance, several fan-in WLP technologies have been developed [2], such as ball on nitride (BON) [8], ball on polymer (BOP) [8-10], and copper post WLP [11]. Although ball on nitride (or ball on I/O) is seldom used in today's applications, it will be introduced first in the following as a benchmark to compare with other WLP configurations.

Figure 2 shows a redistributed bump on nitride (BON) WLP structure, consisting of solder bump and under bump metallurgy (UBM) seated on the thin inorganic passivation [8,10]. In the case of WLPs with a BON structure (or ball on I/O), the added redistribution layer and the passivation layer do not provide additional benefit to solder joint reliability performance since the solder ball is directly connected to the silicon base. In this case, solder balls become the weakest link under thermal cycling loading conditions. It has been reported that this WLP structure is limited to 6x6 array size (or less) at 0.5mm pitch (~3mmx3mm die-size) to meet the thermal cycling reliability requirement [8]. The predominate failure mode has been fatigue crack propagation in bulk solder near solder ball/die interface, as shown in Fig. 3 [12].

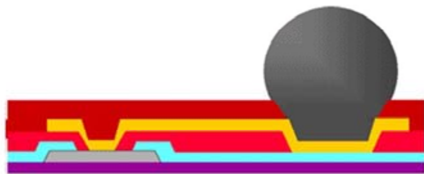

Figure 2. Bump on nitride (BON) stack-up structure

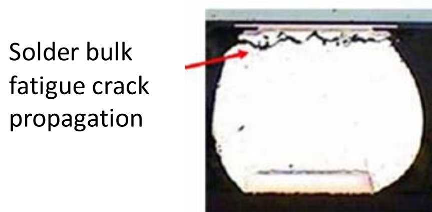

Figure 3. Typical solder bulk fatigue crack propagation in thermal cycling

Fig. 4 shows a schematic diagram of ball on polymer (BOP) WLP structure without UBM layer [2]. The redistribution copper traces allow a process without UBM since the diffusion barrier requirements of the UBM are no longer needed. In the case of the ball on polymer, the bump rests on the polymer film, and thus any stress applied to the solder ball will directly propagate to the underlying polymer film. The common materials for polymer films are polyimide (PI) or benzocyclobutene (BCB), both of which are extremely complaint. Polymer films serve two purposes, passivation for the redistribution layer (RDL) and stress buffer. Because polymer films are very compliant, stresses will be partially 'absorbed' by the films during thermal cycling. As consequences, potential failures might occur at film/copper trace build-up stacks other than in solder balls.

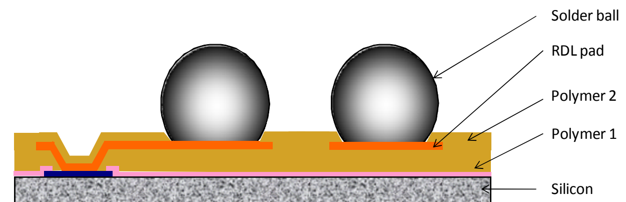

Figure 4. Bump on polymer (BOP) without UBM stack-up structure

Fig. 5 is a schematic of WLP structure for ball on polymer with UBM layer. The UBM now functions only as an adhesion layer that facilitates the bump electroplating process. In this case solder balls also sit on a polymer film layer to avoid a direct connection with the silicon base. Similar to the WLP without UBM, although there may be a concern on failures at film/copper trace stacks, it has been demonstrated that with BOP WLP structures with or without UBM layer (Figs. 4 and 5), the array size can be extended to 12x12 with 0.5mm pitch meeting reliability requirement. In other word, the die size is now extended to 6mmx6mm and the ball count to 144 from the benchmark design of BON WLP of 3mmx3mm and the ball count to 36 [2,8].

