Shichun Qu · Yong Liu

Wafer-Level Chip-Scale Packaging

Analog and Power Semiconductor Applications

Springer

Wafer-Level Chip-Scale Packaging

Shichun Qu • Yong Liu

Wafer-Level Chip-Scale Packaging

Analog and Power Semiconductor Applications

Springer

Shichun Qu
Fairchild Semiconductor
San Jose, California
USA

Yong Liu
Fairchild Semiconductor
South Portland, Maine
USA

ISBN 978-1-4939-1555-2 ISBN 978-1-4939-1556-9 (eBook)
DOI 10.1007/978-1-4939-1556-9
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014946821

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A wafer-level chip-scale package (WLCSP) is a bare die package that offers not only the smallest possible footprints in all IC package forms, but also superior electrical and thermal performance, mostly credited to the direct solder interconnections that are low in electrical and thermal resistance and low in inductance between chip and application PCB it is assembled on. For mobile electronics, where performance needs to be high and size must be small, heat dissipation is limited to the conduction through PCB to the case of the mobile device; WLCSP is the best chip package option that balances the seemingly conflicting requirements.

Sharing the same root with the flip chip package, WLCSP took a bold step forward by placing sufficient size solder bumps on a semiconductor chip and allowing it to be flip mounted directly on an application board. With solder joints taking up a significant portion of chip/PCB CTE mismatching thermal/mechanical stresses, WLCSP has proved to be reliable in mobile-specific reliability tests, such as drop test, bending test, and temperature cycling tests, besides the basic device-specific reliability tests. The robustness of this packaging form is also demonstrated with lasting life of everyday use on billions of mobile consumer electronics devices. With continuous evolvement in the bumping technologies, such as polymer re-passivated bump on pad (BoP), copper redistribution layer (RDL), front side molded copper post on the RDL, aggressive silicon back grinding, advanced solder alloys, and design know-how, WLCSP has expanded the size range from early days under 2–3 mm to 8–10 mm silicon chip size, while at the same time continuously reducing the per unit cost with scaling factors of high volume production in 200 and 300 mm wafer sizes. The availability of package size range and favorable cost structure makes WLCSP a good packaging candidate for a wide array of semiconductor devices, from analog/mixed signal and wireless connectivity chips to opto-electronics, power electronics, and logic and memory chips. Innovations in wafer-level 3D chip stacking further enable WLCSP a viable option for MEMS and sensor chip packaging.

The beauty of WLCSP is the start-to-finish wafer-based processing. It blurs the line between semiconductor wafer fab processes and the backend packaging operations. There is no singulated die packaging operation typically seen in all other types of chip packaging operations. WLCSP packaging operations, including bumping, inspections, and tests, are fully automated from cassette to cassette,

which is known for high efficiency. Also benefiting from half a century of wafer processing know-how is the overall WLCSP packaging (often referred as bumping) yield, which is quite close to 100%. With this in mind, it is not a surprise at all to see even for die fan-out packages, wafer form processing, based on the reconstituted wafers in 200 or 300 mm size, is the preferred approach from start.

WLCSP has appreciated enormous growth in the past decade, largely because of global consumer demand for mobile communication and computing devices. With double-digit market value (wafer bumping, test, and die processing service including back grind, marking, saw, and tape and reel) growth still in sight, WLCSP is one of the most important packaging technologies for packaging engineers of all backgrounds.

It is the purpose of this book to provide readers a comprehensive overview of the general WLCSP packaging technology. It is also the intention of the authors to share specific knowledge of WLCSP in analog and power semiconductors. Advanced WLCSP technologies, such as 3D wafer-level stacking, TSV, MEMS, and opto-electronics applications, are also briefly introduced in this book.

The book consists of ten chapters, with an overview of the demand and challenges for analog and power WLCSP in Chap. 1; Chaps. 2 and 3 cover the basic concepts of fan-in and fan-out WLCSP, bumping process flow, design considerations, and reliability assessment. Chapter 4 is designated for the stackable packaging solutions involving WLCSP. Chapter 5 gets into the details of wafer-level discrete power MOSFET package design considerations. Chapter 6 discusses more on TSV/stack die WLCSP for the integration of analog and power solution. Chapter 7 is all on the critical topics of thermal management, design, and analysis for WLCSP. Chapter 8 continues on the electrical and multiple physics simulation for analog and power WLCSP, with the new progress on electromigraton study of 0.18 μm power technology. Chapter 9 touches on the assembly of WLCSP devices. Chapter 10 wraps up the book with reliability and general testing of WLCSP semiconductors.

Coming up with years of experience in semiconductor packaging, and with focus on wafer-level packaging, the authors attempted to provide well-balanced and yet up-to-date content in ten chapters. We wish this book is a good starting material for young engineers who need to learn the most important of WLCSP technology in a short time. At the same time, we also hope that seasoned engineers find this book good references for them to not only keep up with the rapid technology advancement, but also to help address daily engineering challenges.

San Jose, CA, USA — Shichun Qu
South Portland, ME, USA — Yong Liu

Acknowledgments

The book is impossible without the dedication of Merry Stuber, editor from Springer for timely reminder and coordination of draft submission and critical reviews. Doug Dolan from Fairchild Semiconductor deserves special thanks for taking time performing the primary legal review of all ten chapters. The authors would also like to thank Fairchild Semiconductor for general support of technical publications that led directly to the completion of this book. A few names are mentioned here for the support over the years, Suresh Belani, director for packaging, OS Jeon, senior director for packaging, Dan Kinzer, former chief technology officer, and Paul Hughes, general consults, all from Fairchild. Many coworkers contributed to the data quoted in this book and the author like to take this opportunity to express our sincere thanks as well: Mr. Richard Qian, Mr. Zhongfa Yuan, and Dr. Yumin Liu for simulation support; Dr. Qi Wang for wafer-level power MOSFET and process; Dr. Jun Cai (former Device and Process Senior Member of Technical Staff) and Mr. Andrew Schoenberger (former Fairchild wafer-level process engineer) for the WLCSP ball shearing test; Dr. Yangjian Xu, Mr. Ye Zhang, and Huixian Wu (Zhejiang University of Technology) for the WLCSP ball shearing test and stacking simulations; Dr. Jifa Hao for the wafer-level electromigration test of 0.18 μm power interconnects; Dr. Yuanxiang Zhang (Quzhou University) for building the wafer-level electromigration model; Ms. Jiamin Ni and Professor Antoinette Maniatty (Rensselaer Polytechnic Institute) for solder joint electromigration; Etan Schaham for inspiring discussions of WLCSP packaging challenges; Rob Travis and Dennis Tummy for insights of device reliability and fab process interactions; Mr. Doug Hawks (former Fairchild packaging engineer) for MCSP development; Mr. William Newberry for electrical simulation methodology; Jihwan Kim for the WLCSP drop test; and Steve Martin for support for routine WLCSP research and development activities. Several organizations within Fairchild are also mentioned here for the support WLCSP assembly and tests: Fairchild Bucheon site and Fairchild Cebu site.

Much of the material in this book was derived from previous papers and research notes by the authors. Here the authors like to thank several professional societies that published some of this material and allowed to reproduce some contents in this

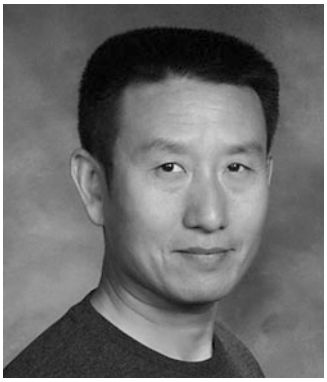
book. They are the Institute of Electrical and Electronic Engineers (IEEE) and its Conferences, Proceedings, and Journals, including IEEE Transactions on Components and Packaging Technology and IEEE Transactions on Electronics Packaging Manufacturing. The authors also appreciate the following Conferences for allowing the reorganization and reproduction of previously published materials: IEEE Electronic Components and Technology Conference (ECTC), IEEE International Conference on Electronic Packaging Technology and High Density Packaging (ICEPT-HDP), and IEEE International Conference on Thermal, Mechanical and Mult-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE).

Lastly, but most importantly, the authors would like to thank their perspective families for the support that made it possible to spend numerous weekends and nights on this book. Shichun Qu likes to thank his wife Shan Huang and daughter Claire Qu, and Yong Liu likes to express his appreciation to his wife, Jane Chen, and sons Junyang Liu and Alexander Liu for their great love and patience throughout the 2+ years of writing this book.

San Jose, CA, USA Shichun Qu
South Portland, ME, USA Yong Liu

Contents

Biography



Shichun Qu started his career in research and development of PTFE-based IC low K dielectric materials at W. L. Gore & Associates' Eau Claire, Wisconsin R&D facility, after receiving a Ph.D. degree in Materials Science and Engineering from SUNY at Stony Brook. Post the 3 M acquisition of the Gore & Associates' IC substrate business, also located in Eau Claire, Wisconsin, he continued working on the manufacturing technology development of organic flip chip/BGA and wire bond/BGA substrates. In the meantime, he was also the lead engineer in the development of high-speed organic substrate dielectric materials: from materials formulation, coating and substrate manufacturing processes development and qualifications to the flip chip assembly process development. The latter work resulted in a very early industrial publication on the management of the flip chip substrate warpage through a differential heating reflow process back in 2003. Shichun Qu joined National semiconductor in Santa Clara, California, in 2007. There he was involved in the development of advanced lead frame package, high-temperature wire bond over pad metallization research and production qualification, and high pin count WLCSP technology research. After joining Fairchild Semiconductor in 2011, he has dedicated most of the time understanding the WLCSP chip/PCB interactions and fine-tuning WLCSP design and bumping process to achieve competitive manufacturing cost by extending the use of low mask count bumping technologies at higher pin count. Besides conventional WLCSP, Shichun Qu also played technical roles in various types of embedded WLCSP power module designs, test, and qualifications.

Besides a Ph.D. in Materials Science and Engineering, Shichun Qu also received a Master of Engineering and Bachelor of Science degree in Engineering Mechanics from Tsinghua University (Beijing, China). He was an assistant professor for four and a half years at the University of Petroleum (Beijing, China) before returning to school for the Ph.D. degree at SUNY Stony Brook and new careers in the USA.

Besides tackling engineering challenges, Shichun Qu enjoys other types of physical challenges. He finished his first half marathon (San Jose Rock n Roll) in 2013 and is eying on full marathon in the next few years. When not practicing long-distance running, he enjoys the time with his wife and daughter and sometimes hiking and biking as a family.

About the Author

Yong Liu has been with Fairchild Semiconductor Corp in South Portland, Maine, since 2001 as a Senior Member Technical Staff from 2008, a Member Technical Staff from 2004 to 2007, and a Principal Engineer from 2001 to 2004. He is now a Fairchild global team leader of electrical, thermal-mechanical modeling and analysis. His main interest area is advanced analog and power electronic packaging, modeling and simulation, reliability, and assembly process. In last a few years he and his team have been working on advanced IC packaging and power modules, modeling and simulation, including the pioneering work on assembly manufacture process, reliability analysis, and the electromigration-induced failures in chip-scale wafer-level packages. He has been invited to give keynote talks and presentations at international conferences Eurosime, ICEPT, EPTC, and universities in the USA, Europe, and China. He has coauthored over 170 papers in journals and conferences and has been granted over 45 US patents in the area of 3D/Stack/TSV IC and power electronic packaging. Dr. Liu obtained his BS, Master, and Ph.D. degrees in Nanjing Univeristy of Science and Technology in 1983, 1987, and 1990, respectively. He once was promoted by breaking rule as a full professor at Zhejiang University of Technology in 1994. Dr. Liu was awarded Alexander von Humboldt Fellowship and studied as a Humboldt fellow at Tech University of Braunschweig, Germany, in 1995. In 1997, he was awarded Alexander von Humboldt European Fellowship and studied as a Humboldt European fellow at University of Cambridge, England. In 1998, he worked as a post-doctor at Semiconductor Focus Center and Computational Mechanics Center, Rensselaer Polytechnic Institute (RPI). In 2000, he worked as a staff opto package engineer at Nortel Networks at Boston. Since he joined Fairchild in 2001, he was awarded the first Fairchild President Award in 2008, Fairchild Key Technologist in 2006 and 2009, Fairchild BIQ award in product innovation in 2005, and Fairchild award for power of pen first place in 2004, IEEE CPMT Exceptional Technical Achievement Award in 2013.

Demand and Challenges for Wafer-Level Chip-Scale Analog and Power Packaging

1

A review of recent advances in analog and power wafer-level chip-scale packaging (WLCSP) is presented based on the development and market demand in semiconductor industry. This chapter covers in more detail how advances in both the analog and power advanced wafer-level package fan-in/fan-out design and 3D integration have co-enabled significant advances in analog and power device capability during recent years. Extrapolating the same trends in representative areas for the remainder of the decade serves to highlight where further improvement in techniques of analog, power switches, and passives can drive continued enhancements in usability, efficiency, reliability, and overall cost of analog and power semiconductor solutions. Challenges of die shrinkage in both wafer-level analog and power semiconductor packaging in next-generation design are presented and discussed.

1.1 Demand for Analog and Power WLCSP

Over the last two decades, analog and power semiconductor technology has made impressive progress, particularly in the increasingly high-power density of monolithic and system multiple function [1–6]. One of the significant achievements of the analog and low voltage power packaging is the WLCSP technology, which is constantly evolving—not necessarily in groundbreaking ways but to meet different increasing demands and subtle variations in material compositions, thickness, metal stack structure, and size reduction for new applications. Figure 1.1 shows the basic WLCSP device applications [7], which include analog, logic, mixed signal, opto, MEMS, and sensors. This book will focus more on the analog and power applications.

WLCSPs are by no means the lowest cost solution available, but its tiny volume and electrical performance benefits are turning it into the "go-to" package for use in the mobile phones and tablets; more and more analog and power managements are going into WLCSP. This accelerates the demand for the WLCSPs [6]. While it is now being produced in larger volume, it should be indicated that there is no

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_1

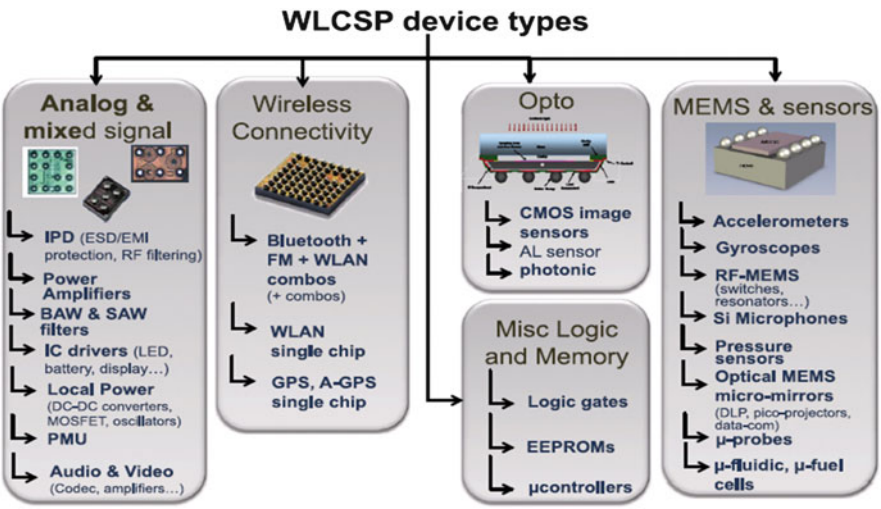


Fig. 1.1 Basic WLCSP device applications [7]

universal standard for WLCSP. There are tremendous variations in WLCSP because each customer has specific application requirements. WLCSP has been primarily applied from small with less than 3 mm^2 chip size to medium size with 3–5 mm^2. Now, it has penetrated to larger die sizes, such as larger than 5 mm^2, increasing the growth rate of WLCSP market. WLCSP technology for I/O counts of less than 50 is considered by most to be mature. There is, however, plenty of activity underway in the industry to extend board level reliability of WLCSP to array size in more than 100 I/O counts with acceptable reliability. In the next 5 years, the WLCSP market growth is expected to climb by 12.6 %, while the compound annual rate (CAGR) shows the trends of the demand in mobile phones and tablets from 2010 to 2016 (see Fig. 1.2) [7].

1.2 Impact of Die Shrinkage

1.2.1 Die Shrinkage Impact

As compared to the development in general wafer-level IC product [8–13], the analog and power wafer-level packaging is far behind due to the high-power density and high reliability requirement. The development of power semiconductor device has begun to aim at 90 and 130 nm technology, while today 180 and 250 nm technologies are beginning to drive significant die size shrinkage as compared to regular 350 or 500 nm power IC technology. As the metal interconnect system inside the die continues to become thinner, current density has significantly increased. The electromigration (EM) issues will grow and new interconnect alternatives will be considered. Current techniques such as the wafer-level solder

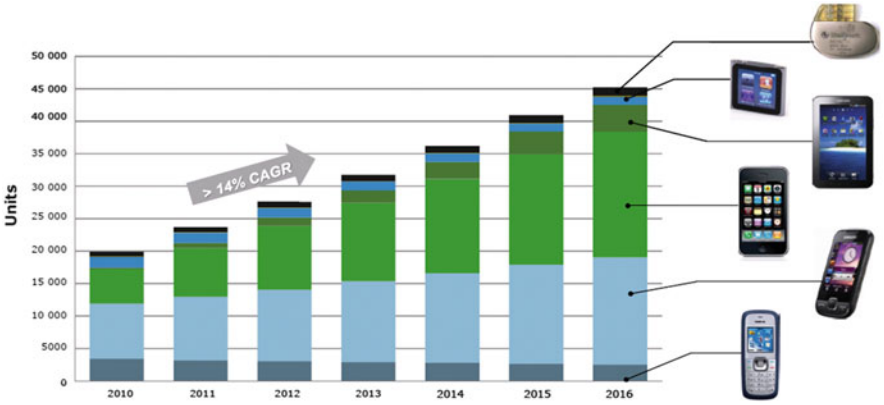


Fig. 1.2 The growth demand of WLCSP for mobile and tablets market [7]

bumping or Cu-stud bumping will meet the challenges of material intermetallic diffusion and mechanical cratering issue for wafer-level Cu-stud bumping. As the die shrinks, the pitch of power wafer-level chip-scale packages (CSP) will move from current 0.5 mm towards 0.4 or 0.3 mm. The heat dissipation will become a very critical and significant challenge. Finding a high efficiency heat dissipation solution is necessary.

1.2.2 Wafer-Level System on Chip Versus System in Package

The power integration devices allow the state-of-art smart power IC with technology such as integration of bipolar, complementary metal–oxide–semiconductor (CMOS) and double diffused metal–oxide–semiconductor (DMOS)-BCDMOS, intelligent discrete power device, and the function integration in both lateral DMOS (LDMOS) and vertical DMOS (VDMOS) for power control and protection as well as other functions. This is so-called wafer-level system on chip (SOC), which is the integration of several heterogeneous technologies—analog, digital, Mosfets, etc.—into a single silicon chip. However, such wafer-level SOC technology often is too expensive and complex. This leads to a wealth of opportunities for system in package (SIP), in which multiple chips with different functions are placed in one package or module [14] which has similar function of SOC but with a lower cost.

SIP has evolved as an alternative approach to SOC for electronics integration because this technology provides advantages over SOC in many market segments. In particular SIP provides more integration flexibility, faster time to market, lower research and development (R&D) or nonrecurring engineering (NRE) cost, and lower product cost than SOC for many applications. SIP is not a replacement for high-level, single-chip, and silicon integration but should be viewed as complementary to SOC. For some very high volume applications, SOC will be the

preferred approach, like a power SOC with the integration of LDMOSFETs and IC controller; the cost of the SOC is not expensive due to larger volume of production as compared to the SIP with two Mosfet die and an IC controller die. In such case, the electric performance of the SOC is clearly super as compared to the SIP. Some complex SIP products will contain SOC components. Wafer-level SIP/stack is one major direction for lower power application.

As the die shrinks, SOC can add more functions, and SIP can include more chips. In SOC, the thermal density will become very high. Determining how to insulate different functions in one chip and how to effectively dissipate the heat through the package will be a challenge [4, 14]. Although the cost of SIP is low, there are challenges due to the assembly of wafer-level multiple chips to wafer. The internal parasitic effects of SIP, like parasitic inductance [15], are higher than SOC. The impact of heat from the power components on the electrical performance of IC drivers will be a concern. To build an advanced SIP which has good thermal and electrical performance with low cost is the largest challenge of wafer-level SIP. Modeling and simulation efforts must be used to support the wafer-level SIP development from design, reliability, and assembly process [16].

1.3 Fan-In Versus Fan-Out

With every new product generation, portable electronic devices such as smart phones or handheld computers have to integrate a growing number of functions within a very confined space. This has been made possible by a significant miniaturization of the chip package. WLCSPs are manufactured before wafer dicing and enable further form factor reduction and allow saving cost particularly when packaging small dice; it can be divided into two categories [17]: fan-in wafer-level packaging and fan-out wafer-level package. True wafer-level packages are inevitably fan-in packages. This means that their contact terminals are all within the footprint of the die. This translates into a severe limitation for adjusting the layout of the contact terminals to match the design of the next-level substrate (printed circuit board, interposer, IC package). Fan-out wafer-level packages represent a compromise between die-level packaging and wafer-level packaging. In both fan-in and fan-out cases, no laminate substrate or epoxy mold compound is needed to connect the die to the PWB (such as underfill). Solder balls are directly attached onto the silicon die and/or the fan-out area. Fan-in WLP fits devices with small dies and relatively low I/O count. Fan-out WLP could manage larger packages and higher I/O count with redistribution layout (RDL) technology for small die with fine pitch. RDL is an interconnect for a fine pitch die to connect a larger pitch fan-out WLP, for example, in an embedded die wafer-level BGA, which is also called eWLB, a solder ball pitch can be maintained that is easily handled on the PWB. The semiconductor wafer is diced and the singulated ICs are embedded in an artificial molded wafer. Within this artificial wafer, the dice are separated from each other by a space which is big enough to allow the desired fan-out RDL to be manufactured

by standard WLP processes. The fan-out WLCSP is a bridge to connect the smaller die with fine pitch to the large pitch of customer PWB.

Fan-in wafer-level chip-scale packaging (WLCSP) is maturing and growing at a relatively brisk pace, and its success appears to be serving as a springboard of sorts for the technology into applications beyond handsets and also accelerating development of other types of wafer-level packages (WLP). So now is a perfect time to take a look at what the industry sees on the horizon for WLP. Will more research and effort be directed towards making cheaper or more reliable fan-in WLCSP, or will these innovation expenses be diverted to fan-out WLCSP or other WLP technologies such as 3D IC stacking with TSVs packaging? Fan-in WLCSPs are established schemes, and development is mainly related to incremental improvement of materials to obtain better thermal cycling performance. There is little value in scaling I/O pitch (narrows down the application space to special, expensive PCB board technologies. Fan-out WLP technology has the potential to reduce package thickness, to be used for next-generation package on package (PoP) packages and passive integration, and opens a broad range of new packaging integration possibilities for future designs. Currently, most analog and power WLCSPs are fan-in-based designs, while analog and power fan-out WLCSPs are still in the early phase of development. New solutions are required to allow for larger die and package size with high reliability. Significant improvements may still be realized within the next 5–10 years. 3D stacking is not in competition with WLP, but 3D WLP stacks are also possible.

1.4 Power WLCSP Development

1.4.1 Wafer-Level Mosfet Compared to Regular Discrete Power Package

Table 1.1 shows the typical development trends of discrete Mosfet package. It gives the representative power transistor package constituent volumetric percentages. As the package develops from Fairchild early DPAK (TO252) through SO8 to Mosfet BGA and Mosfet WLCSP, the molding compound decreases as a percentage of volume, until it reaches zero with the Mosfet BGA package and WLCSP. At the same time, the silicon and interconnect metal increase as a percentage of volume. At DPAK level, leadframe is about 20 % and silicon is about 4 %, while EMC is about 75 %. At WLCSP level, the silicon is about 82 %. There is no EMC in the WLCSP.

Figure 1.3 shows the discrete wafer level—CSP for Schottky diode and vertical Mosfet from the released products of Vishay and Fairchild semiconductor. Those WLCSPs are called fan-in layout.

However, the advantage of EMC is that it can enhance handling and mechanical robustness, as it has in the past. It can provide substantial protection and mechanical integrity to place components across a wide generational range of pick and place equipment. So the EMC today in discrete power package is still useful as a

Table 1.1 Typical discrete power package constituent volumetric percentages towards WLCSP [18]

Package type	Total volume (mm^3)	Approx.% EMC	Approx.% silicon	Approx.% leadframe	Approx.% interconnect
TO-252 (wire)	90	75 %	4 %	20 %	1 %
SO8 (wire)	28	83 %	6 %	10 %	1 %
SO8 (clip)	28	70 %	6 %	20 %	2 %
Mosfet BGA	20	0 %	40 %	50 %	10 %
WLCSP	20	0 %	82 %	0 %	18 %

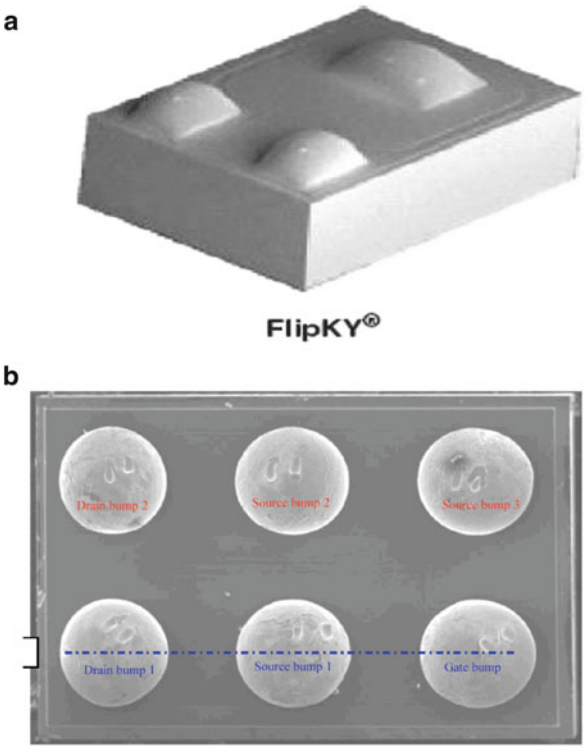


Fig. 1.3 Examples of discrete WLCSP: (**a**) Vishay Schottky WLCSP, (**b**) Fairchild discrete WLCSP

component "encapsulant." For wafer-level power package, the EMC can be used as the redistribution layer substrate material through the molding for the fan-out wafer-level package. That allows for the larger pitch for a smaller shrank die.

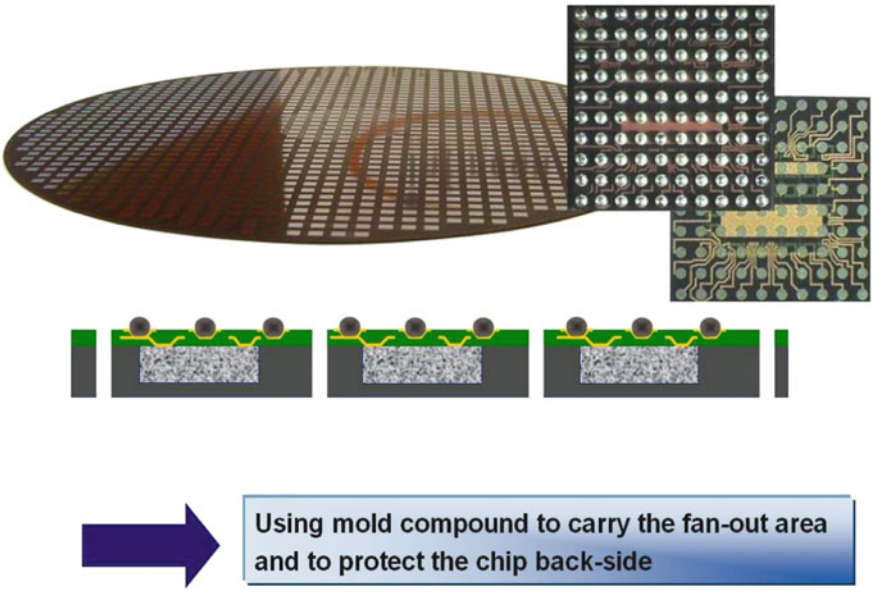


Fig. 1.4 EMC as the redistribution substrate of fan-out wafer-level package in Infineon [13]

Figure 1.4 gives the example of fan-out RDL WLP structure by using wafer-level epoxy molding technology.

1.4.2 Higher Current Carrying Capability

One trend of the discrete wafer-level power package is to increase the current carrying capability per unit area; this is partly due to the customer's request for high current capability and partly due to the die shrinkage. To better manage the thermal performance with the trend to higher current carrying capability in wafer-level power package, there are two approaches: one is to intensify the thermal management requirements from the print circuit board (PCB) level and the other is heat dissipation in multiple directions at the package level which is advantageous for wafer-level discrete power packages. Bonding the wafer-level power discrete package to a metal frame is an effective approach. Bonding die to metal wafer with pre-etched cavity is the wafer-level process to get the wafer-level package with multiple direction heat dissipation. Figure 1.5 shows the examples of multiple direction heat transfer of Fairchild Mosfet BGA and Vishay PolarPAK.

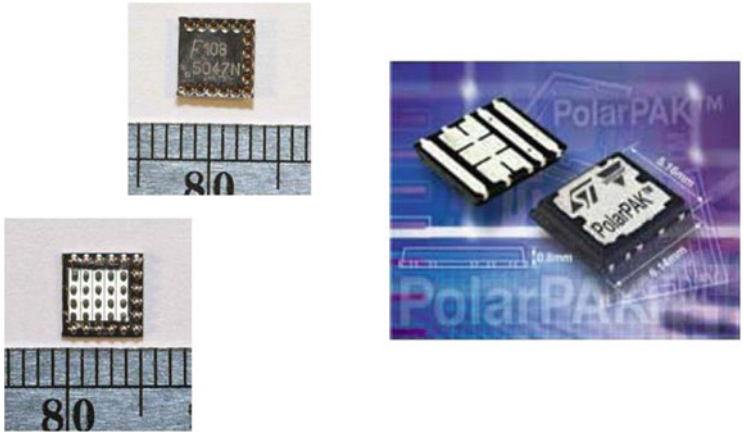


Fig. 1.5 Multiple direction heat transfer packages

1.4.3 Low Rds(on) Resistance and Better Thermal Performance

To get lower Rds(on) and to improve the thermal performance, the wafer-level Mosfet with vertical metal–oxide can be built on a silicon substrate thinned to 7 μm and plated with 50 μm copper as its drain (see [19] Fairchild wafer-level UMSFET). This extremely reduces the Rds(on) resistance and improves the thermal performance. Figure 1.6 shows the internal device structure of the UMOSFET and its comparison with regular Mosfet.

For the wafer-level discrete Mosfet, another trend which obtains the attention in the industry is to move the drain of the Mosfet to the front side of the die so that the drain, source, and gate are at the same side. This would be helpful for the surface mounting application in various PCBs and also for the good electrical performance. Figure 1.7 shows one of the lateral layouts of the drain for a LDMOS WLCSP. Since the drain is in lateral placement, the back metal does not contact the drain directly, so its application limits to lower power and lower voltage area. For VDMOS WLCSP, the trend is to develop the direct connection to the front side by TSV in trench area. The advantage of the direct connection from the back drain to front side is its good electrical performance with lower Rds(on).

1.4.4 Trends in Power IC Packages

For the voltage range at 5–100 V, there are a wider range of inductive loads handled with a monolithic solution and higher level of functional integration in monolithic solution [18]. The most interesting application is the wafer-level integrated system power conversion solution which combines two power switches (the high side and lower side) together with an IC driver. Figure 1.8 shows an example of such a wafer-level power system on chip. There are also integrated advanced digital

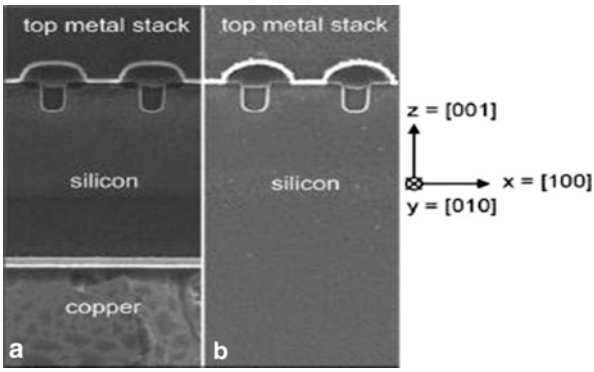


Fig. 1.6 (**a**) Fairchild UMOSFET, (**b**) regular Mosfet [19]

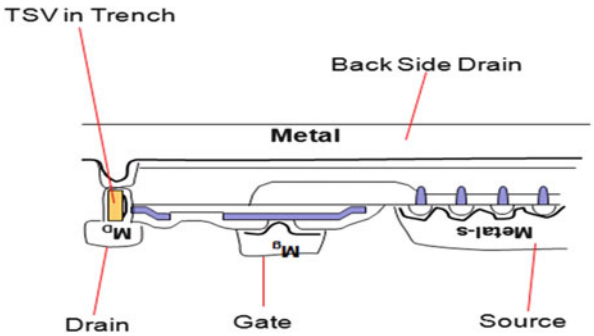


Fig. 1.7 Move the drain to the front side of the Mosfet

control functions for motion which includes "sensorless" positioning and fault detection at the low end and "adaptive" motion control at the high end. For the voltage range at 100–700 V, the next generation of integrated LDMOS structures reaches the limits of Si for breakdown voltage (BVdss) as a function of geometry, which allows a corresponding increase in high-voltage (HV) monolithic power conversion capability (AC–DC) and results in an incremental raising of the limit at which multiple dice are requested in the actual products.

As power die size shrinks, the package footprints shrink as well, and maintaining the thermal transfer capacity at the package level is difficult since the function/unit area of the die is increasing with advanced BCDMOS processes. While the overall package footprint trend is a decreasing area, the thermal dissipation capabilities rely more on the PCB as part of the system. Therefore, insuring mechanical integrity of WLCSP in the form of bare flipped die in conjunction with board level assembly of a heat sink is difficult (see Fig. 1.9).

Instead of air cooling for power wafer-level package, one trend is to build the wafer-level micro-channel on power chip. This can effectively take the heat out of the power chip. Figure 1.10 gives an example of building the micro-channels on both power die active surface and backside. Due to the high efficiency of cooling through the micro-channels, heat sink is not needed anymore. This may

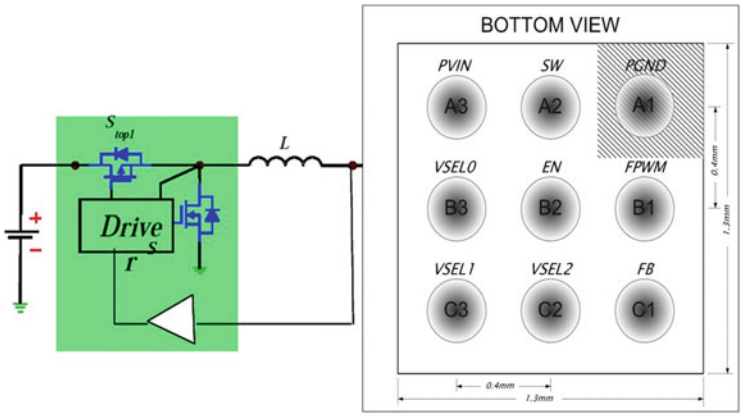


Fig. 1.8 Wafer-level integrated power solution

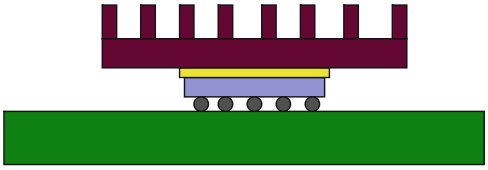


Fig. 1.9 Example of WLCSP which is hard to mount the heat sink

significantly reduce the space of heat sink and remove the noise induced by the fans in the cooling system.

1.4.5 Trends in Wafer-Level Passives

Although the wafer-level passives (resistor, capacitor, and inductor) today are only suitable for very low and tiny power, it is possible for them to integrate with low power BCDMOS or other active IC. Integration of active power switches and passives in wafer level can greatly improve the electrical performance and significantly reduce the parasitic effects. For relative larger power products, like buck converter and DrMos with passives, the development is ongoing. Figure 1.11 gives the development samples of wafer-level inductors for power application [14, 21], which indicates the current level with frequency. One significant advantage of the wafer-level inductor integration is its frequency can reach from several MHz to 100 MHz, which the regular package level and board integration level are hard to get.

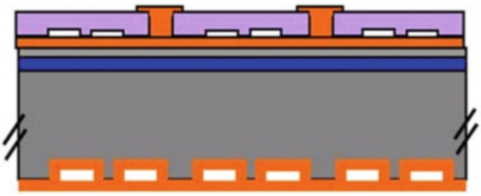


Fig. 1.10 Micro-channels on both active side and backside die [20]

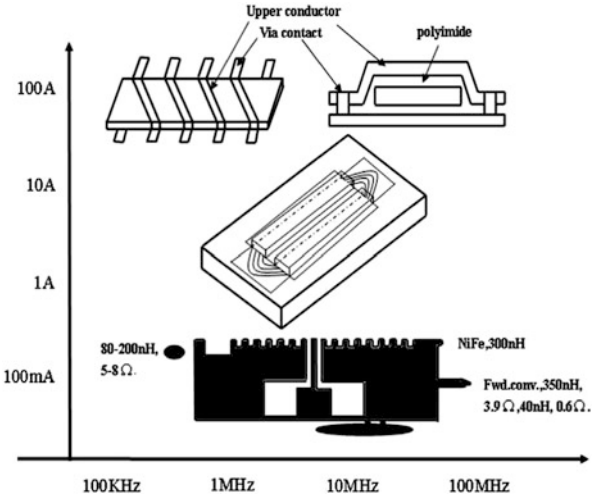


Fig. 1.11 Wafer-level passives

1.4.6 Wafer-Level Stack/3D Power Die SIP

There are two stacking wafer-level methods for power die SIP. One is to stack die on the wafer and the other is to stack two wafers. Figure 1.12 shows the active IC die is stacked on the passive wafer. The power IC die with two Mosfets and an IC driver is bonded on the passive wafer with inductor L. Figure 1.13 shows an example of a wafer-level stacked die package of two Mosfets bonded together with wafer 1 source and wafer 2 drain. The common source (wafer 1)/drain (wafer 2) may be connected to at least one front side by TSVs. This stacking process can be done through wafer on wafer. The advantage of this integration is the very good electrical performance for a half bridge with both N channel and P channel Mosfets for products like liquid crystal display (LCD) backlight inverters. Since the distance between high side die and low side die is very short, this greatly reduces the electrical resistance and parasitic effects.

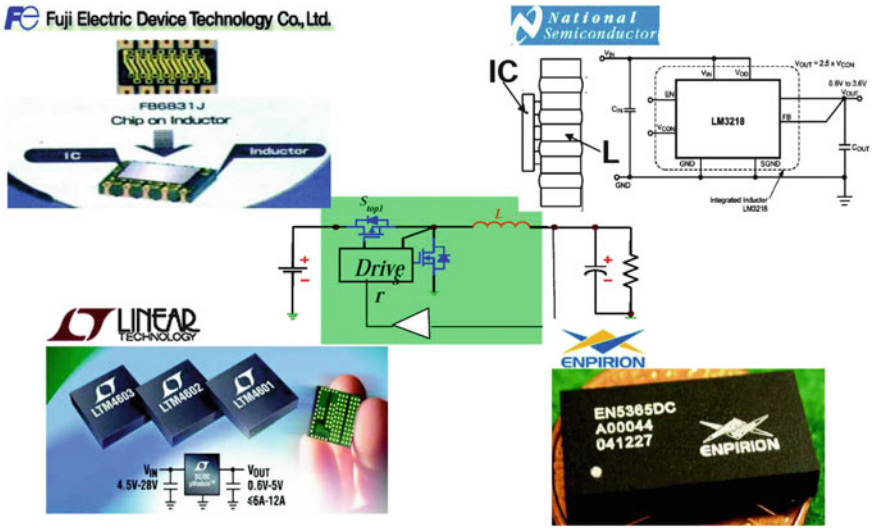


Fig. 1.12 Stacked active die on passive wafer [14]

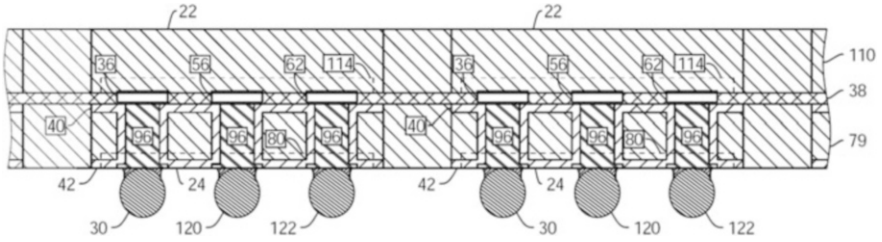


Fig. 1.13 Wafer-level stacked power die package with TSVs by two stacked wafers

1.5 Summary

Today providing energy efficient solutions for various products is becoming increasingly important in our world due to limited energy resources and climate change, especially the fast growth of consumer electronics in both communications and entertainment, industrial power conversion, automotive, and standard power electronic products. Consumers' demand for increased mobility with advanced features and for high efficiency energy solutions has paved the way for a variety of new products, which has driven advances in analog and power electronics technology towards the high-power density design, smart integration that combines the analog, logic, and power together. This chapter discusses the demand of the analog and power WLCSP, challenges and impacts due to the analog and power die shrinkage, fan-in and fan-out wafer-level technology development for analog and

power application, and the power WLCSP development. The development of wafer-level analog and power package is closely related to the development of the analog and power IC technologies. Currently most analog and power WLCSPs are fan-in designs, and the fan-out technology is still under early development for analog and power application. Moving the VDMOSFET drain to front side is a trend of power WLCSP today, which allows the discrete power WLCSP to be used in all the surface mount applications. The trends of power wafer-level IC package are high-power density at die level and small footprint. The trend of advanced wafer-level power IC technology is the integrated solution which combines the analog IC controller and the power Mosfet switches. Wafer-level passives can allow the power integration to obtain the switch high frequency from several MHz to 100 MHz. Power 3D technology in wafer level is a future trend; major efforts may be focused on die-to-wafer and wafer-to-wafer stacking with TSVs.

References

1. Liu, Y.: Reliability of power electronic packaging. International Workshop on Wide-Band-Gap Power Electronics, Hsingchu, April (2013)
2. Liu, Y.: Power electronics packaging. Seminar on Micro/Nanoelectronics System Integration and Reliability, Delft, April (2014)
3. Kinzer, D.: (Keynote) Trends in analog and power packaging. EuroSimE, Delft, April (2009)
4. Liu, Y., Kinzer, D.: (Keynote) Challenges of power electronic packaging and modeling. EuroSimE, Linz, April (2011)
5. Liu, Y.: (Keynote) Trends of analog and power electronic packaging development. ICEPT, Guilin, August, (2012)
6. Hao, J., Liu, Y., Hunt, J., Tessier, T., et al.: Demand for wafer-level chip-scale packages accelerates, 3D Packages, No. 22, Feb (2012)
7. Yannou, J-M.: Market dynamics impact WLCSP adoption, 3D Packages, No. 22, Feb (2012)
8. Shimaamoto, H.: Technical trend of 3D chip stacked MCP/SIP. ECTC57 Workshop (2007)
9. Orii, Y., Nishio, T.: Ultra thin PoP technologies using 50 μm pitch flip chip C4 interconnection. ECTC57 Workshop (2007)
10. Meyer-Berg, G.: (Keynote) Future packaging trends, Eurosime 2008, Germany
11. Vardaman, E.J.: Trends in 3D packaging, ECTC 58 short course (2008)
12. Fontanclli, A.: System-in-package technology: opportunities and challenges. 9th International Symposium on Quality Electronic Design, pp. 589–593 (2008)
13. Meyer, T., Ofner, G., Bradl, S., et al.: Embedded wafer level ball grid array (eWLB), EPTC 2008, pp 994–998
14. Lee, F.: Survey of trends for integrated point-of-load converters, APEC 2009, Washington DC (2009)
15. Hashimoto, T., et al.: System in package with mounted capacitor for reduced parasitic inductance in voltage regulators. Proceedings of the 20th International Symposium on Power Semiconductor Devices and ICs, 2008, Orlando, FL, May, 2008, pp. 315–318
16. Liu, S., Liu, Y.: Modeling and Simulation for Microelectronic Packaging Assembly. Wiley, Singapore (2011)
17. Fan, X.J.: (Keynote) Wafer level packaging: Fan-in, fan-out and 3D integration, ICEPT, Guilin, August (2012)
18. Liu, Y.: Power electronic packaging: design, assembly process, reliability and modeling. Springer, New York (2012)

19. Wang, Q., Ho, I., Li, M.: Enhanced electrical and thermal properties of trench metal-oxide-semiconductor field-effect transistor built on copper substrate. IEEE Electron Device Lett. **30**, 61–63 (2009)
20. Zhao, M., Huang, Z.: Rena, Design of on-chip microchannel fluidic cooling structure, ECTC 2007, Reno, NV, pp. 2017–2023 (2007)
21. Liu, Y.: Analog and power packaging, professional short course, ECTC 62, San Diego (2012)

Fan-In Wafer-Level Chip-Scale Package

2

2.1 Introduction of Fan-In WLCSP

Fan-in wafer-level chip-scale package is the first form of WLCSP. The term "fan-in" comes from the fact that early time WLCSP was originally designed as wire bond devices with bond pads all arranged along the perimeters of semiconductor dies. When converting a perimeter bond pad design into an area array WLCSP, redistribution or "fan-in" technology had to be used.

Over time, WLCSP has become a mature packaging technology with more and more semiconductor devices, especially those for mobile applications, such cell phones and tablets, which are designed into WLCSP from the beginning. As a result of this change, "fan-in" is used less often than the commonly referred WLCSP.

Fan-out WLCSP, which expands the package size beyond the size of silicon and, oftentimes, even breaks the commonly accepted size definition of chip-scale packaging (1.2 $\times$ the die size), is on the opposite end of the wafer-level packaging spectrum. In this case, the reconstituted package is bigger than the silicon die size, so "fan-out" has to be applied to route interconnect from small silicon die to the whole package area. Figure 2.1 illustrated the concept of fan-in and fan-out wafer-level package.

2.2 WLCSP Bumping Technology

Over the past decade, knowledge in solder alloys, solder intermetallic compound (IMC), under bump metallization (UBM), and polymeric repassivation materials and knowledge of field performance and accelerated component level and board level reliability testing all have helped to promote WLCSP from selected applications into mainstream packaging technology. Today, WLCSP is found in bumping technologies that can be categorized in two basic forms: bump on pad (BOP) technology and redistribution layer (RDL) technology. Of the two WLCSP bumping technologies, BOP has the simplest structure that has the UBM directly

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_2

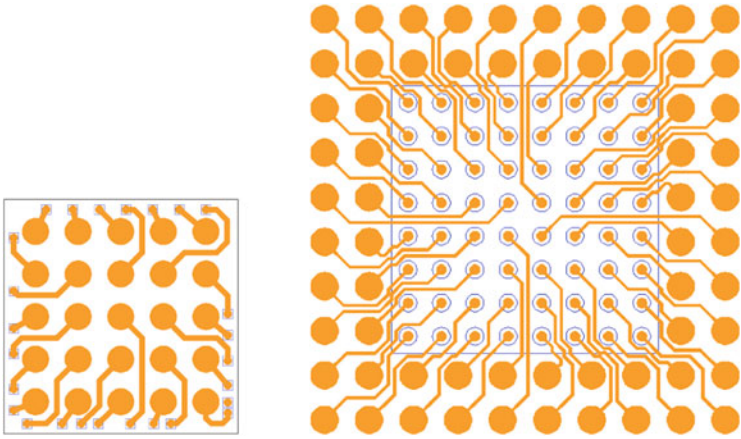


Fig. 2.1 A 5 × 5, 0.4 mm pitch fan-in WLCSP from a wire bond device's perimeter bond pads (*left*) and a one-layer fan-out WLCSP design from a 8 × 8, 0.3 mm pitch device to a 0.4 mm pitch package

bonded onto the chip top metallization. Depending on whether polymer repassivation is applied or not, BOP can be further separated into bump on nitride (BON) or bump on repassivation (BOR) technologies. Figure 2.2 cross-section drawings highlight the fact of bump structure directly bonded onto the chip aluminum pad as well as differences between BON and BOR.

BOP is the most widely adopted WLCSP technology for analog/power devices, which often have limited numbers of solder joints. One of the most distinctive advantages of the technology is the low bumping cost. Also, unlike RDL that forces current flow in the lateral directions, BOP with proper bump arrangement allows more vertical current flow from bump to the semiconductor switch layer—this is often a critical performance factor for the power semiconductor WLCSP. However, as the industry moves towards denser device level integrations and subsequently larger die size with more solder joints to facilitate signal, power, and ground connections, RDL technology becomes inevitable due to solid board level reliability at high pin count. One other benefit of the RDL is that it allows easy reuse of the basic building blocks, since top-level interconnection is realized in the bumping operations. Without the needs of re-layout at the chip level, time to market is much shortened. There are also times when the same semiconductor chip is desired to be offered in different packaging forms, such as WLCSP and QFN package; fan-in RDL design balances both wire bond package and solder bumping package well, which provides an overall low-cost solutions.

Unlike BOP, which has bump/UBM directly anchored to the chip aluminum pad, RDL separates bump/UBM structure from the device surface with a polymer layer, i.e., polyimide (PI) or polybenzobisoxazole (PBO). The addition of this soft, stress-buffering polymer layer is the most distinctive feature of the RDL technology and also the foundation for solid mechanical reliability performance.

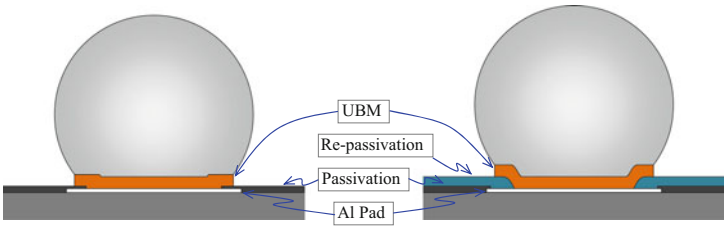


Fig. 2.2 Cross sections of bump on nitride (BON) and bump on repassivation (BOR). Both belong to the commonly referred bump on pad (BOP) WLCSP bumping technology

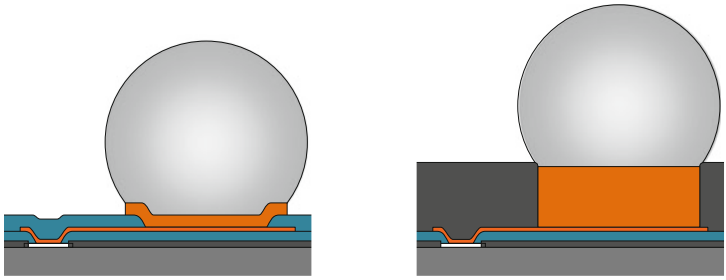


Fig. 2.3 Cross sections of a typical 4 mask RDL bump structure and a RDL + molded copper post WLCSP bump structure. The latter structure provides superior board level reliability performance

Variations of RDL technology exist, all aiming either to improve the mechanical reliability performance of WLCSP or to reduce the cost of wafer bumping. Figure 2.3 gives examples of a typical RDL bump cross section and a more advanced RDL plus molded copper post technology. The molded copper post approach provides superior board level reliability performance because of RDL and copper post that increase standoff height, as well as front side molding materials that has a close CTE match with PCB materials and aggressive thinning of silicon wafers.

2.3 WLCSP Bumping Process and Cost Considerations

When it comes to the selection of bumping technologies for specific devices, many factors come to play. Electrical and thermal factors are always the top considerations. Reliability of WLCSP, especially board level reliabilities, i.e., drop, TMCL, and bending test, also has to live up to the general industry standards or customer-specific standards. In the very competitive mobile computing world, costs often come as one of the key factors to be considered, and because of this, it is beneficial to understand the process flow and basic cost structures of various bumping technologies.

Mainstream WLCSP bumping all adopts additive plating pattern formation process, without exceptions. In the plating process, photomasking layer is used to

Table 2.1 Major bumping process step comparisons

Step	BON	BOR	RDL	Molded Cu post[a]
1	Seed layer sputter	PI coat	PI coat	PI coat
2	Resist coat	PI expose[b]	PI expose[b]	PI expose[b]
3	Resist expose[b]	PI develop	PI develop	PI develop
4	Resist develop	PI cure	PI cure	PI cure
5	UBM plate	Seed layer sputter	Seed layer sputter	Seed layer sputter
6	Resist strip	Resist coat	Resist coat	Resist coat
7	Seed layer etch	Resist expose[b]	Resist expose[b]	Resist expose[b]
8		Resist develop	Resist develop	Resist develop
9		UBM plate	RDL plate	RDL plate
10		Resist strip	Resist strip	Resist strip
11		Seed layer etch	Seed layer etch	Dry film laminate
12			PI coat	Dry film expose[b]
13			PI expose[b]	Dry film develop
14			PI develop	Cu post plate
15			PI cure	Dry film strip
16			Seed layer sputter	Seed layer etch
17			Resist coat	Front side mold
18			Resist expose[b]	Mold cure
19			Resist develop	Mechanical buff
20			UBM plate	Cu etch back
21			Resist strip	
22			Seed layer etch	

[a]Molded copper post only takes three masking steps. However, the prolonged copper post plating and molding operations add to the process complexity and cost
[b]Masking step—more masking steps often imply higher overall bumping cost

define plating patterns and is considered a main cost adder in WLCSP bumping operations. As a matter of fact, overall WLCSP bumping costs can often be measured by the number of masking operations—the more masks are needed for a specific bumping technology, the higher the bumping cost is. Using this relationship, it is not hard to find that bump on pad (BON and BOR) have clear advantages in bumping cost. On the other hand, the molded copper post technology, though requiring only three mask operations, commands the highest bumping costs due to slow copper post plating, additional molding and post planarization operations, and overall process complexity.

Table 2.1 highlights the major process steps for four most commonly used WLCSP bumping technologies. From BON to molded copper post, WLCSP board level reliability is improved alongside the bumping operation cost increase. For molded copper post, it is actually one mask less than RDL. However, additional copper post plating, front side molding, and mechanical buffing for post top planarization are added to the overall cost of this unique bumping technology. Process flow of the molded copper post is illustrated in Fig. 2.4 to help understand the bumping process.

There is natural drive to extend low-cost bumping technology to highest possible pin count, due to significant cost increase from low mask count bumping

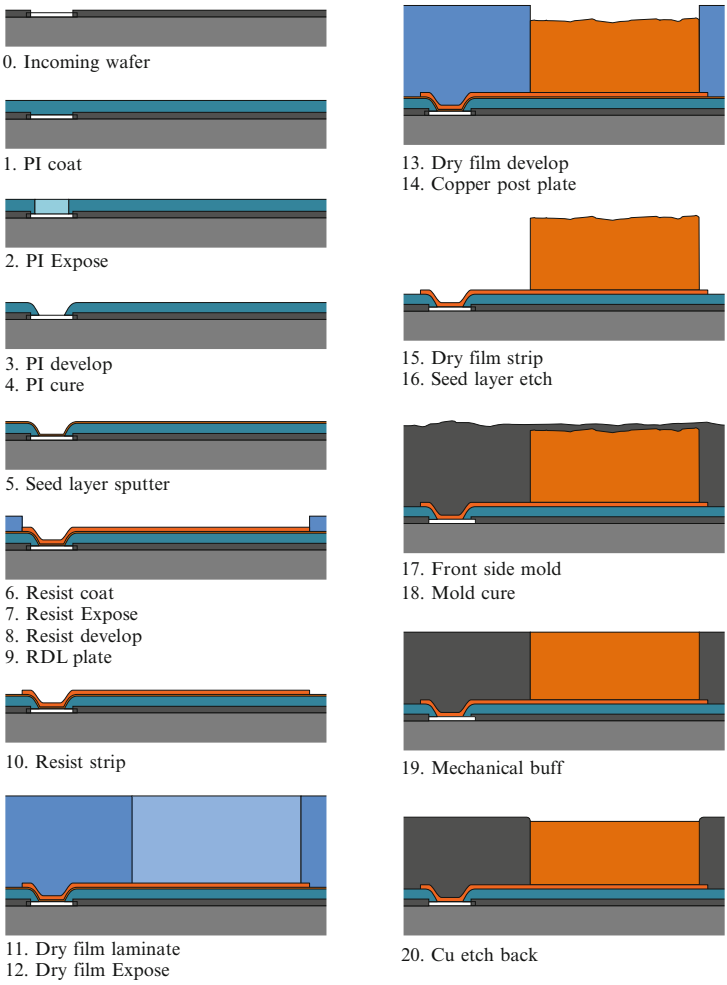


Fig. 2.4 Process flow of molded copper post bumping technology. Flux print, solder ball drop, and reflow post copper etch back are not shown

technology to high mask count bumping technology. For example, though BOP is thought a bumping technology for low pin count WLCSP, it is not uncommon to see it being used in typical high pin count applications that were once though only possible with RDL technology. Evidently, optimization of on-chip metal/via stack, polymer repassivation materials, UBM metal stack and size rules, as well as underfill materials is necessary to ensure that minimum reliability requirements are not comprised. This practice alone can substantially save bumping cost from early adoption of high-cost bumping RDL technology. The same approach applies to every bumping technology transition, i.e., from BON to BOR, from BOR to RDL, and from RDL to molded copper post.

2.4 Reliability Requirements for WLCSP

Though cost saving is a never-ending challenge to packaging engineers, meeting reliability requirements should never be compromised in qualifying specific WLCSP bumping technology. For WLCSP reliability, it typically involves board level drop test, temperature cycle test (TMCL), and bending test. When adopting WLCSP technology on individual device, component level TMCL also must be considered alongside other typical device reliability tests, such operational life (OPL), highly accelerated stress test (HAST), ESD, etc.

In typical WLCSP board level reliability tests, WLCSP components are mounted on the test printed circuit board (PCB) and then subjected to mechanical or environmental stress—in drop test, it is the mechanical impact on the PCB mounted on a test fixture during drop, and in TMCL, it is the extreme temperature that induces that thermal mechanical stress from the mismatch of CTE of silicon (typically 2 ~ 3 ppm/°C) and PCB (typically 17 ppm/°C). Understanding the stress distribution and typical failure modes is one of the most important in designing the WLCSP test and production device.

2.5 Stress in Drop Test

Test method described in JEDEC standard JESD22-B111, Board Level Drop Test Method of Components for Handheld Electronic Products, is the most referenced and widely accepted industrial standard for WLCSP board level drop performance assessment. The board itself is a 132 mm × 77 mm rectangle with four corner mounting holes spaced by 105 mm × 71 mm (Fig. 2.5). Also PCB stack has eight copper layers and seven dielectric layers plus solder mask layers on both sides.

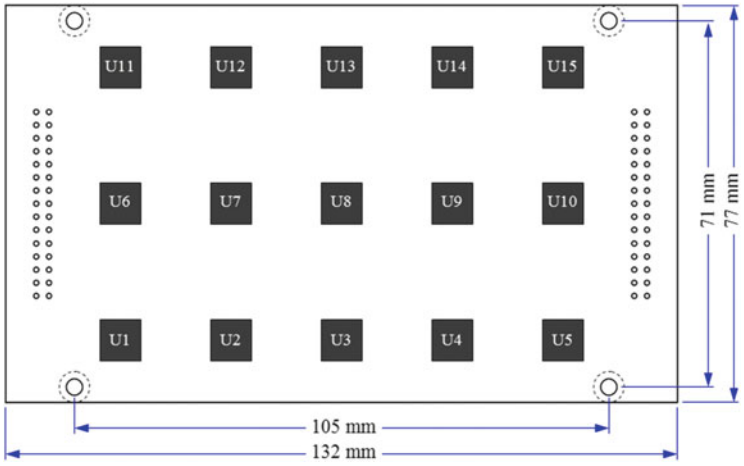


Fig. 2.5 JEDEC drop PCB with 15 components mounted on

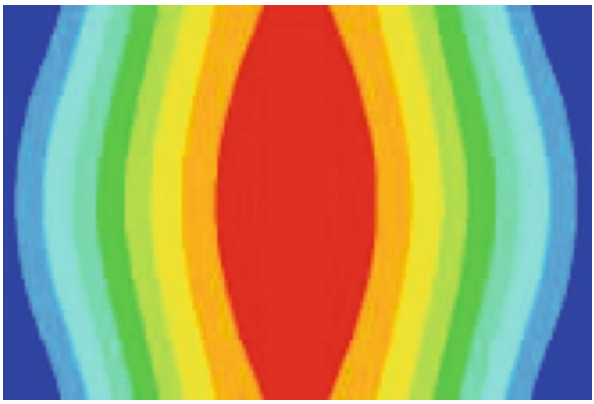


Fig. 2.6 Bending mode of drop PCB in the drop test, assuming use of 4 mounting screws to hold board onto the drop fixture

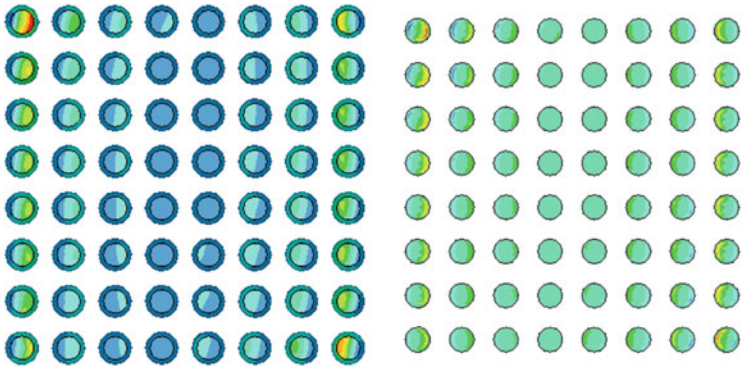


Fig. 2.7 First principle stress S1 distribution in solder joints on WLCSP chip side (*left*) and normal stress Sz in BOP aluminum pad. Simulated drop PCB has eight copper layers and a length-width aspect ratio of 1.71

Copper layer thickness and area cover as well as dielectric layer thickness and materials are specifically defined in the JEDEC standards.

In drop test, PCB bends in length and width directions, with bend in length direction dominating bend in width direction (Fig. 2.6). The bending of PCB creates stress in solder joints of WLCSP because silicon is small and rigid and does not comply well with the PCB.

Both numerical simulation and failure analysis of drop failed WLCSP components have revealed the high stress locations on a WLCSP that soldered on a drop board. Across a WLCSP component, the stress distribution is more or less one dimensional, with bumps on the sides perpendicular to the length of the drop PCB (main bending direction) seeing the most of the stress (Fig. 2.7). Depending on the mounting position of the component and geometry of the test board, some

deviation from ideal one-dimensional stress distributions is anticipated. Yet not surprisingly, highest stress is found at the corners.

The dominant one-dimensional nature of stress distribution is a distinctive characteristic of the drop stress and needs to be put into serious consideration when designing WLCSP device or test chips or laying out PCB with WLCSP devices. For actual WLCSP devices, it is all about avoiding the highest stress locations for critical routing, and for WLCSP test chips, it is all about understanding the effects of the highest stresses. It is especially important with non-square WLCSP chip design. For rectangular WLCSP, aligning length side with major bending direction of PCB could result unnecessary stress in a drop incident and should be avoided whenever possible. However, rectangular WLCSP test chip is often purposely aligned to the major bending direction of the drop PCB in order to understand the worst case stress implication.

2.6 Stress in TMCL

Unlike the stress distribution in drop test, TMCL stress arises from mismatch of coefficient of thermal expansion (CTE) of silicon (typically 2 ~ 3 ppm/°C) and PCB (typically 17 ppm/°C) and is more along the radius from the die center stress neutral point (Fig. 2.8). TMCL highest stress occurs at low temperature extremes (maximum temperature deviation from solder solidifying point where silicon and PCB are coupled together).

2.7 High Reliability WLCSP Design

For reliable WLCSP, it is important to know the locations of highest stress point and attempt to avoid routing trace along the high stress directions. From stress simulations, it is also evident that the stress on corner solder joints exceeds the

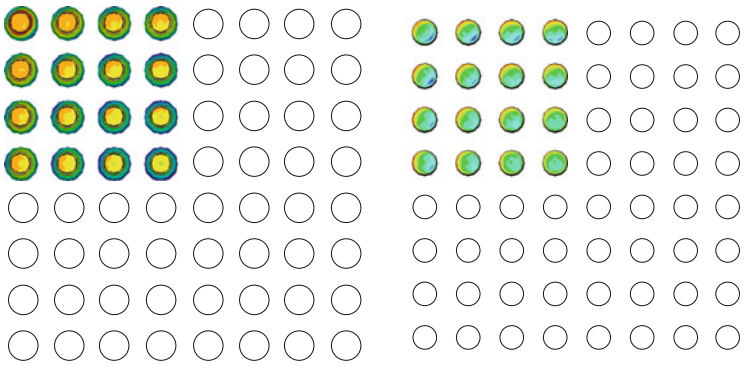


Fig. 2.8 Von Mises stress (Svm) distribution in solder joints on WLCSP chip side (*left*) and first principle stress S1 in BOP aluminum pad. Simulated TMCL PCB has eight layers of copper and aspect ratio of 1.71

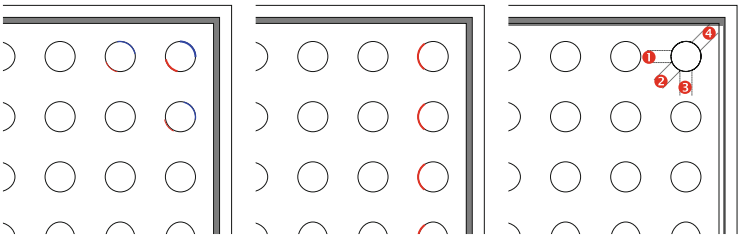


Fig. 2.9 High stress on WLCSP in TMCL (*left*), in drop (*middle*, assuming major bending is in *X* direction), and directions of traces that should be avoided (*right*): direction 1 and 3 correspond to potential high stress in drop and direction 2 and 4 correspond to potential high stress in TMCL

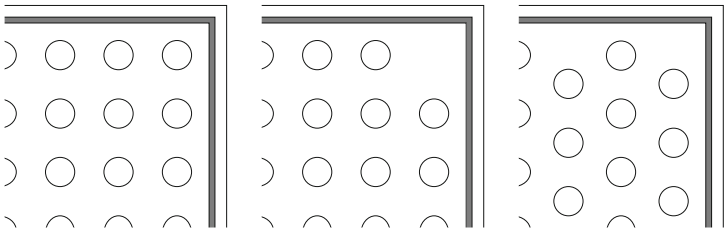


Fig. 2.10 Full array WLCSP (*left*), WLCSP without corner joints (*center*) and a staggered WLCSP design that features two corner joints

stress on neighboring solder joints by large margin. So to reduce the risk of early failures on WLCSP, it is a common practice to remove the corner solders for high pin count WLCSP. Figure 2.9 below gives examples of high stress locations on a WLCSP and non-preferred corner routing directions. Figure 2.10 shows sample designs of WLCSP with corner solder joints removed.

Besides attention to the corners, WLCSP engineers should also be aware that actual failures may occur at locations other than the corners. For example, BoP WLCSP could see crack through silicon occurring at non-corner location due to insufficient bump metal thickness, properties of polymer repassivation layer, weak under bump metal layers, via structure and dielectrics, etc. For RDL WLCSP, fatal failure like through RDL trace crack might occur at non-corner locations due to not optimized RDL routing, insufficient RDL copper thickness, as well as polymer materials properties, and layer thickness, etc. General guidelines can be given, yet there is no replacement of the intelligent work by the packaging engineers.

2.8 Test Chip Design for Precise Reliability Assessment

Test chip is an irreplaceable element in WLCSP technology development. Proper designed test chip can help select the most cost-effective, yet reliable, WLCSP solutions for a given package size, height, and pitch requirements. Specifically

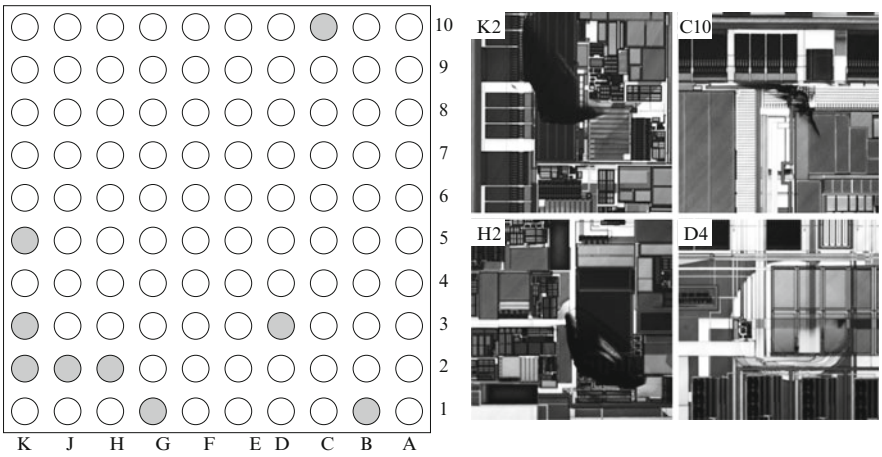


Fig. 2.11 Through silicon cracks found on this 10 × 10 BOP WLCSP at non-corner locations due to localized stress effect on multilayer chip stack. Shown are four example images done with laser scanning microscopy. Sites found silicon cracks are highlighted in *shade* in the *left side map*, though not necessarily on the same chip

designed WLCSP can also help to reveal BoP package/chip metal stack interactions and provide design guidelines for WLCSP chip upper metal layer design. For RDL WLCSP, thoughtfully designed test chip helps answer questions related to RDL copper thickness, trace width/orientation, as well as polymer layer material properties and thickness. While it is nearly impossible to detail out every single aspect when considering test chip design, some general guidelines are provided here for WLCSP engineers:

1. Multilayer versus single metal layer test chip design
   Single metal layer test chip design was the first used in the early time of WLCSP technology development, when chips were typically small with a few solder joints, and silicon technology was still the traditional aluminum metallization with oxide dielectrics. The main concern was the solder joint survivability in the environmental stress tests. As the silicon technology advances, copper metallization, finer metal lines, dense vias, and low-κ dielectrics with built-in porosity are increasingly used to achieve the desired performance at higher frequencies. At the same time, chip size grew rapidly due to demand for higher level integrations. In pursuing of higher pin count WLCSP with advanced semiconductor dielectrics and metallization, failure mode that was not observed before on older, smaller WLCSP, such as crack through multilayer silicon, was recorded. Figure 2.11 shows such failure mode identified on a 100 solder bump WLCSP through a few hundred temperature cycles. The specific failure location cannot be well explained by the traditional thermal expansion mismatch theory due to non-corner nature. So the only reasonable interpretation is that multilayer metal/low-κ dielectrics structure contributes to the through silicon

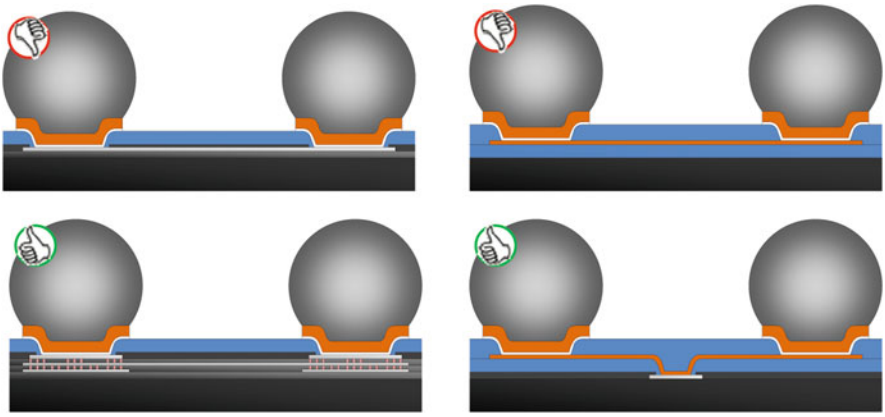


Fig. 2.12 Multi-on-chip-metal-layer BOP test chip is recommended, while single-on-chip-metal-layer RDL is sufficient for the intended investigations of WLCSP board level reliability performance

crack at not-highest stress location in early cycle. In the preproduction board level reliability study, simple metal layer daisy chain test chip was used, and there were no indications of the particular failure mode. All test showed satisfied reliability life, and TMCL failure mode was dominantly solder joint crack under cyclic temperature conditions. So the lesson learnt from this case is that in order to clear all potential failures at the final product stage, test chip should have the same or similar multilayer metallization, via structures and dielectrics; this is especially true for BOP-type WLCSP, since in this technology, UBM is directly anchored on the chip top metallization. For RDL WLCSP, stress is more exerted on the RDL copper and UBM structures, as well as polymer layers, and rarely goes beyond. So for RDL WLCSP test chips, single metal layer test chip is an appropriate, more cost-effective approach. Figure 2.12 illustrated the concept of this multilayer BOP and single-layer RDL test chip design, with references of undesired stack up for BOP and RDL test chip. As discussed above, having a multi-on-chip-metal-layer BOP is very important to catch fatal failure mode like silicon crack. For RDL test chip, all "floating" RDL design can result in "exceptionally" well board level reliability performance due to less restricted RDL and metal stack under the solder joints. So the design should be avoided by all means.

2. Scalable test chip design
   It is not uncommon to see test chips with the exact pin count, similar layout as the final targeted products. Yet it is more cost-effective to adopt a scalable test chip design for rapid technology development or verification. The concept of scalable test chip design is illustrated in Fig. 2.13. Here a basic 2×2 WLCSP unit with uniform within die and across die pitch is laid out on the whole wafer. When specific pin count WLCSP test chip is needed, such as 6×6, 8×8, or 10×10, a different wafer map is applied to dice the wafer into the desired size.

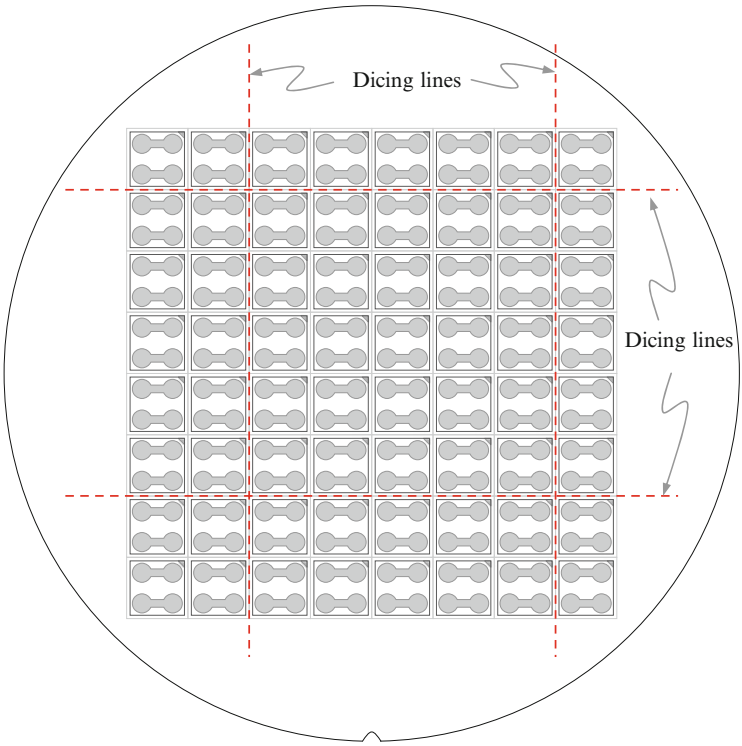


Fig. 2.13 A modular 2 × 2 daisy chain design is illustrated here; also dicing lines for a final 8 × 8 test chip are highlighted in *dotted lines*

Limitations exist with this modular test chip design. First, it is limited to the even number pin count. Odd number could be achieved in one direction, such as base 2 × 1 unit designs; it is not feasible in both directions. Second, the concept works best when wafer fab yield and bumping yield is high. It could be challenging creating wafer maps if yield at either step is low. Fortunately high-yielding fab and bumping could always be expected for test chips due to easy metal layout. Third, the modular design test chip will likely end up with discontinued passivation (SiN) and repassivation (polymer) coverage due to saw street within the final die area. This is quite different from any regular WLCSP chip, which always features continuous passivation (SiN) and repassivation (polymer) across the whole die area and only terminates in the saw street around the die perimeters. Some stress changes are inevitable, but simulation confirmed that the changes are minimum and should not shift the overall chip performance in the typical board level reliability tests, such as drop and TMCL.