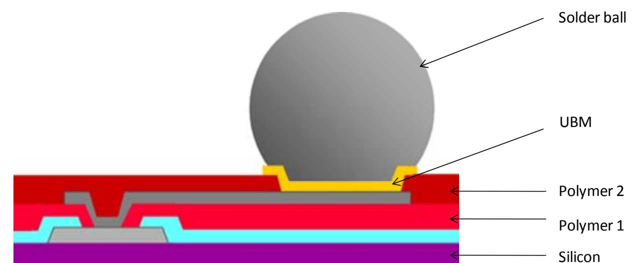

Figure 5. Bump on polymer (BOP) with UBM stack-up structure

Fig. 6 shows a schematic of a copper post WLP structure. Thick copper pillars (~70 µm) are electroplated, followed by an epoxy encapsulation. In this case, solder balls rest on the copper post. The standard process uses 100µm thick photoresist to form ~70 µm copper pad and ~35 µm tin plating. The copper post WLP can also incorporate with redistribution layer, as shown in Fig. 6. In the case of a copper post WLP even without redistribution layer, it has been demonstrated that the copper post structure has superior thermo-mechanical reliability performance. The array size can be extended to 12x12 with 0.5mm pitch [11].

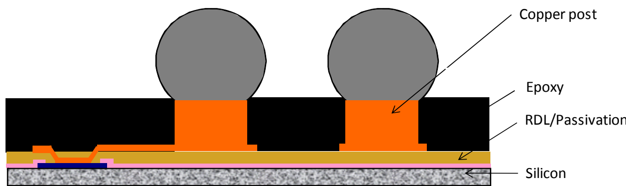

Figure 6. Copper post stack-up structure

There are also other fan-in WLP strictures, such as double bump WLPs and compliant layer process [13,14], which can be applied for large array applications. The underlying mechanism to improve thermal cycling reliability of fan-in WLPs is to make the WLP structures more flexible so that the stresses transmitted to solder balls can be reduced [2].

2.2 Reliability Analysis and Failure Mode

To understand the mechanism of reliability for various WLP structures, first, the attention is confined to solder ball fatigue failures under thermal cycling conditions. Finite element modeling is performed to investigate the accumulated inelastic strain energy density at solder ball/chip region subjected to -40°C and 125°C thermal cycling [2]. Fig. 7 plots the per-cycle inelastic strain energy densities for the four fan-in WLP structures, for a 12x12 array package with 0.5mm pitch. It can be seen that, compared to the ball on nitride structure, all other three structures: BOP without UBM, BOP with UBM, and copper post WLP, show more than 30% reduction in terms of the accumulated inelastic strain energy density per cycle. This means that, with the incorporation of a dielectric polymer film or an encapsulated copper post layer embedded in an epoxy, between solder balls and chip, the stresses in solder joints can be reduced significantly compared to a 'rigid' ball connection as in a BON configuration.

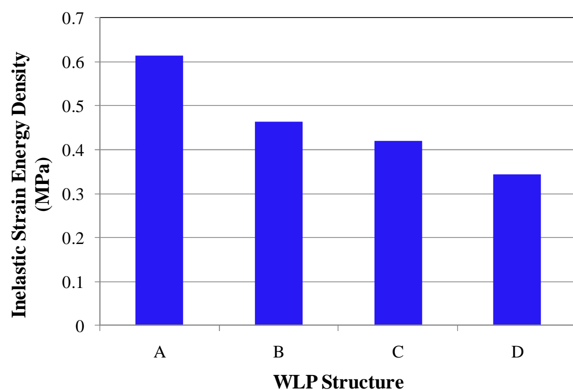

Figure 7. Inelastic strain energy density for different WLP structures: A. BON; B. BOP without UBM; C. BOP with UBM; and D. Copper post [2].

For ball on polymer structures, the extreme compliance of the polymer film is attributed to be the reason for thermal-mechanical performance improvement in solder joints. The Young's modulus of polyimide film is 1.2GPa, which is one order lower than the modulus of solder alloy (50GPa for SAC305). The polymer film creates a 'cushion' effect to reduce the stresses transmitted to solder joints. Studies have shown that the coefficient of thermal expansion (CTE) of the polymer film has insignificant effect on solder joint stresses provided that the polymer film modulus is extremely low.