3. Daisy chain
   A test chip with solder bumps is only half of a daisy chain that allows monitoring solder joint interconnection failures in board level reliability testing. The other half of daisy chain has to be completed by the PCB design. In the early time of

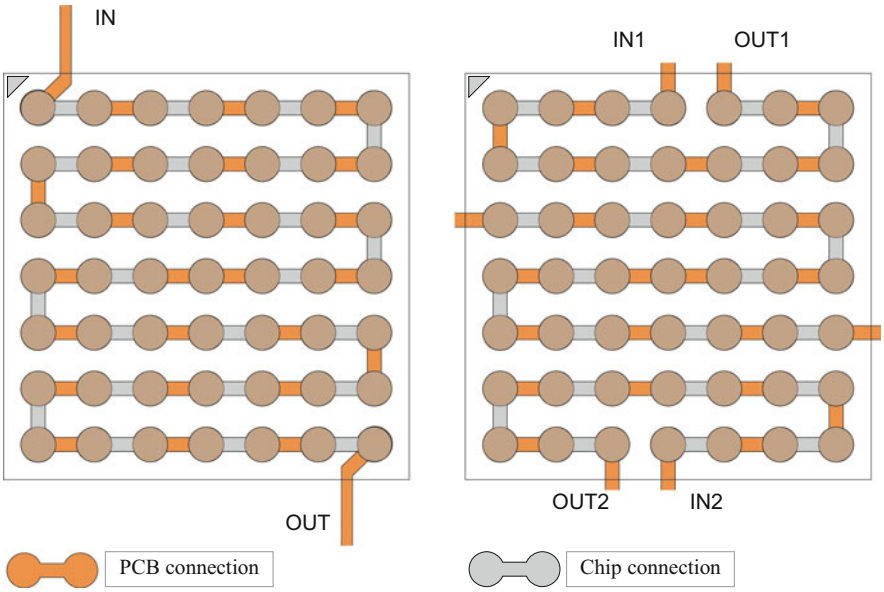


Fig. 2.14 A single daisy chain design and a split daisy chain design for an identical 7 × 7 WLCSP test chip. The split daisy chain design only continuously monitors the *top* and *bottom nets* and leaves the third net in the *middle* for manual probing only

WLCSP, daisy chain often meant an electrical path that allows testing of every single solder joints on the chip. Exceptions exist; however, it only applies to the center solder joint, which sometimes was left unmonitored/untested. The argument is pretty straightforward; WLCSP chip center is a stress neutral point and is the least likely failure location.

With introduction of multilayer test chip design, where on-chip connections are made through vias and inner metal layers, and to combine with higher pin count, higher daisy chain resistance is anticipated, and it could be too high for event detector to function properly if every solder joint is connected in a single daisy chain. In this case, daisy chain design with only selected perimeter solder joints or just corner solder joints being monitored/tested becomes a sensible approach. Figure 2.14 illustrates a more traditional daisy chain layout and a split daisy chain design with only top and bottom two rows being connected for continuous monitoring during board level reliability tests. All solder joints in the center section will only be checked when test stopped.

There are concerns about this aggressive daisy chain layout. One is about missing registrations of solder joint failures that is not at locations being continuously monitored. The supporting evidence of this concern is described in the previous section (Fig. 2.11), where non-corner balls failed in TMCL. To address this concern, one has to look into the differences between a test chip and a real functional chip: test chip typically has uniform on-chip layer stack from bump to bump, while real chip typically has different on-chip layer stack from

bump to bump. So for a real chip, it might have non-corner first failures due to localized weak on-chip layer stack, but for a test chip, it will always fail first at corners, as predicted by the numerical models, regardless of the particular failure modes.

Other constraining factors of the split daisy chain design include the needs of more event detector channels to monitor the added test nets, complexity of test PCB design, and excess in data analysis. However, none of the shortcomings can resist the benefits of employing the split daisy chain design.

In addition to lower the perspective chain resistance at high pin count with multiple on-chip metal layers, the split daisy chain design allows data-based fault isolation without the need of manual probing and guess work that is too often associated with the single daisy chain design. For example, in a single daisy chain design, when a failure occurred, it can be sure it is one of the four corners. However, it is often hard to determine which corner is with manual probing, because failures were almost always first recorded during drop impact or at extreme temperatures. When probing at room temperature and without bending of PCB, the early crack could be close and make electrical test nearly impossible to confirm. With split daisy chain design, when failure occurred, it will be certain whether it is on top or bottom portion of the daisy chain so that FA can start from that side without the need for manual probing. In situations where drop or TMCL has to continue after initial failures occur, this capability, in combination with proper FA techniques, becomes critical in determining the true causes of the failure.

4. Silicon thickness, BSL and front side molding
   Silicon thickness and backside laminate (BSL) is known to have direct influences on WLCSP board level reliability performance. In a simulation study of silicon thickness and BSL effects, six cases were modeled and the conclusions are somewhat interesting (Fig. 2.15).

   First of all, the simulation confirmed that thinner silicon does make solder joint stress lower in drop. Table 2.2 clearly indicates the trend of stress reduction in UBM or aluminum pad from 378 μm silicon thickness to 292 μm and 200 μm

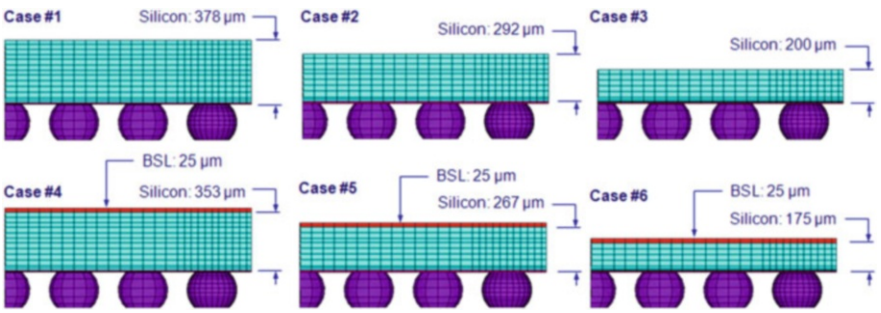


Fig. 2.15 Simulation case study of the effect of silicon thickness and backside laminates on WLCSP board level reliability performances

Table 2.2 Stress in UBM and aluminum pad in BLR drop test

Simulation case	#1	#4	Δ	#2	#5	Δ	#3	#6	Δ
Die thickness (μm)	378			292			200		
Si thickness (μm)	378	353	–	292	267	–	200	175	–
BSL thickness (μm)	–	25	–	–	25	–	–	25	–
S1 in UBM	633.9	632.9	−0.16 %	625.8	621.5	−0.69 %	596.8	581.8	−2.51 %
Sz in UBM	585.8	584.3	−0.26 %	575.9	571.0	−0.85 %	543.0	525.7	−3.19 %
S1 in Al Pad	302.1	301.6	−0.17 %	298.7	297.3	−0.47 %	292.4	291.5	−0.31 %
Sz in Al Pad	278.8	277.5	−0.05 %	271.9	269.0	−1.07 %	255.2	247.8	−2.90 %

Table 2.3 Stress in UBM and aluminum pad in BLR drop test

Simulation case	#1	#4	Δ	#2	#5	Δ	#3	#6	Δ
Die thickness (μm)	378			292			200		
Si thickness (μm)	378	353	–	292	267	–	200	175	–
BSL thickness (μm)	–	25	–	–	25	–	–	25	–
First fail (cycles)	521	479	−8.06 %	592	538	−9.12 %	694	631	−9.08 %
Char life (cycles)	848	779	−8.14 %	963	875	−9.14 %	1128	1026	−9.04 %

silicon thicknesses in drop test. BSL appears to help reducing the drop stress as well, though it can be insignificant when silicon is thick.

The TMCL simulation predicates first fail cycle and characteristic cycles of the modeled cases. It is quite evident that thinner silicon boosts the cycle life significantly when comparing between different die thicknesses. However, the BSL contribution is opposite as predicated by the model, and it is not as commonly believed statement that BSL helps WLCSP reliability (Table 2.3).

Knowing the effect of silicon thickness and BSL will help make decisions when designing test chip and selecting all the parameters. Since the main role of a test chip is to confirm the reliability performance and reveal all potential risks surrounding a particular WLCSP technology approach, it is generally desired to test the worse case conditions, which means thick silicon thickness, if possible, for a test chip.

5. PCB trace orientation

PCB trace orientation next to the soldering pad is as important as daisy chain design itself. Many experimental demonstrations and numerical simulation have confirmed that improperly oriented trace can result in trace crack in drop tests and skew the test data that supposedly should only include solder joint-related failures. FA to confirm or eliminate copper trace failure mode can be time consuming—it normally involves lapping through layers of PCB copper and dielectrics. So the best practice is to make trace orientation as robust as possible.

Syed et al. [3] have detailed analysis that can be found in the references. Basic guideline is to avoid laying out fan-out traces from the corners or in the direction that is parallel to the major bend direction of the drop test PCB. Normally that means the length direction of the PCB. In case trace has to come out from the non-preferred side or copper pad, a 45° angle is recommended to orientate the trace towards the center line of the PCB copper pad array. Figure 2.16 gives examples of copper trace crack and shows simulation results of the Syed's work.

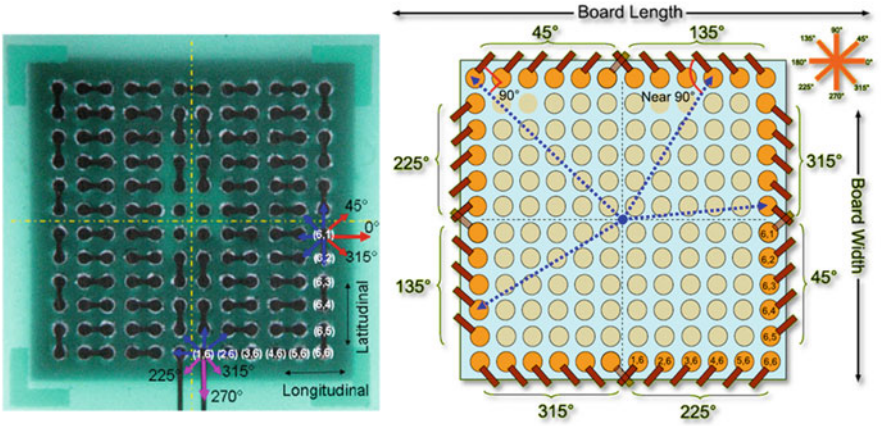


Fig. 2.16 Angle definition and preferred fan-out trace orientation based on the simulation study and experiment verifications. According to Syed et al., the preferred trace directions lower the accumulated plastic strain in copper trace and thus lower the risk of failure in copper trace. Reproduced with permission from [3]

6. Depopulated array
   Full array WLCSP has PCB-induced stress concentrated on the corner solder joints, and oftentimes it is related to early failures. In case a boost of performance is needed and chip design does not need all the solder joints, corner solder joints can be omitted (Fig. 2.9), and highest stress could be reduced by sharing load with two near-corner solder joints. The effect is confirmed by simulations. To verify the benefits, a dedicated daisy chain has to be designed and tested with a full array reference.

2.9 BOP Design Rules

1. Aluminum pad size and geometry
   BOP WLCSP typical design rule has UBM enclosed by aluminum pad, with margins of processing tolerances being considered. Also in the consideration is the solder ball size used in the WLCSP ball drop operation. Bigger solder balls were often set on bigger UBM and aluminum pads and is often favored for improved board level reliabilities. Figure 2.17a illustrated the relationship between aluminum pad, PI open, and UBM. Solder ball size after ball drop and reflow is also given as reference.

   The potential issue for UBM bigger than aluminum pad is the stress concentration on traces connecting the aluminum pads. Figure 2.17b shows an early TMCL trace crack under a UBM that is 5 μm beyond the perimeter of aluminum pad. Even separated by a 5 μm PI repassivation layer, the stress in aluminum is high enough to crack the trace and nitride passivation.
2. Passivation open
   Wafer fab passivation open is typically enclosed by aluminum pad. Actual overlap over aluminum pad depends on fab process. However, 2.5–5 mm overlap is common. Unlike aluminum pad, which might have circular, octagon, or

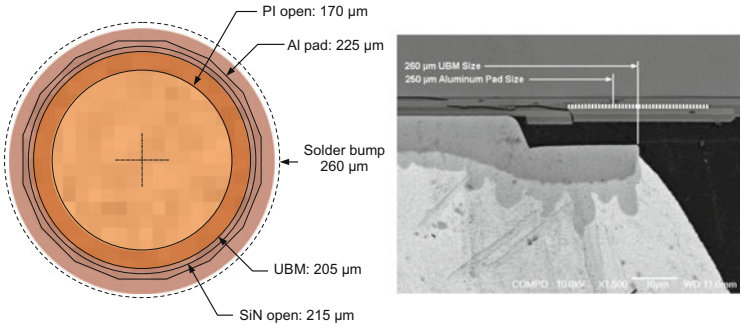


Fig. 2.17 *Top view* of an example UBM, PI repassivation, passivation and aluminum pad stack up, and enclosure relationship. On the *right* is the a FA image of an early TMCL fail on the 8×8 array with UBM size (260 μm) larger than the aluminum pad size (250 μm). The extension of UBM beyond perimeter of aluminum pad is thought as the root cause of the early failure

hexadecagon shape, fab passivation open for BOP WLCSP is always circular. However, it is not uncommon to see rounded square passivation open for RDL contact pads, so long it meets the basic overlapping rule requirements.

3. Polymer repassivation, via open and sidewall angle
   The circular open in polymer repassivation defines contact area between UBM and underlying chip layers. Poor drop and TMCL performance can often be found on WLCSP with insufficient UBM/bump aluminum bonding. Also proper sidewall slope of polymer repassivation ensures seeding metal layer coverage in the sputtering chamber, which further ensures uniform current carrying in UBM electrolytic plating. Due to the same seed metal coverage consideration, polymer repassivation typically covers the wafer fab passivation, which oftentimes does not have the desired sidewall slope for seeding metal coverage. However, in the case of electroless NiAu (ENIG) or NiPdAu (ENEPIG) UBM, where no sputter seed is needed, polymer repassivation, if applied before ENIG or ENEPIG, is pulled back to the fab passivation top to avoid corner lifting of polymer repassivation in the electroless plating process. Oftentimes, the lifting of repassivation terminated on top of aluminum surface gets started in the aluminum pad pre-clean step already. Figure 2.18 illustrated the difference between polymer repassivation open in the sputtered/plated UBM flow and electroless plating flow.
4. UBM size and metal stack
   UBM size defines the minimum solder bump cross section on the component side, which directly relates to the TMCL performance. There was referred 80 % (of solder ball size) rule of UBM size, which is believed to well balance the lateral size and solder joint standoff height. Practically, UBM size often breaks the 80 % rule and often are on the large end. For example, for 0.4 mm pitch WLCSP, the most often used solder ball size is 250 μm, and the 80 % rule suggests a UBM of 200 μm in diameter. In practice, it is not uncommon to see UBM size in the range of 230–250 μm. Though standoff height is reduced on larger UBM, the increase of solder joint cross-section size and PI open size seems to perform better in BLR than smaller UBM with higher standoff height. One other practical consideration of UBM size is the solder bump height. In case

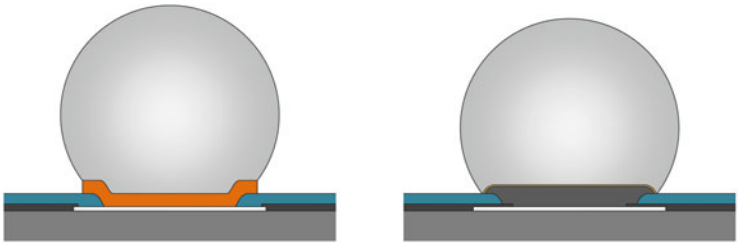
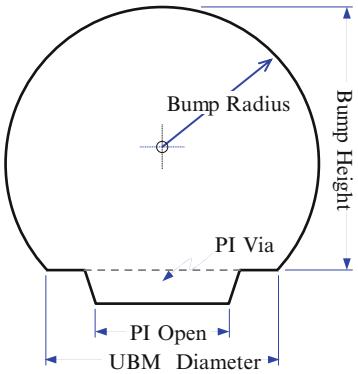


Fig. 2.18 Illustration of polymer repassivation open in the sputtered/plated UBM flow and electroless plating (ENIG) UBM flow. In the sputtered/plated UBM flow, polymer open is inside the fab passivation perimeter; while in the ENIG flow, polymer open is outside the fab passivation perimeter

Table 2.4 Solder bump height of 250 μm and 200 μm solder balls dropped and reflowed on various UBM sizes, assuming the same overlapping rules

UBM size (μm)	200	215	230	245	260	200	215	230
Solder ball	250 μm diameter					200 μm diameter		
Bump dia. (μm)	256	259	262	267	274	215	222	232
Bump height (μm)	208	201	194	187	179	147	139	131

a low profile WLCSP is needed, and with practical limitation of silicon back-grinding thickness, making UBM bigger or dropping a smaller solder ball on the same size UBM is a cost-effective approach for low profile bump and WLCSP. Table 2.4 below exemplifies the simple relationship between UBM size and bump height using the simple yet realistic truncated sphere model on the right, giving otherwise similar overlapping rules and polymer repassivation thickness.



5. Solder alloy
   Among the critical factors for WLCSP board level reliabilities, solder ball in the middle of the very dissimilar silicon and PCB planes plays the most important role in defining the reliabilities. For mostly consumer-oriented application, lead-free solder is the only choice for WLCSP bumping. With many variations of lead-free solder alloys, high/low silver SAC alloys are most widely used on WLCSP. Generally it is believed that high silver alloys are better for TMCL due to higher tensile strength and elongation at break and low silver alloys are better for drop due to less IMC growth. However, while the statement is true with numerous published paper and convincing experimental data to support, packaging engineers need to be aware that specific rules apply best to a planar Sn/Cu interface, where crack propagated through IMC is the driving failure mode in drop test for high Ag alloy (Fig. 2.19a). To improve without drastically changing the overall UBM structure, low Ag or low Ag with dopant that reduces the Sn/Cu IMC growth is the most straightforward solution (Fig. 2.19b). In the cases where polymer repassivation is applied, the Sn/Cu interface is no longer flat, and the drop failure mode is dominated by the crack propagated through bump

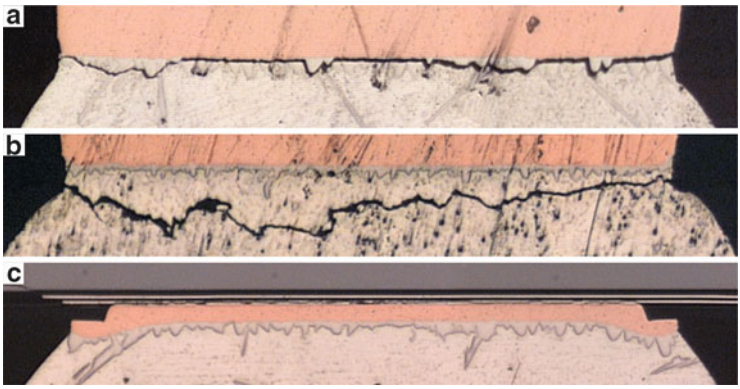


Fig. 2.19 Cross section (**a**) shows crack propagated though the Sn/Cu IMC when SAC405 was used. Cross section (**b**) shows crack propagated through bulk solder when SAC1205N is used. The IMC size and structure is quite different between these two solders on the otherwise identical copper UBM. Cross section (**c**) shows crack propagated through the under bump aluminum pad with PI repassivation, gull winged (non-flat) UBM, and SAC405

aluminum layer (Fig. 2.19c); high silver solder can perform better than low silver solder even in drop test. In case like this, selection of solder alloy becomes easy—high silver solder is the general selection for robust board level reliability performance.

2.10 RDL Design Rules

There are different approaches of RDL WLCSP: some aim at saving cost, while some aim at boosting reliability performance. One thing in common is that all utilizes pattern electroplated copper to redistribute chip connection to solder bump area array. Figure 2.20 highlights four RDL approaches, with (a) and (b) representing two typical three mask RDL solutions and (c) and (d) representing more costly RDL bumping approaches. In case solder ball directly dropped on RDL copper (Fig. 2.20a), the RDL copper layer has to be thick enough to survive 3× reflow without all copper being consumed by tin in lead-free solder ball. As far as mechanical reliabilities are concerned, three mask RDL approaches seem to be on the lower side than four mask RDL or molded copper post approaches. Molded copper post WLCSP technology is believed to be the most robust due to existence of RDL on polymer, copper post that extends the standoff height and aggressive backgrind of silicon that is only possible with front side molding.

1. Aluminum pad size and geometry
   RDL features small aluminum pads that provide connections from chip to RDL trace and bigger size copper pad. It is not restricted to special geometries, so long as pad size is bigger enough for polymer passivation via open, plus process margin. However, for better reliability, on-chip aluminum pad location needs to

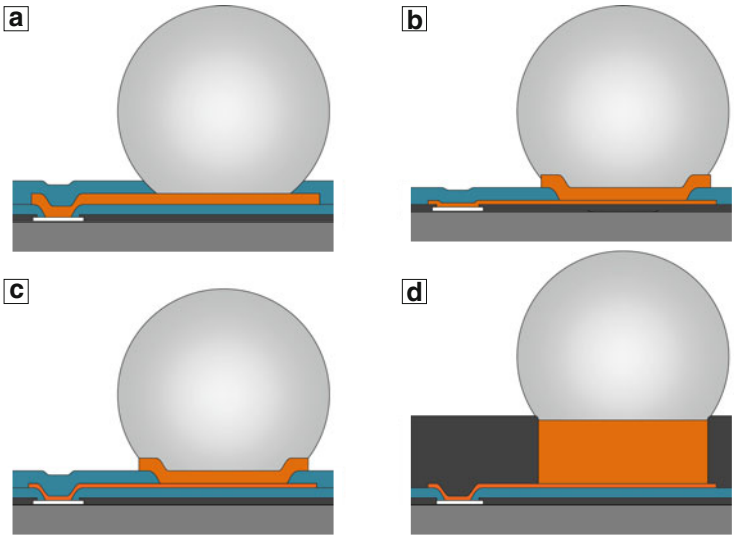


Fig. 2.20 Cross sections of a typical 3 mask RDL bump structure (**a** and **b**), 4 mask RDL bump structure (**c**), and a 3 mask RDL + molded copper post WLCSP bump structure (**d**)

be thought carefully, because it often decides the RDL copper orientation that in turn affects reliability of the WLCSP. More about trace orientation will be discussed later.

2. Passivation open
   There is no special requirement for passivation open; it could be circular, square, or even rectangles. The only consideration is the size, which has to be big enough to allow the PI1 to open completely encompassed with the passivation open. Process margin needs to be factored in the size calculations.
3. Polymer and polymer layers
   Choice for RDL polymer is abundant. For RDL WLCSP, the most common polymer of choice is polyimide. Yet there are reports claiming improved board level reliability performance due to the use of low modulus polymer materials, such as PBO.

   Regardless of the shape of the aluminum pad, polymer open in layer one and layer two is often circular (see PI via example in Fig. 2.21). Depending on the polymer layer thickness, imaging tool setting, and other process conditions, there is limitation on how small the polymer open can be. While going to smaller size is not all impossible, 35 μm minimum polymer via size is typical for most wafer bumping service providers.

   The second layer polymer open defines the contact area of UBM to copper pad or base solder to the thick RDL copper when there is no dedicated UBM. This size is found critical for reliability performance. Just like the BOP case, generally larger size gives better performance than smaller size, even with some loss of standoff height.

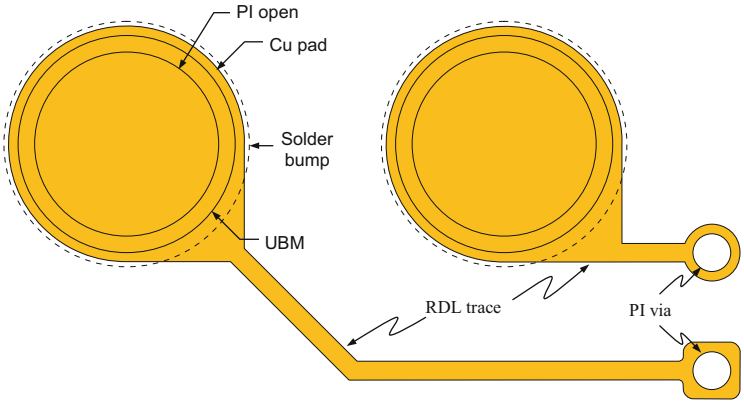


Fig. 2.21 Examples of RDL traces and copper pads

4. RDL trace and pad
   Copper RDL often serves as a routing layer besides improving WLCSP board level reliabilities. The pattern (line) formation process of RDL is very different from that of the silicon backend process. For RDL layer, copper lines and pads are pattern (additively) plated, while for silicon backend process, lines are subtractively etched on the thin aluminum layers. The line and space rules for on-chip interconnect are often from submicron to less than 5 μm range, depending the aluminum thickness and process optimization. For RDL with 3–5 μm copper thickness, it is typical to see line and space rules in the 15 μm range. So routing in RDL copper is often less dense than the on-chip interconnection layers.

   Similar to BOP WLCSP, it is desired to design the copper pad to fully enclose the UBM, with considerations of registration errors. Circular copper pad is often required, and teardrop transition from pad to trace is recommended for thin copper traces to avoid stress concentration at the pad/trace neck area. However, for traces as wide as 35 μm or up, teardrop transition can be omitted to allow more design flexibility. Figure 2.22 shows a right angle teardrop design for a fine RDL trace and a wide trace no teardrop design. The right angle teardrop is easy to layout in CAD environment and is the most often used teardrop shape.

   Sharp angles less than 90° should be all avoided in the RDL design, because they cause stress concentration at the angle tip, and it is more problematic when close to copper pad.

   For RDL to play its full potential for board level reliability gains, addition of connected RDL traces and vias for current carrying should be carefully examined, along with consideration of trace direction for the corner and immediate neighboring locations. High stress map illustrated in Fig. 2.9 should be referenced, and layout should avoid, if all possible, routining RDL traces in certain directions to the corner copper pads. The restrictions could be relaxed for non-corner solder joint locations; layout engineers should be aware of general rules for additional traces and vias to better balance mechanical reliability with electrical performance.

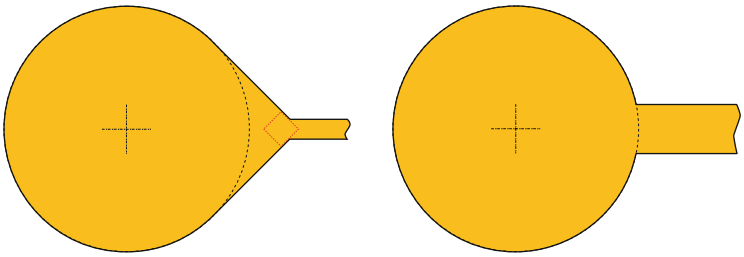


Fig. 2.22 Right angle teardrop design and a wide trace no teardrop design

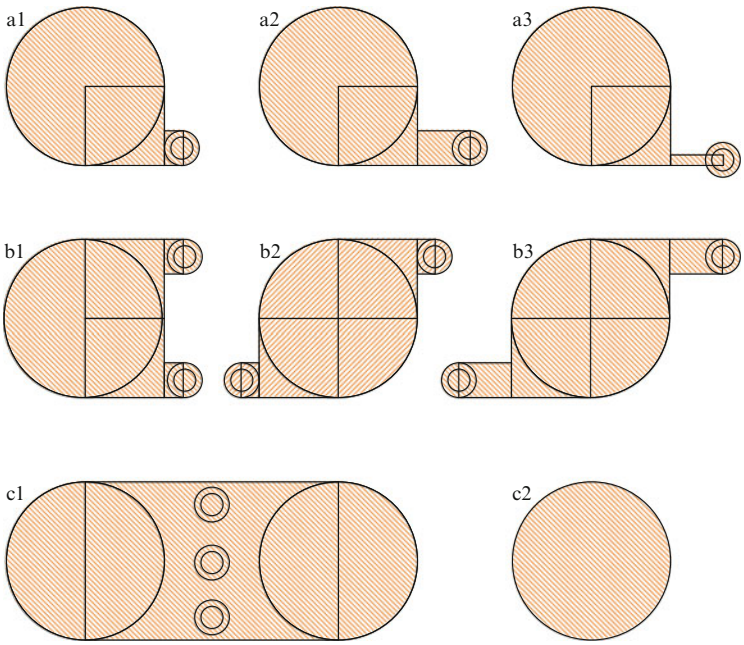


Fig. 2.23 Examples of RDL layout: (**a**) thinner and longer RDL trace is preferred to reduce the restrictions on copper pad movement; (**b**) in the two vias and two-trace case, center symmetric design is generally preferred over same side design for more flexible copper pad under stress conditions; (**c1**) is the worst case RDL trace and via layout. In this design, the copper pad under solder joint is all constrained from any possible movement when stress is applied; (**c2**) is the ideal copper pad layout with no RDL traces attached to the copper pad

In general, best solder joint reliability is achieved when there are little constraints to solder joints and copper pad on the die side. The best case scenario is when solder joint is "floating freely" on the polymer layer without any copper RDL connected. However, the best case scenario will most likely not happen in reality, unless the specific solder joint is purely mechanical or dummy joint. Therefore, RDL constraints will exist on copper pads. There are also cases where more than one RDL traces are connected to the same copper pad. In these cases, some preferences apply (Fig. 2.23). However, one should always keep in mind

the general rules show its best results in high stress situations, such as corner or edge solder joints. As soon as it is out of the most critical zones, there is more freedom to layout RDL for not mechanical reliabilities, but electrical performances.

2.11 Chapter Summary

Fan-in WLCSP technology is introduced in good details in this chapter, along with brief introduction of fan-out WLCSP. Fan-in WLCSP, or mostly referred as WLCSP, is a widely adopted packaging technologies in the consumer electronics market and is proliferating into other application fields. Many variations of the WLCSP technologies were developed over the years, each with its own advantages and specific application targets. There is no one-fits-all WLCSP technology that offers low-cost and high reliabilities at the same time. So understanding the available options and pros and cons of various WLCSP bumping technologies is important for packaging engineers to make the reliable, high-performance, yet cost-effective, decisions for specific device applications. In this chapter, basic flow of fan-in WLCSP bumping options, such as BON, BOP, RDL, and molded copper post, is described and compared in terms of reliability and cost factors. In the process of identifying suitable technologies, or the developing new technologies, precise assessment of WLCSP board level reliability performance is one of the key elements. Test chip design, as well as test PCB layout, has to be thought through completely to best avoid non-WLCSP-related failures mixed into the test results and at the same time help to isolate failure locations for proper FA.

BOP and RDL design rules are introduced as well in this chapter. However, one should always bear in mind that rules could be changed as we gain more knowledge of WLCSP and no rules could replace intelligent judgment by experienced and knowledgeable engineers.

References

1. Novel embedded die package technology tackles legacy process challenges. CSR Tech Monthly
2. Edwards, D.: Package interconnects can make or break performance. Electronic Design, Sept 14 (2012)
3. Syed, A., et al.: Advanced analysis on board trace reliability of WLCSP under drop impact. Microelectron. Reliab. **50**, 928–936 (2010)
4. Fan, X.J., Varia, B., Han, Q.: Design and optimization of thermo-mechanical reliability in wafer level packaging. Microelectron. Reliab. **50**, 536–546 (2010)
5. JESD22-B111, Board level drop test method of components for handheld electronic products
6. JESD22-B113, Board level cyclic bend test method for interconnect reliability characterization of components for handheld electronic products
7. IPC-7095, Design and assembly process implementation for BGAs

Fan-Out Wafer-Level Chip-Scale Package

3

3.1 Introduction of Fan-Out WLCSP

The concept of fan-out is hardly new for semiconductor packaging. Ever since the early days of semiconductor industry, fan-out scheme that expands the tight lead pitch on semiconductor to coarse lead pitch on package is the dominant form in all chip packages, for example, leadframe package fan-out from chip to leads via bonding wires and flip chip package fan-out from chip to BGA via inner metal layers in the substrate (Fig. 3.1a, b).

What makes fan-out WLCSP stand out, in comparing with all other packaging technologies, is that fan-out WLCSP fully adopts semiconductor wafer bumping technologies, along with unique wafer molding process and all processing steps are done in the familiar circular wafer form that is either 200 or 300 mm in diameter. Modern semiconductor-like high bumping yield and fine line capabilities are therefore expected for fan-out WLCSP, taking the advantage of wafer fab manufacturing process. On the other side, fan-out WLCSP also carries the overhead of semiconductor cost structure, which always comes in light as soon as fan-out discussion is on the table.

Fan-out concept came to life as a result of proliferation of mobile device, such as cell phone and tablet computers, and accompanying demands for semiconductor miniaturization and cost reduction. Originally, fan-out was developed as a substrate-less embedded chip package that offers smaller form factor, lower cost, and higher performance than the widely adopted substrate-based wire bond BGA and flip chip BGA packages. It is soon realized that in certain circumstances, even traditional WLCSP could take the advantages of fan-out WLCSP approach, if the die size can be aggressively shrunk without the minimum 0.4 or 0.5 mm bump pitch limitations for the purpose of high-yielding surface mounting. Over the years, rapid advancement in semiconductor technology has allowed shrinkage of the silicon die size via the progress of every generation of fab technologies. Technically, there is little hurdle to make certain WLCSP device at 0.3 mm or even finer pitch from modern semiconductor manufacturing—this is especially true when transitioning

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_3

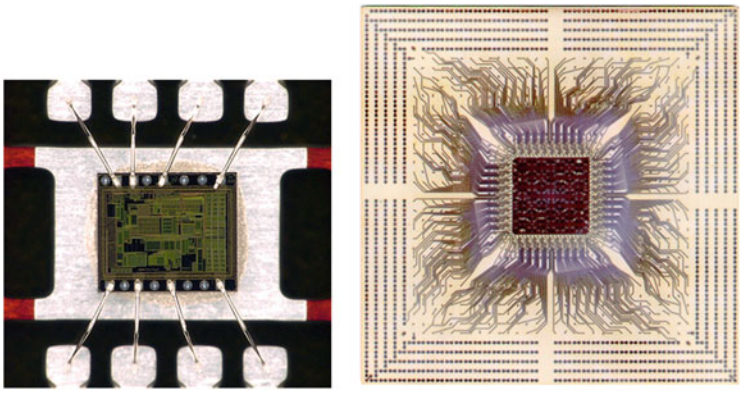


Fig. 3.1 Examples of fan-out in semiconductor packages: (**a**) fan-out from 70 μm on-chip terminal pitch to 0.40 mm leadframe pitch via bonding wire in a DFN (dual flat no-lead) leadframe; (**b**) fan-out copper layer inside a 40 × 40 mm flip chip BGA substrate from 0.18 mm on flip chip terminal pitch to 0.8 mm BGA terminal pitch

from older fab technology to newer fab technologies for analog/power semiconductors. At the same time, advancement of PCB technology has been relatively slow, for example, though 0.4 mm pitch WLCSP has become mainstream, finer pitch WLCSP, such as 0.35 mm pitch and 0.3 mm pitch, is yet to happen at a large scale, due to difficulties in high-yield manufacturing of fine pitch PCB. So to fully utilize the advanced semiconductor wafer technology and save on per unit cost, which means realizing the same functionality with smaller silicon size, fan-out from the reduced wafer terminal pitch to commonly accepted, manufacturing-worthy PCB pitch could be a cost-effective approach. The thought is that with sufficiently reduced silicon die cost, additional fan-out WLCSP cost can be justified and still yield overall reduced packaged device cost for the same or similar functionalities.

To better understand fan-out WLCSP, it is beneficial to review the process of one layer fan-out WLCSP with a typical RDL WLCSP bumping process, since both adopts two polymer passivation layers and copper redistribution technology, flux print, ball drop and reflow, as well as wafer form laser marking, die singulation, tape, and reel as the finishing steps. In Fig. 3.2, two processes are compared side to side in two columns. Processing steps taking place on the same production platforms are listed in the unshaded cells, while unique processing steps for fan-out process are highlighted in the shaded cells. Graphical schematics are also provided in Fig. 3.3 to assist. Readers should be advised though, the inspection steps, important in WLCSP and fan-out WLCSP bumping, are not listed in the table since when and what to inspect is more or less manufacture dependent.

It might be illusive from Fig. 3.2 that the difference between fan-out WLCSP and 4 mask RDL bumping is subtle—only one extra wafer probing step before pick and place of known good die (KGD) and wafer molding. However, the wafer molding step is a multistep process including adhesive coated temporary carrier

RDL WLCSP	
1	Polymer 1 coat
2	Polymer 1 image/develop/cure
3	RDL seed layer sputter
4	Resist coat
5	Resist image/develop
6	RDL copper pattern plate
7	Resist strip
8	Seed layer etch
9	Polymer 2 coat
10	Polymer 2 image/develop/cure
11	UBM seed layer sputter
12	Resist coat
13	Resist image/develop
14	UBM pattern plate
15	Resist strip
16	Seed layer etch
17	Flux print
18	Solder ball drop
19	Solder reflow
20	Wafer probe
21	Wafer backgrind
22	Back side laminate
23	Laser mark
24	Wafer saw
25	Tape and reel

Fan-out WLCSP	
1	Wafer probe
2	Wafer back grind
3	Wafer saw
4	KGD pick and place
5	Wafer mold*
6	Polymer 1 coat
7	Polymer 1 image/develop/cure
8	RDL seed layer sputter
9	Resist coat
10	Resist image/develop
11	RDL copper pattern plate
12	Resist strip
13	Seed layer etch
14	Polymer 2 coat
15	Polymer 2 image/develop/cure
16	UBM seed layer sputter
17	Resist coat
18	Resist image/develop
19	UBM pattern plate
20	Resist strip
21	Seed layer etch
22	Flux print
23	Solder ball drop
24	Solder reflow
25	Wafer probe
26	Laser mark
27	Wafer saw
28	Tape and reel

Fig. 3.2 Process flow comparison between RDL WLCSP and fan-out WLCSP. (*Asterisk*) Wafer molding is a multiple processing step as highlighted in Fig. 3.3 step 5a through 5c

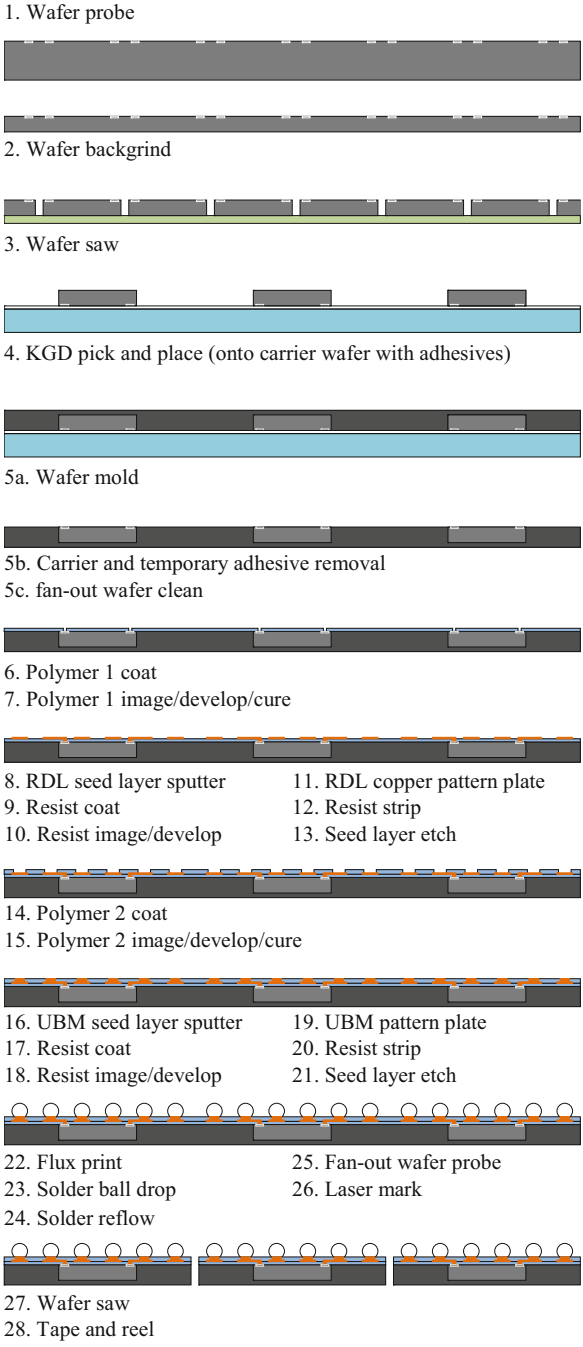


Fig. 3.3 Process flow for a typical fan-out WLCSP

bonding, wafer molding, carrier and adhesive removal, clean and inspection to record die locations, rotational angles, etc. These extra steps and processing challenges add significant costs to the fan-out WLCSP and make it a viable production option to displace WLCSP only when significant wafer cost saving is realized by redesigning a larger chip that is suitable for traditional WLCSP into a much smaller silicon size that, even with additional fan-out cost, the final package is still cost-effective.

Substantial overlap of application space exists among WLCSP, fan-out WLCSP, and BGA packages, when package body size and bump count are being considered (Fig. 3.4). In reality, other factors, such as cost, bump pitch, package height, bump layout, time to market, etc., also need to be thought about in selecting the right packaging solutions for specific devices. As stated before, fan-out WLCSP is only a sensible choice over WLCSP in unique cases when overall packaging cost and reliability outperforms the later, but for traditional wire bond or flip chip BGA packages, fan-out WLCSP could be a good alternative approach without separate wafer bumping, substrate build, flip chip attach/reflow, or wire bond and over-molding processing steps.

Fan-out WLCSP integrates semiconductor packaging into a bumping process of the reconstituted wafers and provides die and system packaging solutions in attractive low profile packages. Fan-out also addresses the limitations associated with the existing packaging technologies by eliminating wire bonds, package substrates, and flip chip bumping. Figure 3.5 provides illustrative cross-sectional view of fan-out WLCSP alongside wire bond and flip chip BGA packages. It shows that without additional substrate thickness and spaces for wire loop and flip chip/substrate solder interconnections, fan-out WLCSP is inevitably a lower profile package solution than BGA packages. Besides thickness, fan-out simplifies package assembly and is more cost competitive than typical BGA package.

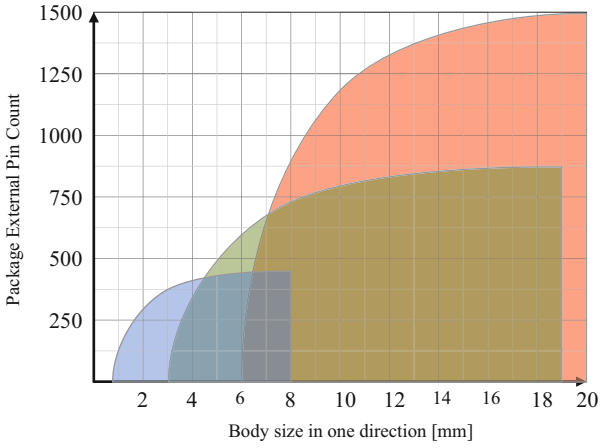


Fig. 3.4 Application space of WLCSP, fan-out WLCSP and BGA packages. Overlaps are evident

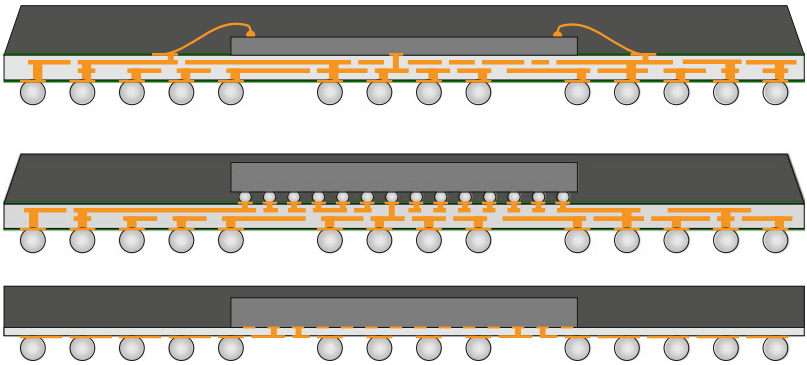


Fig. 3.5 Cross sections of wire bond BGA package, flip chip BGA package, and fan-out WLCSP. Fan-out WLCSP can achieve thin package profile easier than both BGA package due to elimination of substrate, bonding wires, and flip chip solder joints

Advanced wafer manufacturing processes utilizing low-**κ** interlayer dielectrics pose serious challenges to the packaging engineers due to the naturally weak material properties. Fan-out WLCSP, which avoids excessive mechanical stresses that low-**κ** semiconductor devices have to endure during wire bonding and flip chip die attaching process, seems to offer a good alternative to BGA packages as well.

3.2 High-Yielding Fan-Out Pattern Formation

BGA substrate manufacturing and fan-out WLCSP both adopt additive pattern plating process to form conductor (copper) lines. Yet fan-out WLCSP achieves higher routing density than typical substrate, for example, 10 μm line/space for fan-out WLCSP versus 25 μm line/space for BGA substrate. This major difference allows fan-out to often use one routing layer to achieve the same functionality that requires two or more layers in a typical substrate. Two main factors play here to enable the fine line capability on fan-out: (1) ultrasmooth and ultrathin sputtered adhesion/seed metal that allows easy etch clean of narrow spaces between fine lines without comprising the interfacial adhesion and (2) semiconductor wafer processing tool with precision control of fluid flow that allows much more uniform plating and etch across a 200 or 300 mm diameter. In contrast, typical substrate process takes place on a panel size that could be more than 600 mm (24 in. on a side, with plating seed layer that relies on microscopic mechanical interlocking structure for acceptable adhesion. The mechanical locking structure built on substrate seed layer makes it harder to clean between lines when space is tight. What is more, the substrate manufacturing plating and etching tool, though seeing continuous improvement over the years, is still far less controlled to the level that is necessary for the uniformity achieved on a typical single wafer processing tool. Figure 3.6

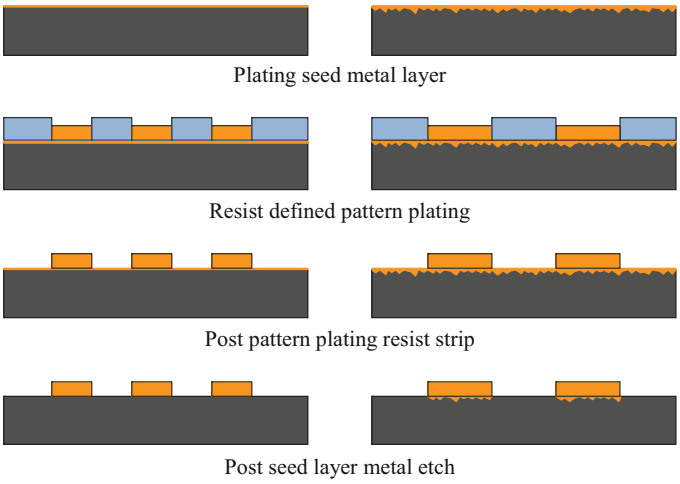


Fig. 3.6 Comparison of pattern formation on fan-out WLCSP and typical organic chip package substrate. Distinctive differences are seed layer metals: on fan-out WLCSP, it is sputtered seed metal with smooth bonding interface; on substrate, it is a rough mechanical locking structure built in the seed metal

illustrates the seeding layer (adhesion layer) differences in pattern formation between fan-out WLCSP and traditional BGA substrate.

Nevertheless, fan-out WLCSP expands the envelope of the traditional WLCSP and opens the door for multi-chip packaging, passive/chip stacking for system in package (SIP), and many other 3D packaging concepts. So it is not a surprise to see fan-out remains a hot topic at different technology venues. Besides wafer-level fan-out that adopts semiconductor wafer manufacturing technologies, fan-out utilizing traditional PCB manufacturing technology and infrastructure is also in the work, with promises of offering cost-effective approaches for multi-chip packaging and SiP, as well as enhanced thermal performance.

3.3 Redistributed Chip Package and Embedded Wafer-Level Ball-Grid Array

Redistributed chip package (RCP)[1] was first announced in 2006. Embedded wafer-level ball-grid array (eWLB)[2] was released to news in 2007. The two technologies share similar concept and much of the basic process flow and are the most referred fan-out WLCSP technologies. Comparing to typical eWLB package cross sections,

[1] Redistributed chip package (RCP) is a proprietary packaging technology developed by Freescale Semiconductor.

[2] Embedded wafer-level ball-grid array is a proprietary packaging technology developed by Infineon Technologies AG.

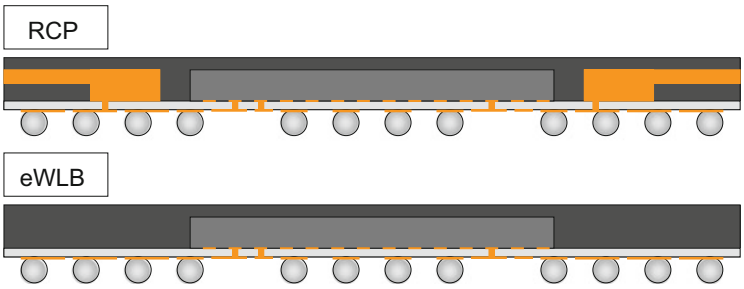


Fig. 3.7 Cross section of RCP (Freescale) and eWLB (Infineon) fan-out WLCSP. The RCP cross-section features an embedded copper ground plane, which is unique for the specific fan-out WLCSP

RCP might feature an embedded copper ground plane that is not seen in a eWLB (Fig. 3.7). The embedded copper plane with window opens for semiconductor devices or integrated passives in RCP is said to help limit the chip movement during wafer molding and provide device electromagnetic (EM) shielding, rigidity for the final chip packages, etc. In reality, the addition of embedded copper plane in RCP adds up to the packaging materials and processing cost while helping the manufacturing yield by limiting chip movement during molding process. Overall benefits vary depending on the materials/process selection, die and package design, performance requirements, and other manufacturing cost factors.

3.4 Fan-Out WLCSP Advantageous

Fan-out WLCSP is suitable for both highly sensitive analog devices and digital platforms. The technology is compatible with small or larger package sizes. Fan-out can accommodate single and multiple routing layers to optimize package size, performance, die size range of I/O, and cost. Key advantages of fan-out WLCSP include (1) improved electrical performance resulting from shortened routing distances and reduced contact resistance; (2) reduced cost due to smaller die size that is only enabled by the advanced semiconductor manufacturing technologies; elimination of substrate, wire bond or flip chip interconnections, over-molding, etc.; assembly processing steps; and large wafer format batch processing that is significantly larger than typical strip format for wire bond and flip chip assembly; (3) reduced assembly stress which makes it suitable for packaging low-κ dielectrics that becomes increasingly common on modern semiconductor dies.

In electric performance improvement by the fan-out WLCSP, major performance enhancement comes from the replacement of bond wires, and flip chip solder interconnects with copper metalized via contacts. Fewer routing layers, smaller package size that is achieved by aggressive line, and space rules also contribute to the lower package resistance and inductance. In a study by Infineon,

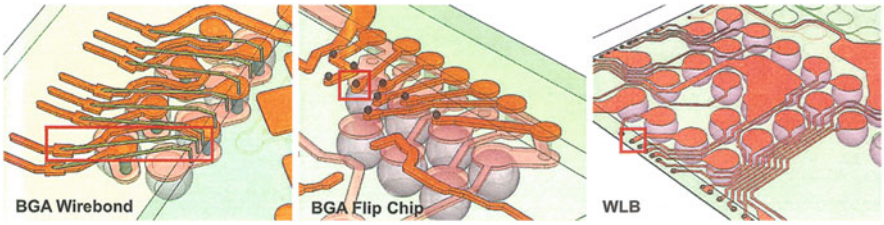


Fig. 3.8 Electrical modeling on wire bond BGA, flip chip BGA, and eWLB fan-out WLCSP. Chip to package interconnections are highlighted in *red squares*/*rectangle*

Table 3.1 Electrical modeling of chip to package interconnections

Package	Wire bond	Flip chip	Fan-out
R at DC	76 mΩ	7.5 mΩ	3.2 mΩ
R at 5 GHz	375 mΩ	41 mΩ	15 mΩ
L	1.1 nH	52 pH	18 pH

Table 3.2 Electrical modeling of chip packages

Package	Wire bond	Flip chip	Fan-out
R at DC	89 mΩ	22 mΩ	23 mΩ
R at 5 GHz	629 mΩ	248 mΩ	91 mΩ
L	1.79 nH	0.95 pH	0.34 pH

electrical peance was modeled on a wire bond BGA package, a flip chip BGA package, and a eWLB fan-out WLCSP, with comparable package functionalities (Fig. 3.8). The advantage of fan-out WLCSP with low overall package parasitics is evident (Tables 3.1 and 3.2).

3.5 Fan-Out WLCSP Challenges

To take most of the advantageous of fan-out WLCSP, such as cost savings and package size reduction, fine pitch, either on semiconductor device or package, is often needed. When pitch is fine, challenges are encountered going through the fan-out process with reasonable yield and cost targets. Two of the most significant challenges for fan-out WLCSP are (1) chip moving during mold process and (2) low process temperature imposed by the molding compound, which limits the choice of polymer repassivation materials for RDL copper to a few.

Compression molding, either liquid resin based or dry powder/granular based, is adopted for fan-out WLCSP wafer molding due to the desire to minimize hot resin lateral flow that is common in the transfer molding. Dry resin-based molding compound has the advantages of long shelf and floor life, while liquid resin has the advantages of low viscosity and good narrow space-filling capability. Regardless what is the use, molding compound, like all other thermal set resins, has volume shrinkage when cross-linked (in molding and curing stage). Also thermal

shrinkage of molding compound (CTE > 8 ppm/°C) from curing temperature is always more than that of silicon (CTE 2–3 ppm/°C). Molding compound dimensional changes from curing volume shrinkage and thermal shrinkage from molding temperature to room temperature will affect silicon die location in the reconstituted fan-out WLCSP wafers. Die shift and/or rotations from initial placement position is inevitable, which poses immediate challenges to the subsequent bumping steps following the wafer molding.

For typical wafer bumping, standard imaging tool, mostly stepper or aligner, is used to precisely define each bumping layer, such as polymer repassivation layer, RDL and UBM, etc. Exposure typically involves an array of individual dies (for stepper), if not the whole wafer (aligner). While it is possible for either imaging tool to adjust to slight rotation and shift across the exposure field, neither tool is designed to compensate shift or rotation of individual dies within the exposure field. Furthermore, flux printing and solder ball drop also rely on stencils that are made to match a single wafer layout. Die shift in fan-out WLCSP, if not controlled under certain threshold value, makes it hard for fine pitch interconnection registration, flux print, and well-wetted solder bumps. In the worse scenario, there could be no reliable interconnection forming at all at the severely misregistered die site. So controlling or managing the die shift is a must to allow the use of semiconductor wafer imaging tool for high-yielding fan-out WLCSP wafer bumping.

It is understandable that die shift requirements vary depending on the pitch and size of the design features, with coarse pitch and larger feature size tolerating more die shift than fine pitch designs without compromising bumping yield. In the example given in Fig. 3.9, a comparison is made between a case with no die shift and another with 30 μm die shift in one direction. Clearly in the second case, the 30 μm via contact from copper RDL to the chip metal layer with 75 μm pitch is on the borderline of shorting the two adjacent pads on the silicon chip. For fan-out WLCSP, which sees more adoption benefits at fine pitches, characterizing the die shift and searching for solutions to minimize or eliminate die shift becomes one of the most critical elements to ensure fan-out is a manufacturing-worthy process.

Extensive researches have been carried out to understand the die shifting phenomenon through fan-out WLCSP wafer molding and mold compound curing steps. A widely accepted methodology in the die shift study is to map out individual die shift across the whole reconstituted wafer. In a publication authored by Sharma et al. [3] in 2011 entitled "Solutions Strategies for Die Shift Problem in Wafer Level Compression Molding," die shift is studied using test chips with built-in cross targets, as shown in Fig. 3.10. Die shift is clearly reviewed when stacking the same size cross targets in the overlay copper layer on top of the cross target built in the aluminum. Further studies of die shift followed in three categories: (1) die shift on regular adhesive tape only, (2) die shift on adhesive tape with silicon carrier, and (3) die shift with silicon carrier and low CTE and low shrinkage molding compound. As reviewed in series from Fig. 3.11a–c, die shift can be dramatically influenced by the existence of silicon carrier and molding compound material properties. When there is no silicon carrier attached to the single-sided adhesive tape for KGD placement, the expansion of adhesive tape (CTE > 20 ppm/°C) at

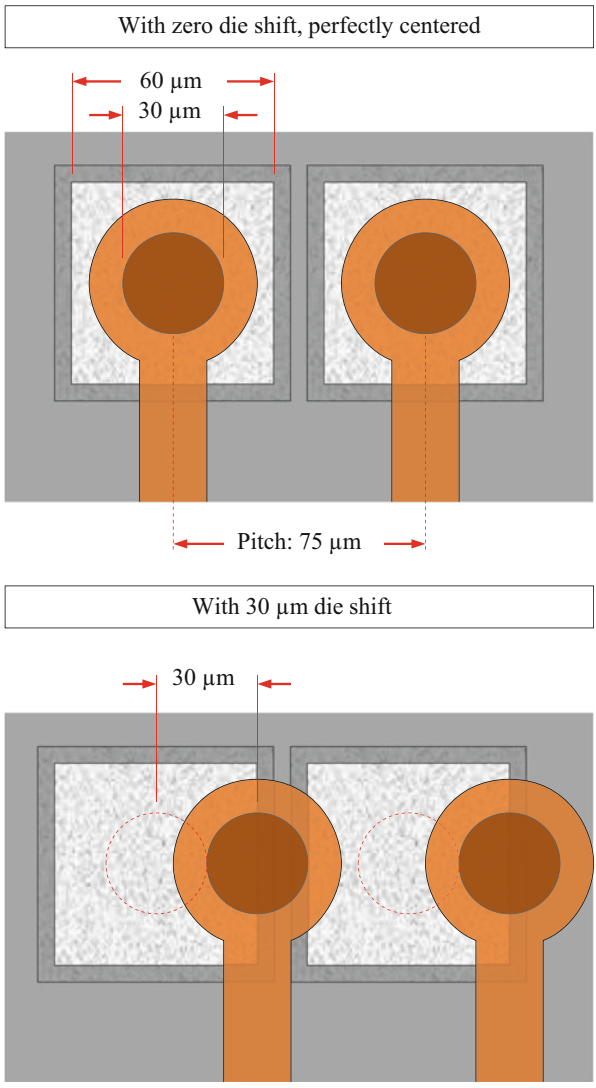


Fig. 3.9 Die shift and influence on copper RDL layer to chip metal layer registration

molding temperature (>150 °C) is not fully canceled out by the molding compound (CTE > 8 ppm/°C) thermal shrinkage. So overall result is die shifting away from the center point (Fig. 3.11a). When a silicon carrier wafer is used with KGD attached to it via a double-sided adhesive tape, thermal contraction from silicon carrier (3 ppm/°C) and adhesively attached dies are less than the thermal shrinkage of molding compound; overall effect after mold compound cure and removal of silicon carrier is die shifting towards the wafer center (Fig. 3.11b). Low 9 shrinkage and low CTE molding compound just make that shift even less significant (Fig. 3.11c), but the shifting direction is similar to the case with silicon carrier.

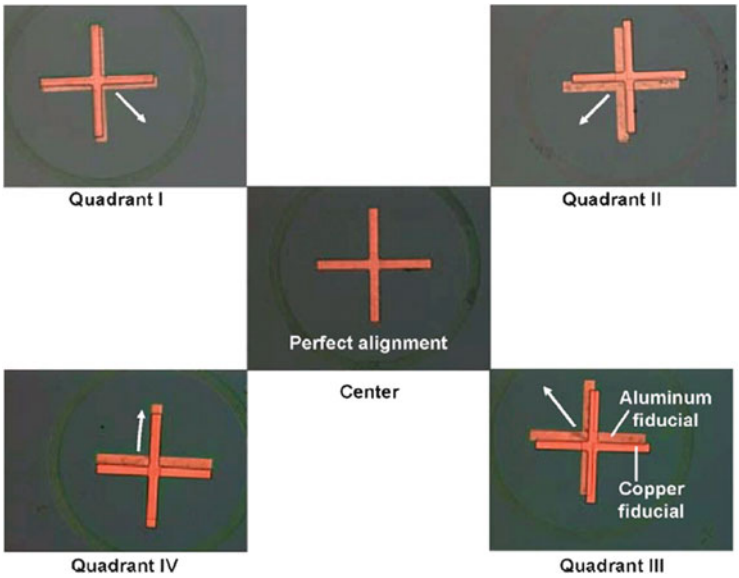


Fig. 3.10 Die shift through fan-out molding process revealed by the cross target method. Here cross targets built on the chip and cross targets build in the bumping layer, at 5 locations across a 200mm fan-out wafer, show the directions of die shift through the molding/curing process

Knowing the die shift exit in fan-out WLCSP wafer molding/curing process, it is more or less a matter of managing it or, to be more specific, compensating it so as to reduce or eliminate the die shift in the final finished reconstituted, molded fan-out wafers. The thinking is fairly straightforward: if a die is known to move in one direction during wafer molding, then it should be placed in the opposite direction away from the targeted, finishing position. This is what was exactly done by Sharma et al. Three scenarios were investigated: (1) no shift compensation, (2) full shift compensation based on the shift measured from case #1, and (3) half shift compensation. Figure 3.12 shows detailed die shift map of the three investigated scenarios. It is interesting, yet not surprising, to see that half shift compensation yielded the best results of all, since die shift is not exactly linear from center to edge (Fig. 3.11a–c) and also die shift applied is not measured from the exact location a shifted die is placed.

Table 3.3 summarizes the results that are graphically presented in Fig. 3.12. The trend is clear between non-compensated case and compensated case. It is also evident that optimized die shift compensation can reduce the magnitude of average die shift across a 200 mm wafer to less than 30 μm, a level that is considered acceptable for 75 μm pitch applications. For even tighter pitch designs, more thorough study is required. Iterations of trial and errors are necessary for ultimate process optimization.

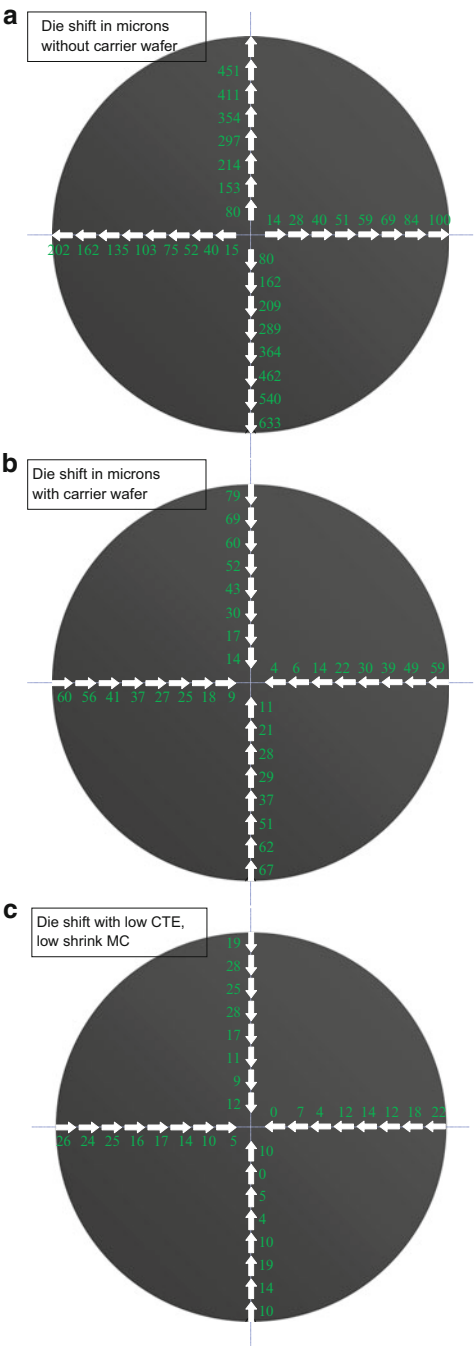


Fig. 3.11 Fan-out WLCSP die shift in the diagonal directions across 200 mm wafer size, all measured in micrometers: (**a**) die shift without silicon carrier, (**b**) die shift with silicon carrier, (**c**) die shift with silicon carrier and low CTE, low shrinkage molding compound

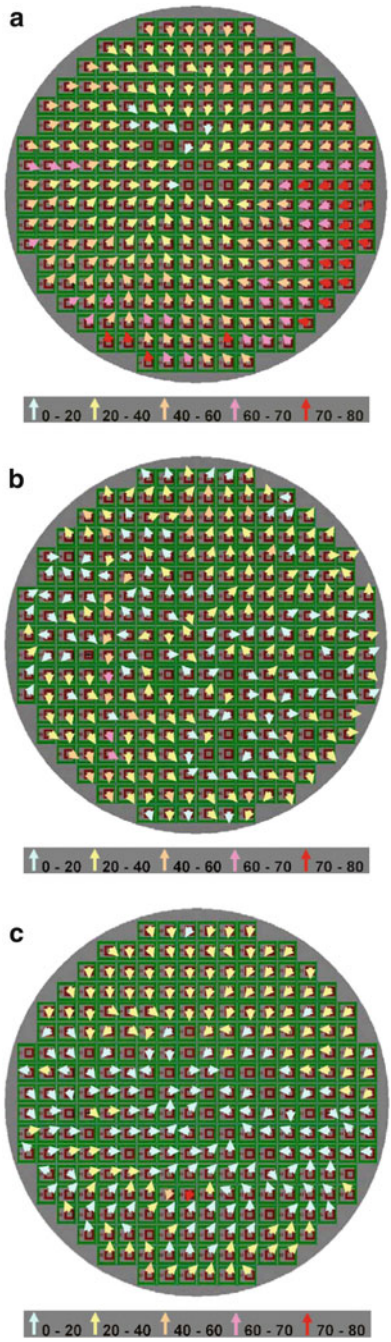


Fig. 3.12 (**a**) Die shift mapping after molding, with no shift compensations. Dies with no *arrows* are locations where no measurable die shift could be observed. All the dies tend to shift inward. The magnitude of shift is somewhat proportional to the distance from the wafer center. (**b**) Die shift mapping after molding with 100 % pre-shift during pick and place process. If shift at certain die location measures as (a, b) in x, y coordinates, then 100 % pre-shift corresponds to deliberate misplacement during pick and place by (−a, −b). (**c**) Die shift mapping after molding with 100 % pre-shift during pick and place process. If shift at certain die location measures as (a, b) in x, y coordinates, then 100 % pre-shift corresponds to deliberate misplacement during pick and place by (−a, −b)

Table 3.3 Fan-out WLCSP die shift comparison

	No die pre-shift		100 % die pre-shift		50 % die pre-shift	
Die shift range	Cumulative %	Average shift (μm)	Cumulative %	Average shift (μm)	Cumulative %	Average shift (μm)
<20 μm	6 %	15 ± 9	41 %	14 ± 7	55 %	12 ± 6
<40 μm	38 %	28 ± 10	89 %	24 ± 9	99 %	19 ± 10
<60 μm	81 %	40 ± 14	99 %	26 ± 12	99.8 %	19 ± 10
<70 μm	92 %	43 ± 16	100 %	27 ± 12	100 %	19 ± 10
<80 μm	100 %	45 ± 17				

Fan-out WLCSP die shift depends heavily on die to resin area ratio, since die shift is induced by molding compound resin curing volumetric shrinkage and the post cure thermal contraction from high curing temperature to room temperature. It is not easy, if not all impossible, to develop one simple formula for all different die size and fan-out WLCSP designs. Furthermore, molding compound resin materials, percent and type of filler loading, will also affect the shrinkage and CTE of the final materials. Adhesion of tape to carrier and chip to tape, especially at elevated molding temperature, plays yet another role in die shift management.

Besides die shift during wafer molding and curing, copper RDL fan-out process has to be reestablished from the standard WLCSP RDL process. For standard WLCSP, polyimide, such as HD-4100 series from Hitachi Chemical, is one of the most widely adopted polymer repassivation materials. Once cured at close to 350 °C, the materials are fully compatible to the copper RDL and exhibit good adhesion to the underlying materials, such SiN, PI, and copper, as well as good mechanical strength and resistance to film crack on top of the well-known chemical resistance. The most widely used WLCSP polyimide materials are solvent developed, meaning an organic solvent is used to develop away the polymer in the areas it was not needed. While solvent development is never an issue for silicon wafers, it is not desired for the epoxy-based fan-out molding compound. Besides the solvent concerns, high curing temperature of repassivation polymer for typical WLCSP is beyond the survivable temperature range for fan-out molding compound. Since fan-out is more suitable for system in package (SIP) solutions, the high processing temperature also imposes practical circuitry survivability concerns to the embedded memory in the high functional ICs and ICs featuring low-κ dielectric materials, which are themselves temperature sensitive. So for fan-out WLCSP, solvent-free, low curing temperature polymer repassivation material is a non-compromising requirement. Solvent-free materials are also desired with many manufacturing operations focusing more on environmental stewardship and looking every way to reduce their use of organic solvents.

This desire to reduce the use of organic solvents led to the introduction of polybenzoxazole (PBO)-based dielectric materials. These were processed with an aqueous-based developer; in fact, the same one is used for photoresists. PBOs have similar properties to polyimides, but while they cannot hold up to high processing

temperatures compared to polyimides, they tended to fully cure at lower temperatures and exhibited properties that helped RDL WLCSP survive drop tests required for handheld devices. For fan-out WLCSP application, a further modified PBO formulation has been developed that cures at as low as 200 °C. There exist trade-offs, such as reduction in the mechanical and chemical resistance properties of the dielectric, but the low cure PBO provides a workable path forward for fan-out WLCSP with development of a suite of process chemicals that allows a robust process for volume manufacturing.

3.6 Reliability of Fan-Out WLCSP

Fan-out WLCSP subjects to the same reliability requirements set for semiconductors in handheld devices. Board level reliability, such as drop and temperature cycle (TMCL), is the most asked question concerning the fan-out technology. Fortunately for fan-out, which features smallest possible silicon surrounded by molding compound that has material properties closely matched to the PCB dielectric materials, reliability always outperforms the comparable size WLCSP.

In a simulation study of 16 × 16 bumps, 0.5 mm pitch fan-out WLCSP with 3 × 3 bumps under silicon, Fan et al. [8] clearly demonstrated that maximum inelastic strain energy density, which corresponds to the board level TMCL life, occurs at the corner of silicon chips and decreases sharply even with DNP (distance to neutral point, aka center point of WLCSP) increase into the outer molding compound area (Fig. 3.13). In the molding compound area, as soon as it is away from the silicon, the inelastic strain energy density does not vary with the DNP, since the CTE of the molding compound is closely matched to that of PCB materials (both are epoxy materials with either glass particles fillers or woven glass cloth reinforcement).

So for fan-out WLCSP, the stress and distribution, and thus package board level reliability, is determined by the size and location of the embedded silicon. High stress area is at the corners and edges of the silicon. Package size that grows beyond silicon size for fan-out bumps does not affect the package life the same way as typical all silicon WLCSP, which shows a simple relationship between DNP and inelastic energy density in solder joint (Fig. 3.14). When designing fan-out WLCSP, by purposely avoiding bumps in the high stress area, robust reliability performance of fan-out WLCSPs should be expected.

Standard WLCSP like component reliability performance has been reported for the fan-out WLCSP, such as pass of MSL conditions, HAST, HTS, and TMCL. Area of reliability interest is mostly the silicon/molding compound neighboring area and narrow space between RDL copper traces, due to CTE mismatch between silicon and molding compound and low curing temperature polymer, respectively. However, both seem to be robust enough to survive the common requirements applied to WLCSP.

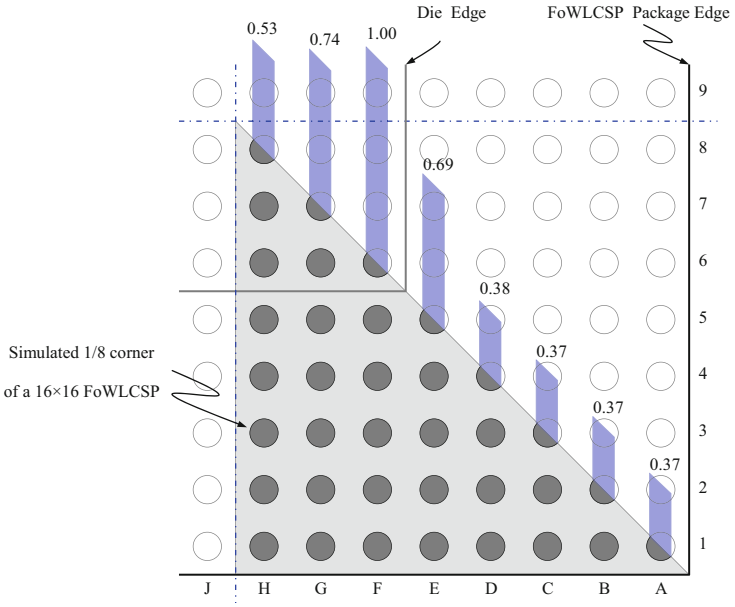


Fig. 3.13 Normalized inelastic strain energy density along the diagonal of a 16×16 bump fan-out WLP package with a 3×3 mm, 0.5 mm pitch silicon die size embedded in the middle of the package. Higher inelastic strain energy density prompts earlier TMCL fails

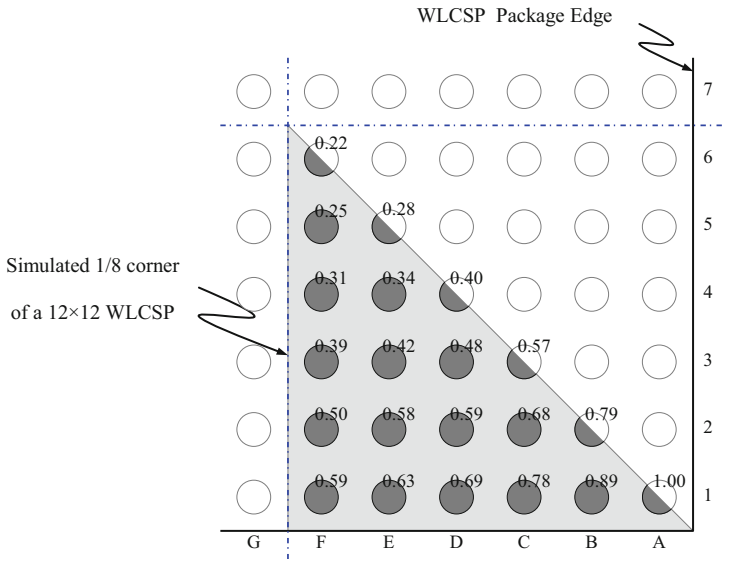


Fig. 3.14 Normalized inelastic strain energy density per cycle for the 1/8 part of a 12×12 bump WLCSP. Higher inelastic strain energy density prompts earlier TMCL fails

3.7 Fan-Out Design Rules

Aside from wafer molding and challenges of handling the molded fan-out wafers, fan-out bumping process shares much with the long-standing RDL process. The design rules for fan-out are essentially the same as RDL rules. Line and space rule is the basic matrix for measuring the bumping technology. It is often seen in various development road map paths forward from 15 or 20 μm L/S to 10 or even 8 μm L/S. In reality, due to the weak photoresist adhesion to the sputtered copper plating seed layer and with process bias being considered, additive copper plating RDL pattern formation process has a practical line and space limit close to 10 μm. Anything beyond it requires innovation in materials and substantial engineering work. It is never a simple matter of scaling down to finer geometries.

Fan-out WLCSP does have its own unique design requirements that are not presented in the WLCSP design book. For example, special requirements for die and package thickness exist for the balance between wafer warpage and robustness of automated wafer processing, placement of silicon in the fan-out package for the best electrical or thermal performance, etc. Minimum distance between die edge and fan-out package edge is another fan-out only requirement for damage-free wafer dicing. A minimum die size adder of 0.6 mm (0.3 mm per side) is often needed along all the edges of silicon die. However, in practice, to take most of the advantages from fan-out, it is desirable to have minimum silicon size in a maximum allowable package body. So the minimum die edge to package edge distance requirement is mostly applicable in multi-die fan-out packages.