On the other hand, for copper post WLP structure, the beneficial effect comes from the larger CTE of copper post and epoxy, which are typically 17×10^{-6}/°C and 20×10^{-6}/°C, respectively. The combined silicon chip and epoxy/copper post stack-up can be thought of as a 'molded' die with an effective CTE (in Fig. 8), which will be significantly greater than the CTE of the silicon die itself (2.6×10^{-6}/°C). This results in a significant reduction of the stresses on solder joints. For a copper post WLP, the redistribution layer (RDL) may be incorporated if needed. However, It has been found that the effect of the polymer film in copper post WLP is not as effective as that in BOP structures [2]. This indicates that for the copper post WLP structure, the dominant effect to reduce the solder joint stresses is due to the larger CTE of copper and epoxy. The modulus of copper/epoxy and the appearance of the RDL (polymer film) are of the secondary effect in solder joint reliability improvement.

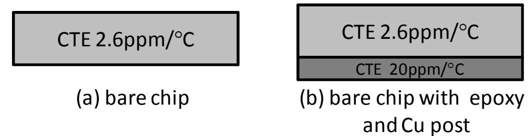

Figure 8 Effective CTE increase of the 'molded die' in copper post WLP

As shown in Fig. 3, the predominant failure mode for BON WLP structure is solder bulk fatigue crack propagation under thermal cycling. However, for BOP and copper post WLP structures, the copper interconnect reliability might be more important than solder joint reliability. For example, for a BOP WLP, studies have shown that the failures were predominantly on copper RDL trace cracks at the component side under drop and thermal cycling test [12,15,16]. The copper RDL failures were found along the 2 outer rows of I/O in a JEDEC board set up subjected to drop (Fig. 9). Fig. 10 shows an example of Cu/UBM delamination using dye & pry, which correlates to cross sectional pictures. This means that the failure mode in BOP WLPs may shift to wherever the weakest link of the system from solder ball regions. Nevertheless, the overall reliability performance of BOP WLP structures has been greatly improved as compared to the BON WLP structure.

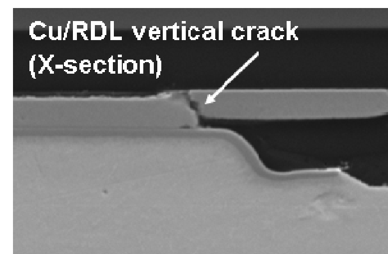

Figure 9. Copper RDL trace failure under drop test [12]

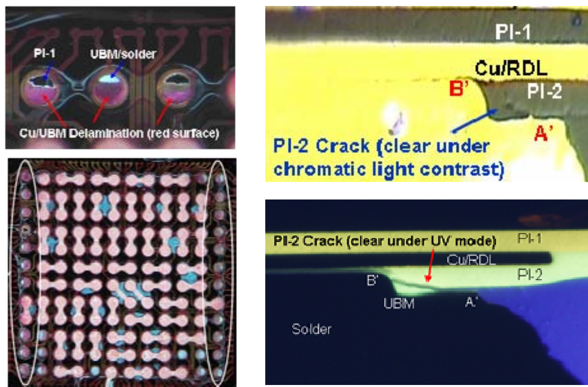

Figure 10. Polymer film/UBM delamination [15]

For wafer level packages, PCB is considered as 'part' of the package since one cannot decouple the PCB from the WLP. PCB design plays an important role to assess the reliability of WLPs. With the conventional JEDEC board test set up and design, PCB trace cracks were often observed at locations near the outer row balls. All failures occurred in the PCB traces approach in the longitudinal direction under drop test (Fig. 11). This is because the traces in the longitudinal direction are suffered more mechanical stresses than other directions under drop. After the PCB trace direction was changed from a longitudinal routing to trace latitudinal routing, Cu trace failures in PCB can be eliminated [16]. It is also important to use low-CTE PCB board, which will improve the WLP reliability under thermal cycling significantly [2].

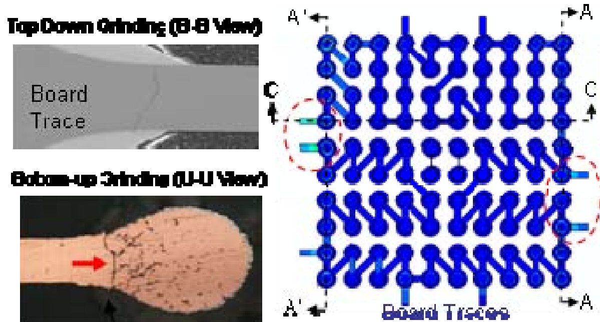

Figure 11. PCB copper trace failures [16].