3.8 Future of Fan-Out WLCSP

Fan-out WLCSP remains an active development area for packaging engineering. Early work in fan-out WLCSP focused more on understanding the process, seeking of engineering solution, and improving manufacturing yield. Late focus shifted towards expanding the application front. One related development is the multilayer system in package (SIP) fan-out package with one or more dies or even passives embedded or surface mounted on the back of the package. The addition of more routing layers is necessary for complex routing that is typical for SiP packages and allows utilization of the whole package area for a fully populated bump array to achieve enhanced system level interconnections. This is quite an advantage compared to a single-layer fan-out, which typically features solder bumps out in the fan-out area or, in some cases, with some bumps in the center, under the die area (Fig. 3.15). Interestingly this later design is actually a combination of a traditional fan-in WLCSP and modern-day fan-out chip package technologies in one.

There will be economic factors that have to be considered for multilayer fan-out. To date, up to four-metal-layer fan-out has been demonstrated by leading bumping subcons. As the industry moves towards high level of system, subsystem integration, multilayer and multi-chip fan-out will find out more sweet niches where cost, performance, and package form factors all converge to the preference of fan-out

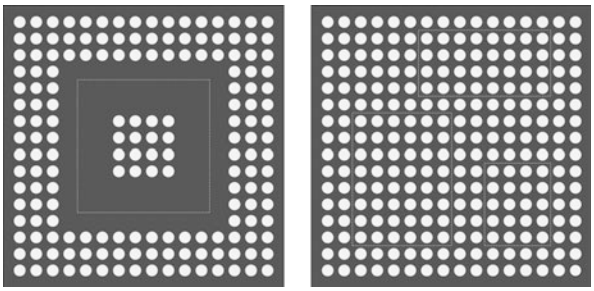


Fig. 3.15 *Bottom view* of a typical single-chip/single-layer fan-out WLCSP with bumps in the center and surrounding the die (fan-out area). However, it is not uncommon to see no solder bumps under the embedded die (*dashed square*) area, due to pitch and trace routing limitations. However, for multilayer/multi-chip fan-out, it is common to see fully populated bump arrays due to routing layers and via arrangement

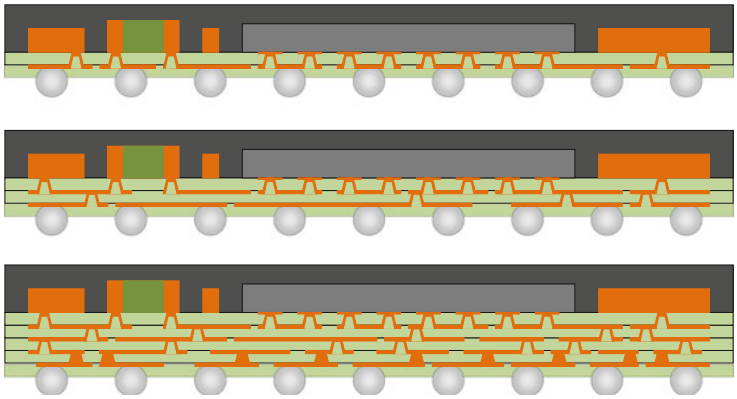


Fig. 3.16 Schematic cross sections of single-metal-layer, two-metal-layer, and four-metal-layer fan-out SiP packages. Fan-out with embedded group copper plane is used as examples. Window open in the copper plane allows passive integration in the package itself

packaging solutions. Figure 3.16 illustrated the one-metal-layer, two-metal-layer, and four-metal-layer fan-out packages. Silicon with embedded ground plane (RCP technology) and passives is shown to highlight the unique capability of multilayer fan-out for package level system integration.

With the same driving force for high-level integration, 3D stack up of fan-out WLCSP seems inevitable using through-mold vias (TMV) and package backside circuitry. It is a widely accepted TMV, which is typically laser drilled through the molding compound in the fan-out region, and can be made much more cost-effective than through-silicon via (TSV), which typically involves the Bosch process, meaning repeated etch/passivation cycles for controlled via sidewalls. Once the TMV via is metalized, one or more layer backside circuitry in fan-in fashion can be added using the similar front side RDL approach. This, together with

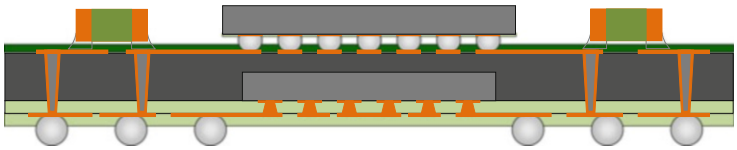


Fig. 3.17 A 3D package built on the base fan-out WLCSP. Through-mold vias (TMVs) make it possible to use the backside for additional circuitry for stacking other semiconductor devices and passives

the flexibility of fan-out packages incorporating one or more dies inside the package, gives 3D fan-out WLCSP unparalleled advantages over interposer-based 2.5D or 3D silicon stack with TSVs. In Fig. 3.17, a fan-out package with TMVs, backside circuitry, and surface-mounted second die and passives is used as an example of 3D fan-out packages. More innovative 3D packaging concepts exist and are constantly created building on the same base fan-out WLCSP. However, the approach is that it is the responsibility of package engineering to provide semiconductor technology with a well-balanced cost/performance solution.

Cost reduction is a constant challenge for electronic packaging engineers beside overall performance in electrical, thermal, and reliability. Fan-out WLCSP, since its infancy, has been always under scrutiny for mostly cost reasons. This continuous pressure on fan-out pushed the technology moving from 200 mm wafer fan-out to 300 mm wafer fan-out to achieve the needed economic scaling factor. However, moving beyond 300 mm requires different thinking.

With thinking of boosting productivity and reducing the cost at the same time, panel-based fan-out package sheds some light from different perspective. In the example shown in Fig. 3.18, a 200 mm and a 300 mm wafer can only produce 33 and 89 units of 25 × 25 mm fan-out packages, respectively, while a 450 mm square panel can produce 225 units of fan-out packages of the same size. Fan-out panel area utilization, even with 25.4 mm edge keep-out zone in the example, is a respectable 70.1 %, which is in the middle of 200 mm wafer's 66.3 % and 300 mm wafer's 79.5 %, assuming uniform 5 mm edge exclusion and 10 mm flat/notch height.

Like the variety forms of IC packaging solutions, panel-based fan-out also takes various approaches for the sake of manufacturability and cost management—some adopts organic substrate manufacturing technologies and low-cost structure; others take a hybrid approach of PCB lamination and carrier-based TFT-LCD panel pattern processing for better layer registration accuracy and fine line capability. The two carrier TFT-LCD panel fan-outs also allow thin package profile that has become so important for mobile application and 3D stacked SiP packages. Figure 3.19 shows the basic process flow of the hybrid TFT-LCD panel-based fan-out. While carrier 1 is also presented in wafer fan-out, carrier 2 helps large panel size warpage control and fine line RDL processing.

Panel fan-out adopting PCB fab technologies typically uses less aggressive line and space rules than the wafer-level fan-out, i.e., 20 μm/20 μm or more L/S for PCB

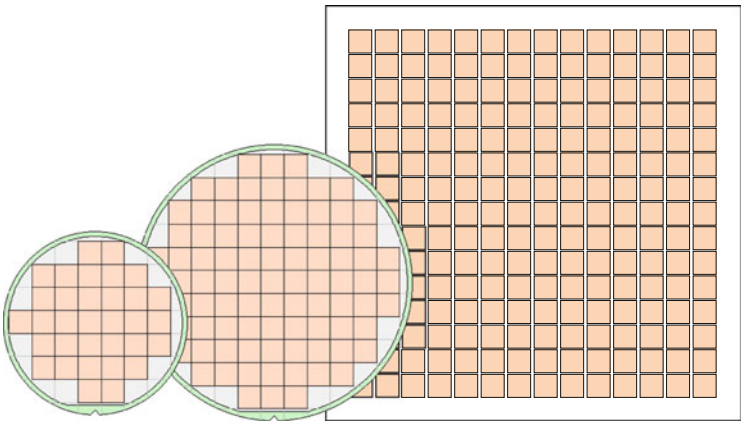


Fig. 3.18 Area utilization examples of 25 mm × 25 mm fan-out packages on a 200 mm wafer, a 300 mm wafer, and a 450 mm square panel. Wafer edge clearance is set at 5 mm and flat/notch height is set at 10 mm. Panel edge exclusion area is set at 25.4 mm (1 inch), which is typical for panel processing

panel fan-out and 15 μm/15 μm or less for wafer fan-out. Die movement during encapsulation, on the other hand, is less in panel fan-out than the wafer-level fan-out, mainly because of the use of sheet dielectrics vacuum lamination process, which requires very limited molten resin flow in lateral directions during encapsulation and provides shrinkage constraining resin cure environment. Panel dielectric encapsulation curing shrinkage still exists but is typically orders of magnitude less than that of fan-out WLCSP, which often averages more than 10 μm across a 200 mm fan-out wafer.

Unlike the exclusive use of spin-on low temperature cure aqueous polymers for wafer fan-out repassivation, panel fan-out has options of using a wide array of dielectric materials with controlled properties for better shrinkage and warpage management. For example, B-staged dielectrics (partially cross-linked prepregs or RCC) can be used for panel fan-out repassivation before copper RDL. It can be highly filled materials with CTE closely matched to the panel molding compound, and that helps to reduce large-size panel warpage, especially for the most often used single-sided fan-out package design.

With less die movement, it is possible for panel fan-out with wider line/space rules to actually achieve similar, if not better, routability as wafer-level fan-out with more aggressive line/spaces rules. Cost of panel fan-out is another attraction due to much improved productivity of large-size square/rectangular panel processing over the limitation of 200- or 300 mm circular wafer sizes. In a case study comparing fan-out chip package with a 6 × 6 mm chip encapsulated inside a 9 × 9 mm one layer RDL fan-out package with 356 I/Os, a total of 1,862 (49 × 38) gross packages can be built on a 2.5G TFT-LCD panel (370 × 470 mm)-based fan-out. Besides 1.5 × package count increase from wafer fan-out (~738 package for 300 mm wafer),

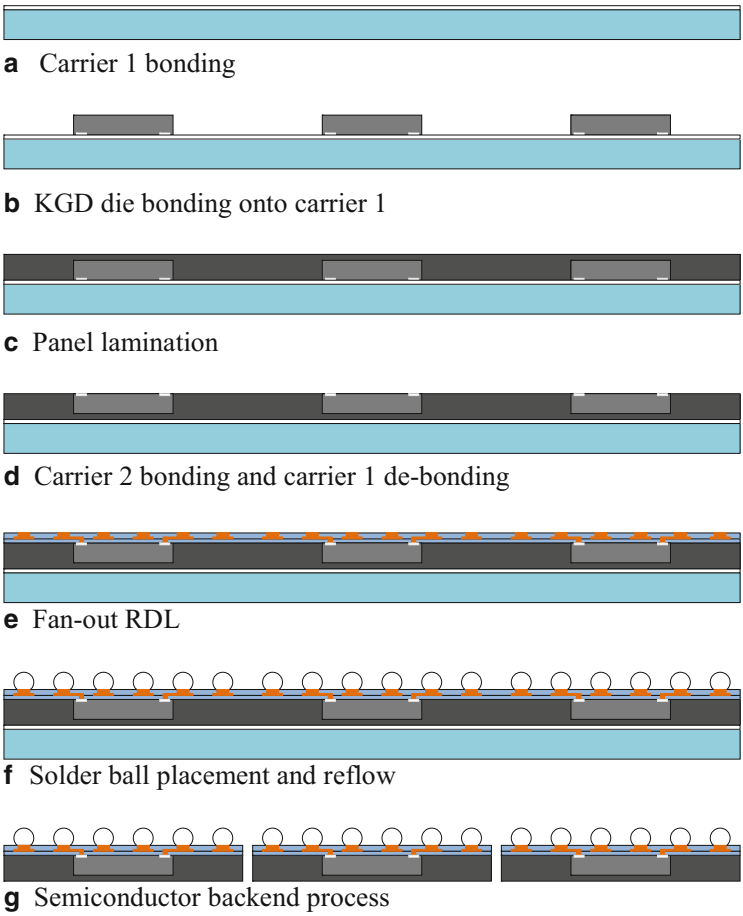


Fig. 3.19 Process flow for a typical fan-out WLCSP

there is also a huge improvement in area usage from less than 85 % for 300 mm wafer fan-out to 95 % of panel fan-out.

Rooted in the substrate technology, PCB panel fan-out also offers the flexibility of the front side fan-out with backside metal plane for improved the thermal and EMI shielding. It is also possible to interconnect the front and back metallization for 3D stacking (SiP or PoP), just like the wafer-level fan-out (Fig. 3.7). According to the early players in the panel fan-out, 2- or 3-layer panel fan-out can readily replace typical 6-layer substrate at significant overall cost reductions at the same time, with no need of solder bumps or wire bond for package chip interconnection, package substrate, and assembly process.

To summarize, fan-out chip package, either in wafer form or panel form, is a substrate-less embedded chip package that offers a low-cost, high-performance,

integrated alternative to wire bond BGA and flip chip BGA packaging. In the fan-out processing, semiconductor devices are encapsulated into wafer or panel form, while routing of signals, power, and ground is built directly on reconstituted wafers or panels. The fan-out chip package provides low-cost packaging solutions by consolidating separate wafer bumping, substrate build, flip chip or wire bond chip assembly, package over-molding, and BGA ball attach all into one highly efficient wafer or panel format processing. The package is inherently lead-free without chip/substrate solder joints. Low on-chip stress due to copper RDL also makes it friendly for semiconductors utilizing ultra-low-κ dielectrics.

Fan-out WLCSP technology is low profile with the elimination of bonding wires, substrate, and flip chip bumps. 2D multi-chip packaging is also straightforward with the flexibility of multilayer RDL. The low profile nature and multi-chip capability make fan-out an excellent platform of high-level 2D and 3D heterogeneous system integration. With solid package thermal, electrical, and reliability performance, fan-out is no doubt a versatile package technology that could be found across various package configurations, including single die, multi-die, and 2D and 3D systems in package (SIP) for various challenging applications, including high-frequency RF modules, high efficiency power management, low-power MCUs, as well as optical sensors/MEMS. As a matter of fact, fan-out interconnect along with the material compatibility and process capability has enabled novel SiP solutions not possible in more traditional packaging technologies or systems on chip (SOC). Yet, greater success of fan-out is expected following the development of lighter, smaller, and faster electronics that will change our way of living in the coming years.

References

1. Keser, B., Amrine, C., Duong, T., et al.: Advanced packaging: the redistributed chip package. IEEE Trans. Adv. Packaging **31**(1), 39–43 (2008)
2. Ramanathan, L.N., Leal, G.R., Mitchell, D.G., Yeung, B.H.: Method for controlling warpage in redistributed chip package, United States Patent US7950144B2, May (2011)
3. Sharma, G., Kumar, A., Rao, V.S., et al.: Solutions strategies for die shift problem in wafer level compression molding. IEEE Trans. Compon. Packaging Manuf. Technol. **1**(4), 502–509 (2011)
4. Hasegawa, T., Abe, H., Ikeuchi, T.: Wafer level compression molding compounds. 62nd Electronic Components and Technology Conference (ECTC), San Diego, CA (2012)
5. Itabashi, T., Dielectric materials evolve to meet the challenges of wafer-level packaging. Solid State Technol., November 01 (2010)
6. Iwashita, K., Hattori, T., Minegishi, T., Ando, S., Toyokawa, F., Ueda, M.: Study of polybenzoxazole precursors for low temperature curing. J. Photopolym. Sci. Technol. **19**(2), 281–282 (2006)
7. Hirano, T., Yamamoto, K., Imamura, K.: Application for WLP at positive working photosensitive polybenzoxazole. J. Photopolym. Sci. Technol. **19**(2), 281–282 (2006)
8. Fan, X.J., Varia, B., Han, Q.: Design and optimization of thermo-mechanical reliability in wafer level packaging. Microelectron. Reliab. **50**, 536–546 (2010)
9. Olson, T.L., Scanlan, C.M.: Adaptive patterning for panelized packaging, United States Patent Application Publication, No. US2013/0167102 A1

10. Oh, J.H., Lee, S.J., Kim, J.G.: Semiconductor device and method of forming FO-WLCSP having conductive layers and conductive vias separated by polymer layers, US Patent 8,343,810B2, Jan. 1 (2013)
11. Braun, T., Becker, K.-F., Voges, S., et al.: From wafer level to panel level mold embedding, 63rd Electronic Components and Technology Conference (ECTC), Las Vegas, NV (2013)
12. Liu, H.W., Liu, Y.W., Ji, J., et al.: Warpage characterization of panel fan-out (P-FO) package, 64th Electronic Components and Technology Conference (ECTC), Orlando, FL (2014)

Stackable Wafer-Level Analog Chip-Scale Package

4

4.1 Introduction

Stacking is a trend that has brought and still brings excitements to the semiconductor packaging societies. While the main driving force for the stacked package has been high level of integration and smaller package footprint, improvement of electrical performance due to the shortened path of signal transmission and power distribution is also frequently referred. Though it is relatively rare, improvement of overall thermal performance is occasionally mentioned as well. More frequently, heat dissipation is one of the mostly concerned areas for the stacked package. Cost of manufacturing the 3D structure seems to be the main hurdle for wider adoptions of the more aggressive 3D package concepts.

Moore's law has been known to drive the semiconductor chip scaling (and performance) since it was first described in Gordon E. Moore's 1965 paper. Stacking of semiconductors, either in die or package form, is an alternative approach to higher level integration, and sometimes it is known as system in package (SIP).

Unlike system on chip (SOC) integrating all functions into a single integrated circuit, the stacking approach takes singulated IC chips or packages and stacks them up in the vertical direction to realize integration in a small package footprint. Chips communication in the stacked package uses off-chip signaling, much as if they were mounted in separate packages on a normal circuit board. While die-level stacking brings the ultimate benefits of the smallest form factors and performance enhancements, package from stacking (package on package, or PoP) remains an attractive option for 3D semiconductor packaging. Regardless of the approaches, critical elements for 3D stacking are (1) aggressive thinning of silicon and/or low profile package and (2) formation of vertical interconnection, which includes basic choices of wire bonding, through-package (substrate) vias or through-mold vias (TMV), through-silicon vias (TSVs), and through-glass vias (TGVs).

Many options exist depending on the nature of the substrate vias have to be through, such as wet or dry chemical etch, photo-defined vias, and laser-drilled vias.

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_4

Mechanical drilling could also be used in the limited designs, i.e., course pitch, large vias size, etc. Conductive material via fill is required following via formation. CVD or PVD metal deposition, electroless, or electrolytically plated copper are the common choices for via metallization. For signal, power, or even thermal connection between the stacked layers, solder jointing is the dominant form, though anisotropic conductive adhesives can also find use in low power devices. Low temperature direct metal-to-metal joining is also being heavily researched.

Embedding WLCSP into package substrates offers yet another path for 3D packaging. Not only passives can readily stack up within or on the surface of the substrates where WLCSP is embedded in; multi-dies can also be stacked within the substrate based on the same foundation technology. Potential productivity boost associated with large-size PCB panel process is one attractive point for the embedded technology, aside from the competitive cost projections from the relatively low-cost PCB infrastructure.

Paths for high-level integration and approaches for 3D stacked packaging will be reviewed first with some details into the manufacturing process and pros and cons of each approach. Building blocks of 3D IC are to be summarized before more context is devoted to WLCSP-related 3D stacking, such as embedded WLCSP modules, stacking of fan-out WLCSP, and, ultimately, stacking of WLCSP. Handling of ultrathin substrate and micro via formation as well as metallic joining of two or more separate substrates will also be discussed, because those two are the foundations of various 3D WLCSPs.

4.2 Multi-chip Module Packages

The needs for high level of package integration were originally driven by the performance gain from the shortened electrical path connecting the neighboring parts—could be either semiconductors or passive components. Long before die or package stacking, practices of high-level integration took the form of multi-chip modules (MCM). MCM are specialized packages where multiple semiconductor dies and/or other discrete components are packaged onto a unifying substrate, facilitating their use as a single package. The MCM itself is often referred to as a "chip" in designs, illustrating its integrated nature.

Multi-chip packaging is an important facet of modern electronic miniaturization. Multi-chip modules come in a variety of forms, depending on the complexity and development philosophies of the package designers. The most significant form is often a fully custom chip package integrating multiple dies on a high-density interconnection (HDI) substrate. MCMs are frequently classified according to the technology used to create the HDI substrate, i.e., MCM-L, MCM-D, and MCM-C, for laminated, deposited, or ceramic substrate, respectively.

Intel® Pentium® Pro is a good example of early generation MCM, when wire bond and ceramic pin grid array (PGA) was still the dominating package technology. Pentium Pro (up to 512 k L2 cache) is packaged in a ceramic multi-chip module (MCM). The MCM contains two underside cavities in which the

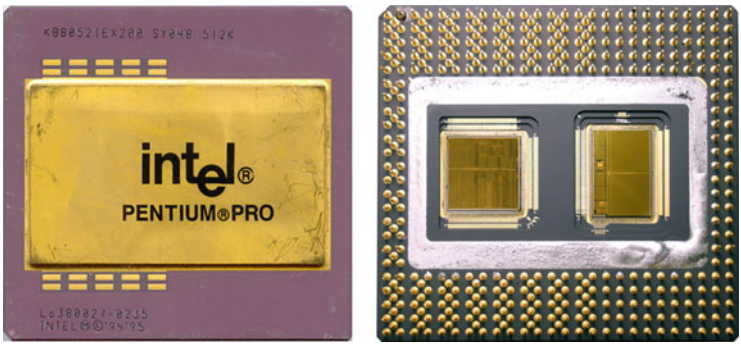


Fig. 4.1 Intel® Pentium® Pro top view of heat sink and bottom view PGA and solder ring (after removal of cap plate) and two dies with gold wire bond

microprocessor die and its companion cache die reside. The dies are bonded to a heat slug, whose exposed top helps the heat transfer from the dies directly to the heat sink that is attached on the package top. Cooling can be further enhanced with outside heat sink. The dies are connected to the package using multitier gold wire bond. The cavities are capped with a ceramic plate. The MCM has 387 pins, of which approximately half are arranged in a pin grid array (PGA) and half in an interstitial pin grid array (IPGA) (Fig. 4.1).

Advantages exist using the MCM approach versus early day's individually packaged semiconductors, including small footprint, improved electrical performance, decreased development time, reduced risk of design errors, and simplified bill-of-materials and product management, etc. The lower overall system cost came from the following factors and is typically enough to offset the higher unit cost of MCMs:

- Reduced production costs: MCM allows fewer components to mount and thus results in a smaller PCB with fewer layers. In many cases, it is possible to reduce two or more metal layer in the system PCB.
- Reduced BOM (bill-of-materials): all components included in the MCM are sourced by MCM supplier in high volumes and potentially better rates. It is also an easier logistic than buying all the components from individual suppliers separately
- Increased production yield: this is simply because of fewer components in assembly.

An example of using MCM to simplify system design is illustrated by Acme systems' FOX Board embedded Linux system that finds applications in Internet gateways, access control equipment, industry automation controllers, etc. With 100 % hardware and software compatibility, board with MCM package, ETRAX100X MCM 4 + 16, that includes the Axis Communications ETRAX 100LX CPU, 4 MB flash, 16 MB SDRAM, Ethernet transceiver, etc., is far less

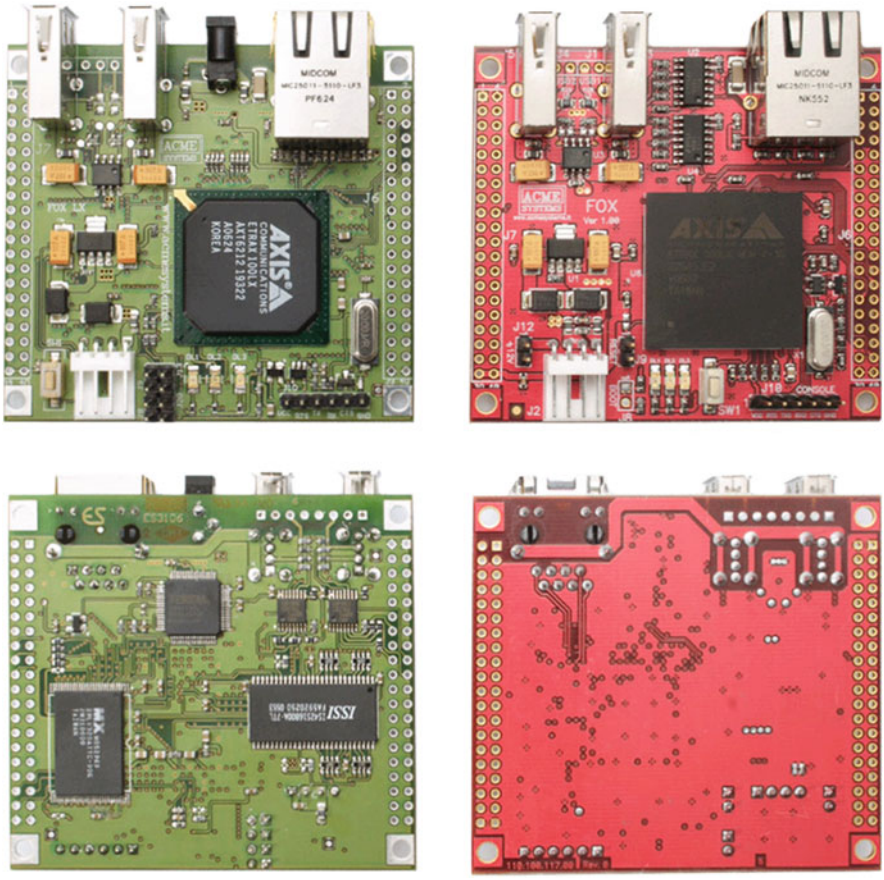


Fig. 4.2 Acme systems' the FOX Board embedded Linux system. The *left side system* is the design with the individually packaged chip set, while the *right side system* is built with the MCM module, which combines memory and Ethernet transceiver in one package

complex than board without MCM, but the discrete version of the same chip set is included in the MCM module. It is arguably, however, to say that it is a simple shift of design complexity and cost center from a system into a subsystem, with all the benefits listed above (Fig. 4.2).

Axis® ETRAX 100LX MCM is technically a fully functional Linux computer on a single chip that allows the build of small and cost-effective embedded devices. The MCM uses high-density packaging (HDP) technology that enables the integration of bare dies (i.e., unencapsulated chips such as ETRAX 100LX, SDRAM, flash) and other components (like resistors) to provide smaller, lighter, and yet cost-effective products. The MCM is built around the ETRAX 100LX system-on-chip processor, with all mandatory components for building a networked device, such as 4MB of flash memory, 16 MB of SDRAM memory, Ethernet transceiver, reset

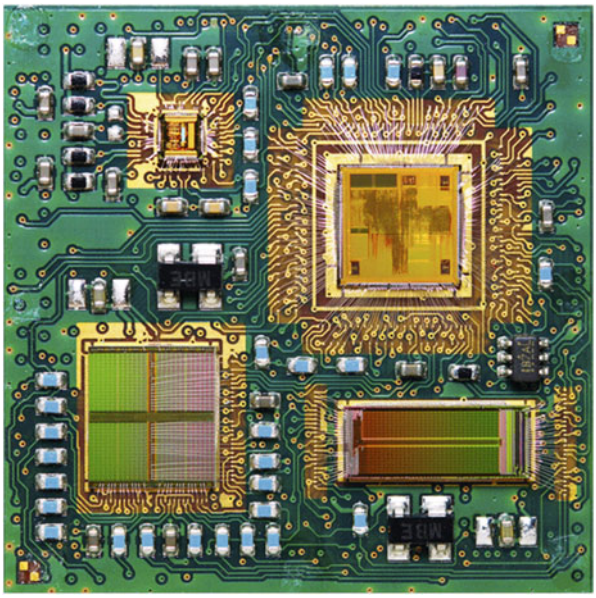


Fig. 4.3 An unmolded ETRAX 100LX MCM modules showing memory chips, Ethernet transceiver, reset circuitry, and passive components

circuitry, and about 55 passive components (resistors and capacitors). The MCM includes enough Flash and RAM for many designs. It is also possible to add more Flash memory and SDRAM outside of the multi-chip module. The only mandatory components outside the MCM are a 3.3 V power source and a 20 MHz crystal oscillator.

Package of the MCM module (ETRAX100X MCM 4 + 16) is 256-pin plastic ball grid array (PBGA) package at 27 × 27 × 2.76 mm, with power dissipation (outputs open) rated at 900 mW typical and 1,100 mW maximum. The MCM package shares the same footprint of the CPU only package (ETRAX 100LX, 27 × 27 × 2.15 mm), just slightly thicker and with slight more power dissipation (outputs open, 350 mW typical and 610 mW maximum) (Fig. 4.3).

4.3 Stacked Die Package and Stacked Package

Die stacking within a package is often referred as chip stack MCM, in which dies are not arranged side by side as in a traditional MCM, but stacked up one on top of the other. Giving the progressive nature of all technology advancement, it should not be a surprise to see die stacking first happened on DRAM chips, using gold wire bond technologies.

Stacking die is an economical method to increase the density of PC memory, especially when combined with long-standing gold wire bond technology. Advancement in wafer thinning and handling of ultrathin (1 mil or 25 μm thick) silicon chips, low profile (oftentimes reverse) wire bonding, and low wire sweeping

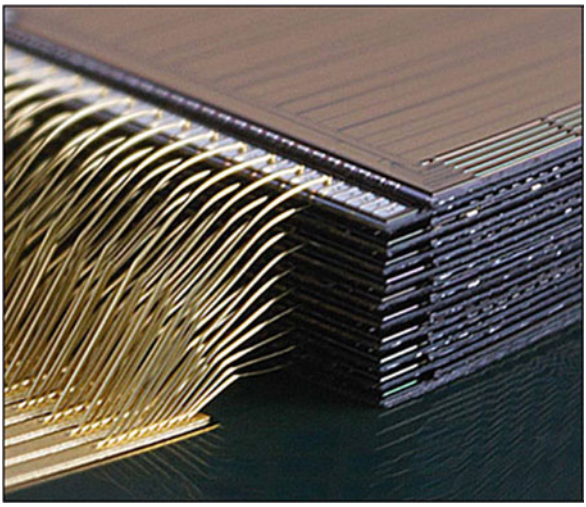


Fig. 4.4 1.4 mm MCP with 20 stacked dies (*source*: Elpida Memory)

resin injection into the narrow gaps between the stacked dies are all building blocks for the adoption of the technology into volume production of memory modules, where interconnection is relatively simple and not supersensitive to the parasitics of the long gold wires. High-yielding and cost-effective production of 1.4 mm overall package height MCP with an amazing 20 dies stacked inside was demonstrated in early 2007 (Fig. 4.4).

Limitations always exist for the wire bond stacked chip MCP technology—gaps between layers are always needed to leave room for wire loop. Horizontal spacing on the package substrate hundreds of microns wide is also required for the die-connecting wires. High level of precision is a must, because a single wire short in hundreds results in expensive failures of the whole module, and it is difficult to find and even more difficult to correct. Low sweeping molding compound(s) and process are just other basics that enable the wire bond stacked MCP.

Hybrid stacked die package with a combination of flip chip and wire bond technology was also feasible as disclosed in various literatures. Volume production, which is highly sensitive to cost/performance balance in consumer electronics, seems to be less clear, though. Drawings in Fig. 4.5 highlight the basic concept of wire bond MCP and flip chip/wire bond MCP.

Alternative to chip stack package, package on package (PoP) offers similar benefits of space savings by stacking individual package on top of another finished package and yet avoided concerns with the stacked chip package technologies, such as known good die (KGD), in-process damage, complexity of assembly and test (for logic and memory SiP), etc. Early PoP adoption is also DRAM modules, from the TSOP package stacking to the once ubiquitous μZ™ ball PoP developed by Tessera (Fig. 4.6).

Package on package (PoP) allows complex mixed-technology functions being produced with high yield, since the semiconductors are individually prepackaged

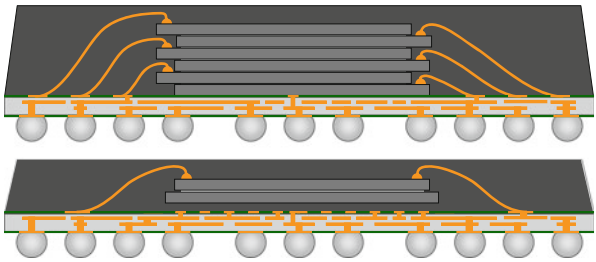


Fig. 4.5 Substrate-based wire bond MCP and flip chip/wire bond MCP

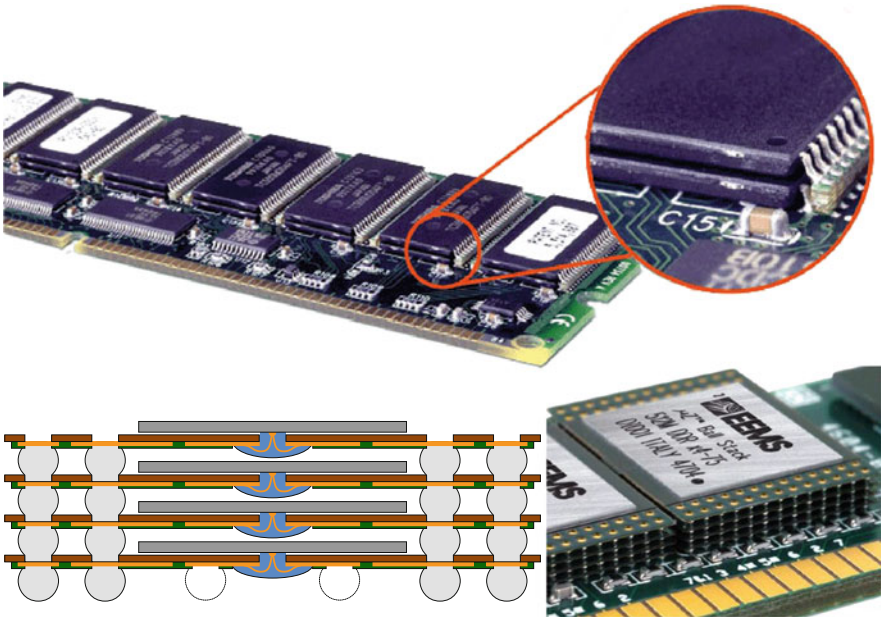


Fig. 4.6 *Top*: TSOP PoP on a memory module. *Bottom*: schematic cross section of Tessera μZ™ ball PoP and actual stack up of μZ™ ball PoPs on a single in-line memory module (SIMM)

and tested before joining and that is proven to provide predictable results for a broad range of SIP applications. Two widely used configurations are (1) pure memory stacking (two or more memory only packages are stacked on each other) (Fig. 4.6) and (2) mixed logic–memory stacking (logic (CPU) package on the bottom, memory package on top). The logic package is on the bottom because it needs many more BGA connections to the motherboard.

The most common PoP applications utilize package sections designed around the JEDEC standard array packaging format. Substrate BGA package seems to be a natural fit for PoP, because its basic function is to provide top (chip side) to bottom (BGA side) interconnection and it is relatively easy to utilize the top perimeter area, which is typically least used, for the additional memory/logic interconnections.

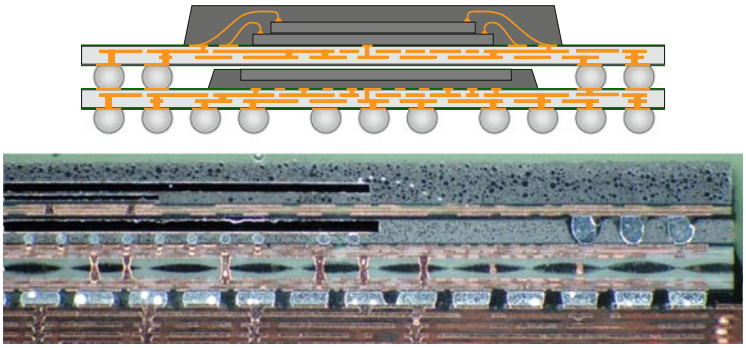


Fig. 4.7 *Top*: schematic cross section of PoP with memory chip stack MCM on top of flip chip BGA package. *Bottom*: cross section of Apple® A7 processor. Overall PoP is 14 × 15.5 × 1.0 mm with 1330 BGA bumps at 0.4 mm pitch. *Top package*: 1 GB LPDDR3 with 456 perimeter bumps at 0.35 mm pitch. *Bottom package*: dual core ARM CPU with integrated GPU and L1, L2, and L3 cache flip chip SoC. Flip chip bump pitch is 150/170 μm (Photo source: Prismark/Binghamton University)

While it is possible to have wire bond memory chip stack package on top of wire bond BGA package, the use of flip chip technology in the bottom package provides more desired flexibility for the PoP stack. For the bottom BGA package, replacing the wire bond die-level interconnect with flip chip increases the X/Y space available for either an increased number of top-to-bottom connections or for a larger processor chip, because the keep-out areas for wire bonds are not needed for flip chip. Flip chip bottom package with underfills might also forego the overmold without sacrificing reliability that could further reduce the gap height between the top and bottom packages and allow the use of smaller solder balls on a tighter interconnect pitch. Overall package height could also be better controlled, which is beneficial for the mobile electronics applications. Figure 4.7 illustrated the concept of the flip chip BGA bottom package PoP; also in the figure is a real-world example of the PoP package, which is an Apple® A7 microprocessor/memory PoP stack.

4.4 Three-Dimensional IC

3D IC is a concept of ultimate chip-level integration, in which two or more layers of active electronic components are integrated both vertically and horizontally into a single circuit. The name itself implies all wafer fab stacking up before packaging is involved, and the semiconductor industry has been pursuing this technology for a long time in many different forms. Yet it is a dream coming to reality. Meanwhile, 3D packaging, which has come a long way from package stack, wire bond die stack, to TSV die stack, seems to be closer than ever to the concept of true 3D IC. Today, TSV die stack chip with direct metal–metal bonding is approaching zero gaps between active layers of semiconductor chips. It might be arguable whether die stack with TSV is considered an integration of chips into a single circuit; however,

3D IC concept has been more than ever widely spread since TSV becoming real in the late 2000. Including TSV chip stacking, there are essentially two ways to build a 3D IC:

1. Monolithic 3D IC
   Monolithic 3D IC build electronic components and their connections (wiring) in layers on a single semiconductor wafer, which is then diced into 3D ICs. There is only one substrate and no through-silicon vias for interconnection between layers. A recent breakthrough overcame the process temperature limitation by partitioning the transistor fabrication to two phases: a high temperature (>800 °C) process phase (ion implant and activation) to be done before the layer transfer and a low temperature (<400 °C) process phase (etch and deposition) to be done after layer transfer. Layer transfer utilize ion-cut with hydrogen implant, also known as smart cut, to transfer a thin (<100 nm) single crystal layer on top of the bottom (base) wafer. Layer transfer has been the dominant method to produce SOI wafers for the past two decades. Multiple thin (10 ~ 100 nm) layers of virtually defect-free silicon can be created by utilizing low temperature (<400 °C) oxide bond and cleave techniques, and placed on top of active transistor circuitry. This is then followed by finalizing the transistors using low temperature etch and deposition processes.

   Monolithic 3D IC is a true 3D IC technology and has been researched at Stanford University under a DARPA-sponsored grant and is now promoted by Monolithic 3D Inc. Figure 4.8 highlights the top-level integration flow. Details of finishing up level circuits are omitted here.
2. TSV 3D IC
   Depending on when to bond chip layers, three process flows can be used for TSV-based 3D IC:

   Wafer on wafer (WoW): component layers to be stacked up are built on separate semiconductor wafers, which are then aligned, bonded, and diced into 3D ICs. Each wafer may be thinned before or after bonding. Vertical connections are either built into the wafers before bonding or else created in the stack after bonding. Through-silicon vias (TSVs) pass through silicon

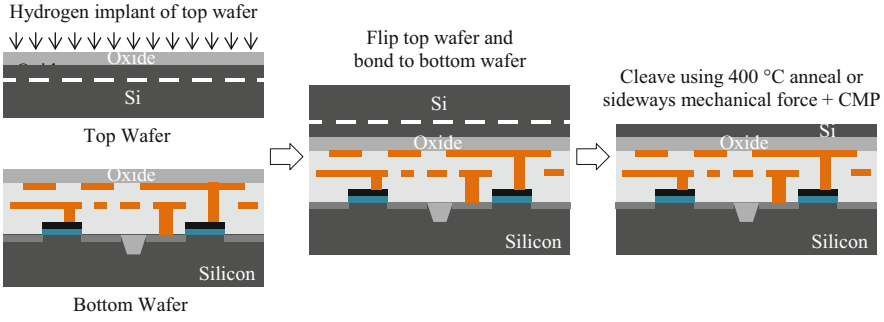


Fig. 4.8 Monolithic 3D IC ion-cut and layer transfer process flow

substrate(s) between active layers and/or between an active layer and external bond pads. Wafer-on-wafer bonding can reduce yields, since if any 1 of N chips in a 3D IC are defective, the entire 3D IC will be defective. Moreover, the wafers must be of the same size, but many exotic materials (e.g., III–Vs) are manufactured on much smaller wafers than CMOS logic or DRAM (typically 300 mm), complicating heterogeneous integration.

Die on wafer (CoW): component layers are built on separate semiconductor wafers. Upper layer wafer(s) is thinned and diced, while bottom wafer remains in the wafer form; the singulated dice are aligned and bonded onto die sites of the bottom wafer. As in the wafer-on-wafer method, thinning and TSV creation can be performed after bonding, though it is often prior to chip to wafer bonding. For CoW, only KGD die will be stacked, thus minimizing potential yield loss that is inevitable for WoW.

Chip on chip (CoC): component layers built on separate wafers are thinned and diced. They are then aligned and bonded sequentially on one package substrate. Thinning and TSV creation should be done before bonding. Similar to CoW, only KGDs are used for CoC build. Moreover, each die in the 3D IC can be binned beforehand, so that they can be mixed and matched to optimize power consumption and performance (e.g., matching multiple dice from the low power process corner for a mobile application).

4.4.1 Through-Silicon Via

From MCM to die stack and package stack packaging, the next move is through-silicon via (TSV) 3D chip packaging. TSV makes possible the shortest interconnection paths between chips and enables fastest transmission with minimum signal losses. In the case of data transmission between CPU and memories, or between flash and controllers, this high speed is of particular importance. In mobile electronics where miniaturization and functionality are always being sought after, TSV enables the smallest SiP packages and is driving the market acceptance of the technologies.

By the name, through-silicon via means electrical conductive path(s) from front (active) side, through the silicon to the back side. TSV enables vertical stack of semiconductor chips with minimum gaps (3D IC) that was only possible previously with use of substrate or wire bond. The short interconnect path TSV enabled is even more advantageous than the traditional system-on-chip (SOC) design, which has to be laid out in the 2D side-by-side fashion on the same semiconductor substrate. With the advancement of the TSV technology, which is also the foundation 2.5D integration, heterogeneous semiconductors can possibly be combined into a space-saving package to truly realize the system-in-package concept.

Via formation, via filling, wafer thinning, and TSV chip/wafer bonding are the basic building blocks of the TSV technology. Design automation, assembly, and tests are also areas where TSV is facing challenges. Cost and productivity, which are directly associated with the slow process of TSV formation and filling, seem to

be the main hurdles for the wide adoption of the TSV technology into mainstream production.

4.4.2 TSV Formation

TSV can be formed with laser drilling, Bosch deep reactive ion etching (DRIE), cryogenic DRIE, or various isotropic or anisotropic wet chemical etch. Typical TSV size range is 5 ~ 100 μm diameter and 10 ~ 100 μm in depth, depending on the design and applications. Size uniformity, throughput, and sensitivity to via cleanest all affect the selection of TSV formation process.

- Laser drill via formation
  Laser micro-hole drilling technology started in the mid-1980. It works through melting and vaporization (also referred to as "ablation") of the materials through absorption of energy from a focused laser beam. Melt expulsion, which arises as a result of the rapid buildup of gas pressure within a cavity created by evaporation, is the preferred process for material removal due to less energy required than vaporization. For melt expulsion to occur, a molten layer must form and the pressure gradients acting on the surface due to vaporization must be sufficiently large to overcome surface tension and to expel the molten material from the hole. On the negative side, melt expulsion creates debris on the via hole sidewalls and surrounds area that has to be cleaned in separate steps. Also when using laser drilling for TSV formation, keep-out zone from via needs to be maintained as well to ensure active devices are not affected. It is hard to make TSV less than 25 μm diameter, and that very much limits the use of laser drilling in only the least demanding TSV scenarios. Tapered sidewalls in the range of 1.3 ~ 1.6° are typical for laser-drilled TSV.
- Bosch DRIE
  Bosch process, also known as pulsed or time-multiplexed etching, is named after the German company Robert Bosch GmbH which patented the process. The process is known for its ability to produce high aspect ratio etch results on silicon substrate of any crystal orientation. Bosch process starts with isotropic plasma etch of silicon through the masking layer opens. It is then switched to high-density plasma to deposit a layer of Teflon-like materials on all the exposed surfaces. Post passivation, anisotropic plasma is used to remove the passivation deposit on the etch surface, and then isotropic plasma etch is followed to start a new etch–passivation cycle. A complete etch process repeats these etching and deposition steps several times to attain deep, vertical etch profiles.

  Inductively coupled plasma (ICP) is the most commonly used form of high-density plasma (HDP) in Bosch process configurations to provide the delicately balanced ions and free radicals. Sulfur hexafluoride (SF_6), which molecules easily break up in high-density plasma, is the source gas to provide free radical fluorine for silicon etching. The sidewall passivation and mask protection is provided by octofluorocyclobutane ($c\text{-}C_4F_8$), a cyclic fluorocarbon that

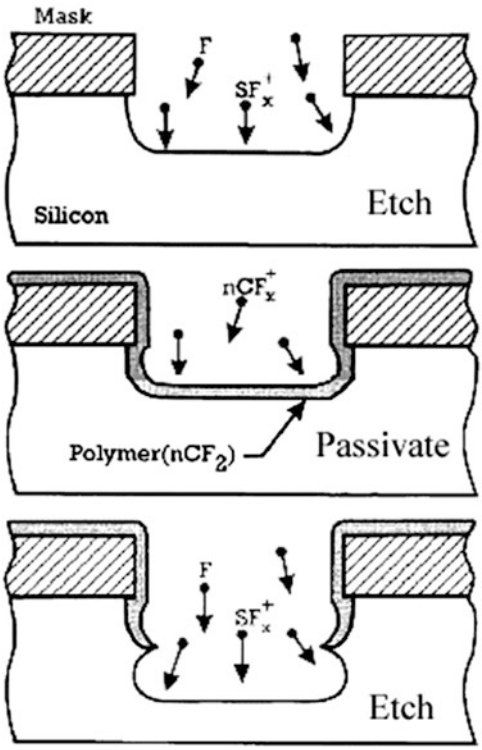


Fig. 4.9 Bosch DRIE process

disintegrates in the high-density plasma to produce CF_2 and longer chain radicals that forms Teflon-like polymer passivation layer that protects the entire substrate from further chemical attack and prevents further etching. During the next etching phase, the directional ions that bombard the substrate attack the passivation layer at the bottom of the trench (but not along the sides). They collide with it and sputter it off, exposing the substrate to the chemical etchant.

Each etch/deposition phase lasts for several seconds. These etch/deposit steps are repeated many times over resulting in a large number of very small isotropic etch steps taking place only at the bottom of the etched pits. To etch through a 0.5 mm silicon wafer, for example, 100–1,000 etch/deposit steps are needed. The two-phase process causes the sidewalls to undulate with an amplitude of about 100–500 nm (Figs. 4.9 and 4.10). The cycle time can be adjusted: short cycles yield smoother walls, and long cycles yield a higher etch rate.

- Cryogenic DRIE
  In cryogenic DRIE, the wafer is chilled to −110 °C (163 K). SF_6 is still used to provide fluorine radicals for silicon etching. Ions bombard upward-facing, unmasked silicon surfaces and etch them away in the form of volatile SiF_4. Instead of using a fluorocarbon polymer, sidewall protection in cryogenic DRIE

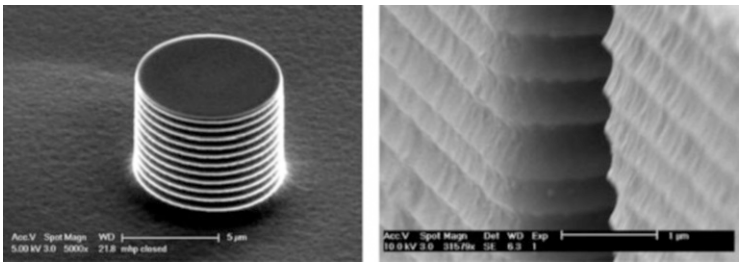


Fig. 4.10 Scallop sidewall of resulted from Bosch process

relies on forming a blocking layer of oxide/fluoride (SiO_xF_y) on the sidewalls (around 10–20 nm thick); this along with cryogenic temperatures slows down the chemical reaction that produces isotropic etching by the fluorine radicals and thus produces highly vertical sidewalls.

Main highlight of cryogenic DRIE is the smooth sidewalls that cannot be achieved by the Bosch process. The smooth sidewall makes it very applicable for special applications, such as molds and optical devices. Unlike the one option of straight sidewalls by the Bosch process, cryogenic etch also allows some fine-tuning of sidewall profile, which provides yet another nice option for certain applications (Fig. 4.11). Primary issue with cryogenic DRIE is that the standard masks on substrate crack under extreme cold plus etch by-products have a tendency of depositing on the nearest cold surface, i.e., the substrate or electrode. Controlling the low temperature throughout the etch process to ensure anisotropic etch is just another thing that has to be considered.

4.4.3 Via First, Via Last, and Via Middle

Before TSV and 3D IC, semiconductor manufacturing flows through two distinctively sectioned areas to completion: front-end wafer processing and backend chip packaging. WLCSP made it possible for backend to process in wafer format, bearing similarities to the front-end semiconductor wafer processing. TSV and 3D IC, however, blurred the lines even further between the traditional front end and backend.

From the function perspective, TSV displaces some interconnect function traditionally realized in IC packaging, either in MCM format or chip/packaging stacking, via wire bond, or flip chip/substrate interconnection. There is no doubt that backend packaging can perform some, if not all, TSV processing functions; major TSV processing will, however, happen as one of the semiconductor front-end wafer processing steps.

Depending on the TSV technology adoption and sequence in manufacturing flow, TSV can be classified in three major categories: via-first, via-middle, and via-last TSV. Via first means TSV is formed before the start of CMOS processing.

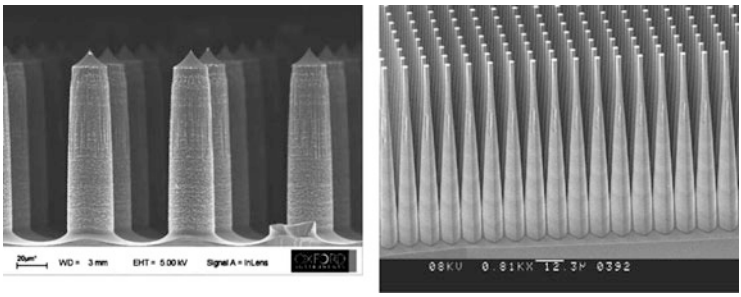


Fig. 4.11 Micro needles created by Bosch process + wet chemical etch (*left*) and baseball bat silicon posts array etched by cryogenic DRIE (*right*)

To survive the subsequent high temperature CMOS processes, polysilicon is the choice of conductive materials for via fill. Via middle means TSV is formed after the CMOS processes, but before interconnection layers. Without the need to survive the high temperature CMOS process, it is possible to use copper for via filling to take advantage of its electrical performance. Tungsten (W) and molybdenum (Mo) via fills are also choices when CTE of copper and voiding in copper via plating in high aspect ratio vias are of concerns. Via last means TSV is formed after the completion of semiconductor wafer processes. Copper is typically the via filling material due to typical large size of via-last TSV. Figure 4.12 illustrated differences in via-first, via-middle, and via-last process flow.

Via first and via middle all happen in the wafer fab, though the function of TSV is interconnection that is traditionally accomplished in backend packaging process. DRIE is the choice for via formation. Size range is typically under 20 μm in diameter, with minimum in the range of 2 ~ 5 μm. Development is underway to submicron size vias. However, it is still far larger than typical line/space seeing on CMOS chip interconnection layers. Typical via depth is 15 ~ 25 μm, depending on the targeted 3D IC needs, for via-first TSV.

Via last means TSV is formed after the completion of semiconductor wafer processes. DRIE is certainly an option for via-last TSV; laser-drilled via seems to be an attractive alternative here due to cost and fast drilling speed. However, size range of laser vias is from 15 ~ 50 μm in diameter. When coupled with position accuracy and nature of sequential processing, laser drilling is better suited for low pin count TSV needs, such as sensors or flash memory. Depending on the wafer back grinding thickness, laser-drilled vias could be as deep as 200 μm with aspect ratio up to 10:1. For via filling, copper plating is the dominant form.

TSV remains an active area for research and development with heavy investment by semiconductor companies, foundries, packaging subcons, and universities. It is the cornerstone of the 3D IC dream that started from the early days of the semiconductor innovations.

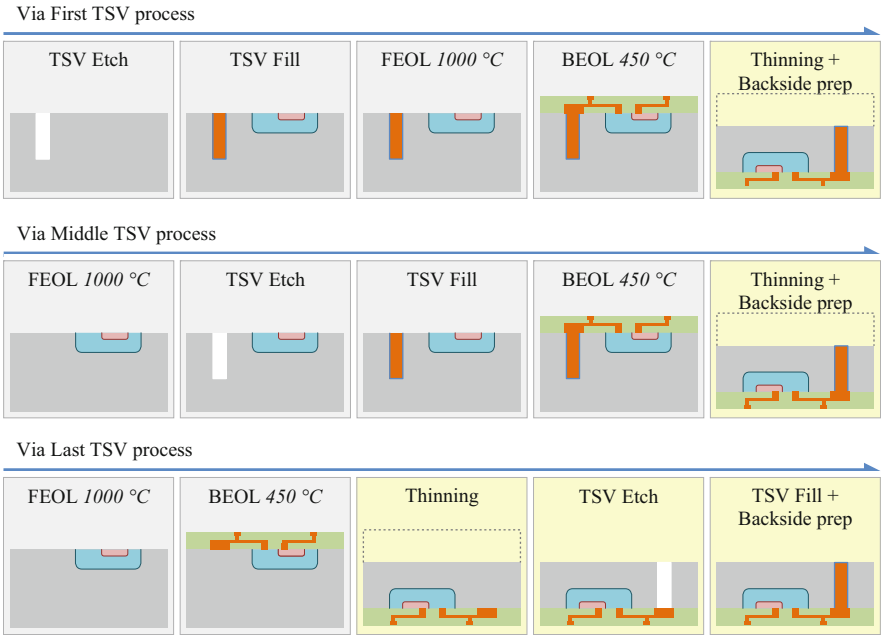


Fig. 4.12 Via-first, via-middle, and via-last TSV process flow. *Gray cells* represent wafer fab processes and *yellow cells* represent process in backend packaging

4.4.4 TSV Fill

Once TSV is formed, insulation layer is deposited on the sidewalls to separate the conductive vias and the silicon substrate. Thermal CVD oxide, silane, and TEOS (tetraethoxysilane)-based PE-CVD oxide or nitride deposited by LP-CVD are the common choices.

After insulation, conductive materials, such as polysilicon, copper or tungsten can be used for via fill. Electroplated copper is the preferred via fill material due to superior electrical conductivity. Bottom-up plating is desired and works nicely in vias of reasonable aspect ratio and diameter. On top of void-free via fill, mismatching CTE of copper (16 ~ 17 ppm/°C) and silicon (2 ~ 3 ppm/°C) is another consideration selecting via fill materials. When via is deep, thermal mechanical stress, which is a function of CTE difference and linear dimensions, could be sufficient to induce cracks in the inner layer dielectrics.

To reduce this stress, low CTE metals, such as tungsten (W, 4.5 ppm/°C) or molybdenum (Mo, 4.8 ppm/°C) can be used for TSV fill. PVD is the mature option for tungsten or molybdenum deposition. Low speed and void-free via fill are the main challenges for using PVD. Laser-assisted CVD, which in theory could deposit tungsten or molybdenum from gas-phase precursors at high speed, shows some promises for large-size-TSV fill. However, the sequential process nature, which very much likes laser via drilling, limits its application to low I/O count TSV needs.

For large diameter TSV, polymeric insulation materials 2 ~ 5 μm thick are also experimented for TSV insulation. The soft polymer insulation layer not only lowers the thermal mechanical stress on the structures surrounding TSVs but also reduces the capacitance due to the thicker insulation layer with lower than oxide dielectric constant. Process compatibility, such as process temperature and duration, has to be ensured for polymer TSV insulation.

4.4.5 3D IC Bonding

After TSV formation and backside via exposure/preparation, layers of semiconductor devices needed to be stacked up and bonded to form the designed 3D structure. Stack up could be done either in wafer form or in die form and so to be referred as (1) wafer-to-wafer bonding, (2) die-to-wafer bonding, and (3) die-to-die bonding. Bonding options are oxide fusion bonding, metal–metal bonding, and polymer adhesive bonding. For metal–metal bonding, it can be either metal fusion bonding or metal eutectic bonding, such as Cu–Sn eutectic bonding.

Oxide bonding

Oxide bonding is a direct wafer bonding method that has been demonstrated on SOI wafers. In the process, LP-CVD oxide is first deposited on the bonding surfaces and polished to atomic level smooth (Ra < 0.4 nm). Wafers are cleaned in H_2O_2 and DI wafer and then spin dried in N_2 environment. Rich Si–OH (silanol) groups will form on the bonding surface after these steps. When two wafers are aligned and bonding surfaces brought together, hydrogen bond will form due to atomic contact of two ultrasmooth surfaces. Vacuum annealing is then followed to drive out H_2O molecule from the condensation reaction of Si–OH and HO–Si from two bonding surfaces and forms covalent Si–O–Si bond to join the two surfaces.

Cu–Sn eutectic bonding

Cu–Sn eutectic bonding can be made at relatively low temperature (150 ~ 280 °C) using low melting tin, either though diffusion or solder fusion, to realize 3D integration with copper TSV. Two bonding systems are widely used: Cu/Sn–Cu bonding and Cu/Sn–Sn/Cu bonding. Completion of the eutectic bonding is marked by the formation of Cu_3Sn alloy, which is a thermodynamically stable binary metal system with melting point at 676 °C. To control the Sn consumption before bonding, thin buffer layer of Au or Ni can be inserted between Cu and Sn. Cu–Sn eutectic bonding has been demonstrated on 50 μm pitch copper TSV.

Direct Cu–Cu bonding

Low temperature copper diffusion bonding is very attractive to 3D IC because it offers the best possible electrical performance and thermal conductivities of all interconnection options. In this process, bonding copper surface is planarized (Ra < 2 nm) and cleaned and/or passivated so as to be free of oxidation. Then the two surfaces are brought in contact with pressure. Low temperature (<350 °C)

annealing is typically applied to promote copper interdiffusion, grain growth, and recrystallization to complete the bonding process. Room temperature Cu bonding with proper surface activation is also reported, though cross bonding surface interdiffusion is limited. Strength of Cu–Cu diffusion bond is found to be influenced by copper oxidation, compress force, annealing ramp temperature, and time. Pitch as fine as 10 μm has been demonstrated. Besides the fine pitch and electrical and thermal performance, strong electromigration resistance is also expected since it is essentially pure copper interconnect with no alloys involved.

Polymer adhesive bonding

Compared to all the bonding methods discussed above, the beauty of polymer adhesive bonding is that it has much relaxed requirements on bonding surface flatness and cleanness. Process of adhesive bonding is also straightforward: first, spin on adhesion dissolved in solvents, preheat to remove solvent and/or partially cure the adhesive, and then align and press two bonding surfaces together under vacuum and cure at elevated temperature to full cross-linking the polymers. There are thermoplastics and thermoset polymer adhesives available. Dry etchable and photosensitive materials are also available for special bonding needs.

Among available polymer adhesives, such as BCB (benzocyclobutene), parylene, and polyimide, BCB is found to provide robust bonding results on various surfaces. The materials also possess excellent chemical resistance and bond strength.

Polymer adhesive bonding is a low temperature bonding method, and it is compatible with typical backend IC processes and packaging/assembly processes. It can be used for 3D integration of heterogeneous semiconductors and providing mismatching stress buffer at the same time. However, adhesive reflow during final curing stage can be a challenge for precision alignment. Coating uniformity and proper partial curing of adhesive are all important steps to ensure wafer stay aligned during final curing.

4.4.6 Integration of TSV 3D IC

Traditional 2D scaling of semiconductor chips has been the driving force improving signal propagation speed. However, scaling from manufacturing and chip-design technologies has become more difficult, in part because of power-density constraints and in part because interconnects do not become faster while transistors do. 3D integrated circuits are set to address the scaling challenge by stacking 2D dies and connecting them in the third dimension. This promises to speed up communication between layered chips, compared to planar layout.

In 2004, Intel presented a 3D version of the Pentium 4 CPU. The chip was manufactured with two dies using face-to-face stacking, which allowed a dense via structure. Backside TSVs are used for I/O and power supply. For the 3D floor plan, designers manually arranged functional blocks in each die, aiming for power

reduction and performance improvement. Splitting large and high-power blocks and carefully rearranging them allowed limiting thermal hotspots. The 3D design provides 15 % performance improvement (due to eliminated pipeline stages) and 15 % power saving (due to eliminated repeaters and reduced wiring) compared to the 2D Pentium 4.

The Teraflops Research Chip introduced in 2007 by Intel is an experimental 80-core design with stacked memory. Due to the high demand for memory bandwidth, a traditional I/O approach would consume 10–25 W. To improve upon that, Intel designers implemented a TSV-based memory bus. Each core is connected to one memory tile in the SRAM die with a link that provides 12 GB/s bandwidth, resulting in a total bandwidth of 1 TB/s while consuming only 2.2 W.

These results of the above experimentations are indeed impressive, and they demonstrate dramatically that 3D integration is more than just stacking memory on top of a processor in order to boost performance. There are clear motivations to start designing processors in the third dimension. But before such chips can reach the mass market, obstacles to manufacturing them must be overcome, and the processor design tool chain has to be overhauled to take into account a whole new family of possibilities and constraints.

3D ICs change the semiconductor in many ways with promises; this includes:

- Footprint: With more functionality fits into a stacked package, it essentially extends Moore's law that defines XY scaling and enables new generations of tiny but powerful devices.
- Cost: Partitioning a large chip into multiple smaller dies can improve the wafer yield and thus reduce the basic fabrication cost. With only known good die for 3D IC, high stacking yield will improve the overall yield of the finished IC.
- Heterogeneous integration: Circuit layers can be built with different processes, or even on different types of wafers. This means that components can be optimized to a much greater degree than if they were built together on a single wafer. Moreover, components with incompatible manufacturing could be combined in a single 3D IC.
- Shorter interconnect: The average wire length reduction is in the range of 10 ~ 15 %, based on the common figures reported by researchers. However, reduction is more on the longer interconnect, which benefits most from 3D stacking. Thus, improvement in circuit delay could be of a greater amount. On the negative side, 3D wires have higher capacitance than conventional in-die wires, so some tradeoffs exist in overall circuit delay reduction.
- Power: Keeping a signal on-chip can reduce its power consumption by 10 ~ 100 times. Shorter wires also reduce power consumption by producing less parasitic capacitance. Reducing the power budget leads to less heat generation, extended battery life, and lower cost of operation.
- Design: The vertical dimension adds a higher order of connectivity and offers new design possibilities.
- Bandwidth: 3D integration allows large numbers of vertical vias between the layers. This allows construction of wide bandwidth buses between functional

blocks in different layers. A typical example would be a processor + memory 3D stack, with the cache memory stacked on top of the processor. This arrangement allows a bus much wider than the typical 128 or 256 bits between the cache and processor.

3D IC technology also carries challenges, including:

- TSV-introduced overhead: TSVs are large compared to gates and impact floor plans. At the 45 nm technology node, the area footprint of a 10 μm × 10 μm TSV is comparable to that of about 50 gates. Furthermore, manufacturability demands landing pads and keep-out zones which further increase TSV area footprint. Depending on the technology choices, TSVs block some subset of layout resources. They occupy either the device layer that results in placement obstacles or, in the worst case, both device and metal layers, resulting in placement and routing obstacles. While the usage of TSVs is generally expected to reduce wire length, this depends on the number of TSVs and their characteristics, as well as design block partitions.
- Testing: To achieve high overall yield and reduce costs, separate testing of independent dies is essential. However, tight integration between adjacent active layers in 3D ICs entails a significant amount of interconnect between different sections of the same circuit module that were partitioned to different dies. Aside from the massive overhead introduced by required TSVs, sections of such a module, e.g., a multiplier, cannot be independently tested by conventional techniques. This particularly applies to timing critical paths laid out in 3D.
- Yield: Extra manufacturing step adds risks for defects and yield losses. In order for 3D ICs to be commercially viable, defects has to under a manageable level.
- Heat: Heat buildup and dissipation within the stack IC must be dealt with innovative solutions. This is a single most critical issue for the stacked IC. Specific thermal hotspots must be carefully managed.
- Design complexity: Taking full advantage of 3D integration requires sophisticated design techniques and new CAD and simulation tools.
- Lack of standards: There are few standards covering TSV-based 3D IC design, manufacturing, and packaging, although the issue is being addressed. In addition, there are many integration options being explored such as via last, via first, via middle; interposers or direct bonding; etc.
- Heterogeneous integration supply chain: Same as for all multi-chip packaging, the delay of one part delays the delivery of the whole product and so delays the revenue for each of the 3D IC part suppliers.
- Lack of clearly defined ownership: It is unclear who owns the 3D IC integration and packaging/assembly and what are the roles of foundry, assembly houses, and the product OEMs.

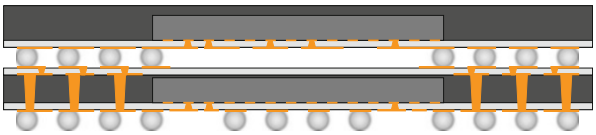


Fig. 4.13 A conceptual PoP based on fan-out WLCSP. Through-mold via (TMV) is drilled through bottom fan-out package to make the front (active side) to back vertical interconnections. Backside rerouting is needed for connecting the top chip. In case of full area array, fan-in routing could be made on the back side of bottom package

4.5 Wafer-Level 3D Integration

Due to unique wafer bumping/packaging process, not all 3D package and 3D IC options are applicable to WLCSP. For example, wire bond chip stack only makes sense when copackaged side by side with WLCSP chips on a module substrate. PoP could be a value-added option for fan-out WLCSP with through-mold via (TMV); it is hardly an option for fan-in WLCSP. Also, TSV would be a hard-to-accept option for general-purpose fan-in WLCSP because of the extra space required and market sensitivity to the cost of WLCSP components.

Being the only bare die IC package, 3D on WLCSP might very well take unique directions that are worth some in-depth discussions. Before that, it is not a bad idea to be refreshed with these unique features of WLCSP: (1) WLCSP is an all wafer process-based package technology; (2) it applies to low IO count (<400 bump) ICs often less than 10 × 10 mm in size, with more typical size less than 6 × 6 mm; and (3) WLCSP often adopts fine pitch (equal to or less than 0.5 mm) solder bump array that is only limited by the available PCB technology. Due to the small size and all silicon nature, 3D approach for fan-in WLCSP is mainly seen in the special applications, such as CMOS image sensors or MEMS package. For fan-out WLCSP, which greatly expand the package size limitation (due to CTE mismatch thermal mechanical stress) to the BGA substrate package size range, 3D concepts that are early developed on BGA substrate packages can be referenced/transferred to the fan-out package, as illustrated in Fig. 4.13, a PoP based on fan-out with TMVs.

4.5.1 3D MEMS and Sensor WLCSP

Long before TSV, 3D on WLCSP was pioneered by Shellcase, Ltd., an Israeli company that engaged in developing, manufacturing, and marketing advanced packaging technologies for microelectronic integrated circuits. In December 2005, Tessera completed its purchase of certain intellectual property and related assets of Shellcase®. This acquisition enabled the company to enter into the consumer optical market.

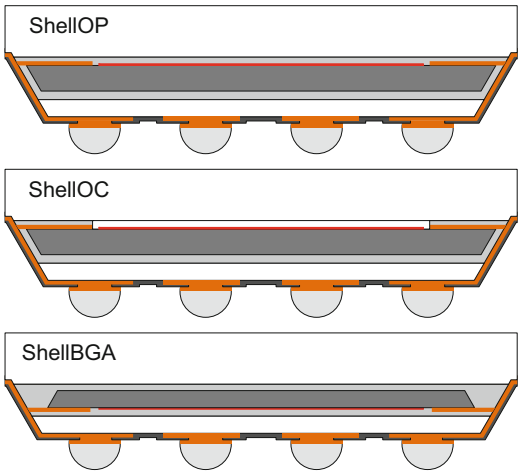


Fig. 4.14 Three basic forms of Shellcase optical package

The key for 3D is the vertical or front (active side) to back interconnections. Shellcase® did it by rerouting on the sloped chip/package sidewalls following exposure of the front side on chip edge contacts. Fan-in redistribution then allows area array to be formed on the back side of the package, leaving front image sensor side protected by a glass shield and free of package level interconnection interferences. There are three basic forms of Shellcase packages in the finished of CMOS image sensors—ShellOP, ShellOC, and ShellBGA (Fig. 4.14)—with each design serving the unique needs for the semiconductor image sensor. ShellOP optical package is the basic form of the family that provides front to back edge routing and full protection of the sensitive active side sensing area. ShellOC added an optical cavity that could boost the light reception with removal of the adhesive between the glass and sensor area. ShellBGA is designed for backside illuminated (BSI) CMOS sensor, in which the light passes through the aggressively thinned silicon substrate to hit on photocells. This design avoids the scattering by the on-chip metal layers and therefore increases the amount of light captured and improves sensor's low-light performance.

ShellOP package process flow is illustrated in Fig. 4.15. ShellOC just requires a special lamination process to create the optical cavity. ShellBGA, on the other hand, takes a quite different sequence to accomplish the packaging. However, the basic concept is not much changed from ShellOP.

Shellcase® process is an all wafer-level packaging process. The process starts with a silicon IC wafer with bonding pad extensions into scribe lanes being adhesively bonded to a glass wafer. For image sensor, optical adhesive is applied in this step. The glass substrate serves subsequently as a mechanical carrier, allowing silicon wafer to be thinned down to 50–100 μm and trench forming beneath the pad extensions. Then a second glass substrate is adhesively bonded, resulting in silicon islands fully encapsulated by the adhesive. At this stage, a compliant polymer layer beneath the future solder bumps is formed, enhancing the package mechanical reliability. A V-shaped dicing blade is subsequently used to

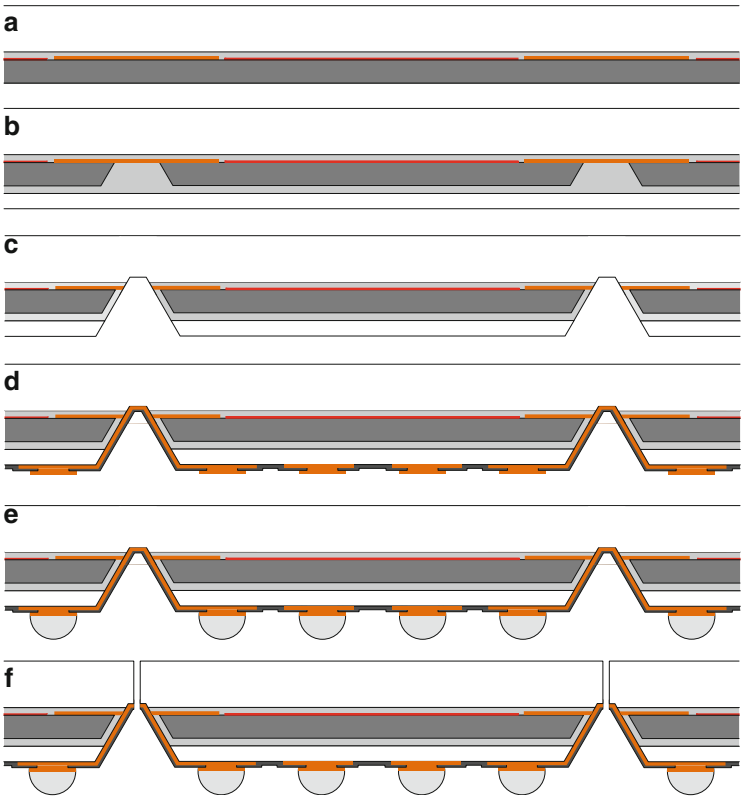


Fig. 4.15 Schematic process flow of ShellOP: (**a**) bonding silicon to glass carrier and backgrind; (**b**) forming V trench and attaching bottom glass; (**c**) V cut through pad metal; (**d**) backside metallization and redistribution + UBM; (**e**) print solder or ball drop and reflow; (**f**) singulation

perform notching within the scribe lane regions. The exposed pad extensions at each die periphery are then redistributed to the area array of solder balls on the bottom glass substrate. This is done by sputtering and patterning of an Al layer, followed by solder mask deposition and solder bump forming using solder paste deposition or attachment of preformed solder spheres. The process is completed by singulating finished wafer into individual dies using dicing saw.

Many improvements have been made to the basic Shellcase® package concept since the IP acquisition by Tessera, a well-known semiconductor packaging IP powerhouse, with main focus on reducing the overall package size and height, as well as more efficient manufacturing technologies. Cost-effective Shellcase® RT low profile package benefits directly from the elimination of the bottom glass and packaging on 300 mm (12″) wafers. Shellcase® MVP adopts TSV for vertical front to back interconnections. Comparing the traditional Shellcase edge connections, TSV package design has fewer restrictions on wafer diameter, bond pad size, pitch, or location, making it directly compatible with the most existing CMOS imagers. The dicing lanes can be as narrow as allowed by the silicon design rules, which help

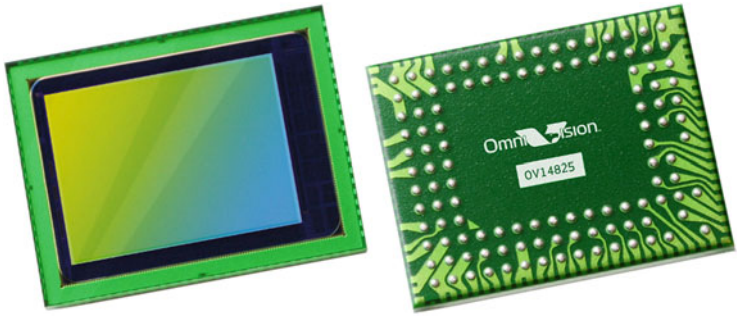


Fig. 4.16 OmniVision® OV14825 image sensor in Shellcase package (*source*: OmniVision® Technology)

to maximize the number of die per wafer and decrease unit cost. The packaged image sensor thickness is approximately 500 μm, suiting it for mobile electronic products that demand extreme thinness.

The OmniVision® OV14825, Fig. 4.16 introduced back in 2010, is a real-world example of image sensor applying the Shellcase® 3D packaging technology. The OV14825 has an active array of 4,416 × 3,312 backside illumination pixels operating at 15 fps in full resolution while delivering full 1080p HD video at 60 fps, using a binning feature to achieve higher sensitivity. In full HD video mode, the sensor also provides additional pixels used for electronic image stabilization (EIS). The sensor is packaged with backside of silicon facing up in a 116-pin chip-scale package (CSP). The unique edge connection of Shellcase package is evident.

Besides image sensor, MEMS is another area where 3D wafer packaging has found unique values. When traditional semiconductor manufacturing might feel challenged, MEMS manufacturing relies on DRIE to form complex 3D mechanical structures for various sensing needs. So 3D comes natural for MEMS, though it still takes effort from MEMS and ASIC designer/packaging to make it happen in the cost-effective fashion.

3D MEMS WLP does not have to be wafer stacking of MEMS and other types of semiconductors using only TSV as interconnects, so long it is wafer format stacking of either chip on wafer or wafer on wafer and processing of interconnection and/or test before singulation into individual packages.

Very much like the proprietary MEMS design, manufacturing and packaging, 3D MEMS WLP can take many unique directions. Figure 4.17 gives one example of a conceptual stack of 3D MEMS package with ASIC flip chip bonded face-to-face to a MEMS die, which is then bumped with solder balls before singulated into individual packages.

For general-purpose WLCSP 3D stacking, cost seems to be a main hurdle for the adoption of the technology. WLCSP device are typically small with uniquely defined applications; urgency of vertical integration is not as high. However, that does not mean lack of innovative activities in the area.