3. Fan-out WLPs

As opposed to a conventional fan-in WLP, fan-out WLPs start with the reconstitution or reconfiguration of single dies to an artificial molded wafer. The fan-out WLP has received increased attention because of the demand for thinner features and increasing I/O count devices. The reconfigured wafer or fan-out solution provides several advantages:

- Reduced package thickness,
- Fan-out capability (for the increased number of I/O),
- Improved electrical performance,
- Good thermal performance, and
- A substrate-less process.

Fan-out WLPs are structurally similar to the conventional ball grid array (BGA) packages, but eliminate expensive substrate processes. Fig. 12 shows a typical process of wafer reconstitution. The "good tested" dies of a silicon wafer are placed face-down onto a carrier with an adhesive tape. The distance (pitch) between the dies on the carrier defines the fan-out area around the chips and is freely selectable. The carrier with the adhesive tape holds the dies in position and protects the active side of the dice during molding. A mold compound is used to combine the placed silicon dies to wafer format in compression mold technique. After this, the reconstituted wafer is released from the carrier system, which can be re-used afterwards. Then the redistribution layer and final bump processes are applied. Fig. 13 is a picture of reconstituted wafer. Several fan-out WLP technologies have been developed, such as embedded wafer-level ball grid array (eWLB) technology, and redistributed chip packaging (RCP) technology [17,18].

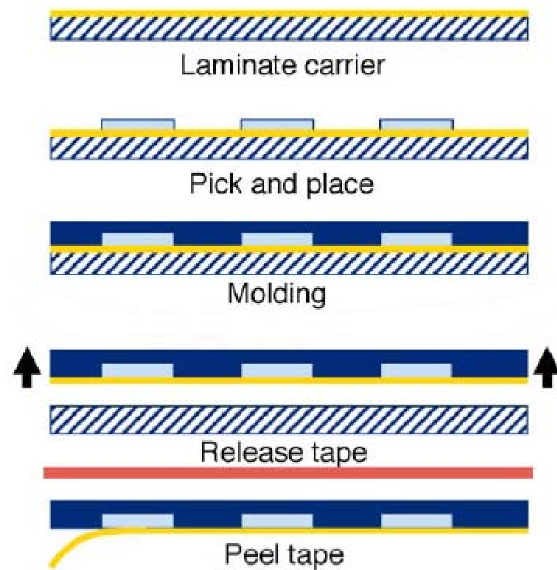

Figure 12. A typical process of wafer reconstitution [17]

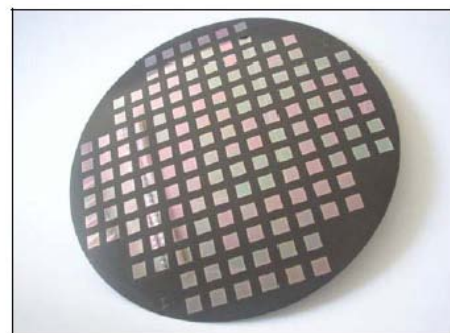

Figure 13. A picture of reconstituted wafer [18]

Since a fan-out WLP is structurally similar to a BGA package, solder joint thermo-mechanical behaviors in a fan-out WLP is similar to a BGA package. The critical solder balls in a fan-out WLP are located beneath silicon chip area, where the maximum CTE mismatch occurs

between the silicon chip and PCB [19]. In Fig. 14, the per-cycle inelastic energy density is plotted against the location of solder balls in a diagonal direction for a 16×16 array fan-out WLP package, in which 6×6 array solder balls are under die area. Fig. 14 shows that outermost ball right beneath silicon die has the maximum inelastic energy density among all balls. This is because the maximum local CTE mismatch is between silicon chip and the PCB. Thus the thermal stresses of solder balls beneath the chip are expected to be higher than the stresses on the outermost solder balls. The results show that fan-out WLP packages can extend the array size greatly while meeting thermo-mechanical reliability requirement.