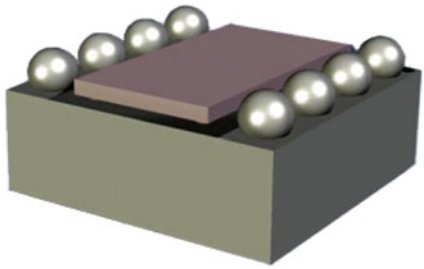


Fig. 4.17 A conceptual 3D MEMS package

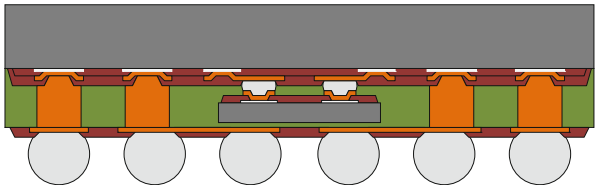


Fig. 4.18 A wafer-level 3D package without using TSVs

Figure 4.18 shows a cross section of a 3D WLCSP concept. Here, there is no need for TSV drilling and plating. CoW interconnection is made with mature mini-bump flip chip attachment. Copper pillar, front side wafer molding and redistributions are all proven technologies on WLCSP, but are integrated to make a unique package solution for special applications (Fig. 4.18).

4.6 Embedded WLCSP

Embedded WLCSP is, as a matter of fact, a 3D package that utilized the basic WLCSP chips and other active and/or passive components. It is not considered 3D WLCSP, since it has nothing to do with wafer format processing. Instead, PCB panel processing is utilized in the embedded WLCSP module production.

There are good reasons to pick WLCSP for embedding. First is that WLCSP bump pitch is just right considering the capability of PCB laser via process. Second, WLCSP with copper UBM is the right metallization for typical PCB seed metal processing (electroless copper plating) and subsequent electrolytic copper plating. Third, copper UBM size on WLCSP is typically larger than 200 μm in diameter, which makes it perfect for laser drill landing pad size. On the other front, back grinding and wafer sawing of UBM only WLCSP wafers down to the thickness range for embedding (>50 μm) are readily available technologies when combining the right tools and processes. All makes WLCSP for embedding an easy move on the silicon side.

There are clear motivations for embedding WLCSP into PCB substrates as well. In a mock-up study of a charger with USB-OTG boost regulator module, the size savings is more than 44 % when moving change design from all surface mount WLCSP and passives to embedded WLCSP plus surface-mounted passives, all

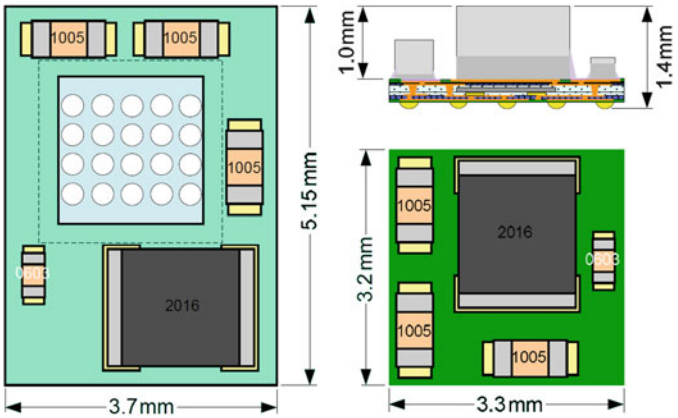


Fig. 4.19 Module size saving is >44 % with changing design from all SMT solutions (*left*) to the embedded WLCSP solution (*right*). Overall package height remains the same due to aggressive thinning of silicon

without changing the over package height and compromising the heat dissipation (Fig 4.19).

Reliability of the embedded WLCSP module is solid. Board level drop and TMCL performance is better than the same size WLCSP, even with the extra weight of the passive components. Primary reasons are as follows: (1) aggressive thinning of silicon and embedding make the effective CTE of the module substrate similar to that of the PCB the module is mounted on for board level testing. So stress on solder is actually less than that of WLCSP mounted on a PCB. (2) Embedding also makes silicon farther from the PCB, which is equivalent to the increasing of standoff height separating the two CTE mismatched elements. Strain and stress is thus reduced on two CTE mismatch elements and interconnections connecting the two. In this case, it is solder joint connecting the module substrate and PCB.

Embedding is an active area for packaging development. Embedded silicon(s) or passives do take up some space that could be otherwise used for through-layer vias and that makes it a limiting factor for 3D module package designs. So embedding might not be a solution for every application. Thermal performance of the embedded modules should also be carefully evaluated when great amount of heat is to be dissipated, though this is less often concerned, due to the existence of thermal paths that consist of buried copper planes and vias, and solder joints, all help passing the heat from the chips to the PCB boards the module is mounted on.

4.7 Summary

Miniaturization, efficiency, integration, and low cost have been driving the development of microelectronic technology to date. Package miniaturization has passed over TSOP, CSP, and WLP and has moved into PoP and SIP with more emphasis of system level integration. 3D packaging has come a long way from MCM to the

stacked package and the stacked die package. 3D IC, which shares the same root desire of maximizing the system functionality at smallest possible footprint, is the ultimate result of years of effort pursuing high-level integrations. TSV, wafer bonding, and extreme wafer thinning are the enabling technologies realizing system in chip 3D IC; with TSV having been leading the technology advancement, electrical/thermal design, wafer thinning and handling, wafer bonding, yield management and test will all influence the general market acceptance of the 3D IC technology.

Future of TSV development calls for big jump in throughput of deep feature etching for MEMS and TSV. The relatively slow speed of current DRIE remains a bottleneck for cost-effective high-volume production. Today's advanced tools may get etch rate no more than 50 μm/min, depending on the exposed area, which is sufficient for automotive devices, but a limiting factor for sensors and semiconductors oriented for the massive consumer electronics. 3D interconnect processes are especially limited, often able to etch TSV at an impractical rate of only a few wafers/hour. With the IC etch market potentially far larger than the MEMS market, higher throughput tools are needed to meet that market's needs. These faster volume production speeds will not only bring down the costs of MEMS devices to expand their use in consumer applications and wafer-level packaging but will also make production of 3D interconnect with TSVs practical. Recent improvements to the basic Bosch process promises significant improvement of throughput, with etch rate up to 100 μm/min.

TSVs are not the only factor to 3D package/3D IC advances. They represent just one part of an unfolding array of materials and processing and packaging/assembly developments. Ultrathin wafer grinding and handling are other critical areas for the success of TSV and 3D IC packaging.

3D ICs require ultrathin wafers. TSV benefits from small vertical dimension in two ways: (1) shortened via drill/etch time and (2) void-free via fill easier to achieve in the small aspect ratio vias. Wafer thinning is either for exposing the vias already formed in the wafers (via first, via middle) or for preparing wafers for via drilling (via last). Compared to convention IC packaging, which only sees thinnest silicon down to 4 mil (100 μm) for wire bond vertical MOSFET, 3D IC stacking requires silicon thickness typically below 100 μm, with more demanding application even down to 30 μm or even 15 μm. At this less than paper thin (>2 mil/50 μm) thickness, semiconductor wafers are becoming transparent (Fig. 4.20). Given the fragility of such very thin wafers, the need arises for highly specialized temporary wafer bonding and debonding equipment to ensure the integrity of the wafer structure, particularly at high processing temperatures and stresses during the etching and metallization processes. After bonding, the wafer undergoes a TSV backside process, followed by a debonding step. These typical steps result in higher yield levels for more cost-effective mass production.

After TSV and wafer thinning, bonding is another important area seeking development investment. Industry consensus is that WoW offers best throughput but requires very high per wafer yield to ensure the overall stacking yield. CoW allows to pick only KGD, thus improves stacking yield. Throughput of CoW is less than WoW, but it has been the process of choice due to yield consideration. With

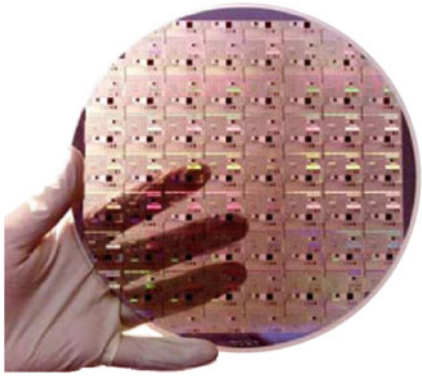


Fig. 4.20 Paper thin silicon wafer is transparent

further reduction of TSV size, alignment could be more challenging in the coming years. Also, bonding is the step where stacking layer interconnection is formed with 3D IC—that is only another reason to pay needed attention to this area.

3D IC is an exciting development area for the semiconductor industry. With mass consumer market moving into mobile and even wearable electronics, there will be more demands for low power consumption, yet more powerful computing capabilities integrated into smaller and smaller package size or end products. The trend for smart and connected electronics will also add more sensing and communication functions into smaller yet more environmental resistant packages. All these suggest that 3D package/IC is just at dawn and future is yet to come. That is great news for the semiconductor packaging society.

References

1. Moore, G.E.: Cramming more components onto integrated circuits. Electronics Magazine, p. 4. (1965). Retrieved 2006-11-11
2. Doe, P.: Bosch: deep etch tools on target for 100μm/min throughput in 2-3 years. Solid State Technology (2013)
3. Allan, R.: 3D IC technology delivers the total package. Electronic Design (2012)
4. Katske, H., Damberg, P., Bang, K.M.: Next generation PoP for processor and memory stacking. ECN, March 2010
5. Solberg, V., Gary, G.: Performance evaluations of stacked CSP memory modules. Electronics Manufacturing Technology Symposium (2004). IEEE/CPMT/SEMI 29th International, August 1, 2010
6. Leng, R.J.: The secrets of PC memory, December 2007. www.bit-tech.com
7. JEDEC Publication 95-4.22. Package-on-package design guide standard
8. Solberg, V.: Achieving cost and performance goals using 3D semiconductor packaging. Solid State Technology, August 1, 2010
9. IFTLE 24 IMAPS National Summary Part 1—3D Highlights. Solid State Technology, November 20, 2010
10. Tummala, R.R., Swaminathan, M.: System on Package: Miniaturization of the Entire System pp. 127~137 (2008)

11. Christensen, C., Kersten, P., Henke, S., Bouwstra, S.: Wafer through-hole interconnects with high vertical wiring densities. IEEE Trans. Compon. Packaging Manuf. Technol. A **19**, 516 (1996)
12. Gupta, S., Hilbert, M., Hong, S., Patti, R.: Techniques for producing 3DICs with high-density interconnect. In: Proceedings of the 21st International VLSI Multilevel Interconnection Conference, Waikoloa Beach, HI, pp. 93~97 (2004)
13. Hsu, D., Chan, J., Yen, C. (2010) TSV manufacturing technology integration. Semiconductor Technology (Taiwan), May 2010
14. Mannava, S.R., Cooper Jr. E.B.: Laser-assisted chemical vapor deposition, US Patent 5,174,826, December 1991
15. Williams, K.L.: Laser-assisted CVD fabrication and characterization of carbon and tungsten microhelices for microthrusters, Acta Universitatis Upsaliensis, Uppsala. ISBN 91-554-6480-7
16. Greenberg, M., Bansal, S.: Wide I/O driving 3-D with through-silicon vias. EE Times, February 2012
17. Euronymous: 3D integration: a revolution in design. Real World Technologies, May 2007
18. Dally, W.J.: Future directions for on-chip interconnection networks, p. 17. Computer Systems Laboratory Stanford University (2006) http://www.ece.ucdavis.edu/~ocin06/talks/dally.pdf
19. Woo, D.H., Seong, N.H., Lewis, D.L., Lee, H.S.: An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth. In: Proceedings of the 16th International Symposium on High-Performance Computer Architecture, Bangalore, India, pp. 429–440, January, 2010.
20. Kim, D.H., Mukhopadhyay, S., Lim, S.K.: Through-silicon-via aware interconnect prediction and optimization for 3D stacked ICs. In: Proceedings of International Workshop on System-Level Interconnect Prediction, pp. 85–92 (2009)
21. SEMI International Standards Program Forms 3D Stacked IC Standards Committee. SEMI press release December 7, 2010
22. Bartek, M., Sinaga, S.M., Zilber, G., et al.: Shellcase-type wafer-level-packaging solutions: RF characterization and modeling, IWLPC, October 2004
23. Humpston, G.: Setting a new standard for through-silicon via reliability, Electronic Design, October 2009
24. Robert, D., Gilleo, K., Kuisma, H.: 3D WLP of MEMS: market drivers and technical challenges. Advanced Packaging, February 2009

Wafer-Level Discrete Power Mosfet Package Design

5

Discrete power device is one of the basic units that fuel the power management and conversion in various applications. Typical discrete products include various diodes, bipolars, metal–oxide semiconductor field-effect transistors (Mosfets), and insulated gate bipolar transistor (IGBTs). As the needs of miniaturization, one trend of the discrete power Mosfet packages is to move the discrete power devices into various wafer-level chip-scale package for better surface amount purpose. Since power levels and power density requirements continue to increase for many types of end equipment such as personal computers, servers, network, and telecom systems, it demands higher performance from the components that make up the power management system. This chapter introduces the design of discrete power package and the analysis of the wafer-level discrete power package performance.

5.1 Introduction and the Trends of Discrete Power WLCSP

Since the discrete product started, most of the power discrete products are molded packages. Typical molded discrete power packages include the three terminal packages, such as small outline transistor (SOT) family [1]; TO family including DAP (TO-252), D2PAK (TO-263); dual-in-line packages, such as small outline (SO) family including thin small outline package (TSOP) family and the thin shrink small outline package (TSSOP) family; and the quad-in-line package such as the quad flat no-lead (QFN) family, power quad flat-pack no-leads (PQFN) family with exposed die-pad for heat sink, and the molded leadless package (MLP) families. Figure 5.1 shows an eight leads SO power package. The eight leads SO package includes a leadframe with a die-attach pad that connects four leads and the drain of the power Mosfet die. The source of the power semiconductor die connects to the three source leads through bond wires. One gate wire connects the Mosfet gate and the gate lead. The whole package is encapsulated with epoxy mold compound (EMC) material. The EMC is a major encapsulated material for the discrete power

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_5

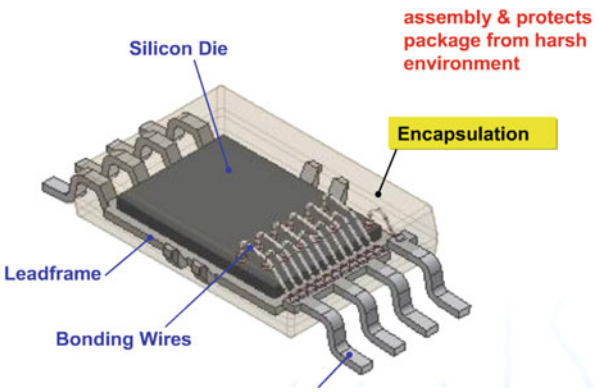


Fig. 5.1 A typical power package SO8

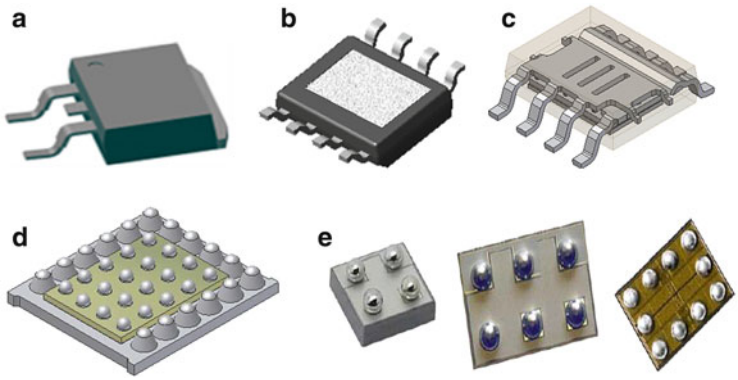


Fig. 5.2 The development evolution of the discrete power package: (**a**) TO-263 package, (**b**) SO8 package with bond wire, (**c**) SO8 package with metal clip, (**d**) Mosfet BGA, (**e**) power Mosfet WLCSP

package. This is because the EMC can provide substantial protection and mechanical integrity to place components across a wide generational range of pick and place equipment. However, one drawback of EMC is that its thermal performance is not good as the metal. Therefore, when larger current density request and smaller size become necessary such as in the portable application, the EMC technology is not enough to satisfy these requirements. To meet such requirement, in the recent year there appears Mosfet ball grid array (BGA) and wafer-level chip-scale package (WLCSP) for discrete power device. Figure 5.2 shows the development evolution of the discrete power device. Figure 5.2a is one of the earliest packages for power device TO-263 which is still widely used today for the discrete device. Figures 5.2b, c are the 8 leads SO power package with and without the bond wires. In the case of SO8 without the bond wires, the package uses the metal clip which improves the electrical and thermal performances. Figure 5.2d shows the Fairchild Mosfet BGA,

which directly attaches the drain of power Mosfet on the folded leadframe without the epoxy molding compound. The drain of the Mosfet can then be found through the pins of leadframe which are located at the front side of the Mosfet due to the folded leadframe design. Figure 5.2e shows the typical discrete power WLCSP in various products. From Fig. 5.2, it can be seen that the designs shown in Figs. 5.2d, e have a very good electrical performance due to reduced die size and package size and its easiness to mount as a surface package. WLCSP has become one of the trends of power discrete packaging due to high integration and automation and cost-effective and excellent electrical performance.

Figure 5.2 also gives the representative discrete power transistor packages development from early DPAK (TO252) through SO8 to MOSFET BGA and MOSFET WLCSP. The molding compound decreases as a percentage of volume, until it reaches zero with the MOSFET BGA and WLCSPs.

5.2 Discrete Power WLCSP Design Constructions

In this section, three typical discrete power WLCSP design constructions are introduced. One is the standard discrete power WLCSP design. The second one is the Mosfet BGA which is the design that integrates the discrete WLCSP with LF to bring the Mosfet drain to the WLCSP front with the same side of gate and source. The third one is copper stud bumping WLCSP (see Sect. 5.4).

5.2.1 Typical Discrete Power WLCSP Design Construction

Figure 5.3 shows the discrete power bump system design construction on a source of a VDMosfet. The backside is the drain of the Mosfet. The gate bumping construction is similar to Fig. 5.3. In Fig. 5.3, the solder bump normally has 150–300 μm height, which includes Pb and Pb-free materials. It connects the Mosfet source to external surface mount application. UBM adheres to metal Al and

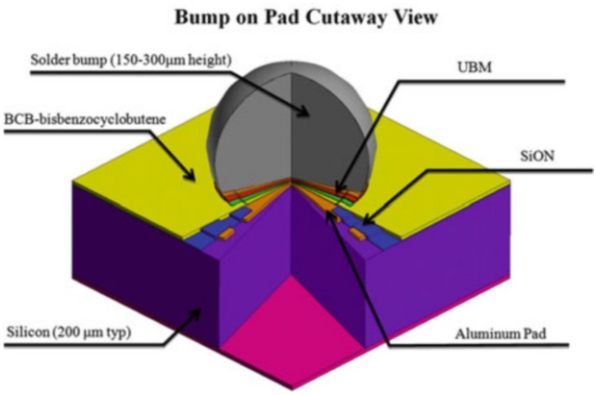


Fig. 5.3 The discrete power WLCSP design construction

provides a solderable interface for solder connection. Bisbenzocyclobutene (BCB) acts as passivation and stress relief layer during thermal excursions. SiON defines the bump pad openings and prevents Al corrosion. Aluminum pad is normally 2.5–5 μm thick and provides the current path from Mosfet silicon source to the bump. The Mosfet silicon normally has 200–300 μm thickness. In this design, the gate and source are in front WLCSP, while the drain of Mosfet is at the backside of WLCSP.

5.2.2 The Power Mosfet BGA

In most cases, the drain of VDMosfet is at the backside of Mosfet as in the standard discrete power WLCSP Mosfet, which is hard for a WLCSP mosfet to work in a surface mount environment. To bring the backside drain of WLCSP VDMosfet to the front active side, a stamped or etched leadframe is used to attach the backside of the WLCSP. It is the concept of Mosfet BGA. The design construction of Mosfet BGA includes the Mosfet die with bumps, leadframe carrier, solder balls, and paste as shown in Fig. 5.4.

The typical assembly process for Mosfet BGA is shown in Fig. 5.5. The first assembly step is to make the fluxing on the leadframe. Then attach the solder ball on it and reflow (Fig. 5.5a). The second step is to dispense the solder paste on the leadframe die-attach pad and attach the Mosfet die (Fig. 5.5b). The third step is to have the strip test followed by the laser marking (Fig. 5.5c). Then the last step is to singulate the Mosfet BGA from the leadframe by punch and to make the tape and reel (Fig. 5.5d). The benefit of Mosfet BGA design is that it provides an approach not only to change the standard discrete power WLCSP to a surface mount solution but also at the same time to dissipate the heat from multiple directions.

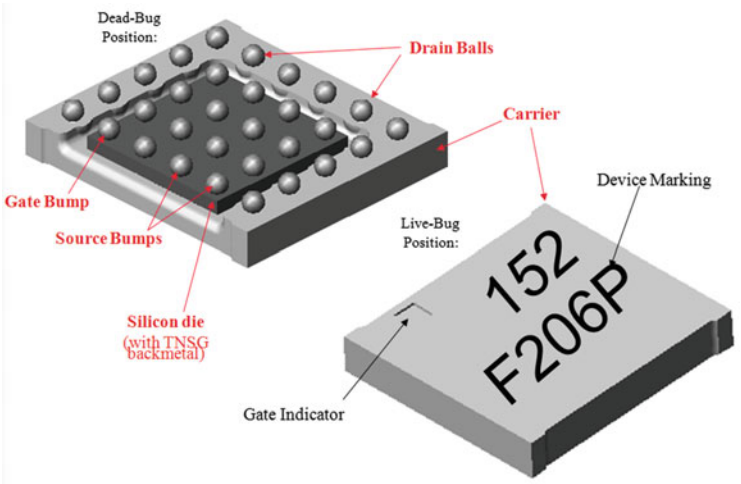


Fig. 5.4 The Mosfet BGA construction

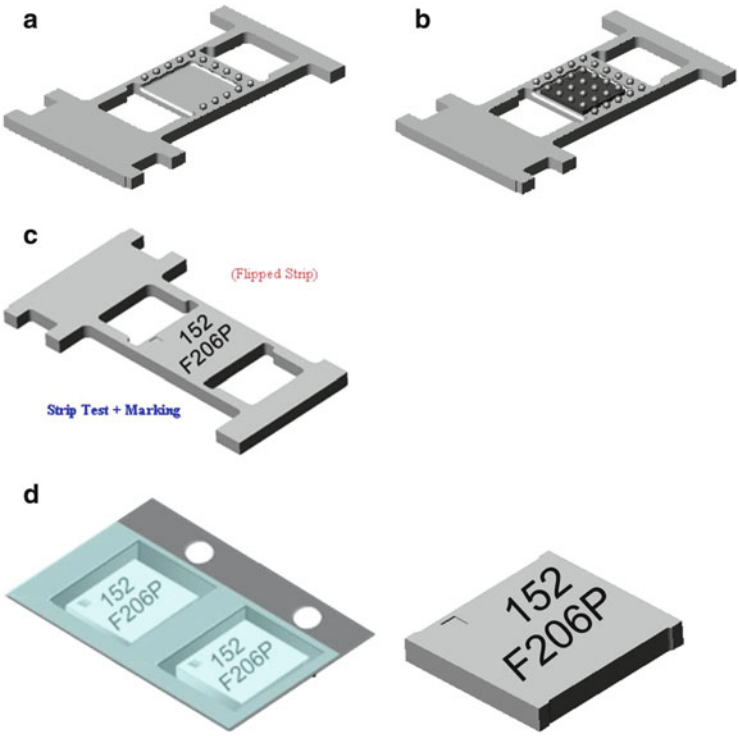


Fig. 5.5 Assembly process for Mosfet BGA. (**a**) Fluxing, solder ball attach and reflow. (**b**) Mosfet die attach and second time reflow, (**c**) strip test and marking, (**d**) singulation, tape and reel

5.2.3 Move the MOSFET Drain to Front Side in Discrete Power WLCSP

For the wafer-level discrete MOSFET, another trend which obtains the attention in the industry is to move the drain of the MOSFET to the front side of the die directly in wafer process so that the drain, source, and gate are at the same side, which would have better electrical performance as compared to the approach Mosfet BGA which uses LF substrate. This would also be helpful for the surface mounting application in various PCBs.

Figure 5.6 shows one of the lateral layouts of the drain for a LDMOSFET WLCSP. Since the drain is in lateral placement, its application limits to relative lower power and lower voltage area as compared to VDMOSFET.

Today, the power technology may develop the backside drain of a VDMosfet to the front side of the Mosfet through silicon via (TSV); example is to connect the backside metal to front through via in trench as shown in Fig. 5.7. Figure 5.8 shows a discrete power WLCSP pin map with all drain, source, and gate on the same front active side. This is very useful for the surface mount applications, especially in mobile.

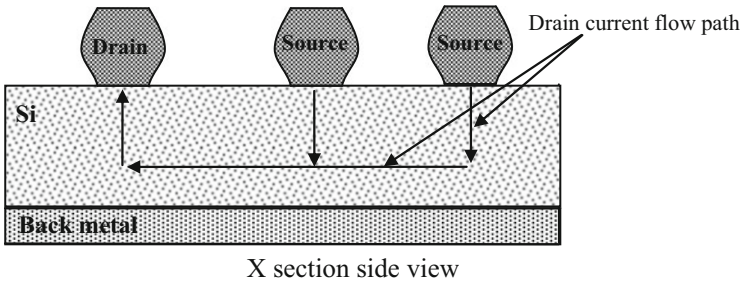


Fig. 5.6 Move the drain to the front side the LDMOSFET

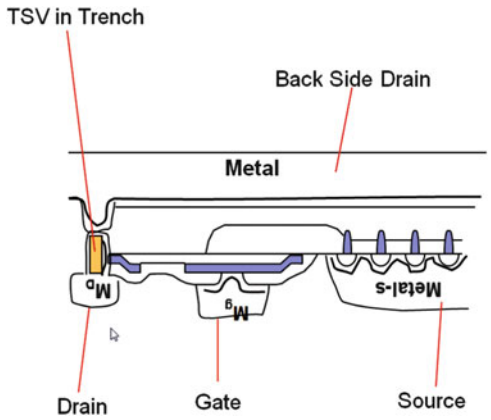


Fig. 5.7 Bring the Mosfet drain to front side

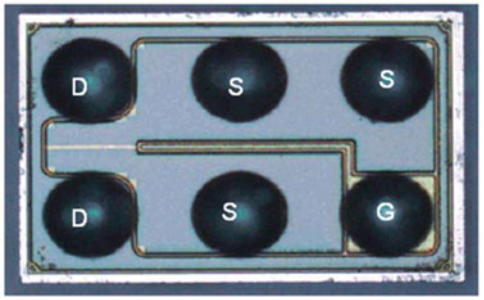


Fig. 5.8 The pin map of 2×3 discrete power WLCSP

5.3 Wafer-Level Mosfet Direct Drain Design

In standard VDMOSFET WLCSP, the drain is at the backside of the WLCSP. To move the Mosfet drain to the front, one approach is to develop the direct connection to the front side. Different from LD Mosfet WLCSP (Fig. 5.6), the direct contact drain design is based on the TSV to connect the backside drain metal of VDMOSFET.

5.3.1 The Construction of the Direct Drain VDMosfet WLCSP

The advantage of the direct connection of the back drain to front side is its good electrical performance with lower Rdson. Because this is a vertical DMOS, the application area may be relatively wider in power range as compared to the structure in Fig. 5.6. Figure 5.9 shows the design construction of the direct drain connection in a discrete power WLCSP, which consists of silicon substrate, thick backside drain metal, TSVs, front drain, and bumps of drain, source, and gate. In Fig. 5.9, the drain has been moved to the front side through the TSVs. The assembly process of the discrete WLCSP includes the following steps: (1) the regular Mosfet active front side process for source and gate; (2) non-through vias process; (3) metal plating process to form the non-through vias and the front drain metal area based on the layout; (4) backside grinding process on silicon substrate to form the silicon through vias; (5) plating the backside drain metal; (6) solder balls dropping on drain, source, and gate; and (7) singulating the wafer to get the discrete VDMosfet WLCSP.

5.3.2 Other Construction of the Direct Drain VDMosfet WLCSP

The VD wafer-level mosfet comprises the silicon substrate with a source and gate metal layers on the front surface of the substrate which connects to source and gate of the power mosfet device and a back metal layer on the bottom of the substrate which is connected to the drain of the VDMOSFET as shown in Fig. 5.10.

Besides the design construction of direct drain VDMosfet WLCSP in Fig. 5.9, there is other construction that contains cavities on the upper surface of the silicon substrate with a depth sufficient to expose the back metal layer. The drain cavities

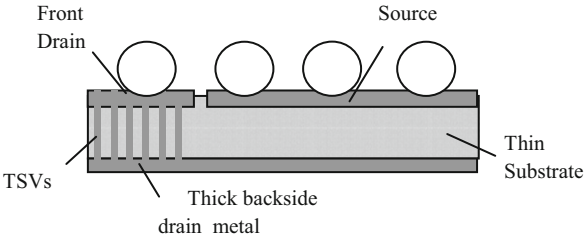


Fig. 5.9 Direct connection of back drain discrete VDMosfet WLCSP

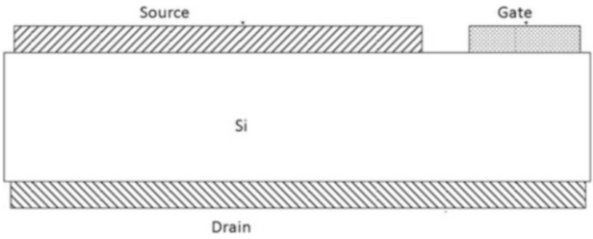


Fig. 5.10 The metal layout of a VDMosfet

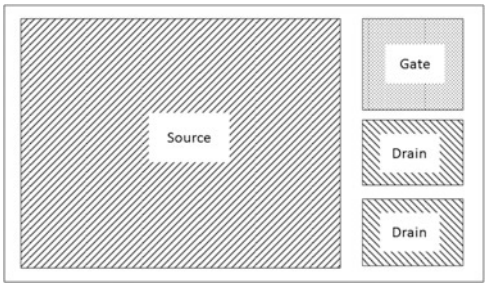


Fig. 5.11 The front layout of a new VDMosfet

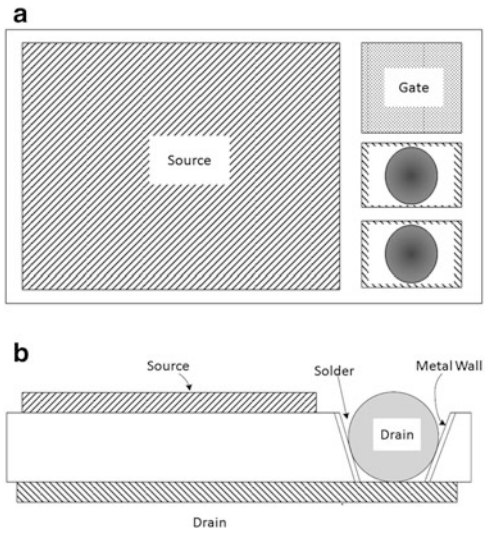


Fig. 5.12 The construction of the new VDMosfet WLCSP. (**a**) The drain cavities filled with solder balls, (**b**) the cross section of the new VDMosfet

layout is shown in Fig. 5.11. In Fig. 5.12, it shows the construction of such design concept: Fig. 5.12a is the front view of the design which includes the source, gate, and the two drain bumps; Fig. 5.12b is the cross section of the new construction. The cavity wall contains a metal wall which has a dielectric layer formed between the metal wall and the silicon substrate. A solder microbump is self-aligned within the cavity so that its bottom contacts the back metal layer (drain) and the sides of the microbump contact the cavity metal wall. Since the solder microbump contacts the

back metal layer, it can operate as a drain for the discrete power VDMosfet WLCSP. Such a configuration of the direct contact WLCSP allows ultrathin device to be made and decreases the on-resistance of the device as well as the parasitic inductance and capacitance. Since most of the source area can directly attach (or mount) on PCB, the heat dissipation capability will also be improved. This is a very promising power technology which shrinks the solution to the thinnest possible; it allows the highly efficient operation with high frequency.

5.4 Power VDMOSFET WLCSP with Cu Stud Bumping

5.4.1 The Cu Stud Bumping Construction on a Power WLCSP

There are two major bumping technologies (see Fig. 5.13): one (Fig. 5.13a) is the regular UBM-based bump method which normally selects high Pb solder material to keep the bump height under reflow, which is the construction as Fig. 5.3. The other (Fig. 5.13b) is the Cu stud bump method which uses the solder material on Cu stud bumping. The UBM-based solder bumping is much more expensive than the Cu stud bumping process. In addition, the Cu stud bumping may use the SAC Pb-free solder materials due to the Cu stud which satisfies the requirement of the customers and is the trend of power WLCSP with Pb-free solder. Figure 5.14 shows the construction of the copper stud bumping structure on the metal of the power chip. It includes the solder ball, Cu stud, SiON passivation, metal aluminum, and the silicon substrate. Figure 5.14a shows the cutaway view of the construction and Fig. 5.14b shows the Cu stud bumps without solders.

5.4.2 BPSG Profile Under the Al Layer During Cu Stud Bumping Process

One of the benefits of Cu stud bumping is the lower cost. However, due to the Cu stud bumping process and the harder material properties of Cu, it would induce the potential cratering on the silicon as well as the crack on the device layer, such as the borophosphosilicate glass (BPSG) profile (see Fig. 5.15).

In Cu stud bumping process, the wire bonder forms a free air ball (FAB) on the tip of a wire protruding from the capillary. The capillary then descends to the work surface and bonds the ball. Instead of moving on to form a wire loop, as in a typical wire bond process, the capillary rises and shears off the wire above the ball before forming a new ball. The process is repeated for as many bumps required on the device. The Cu stud bumping process consists of two distinct steps: one is the Cu wire bonding process and the other is the shearing process as shown in Fig. 5.16. Stud bumping is significantly faster than wire bonding, because all of the looping motions of normal wire bonding are not needed. Incorporating wire bond technology as a part of the flip chip Cu stud bumping process is attractive because existing facilities and infrastructures can be used without the high capital costs required by

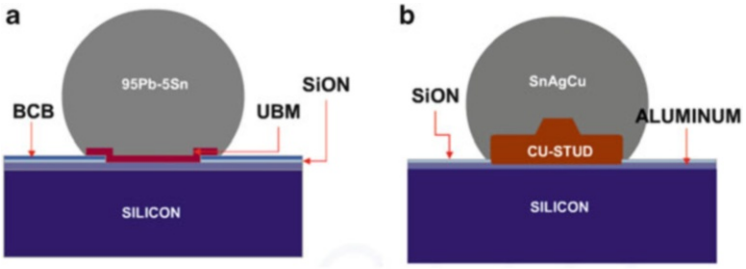


Fig. 5.13 Power WLCSP bumping options: UBM versus Cu stud. (**a**) Regular power WLCSP, (**b**) copper stud bumping with solder

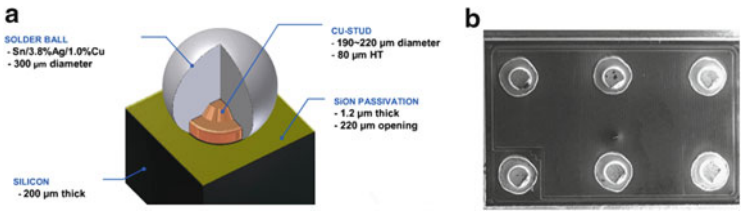


Fig. 5.14 WLCSP Cu stud bumping package. (**a**) The construction of Cu stud bumping, (**b**) the Cu stud layout on the source and gate

more expensive sputter/plating facilities inevitable in the conventional flip chip process. This technology has several advantages: a UBM process is not necessary, bumping cost is low, and fine pitch and chip-level bumping is possible.

During the Cu stud bumping process, wafer cratering may be induced. Cratering means the copper stud comes off and creates a local deformed area in the silicon or interlayer dielectric under the bond pad [2]. The numerical analysis of the Cu stud bonding and shearing process is conducted by using the commercial finite element code ANSYS®. The impact of three BPSG profiles, dome shape, square shape, and M shape, and two FAB diameters, 190 µm and 145 µm, is investigated. Experimental results of the bond crater test are discussed for the wafers with these three different BPSG profiles.

5.4.2.1 Bumping Model [3–5]

A 2D FEA bumping model is shown in Fig. 5.17, which includes power die, Cu wire, and capillary. The power device includes silicon layer, BPSG layer, TiW layer, and Al metal layer (see the local structure under the Al metal shown in Fig. 5.18). During the Cu stud bumping process, an ultrasonic power is applied to the capillary to form the ball bond.

The 2D FEA models with two different FAB sizes 190 µm and 145 µm are illustrated in Fig. 5.17. The silicon structure with three different BPSG profiles, dome shape, square shape, and M shape, is shown in Fig. 5.19, together with the actual BPSG SEM pictures.