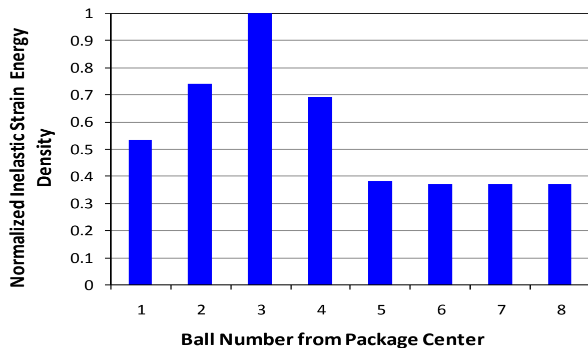

Figure 14. Inelastic strain energy density for a fan-out package [2]

Several challenges appear in fan-out WLP technology [17]. First, the position of the die (after cut) must be maintained from pick & place until the reconstituted wafer is molded. This is important, because the dielectric openings, the redistribution layer and the solder stop openings are applied via photo lithography step. Since the mask aligner is exposing the wafer in one shot, the accuracy of the die position needs to be very high. So the tolerances of the die bonder, the influence of heating and cooling steps and of the chemical shrink of the mold compound needs to be considered.

Another challenge of the reconstituted wafer is the warpage of the artificial wafer. Since the wafer consists of polymer material and silicon, bi-metal effects as well as the different curing behavior of the mold material (chemical shrink, segregation of filler particles) and the ratio of silicon and mold compound in x-, y- and z-direction influence the bow of the wafer, it is necessary to develop a cold dielectric process.

Moisture sensitivity becomes a concern for fan-out WLPs. Encapsulated moisture is a special and unique form of loading that has interactions with mechanical behaviors of mold compound [20-25]. In fan-out WLPs, one of the challenges is to design materials without delamination subjected to moisture loads. When atmospheric moisture is absorbed through mold compound and other polymer materials, it condenses in free-volumes or micro-/ nano- pores in polymer materials and along interfaces. The condensed moisture will vaporize and produce high internal vapor pressure during the reflow process, which is typically completed within a few minutes. The peak temperature during reflow ranges from 220°C to 260°C. Plastic materials become extremely compliant when temperature exceeds their glass transition temperatures. In addition, the interfacial adhesion strength drops substantially. As a result, delamination at weak interfaces may occur due to the combined effects of vapor pressure, thermo-mechanical stresses, hygroscopic stresses, material softening, and adhesion degradation.

4. 3-D Packaging Integration with WLP Technology

Many new developments are currently underway to incorporate 3-D packaging technology with WLP solutions. The IC industry is using WLP for various applications, such as in MEMS and image sensor applications. The MEMS industry has been using wafer level cap since a long time, but it is more wafer level caps than true WLP. The wafer level 3D integration is now becoming a reality for MEMS, as shown in Fig. 15 [3].

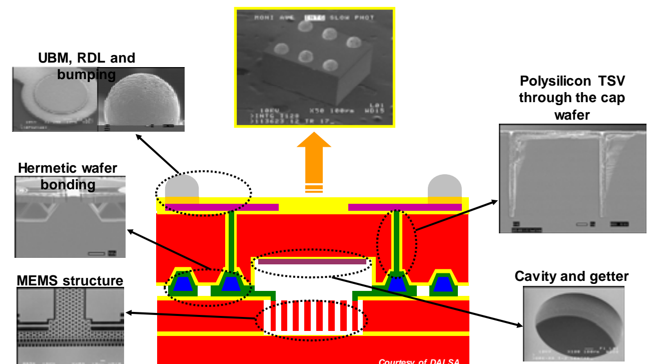

Figure 15. MEME WLP integration

In addition, WLP is being widely used in image sensor applications. TSV technology has been incorporated, as shown in Fig. 16. The main advantage of the TSV technology with WLP in image sensors is to reduce the size of the image sensor module. In this module, the image device is mounted face down on a glass carrier wafer using a patterned adhesive. Next the silicon wafer is thinned using back grinding. The CSCM uses tapered vias followed by conformal deposition of a dielectric layer and a barrier/seed metal-stacked layer for copper electroplating. At the bottom of the vias, openings are formed to allow copper-plated lines to have electrical contact with the sensor interconnection layer. After the final insulating layer is deposited, openings are formed on copper pads, then solder balls are directly attached on the pads during reflow process.

Three-dimensional integrated circuits (3D IC) has been generally acknowledged as the next generation semiconductor technology with the advantages of small form factor, high-performance, low power consumption, and high density integration. TSV and stacked bonding are the core technologies to perform vertical interconnect for 3D integration. For the fabrication approach, there are

three stacking schemes in 3D integration: chip-to-chip, chip-to-wafer, and wafer-to-wafer. Wafer-to-wafer technology can be applied for homogeneous integration of high yielding devices. Wafer-to-wafer bonding maximizes the throughput, simplifies the process flow, and minimizes cost. The drawback for this wafer-to-wafer method is the number of known-good-die (KGD) combinations in the stacked wafers will not be maximized when the device wafer yields are not high enough or not stable. In this case, chip-to-chip or chip-to-wafer will be adopted to ensure vertical integration with only good dies. Considering mass production in future, the chip-to-wafer and wafer-to-wafer technologies have gradually become the mainstream for 3D integration.

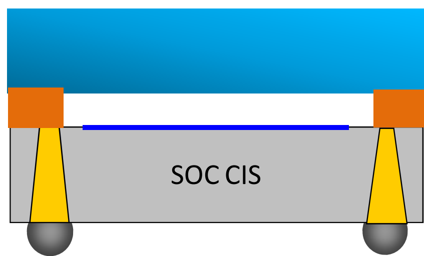

Figure 16. Image sensor WLP with TSV

Wafer-level bonding/bonding/stacking technologies can be further differentiated by the method used to create TSVs: either via-first or via-last. The common definition for via-first and via-last is based on TSVs formed before and after BEOL process. TSV fabrication after the wafers are bonded, using a ''drill and fill" sequence, is definitely via-last approach. Whereas via-first and pre-bonding vialast approaches, building TSVs on each wafer prior to the bonding process are generally more efficient and cost-effective. The leading wafer-level bonding techniques used in 3D integration include adhesive bonding (polymer bonding), metal diffusion bonding, eutectic bonding, and silicon direct bonding [26]. The future development will reply on the full integration of fan-out technology, WLP, and vertical 3D interconnect technology together.

5. Conclusions

Conventional fan-in WLPs are a unique form of packages and have the distinction of being truly die-sized, not "chip-scale'. With fan-out WLP technologies emerging, expensive substrate process can be eliminated. For fan-in WLP, the RDL build-up stacks or copper post/epoxy stacks serve as stress buffer to reduce solder ball stresses significantly. However, the failure mode may shift to the failures in stack-up layers. For fan-out WLP, the die-size is no longer a limiting factor for WLP reliability. Instead, several challenges in wafer reconstitution process arise. Moisture sensitivity becomes a coneren in fan-out WLP development. The integration of fan-out (wafer reconstitution), WLP, and TSV will truly realize system integration in future.

References

1. Fan, X.J. and Liu, Y., Design, Reliability and Electromigration in Chip Scale Wafer Level Packaging, ECTC Professional Development Short Course Notes, 2009.
2. Fan, X.J., Varia, B., and Han, Q. Design and optimization of thermo-mechanical reliability in wafer level packaging, Microelectronics Reliability, 2010, doi:10.1016/j.microrel.2009.11.010.
3. Vardaman, E.J., 2008 Flip chip and WLP trends and market forecasts, 2008, TechSearch.
4. Fan, X.J. and Han, Q., Design and reliability in wafer level packaging, Proc of IEEE 10th Electronics Packaging Technology Conference (EPTC), 834-841, 2008.
5. Rahim, M.S.K., Zhou, T., Fan, X.J., Rupp, G., Board level temperature cycling study of large array wafer level packages, Proc of Electronic Components and Technology Conference (59th ECTC), 898-902, 2009.
6. Varia, B. and Fan, X.J., and Han, Q. Effects of design, structure and material on thermal-mechanical reliability of large array wafer level packages, 2009, ICEPT-HDP.
7. Ranouta, A.S., Fan, X.J., Han, Q. Shock performance study of solder joints in wafer level packages, ICEPT-HDP, 2009.
8. Reche, J.H.J. and Kim, D.H., Wafer level packaging having bump-on-polymer structure, *Microelectronics Reliability*, 43, 879–894, 2003.
9. Kim D-H, Elenius P, Johnson M, Barrett S. Solder joint reliability of a polymer reinforced wafer level package, *Microelectronics Reliability* , 42,1837, 2002.
10. Bumping Design Guide, [online], Available: http://www.flipchip.com/
11. Kawahara, T., SuperCSPs, IEEE Transactions on Advanced Packaging, v.23, No. 2, 2002.
12. Anderson, R., Tee, T.Y., Tan, L.B., Ng, H.S., Low, J.H., Khoo, C.H., Moody, R. Rogers, B., Integrated testing, modeling, material and failure analysis of CSP for enhanced board level reliability, 2008, IWLP.
13. Gao, G., Haba, B., Organesian, V., Honer, H., Ovrutsky, D., Rosenstein, C., Axelrod, E., Hazanovich, F. and Aksenton, Y., Compliant WLP for enhanced reliability, International Wafer Level Package Conference, September 17-19 2007, San Jose, CA.
14. Mitsuka, K.; Kurata, H.; Jun Furukawa; Takahashi, M., Wafer process chip scale package consisting of double-bump structure for small-pin-count packages, ECTC, 572 – 576, 2005.
15. Tee, T.Y., Tan, L.B., Anderson, R., Ng, H.S., Low, J.H., Khoo, C.P., Moody, R., Rogers, B., Advanced analysis of WLCSP Copper Interconnect Reliability under Board Level Drop Test, 10th EPTC Conference Proc., 1086-1095, 2008.
16. Tee, T.Y., Ng, H.S., Syed, A., Anderson, R., Khoo, C.P., Rogers B., Design for board trace reliability of WLCSP under drop test, 2009, EuroSimE.

17. Meyer, T., Ofner, G., Bradl, S., Brunnbauer, M., Hagen, R., Embedded wafer level ball grid array (eWLB), 2008. EPTC, p.994
18. Keser, B., Amrine, C., Duong, T., Hayes, S., Leal, G., Lytle, M., Mitchell, D., and Wenzel, R., Advanced packaging: the redistributed chip package, IEEE Transactions on Advanced Packaging, v.31, No.1, 2008.
19. Fan, X.J., Zhou, J., Zhang, G.Q., Multi-physics modeling in virtual prototyping of electronic packages - combined thermal, thermo-mechanical and vapor pressure modeling, Journal of Microelectronics Reliability, 44, 1967-1976, 2004.
20. Xie, B., Fan, X.J., Shi, X.Q., Ding, H. Direct concentration approach of moisture diffusion and whole field vapor pressure modeling for reflow process: part I – theory and numerical implementation, ASME Journal of Electronic Packaging, 131(3), 031010, doi:10.1115/1.3144147, 2009.
21. Xie, B., Fan, X.J., Shi, X.Q., Ding, H. Direct concentration approach of moisture diffusion and whole field vapor pressure modeling for reflow process: part II – application to 3-D ultra-thin stacked-die chip scale packages, 131(3), 031011, doi:10.1115/1.3144154, 2009.
22. Fan, X.J., Lee, S.W.R., Han, Q. Experimental investigations and model study of moisture behaviors in polymeric materials, Microelectronics Reliability, 49, 861–871, 2009.
23. Fan, X.J., Zhang, G.Q., van Driel, W.D., Ernst, L.J. Interfacial delamination mechanisms during reflow with moisture preconditioning, IEEE Transactions on Components and Packaging Technologies, 31(2), 252-259, 2008.
24. Fan, X.J. Moisture Related Reliability in Electronic Packging, Electronic Component Technology Conference (ECTC) Short Course Notes, 2008.
25. Fan, X.J., Zhou, J., Zhang, G.Q., Ernst, L.J. A micromechanics based vapor pressure model in electronic packages, ASME Journal of Electronic Packaging, 127 (3), 262-267, 2005.
26. Ko,C.T., Chen, K.N. Wafer-level bonding/stacking technology for 3D integration, Microelectron Reliab, doi:10.1016, j.microrel.2009.09.015, 2005.