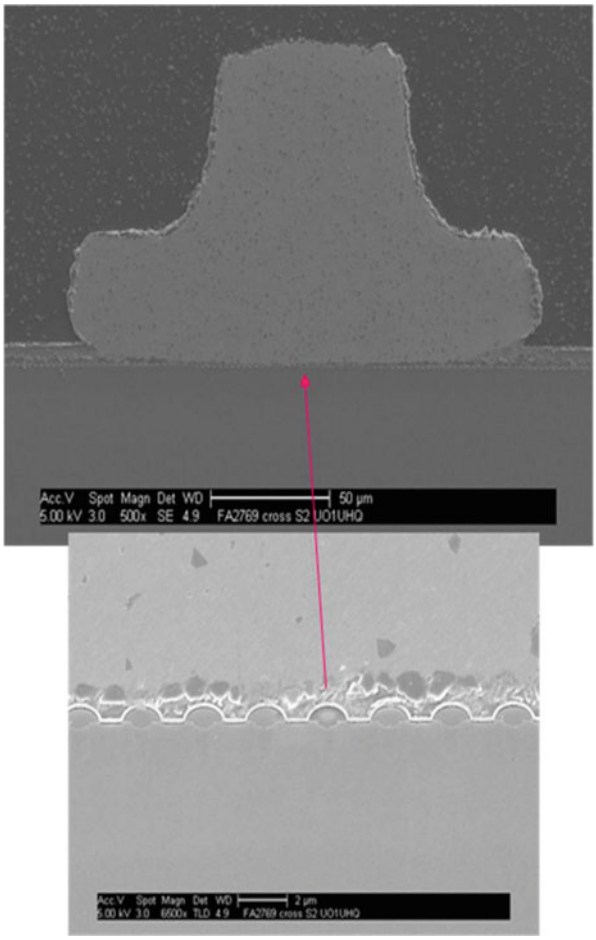


Fig. 5.15 The BPSG profile under the Cu stud bumping

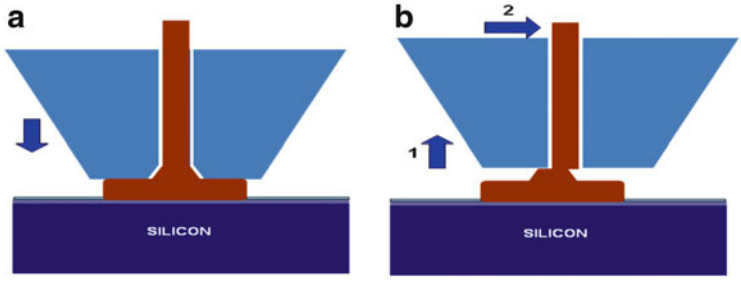


Fig. 5.16 The Cu stud bumping process. (**a**) Capillary moving down for wire bonding, (**b**) capillary lifting and shearing

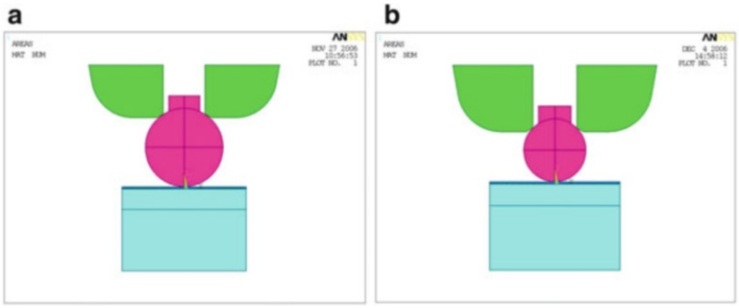


Fig. 5.17 A 2D FEA model including silicon, Cu wire, and capillary with different FAB sizes: (**a**) 190 μm and (**b**) 145 μm

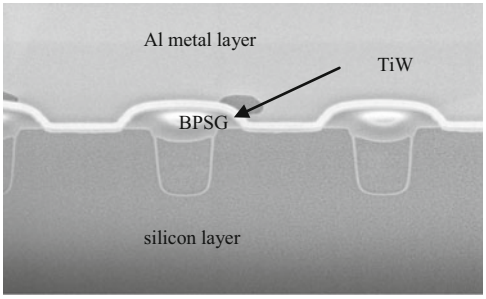


Fig. 5.18 Local structure under Al metal

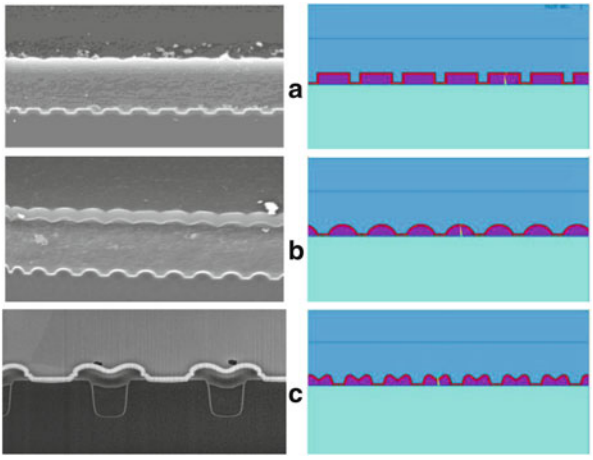


Fig. 5.19 Wafer structure with different BPSG profiles: (**a**) square shape, (**b**) dome shape, (**c**) M shape

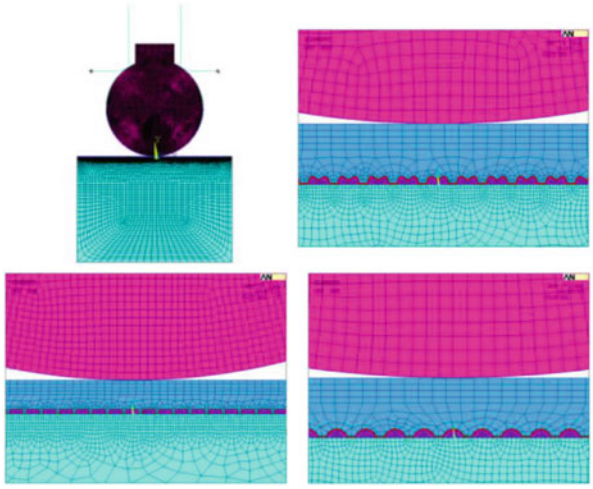


Fig. 5.20 FEA mesh of the bonding model with different BPSGs

The Cu stud bumping process consists of bonding and shearing. In the following sections, the simulation procedure and results discussion are described separately.

5.4.2.2 Bonding Process Simulation

In the bonding process, the capillary vibrates laterally due to the ultrasonic power; at the same time, it presses the FAB downwards to form the ball bond. The mesh of the FEA model with three different BPSG profiles is illustrated in Fig. 5.20.

1. Effect of FAB size on BPSG profile
   The deformation contour of the model after bonding process is shown in Fig. 5.21, with two different FAB sizes 190 μm and 145 μm. The simulation results show that the deformation of 190 μm FAB needs to be larger to obtain the proper final ball bond shape.
   Shear stress distribution of BPSG/TiW layer is shown in Fig. 5.22. It can be seen that smaller FAB size induces less stress in BPSG/TiW layer, thus produces less destruction of these layers during bonding process.
2. Effect of BPSG profile
   The shear stress distribution of BPSG/TiW layer is illustrated in Fig. 5.23 for square and M shape BPSG profile with copper FAB diameter 145 μm. By comparing with Fig. 5.23a, b for dome shape BPSG profile, it can be seen the M shape BPSG induces the largest stress in BPSG/TiW layer after the bonding process.

5.4.2.3 Shearing Process Simulation

During Cu stud bumping process, first the ball bond is formed in the bonding process as described in Sect. 5.4.2.2, and then the capillary rises and shears off the wire above the ball bond, which is regarded as shearing process. The FEA model for shearing process is shown in Fig. 5.24. The shearing height is based on

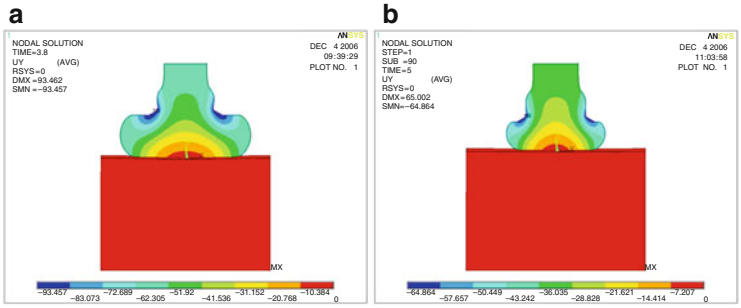


Fig. 5.21 Deformation of the model after bonding process with different FAB sizes: (**a**) 190 μm, (**b**) 145 μm

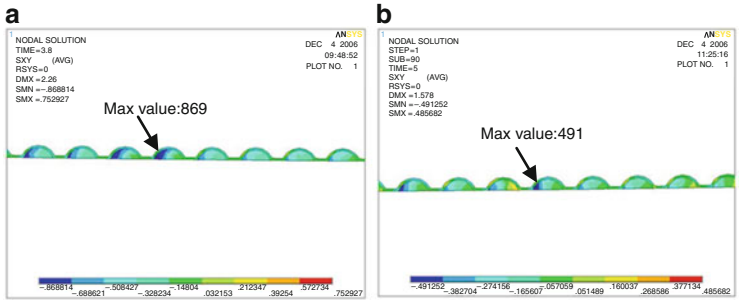


Fig. 5.22 Shear stress distribution in BPSG/TiW layer with different FAB sizes: (**a**) 190 μm, (**b**) 145 μm

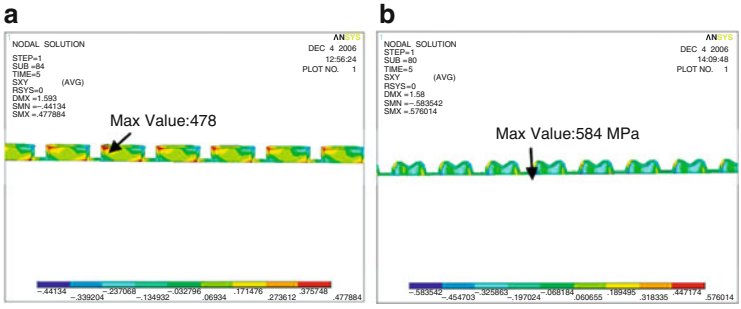


Fig. 5.23 Shear stress distribution of the BPSG/TiW layer for different BPSG profiles: (**a**) square shape, (**b**) M shape

the picture of the actual finished Cu stud shown in Fig. 5.24b. A horizontal displacement 75 μm is applied to the capillary as the shearing load. It is assumed that the ball bond is connected to the Al metal layer ideally, no separation occurred during the shearing simulation.

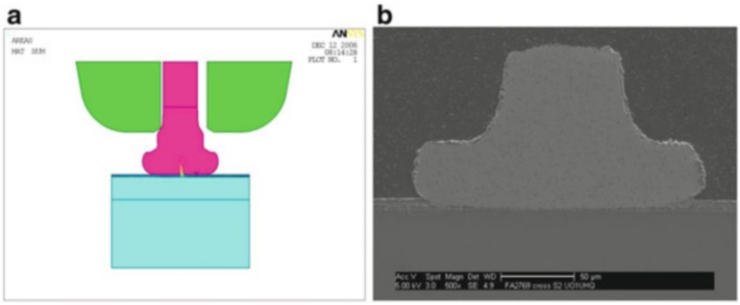


Fig. 5.24 (**a**) FEA model for shearing process, (**b**) actual SEM picture of finished Cu stud

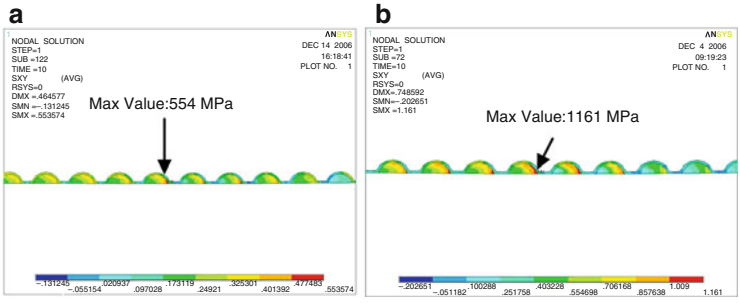


Fig. 5.25 Shear stress distribution in BPSG/TiW layer with different FAB sizes: (**a**) 190 μm, (**b**) 145 μm

1. Effect of FAB size on BPSG
   The shear stress contours of the BPSG/TiW layer with two smashed FAB sizes under shearing process are illustrated in Fig. 5.25. It can be seen that smaller FAB size with diameter 145 μm induces significantly large stress during the shearing process.
2. Effect of BPSG profile
   The shear stresses in BPSG/TiW layer and BPSG layer are illustrated in Fig. 5.26 for square and M shape BPSG profiles with copper FAB diameter 145 μm. By comparing with Fig. 5.25b for the dome shape BPSG profile, it can be found that the shear stress in the dome shape BPSG/TiW layer is the lowest.

5.4.2.4 Experimental Results and Discussion

Experiment is conducted to investigate the BPSG profile impact on the cratering of silicon.

The experiment results are listed in Table 5.1. Crater check is conducted after Cu stud bumping and solder ball reflow for the three settings of the BPSG. No bond cratering occurred for all the cases with different BPSG profiles and setting in Table 5.1. This indicates that with the process used, the BPSG profile might not be sensitive to silicon cratering in the Cu stud process. This result seems consistent with the FEA simulation results, which show that the stresses on the silicon dice

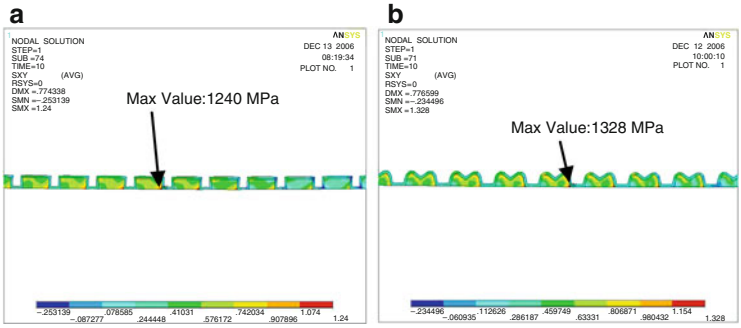


Fig. 5.26 Shear stress distribution in BPSG/TiW layer for different BPSG profiles: (**a**) square shape, (**b**) M shape

Table 5.1 Experimental results for bond crater check

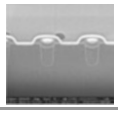
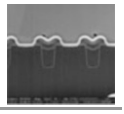
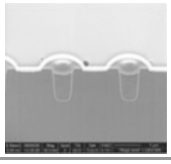
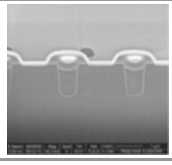
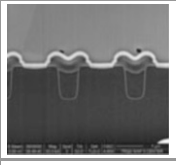
BPSG profile test			
Bond crater test (after Cu stud bumping)	0/128	0/128	0/128
Bond crater test (after solder ball reflow)	0/128	0/128	0/128

Table 5.2 Reliability test result

BPSG profile test			
Cu stud bump shear test	Pass	Pass	Pass
Solder bump shear test	Pass	Pass	Fail
72 h HTSL	0/248	0/244	1/247

with different BPSG shapes do not have significant difference in Cu stud bumping (see Figs. 5.22b and 5.23a, b) with the same FAB diameter 145 μm), while the stress on silicon with M shape BPSG is the largest. However, the reliability tests for the same settings of the BPSG profile have shown the fails in shear test and in 72 HTSL test (see Table 5.2). This indicates that the M shape BPSG is not strong enough in its reliability test even if it can pass the bond crack check after the Cu stud bumping.

5.5 3D Power Module with Embedded WLCSP

This section introduces a 3D fan-out power module that integrates the power discrete WLCSP or mixed with analog WLCSP and the passives. The qualifications of a 2 × 3 and a 7 × 7 embedded WLCSP power module were shared in this chapter.

The 2×3 module was built on a switching voltage regulator with three passive components and one embedded WLCSP (2×3 at 0.4 mm pitch). The chip to module size ratio is 18.1 %. The 7×7 module was built on a 7×7, 0.4 mm pitch WLCSP daisy chain test chip and five passive components. The silicon to module size ratio is 52.4 %. Both 2×3 modules and 7×7 modules were subjected to drop test and temperature cycling test (TMCL). In addition, functional 2×3 modules underwent more device level reliability tests, such as dynamic optional life (DOPL), temperature humidity bias test (THBT), high temperature storage life (HTSL), TMCL, etc. Solid chip to module, passives to module interconnect reliability was demonstrated alongside the robust module board level reliabilities.

With adoption of PCB build-up technology, the embedded WLCSP module offers 3D packaging options that are readily available to customers who demand full functionality in a small package and an attractive cost [6]. Compared to other 3D packaging options, routing flexibility and robust interconnect reliability of the embedded module are distinguishing benefits. More designs taking advantages of this packaging technology are expected in the coming years.

5.5.1 Introduction

As the marketplace continues to demand size reductions in portable electronic products, the need for smaller packaging and smaller subsystem packaging becomes paramount. In order to reduce the overall package size, meaning length and width, 3D stack up is inevitable. The system in package (SIP) technology integrating embedded WLCSP and surface mounted passive components in a single device is a natural choice of the trend. The small 3D SIP footprint provides a stand-alone power supply platform driving power to package density to a new level.

Embedding semiconductors or passives into a PCB is not a new concept, though it has been an active topic in recent publications [7–9]. When the inner layer routing is not a concern, saving of lateral size compared to all surface mounting solutions can be easily realized with the embedded WLCSP modules. In a study of potential size reduction of a module with a 20 pin integrated switch-mode charger and five passive components, module lateral size reduction is estimated to be more than 37 % when changing design from an all surface mounted solution to a module with the WLCSP embedded, without sacrificing the overall package height (Fig. 5.27).

Embedding WLCSP does come with unique requirements for the semiconductor devices. For example, chip to package interconnection is formed via typical PCB blind via process, meaning UV laser via open, copper seed deposition, copper plating, and trace pattern etch. Limited by the PCB via size and layer registration accuracy, minimum on-chip landing pad size is desired to be more than 150 μm in diameter. This pad size requirement makes WLCSP a natural fit, since almost all current WLCSP technology has UBM size larger than the required minimum size for via contact in embedding. Also to make it compatible with standard PCB laser via open process and copper seeding process, on-chip landing pad with copper metallization more than 5 μm thick is necessary to accommodate copper thickness

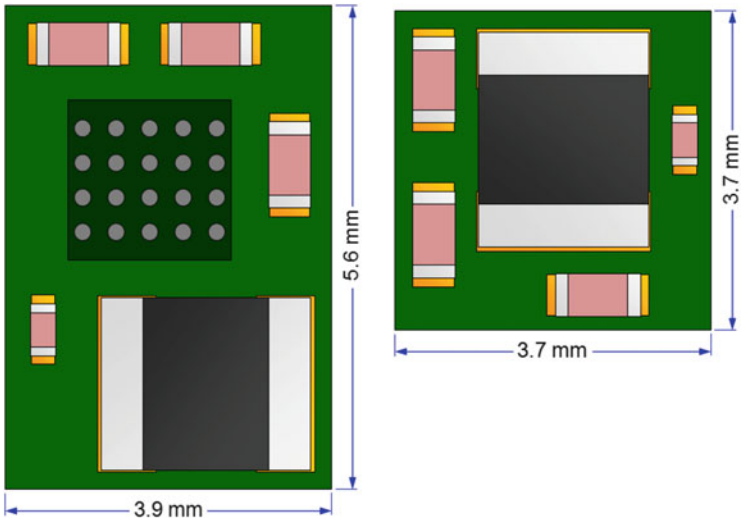


Fig. 5.27 More than 37 % size reduction can be achieved switching from an all surface mount module design (*left*) to the embedded module design (*right*)

loss through laser via open, via preparation, and copper surface clean before electroless seed layer deposition. Polymer repassivation on top of native semiconductor passivation is not a must for embedding. As a matter of fact, the interfacial adhesion between polymer repassivation, such PI or PBO, and typical PCB laminates materials could be worse than the adhesion between SiN passivation and PCB laminates. However, the polymer repassivation is often left intact for an embedding device originated from a WLCSP for minimum process changes. What really makes embedding a challenge is the thinness of the silicon required—for conventional WLCSP, more than 200 μm silicon thicknesses is typical for robust manufacturing operations in wafer back grinding, backside lamination, wafer saw, and tape and reel. For the embedded WLCSP, silicon thickness less than 120 μm or even 50 μm is typically requested in order not to increase the module substrate thickness. The thin silicon thickness needed for the embedding is found challenging for WLCSP back grind and dicing, as well as pick and place in tape and reel.

Once embedded, the reliability of the copper vias connecting the silicon to module as well as the solder joint reliability from passives to module and solder joint reliability from module to PCB need to be fully understood. Since the intended application of the modules is mobile electronics, it is also of particular interest of the module solder joint reliability during board level drop and temperature cycle, as it is believed that the extra weights from the passive components bring additional stress to the module/PCB solder joints, especially in the drop test. All will be investigated in the presented study.

5.5.2 Embedded WLCSP Modules

Two embedded WLCSP modules of different die and module sizes were built for the study of WLCSP embedded module board level reliabilities and component (module) level reliabilities. The first module was built on a six pin (2 × 3) WLCSP synchronous buck regulator with three passives. Traditional component reliabilities (i.e., dynamic operational life (DOPL), temperature humidity biased test (THBT), high temperature storage life (HTSL), temperature cycle (TMCL), etc.) were tested on this module. In order to perform the drop test on this module size, a 2 × 3 daisy chain test chip of the same size replaced the functional silicon to allow continuous monitoring of the chain resistance during drop test. The second module was intended to expand the embedded WLCSP technology to larger size with high pin count. In this design, a 7 × 7 daisy chain test chip was embedded in a module with five passive components surface mounted on the top side. Both modules feature a four-layer substrate construction with silicon being grinded to 50 μm thickness before embedding into the module substrates less than 350 μm thick (Figs. 5.28 and 5.29). To monitor the solder joints between the surface mounted passives and module substrate, capacitors were replaced with inductors of similar sizes and weights to allow the use of event detector during board level reliability tests.

For drop and TMCL test, customer-specific PCB and PCB defined in JESD22-B111 was designed and fabricated to the specifications. Then component modules were mounted and tested. Due to the test board cabling and event detector channel limitations, only one channel per component (through every blind/buried vias and

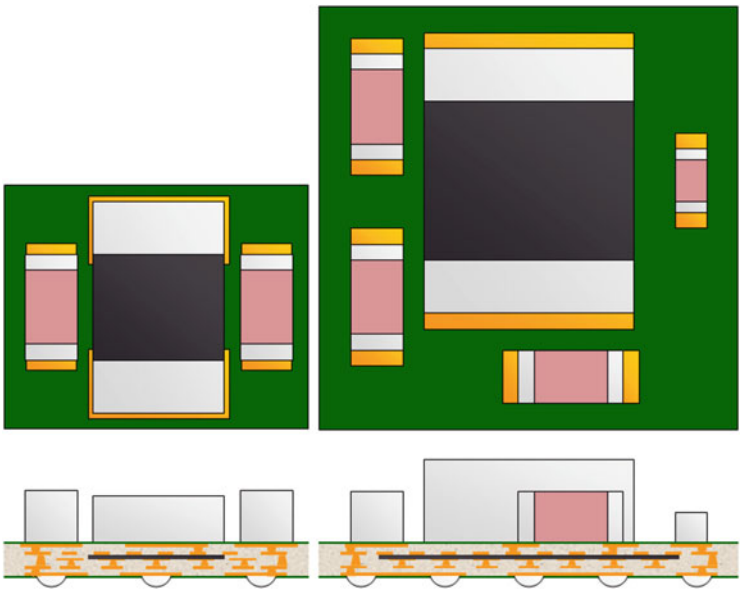


Fig. 5.28 Top view and cross-sectional view of the two studied embedded WLCSP modules

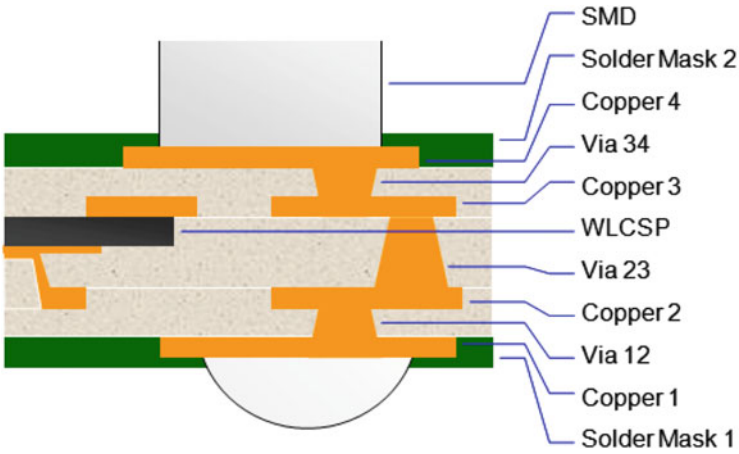


Fig. 5.29 Embedded WLCSP module stack up

all solder connections) was continuously monitored during test. However, the PCB was specially designed so that the individual module component, either WLCSP or passives, can be manually probed at the selected time points to help confirm or isolate failure locations. The schematic of this PCB cell design for the 7×7 WLCSP module is illustrated in Fig. 5.30. For example, though only one event detector channel was used to continuously monitor all interconnections within the module and between module and PCB, the 49 blind vias that connects the module to the embedded WLCSP can be manually probed between PIN C2 and D2. Similarly, when solder between the big inductor and module is in doubt, a manual probe between C1 and D1 can address concerns.

5.5.3 Reliability Tests

Drop and TMCL of the 2×3 daisy chain modules were tested on a customer-specific PCB and under customer-specific test conditions. JEDEC drop test condition B (1,500 G, 0.5 ms duration, half-sine pulse), as listed in JESD22-B110, was applied to the 7×7 embedded WLCSP modules mounted on JEDEC drop PCB. IPC9701 cycle condition TC3 (−40 °C to 125 °C, 52 min/cycle) and cyclic bend test defined by JESD22-B113 (2 mm displacement, 200,000 cycles) were also applied to the drop PCB/module assembly of the 7×7 modules. Table 5.3 summarizes the test results of the board level reliabilities of both the 2×3 and the 7×7 daisy chain modules. Figure 5.31 also gives the Weibull plots of the drop of the 7×7 module, though there were not a sufficient number of failures recorded; when stopped at 1,000 drops, it is evident that the 7×7 daisy chain module performed well in JEDEC drop test. The first failure only occurred after 616 drops, and there are plenty of margins from the minimum requirement of 150 drops.

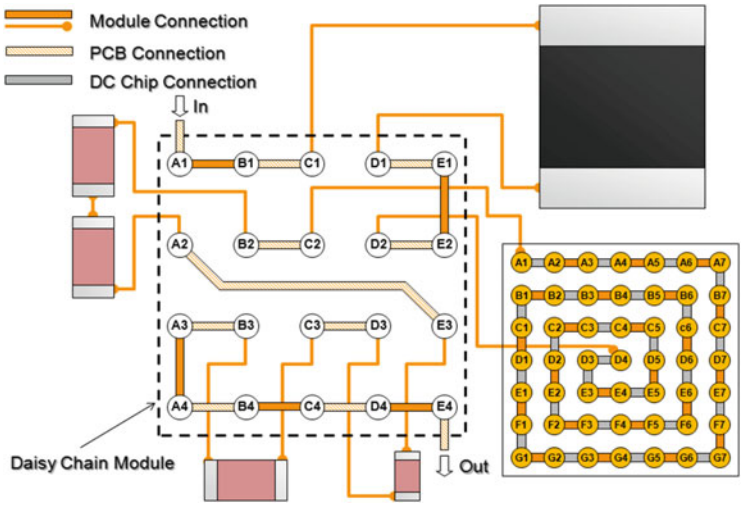


Fig. 5.30 Schematic of module electrical connections, with highlight of manual probe points for individual components of the embedded WLCSP module

Table 5.3 Board level reliability summary of 2 × 3 daisy chain and 7 × 7 daisy chain embedded WLCSP modules

Module	Drop[a]	TMCL[b]
2 × 3 daisy chain	6/108 at 1,000	0/48 at 1,000
7 × 7 daisy chain	17/60 at 1,000	0/60 at 1,000

[a]1,500 G, 0.5 ms, test stopped at 1,000 drops
[b]−45 to 125 °C, test stopped at 1,000 cycles

Failure analysis (FA) was performed on selected failed units in drop test. FA was also performed on randomly picked TMCL units, since there were no recorded failures when test stopped at 1,000 cycles for either module.

On 2 × 3 daisy chain modules, FA found two failure modes on a unit that initially failed at 559 drops and continued to 1,000 drops. The first failure mode was PCB copper trace crack at two corner solder joint locations (Fig. 5.32a–c). This is not a surprise given the fact that the PCB layout has trace parallel to the major bending direction of the drop board. The second failure mode was a typical solder joint crack on the component (module) side (Fig. **5.33**a, b). Based on the previous experience and knowledge in these two failure modes, we suspect the PCB trace crack occurred first and triggered the initial failure event at 559 drops. The solder joint crack happened in the continuation drops from 559 to 1,000 drops. Better drop performance should be expected with change of orientation of the fan-out copper traces that are known to help reduce the odds of PCB trace crack [10, 11]. However, no more effort was put in simply because the 2 × 3 module board level performance is already much higher than the customer requirements.

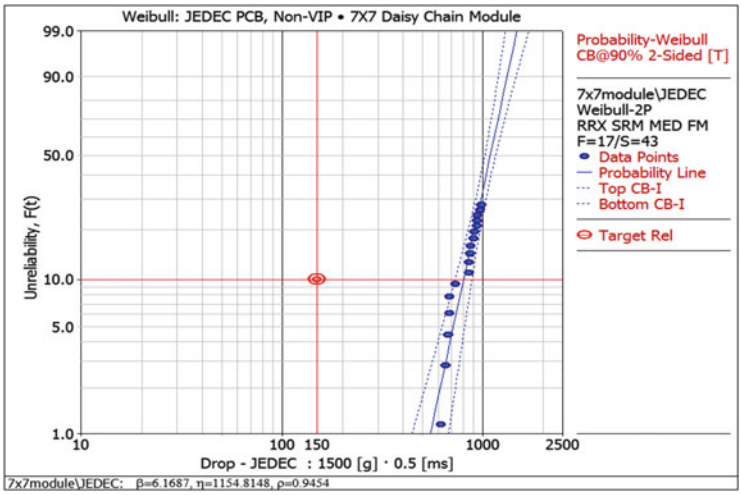


Fig. 5.31 Weibull plot of JEDEC drop test results of the 7 × 7 daisy chain module

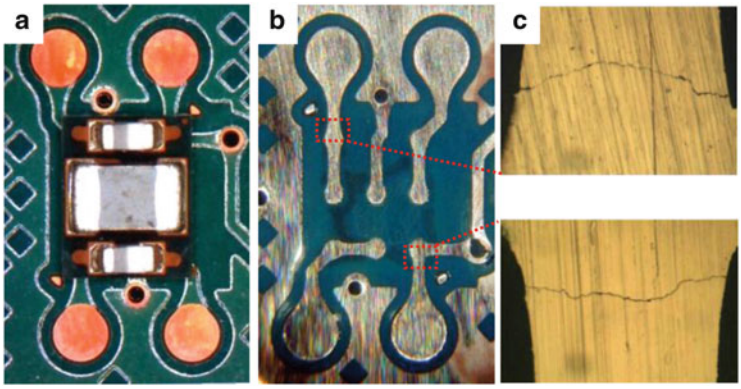


Fig. 5.32 FA on a 2 × 3 module that failed at 559 drops in the drop tests: (**a**) shows the module mounted on the test PCB, (**b**, **c**) reveals the copper trace crack at two corner solder joints

Temperature cycle on the 2 × 3 daisy chain module had zero failure at the point TMCL stopped. Cross sections through all module/PCB solder joints found no crack initiations; cross sections through the buried 2 × 3 via chain and passives/module solder joints found no signs of failure as well (Fig. **5.34**a, b). All confirmed robust TMCL performance.

For the 7 × 7 daisy chain module, test PCB was thoughtfully designed to avoid potential trace crack in drop test (Fig. 5.35). As a result, no PCB trace crack was found in the failure analysis. The only failure mode confirmed was the solder joint cracks and was only recorded at >619 drops.

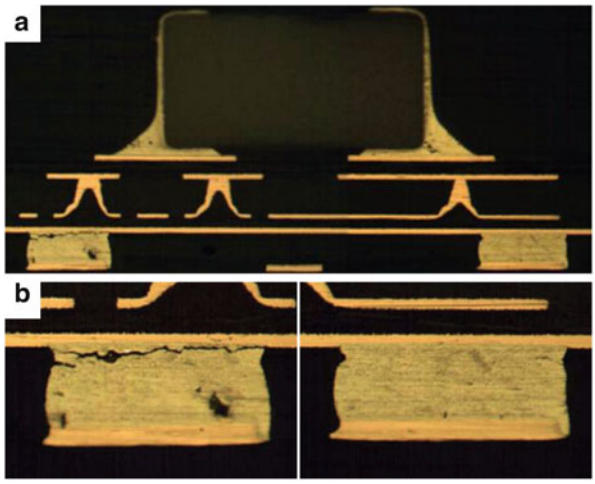


Fig. 5.33 Cross sections of drop failed 2 × 3 daisy chain module: (**a**) overview showing the module/PCB solder joints, buried module via23, and a passive soldered on the module top; (**b**) closeup view of one failed module/PCB solder joint and one passing solder joint on the other corner

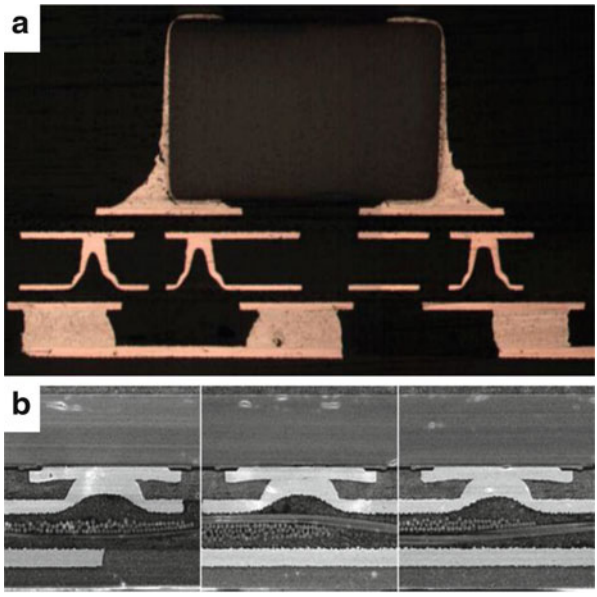


Fig. 5.34 Cross sections of TMCL failed 2 × 3 daisy chain module: (**a**) overview showing the module/PCB solder joints, vias between copper layer 2 and 3, and solder joints between a passive and module; (**b**) closeup view of the blind vias connecting the chip to the module

Figure 5.36 shows solder joint failures on two early drop failed components, as well as crack initiations on adjacent solder joints. Other than the undesired solder joint shape on the first failed module (failed at 619 drops), the crack initiations location match what we typically see on WLCSP.

Temperature cycle from −40 °C to +125 °C on the 7 × 7 daisy chain modules had zero failure when test stopped (1,000 cycles). Cross sections on randomly selected units found no abnormities at the passive/module solder joints and no signs of concerns along the 49 blind vias that connect the embedded chip to the

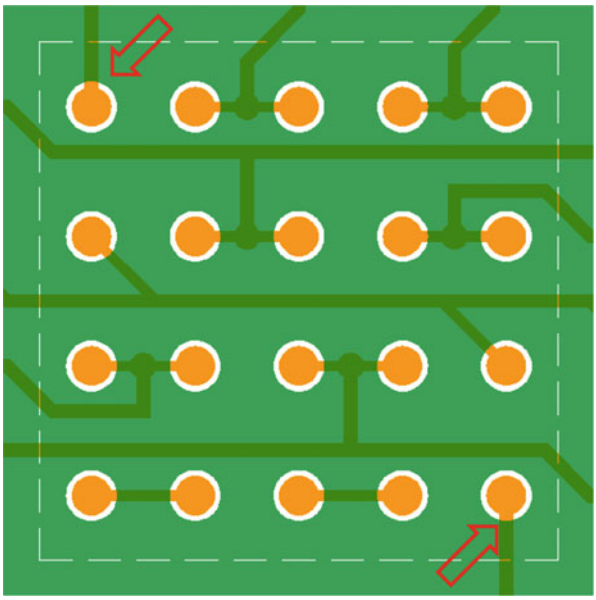


Fig. 5.35 PCB unit layout for the 7×7 daisy chain modules. Traces highlighted by the *arrows* were orientated perpendicular to the major bending directions so as to avoid trace crack in drop test

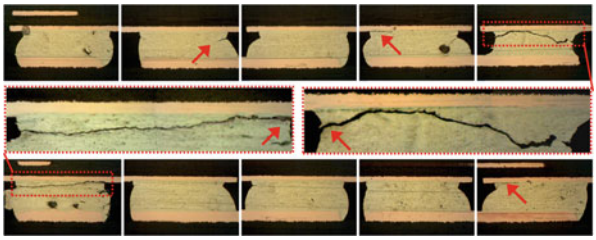


Fig. 5.36 Cross sections of the drop failed 7×7 daisy chain modules. *Top*: component failed at 610 drops on JEDEC drop test PCB; *bottom*: component failed at 731 drops on JEDEC board. *Arrows* highlight the solder joint crack initiations

module (see Figs. 5.37a, c). Early crack initiations, however, were found on two corner solder joints between the module substrate and test PCB (Fig. 5.37). Comparing to the 2×3 module, which showed no signs of module/PCB solder joint crack initiations, it clearly indicates the solder stress level is higher on the 7×7 modules than the 2×3 modules. Another interesting observation is that the crack initiations started from the inner side of the solder joints on the component side. This is, again, quite different from conventional WLCSP, where TMCL crack initiations start from the outer side of the solder joints. One possible explanation of this unusual crack initiation is that the TMCL stress, which is induced by the mismatch of thermal expansion of the component and the PCB, is different from

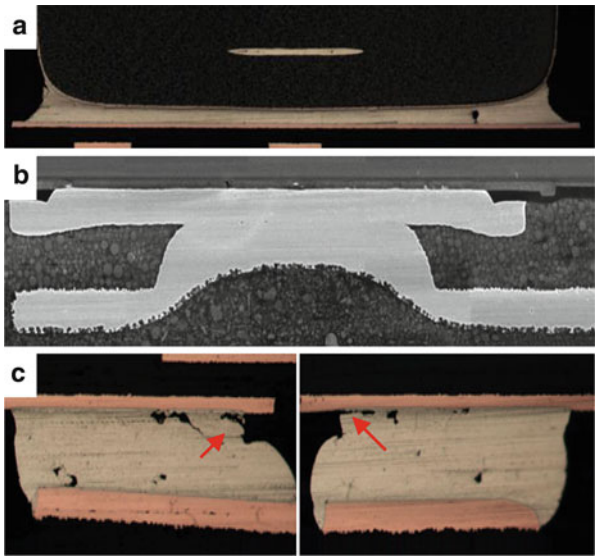


Fig. 5.37 Cross sections of a 7×7 daisy chain module post 1,000 TMCL cycles: (**a**) cross section through a passive showing no solder crack; (**b**) cross section show no damage to the buried vias connecting the module; (**c**) two corner module/PCB solder joints show signs of crack initiation after 1,000 cycles

that of WLCSP. For WLCSP, low CTE of silicon (2–3 ppm/°C) and high CTE of PCB (17 ppm/°C) always initiate solder joint cracks from outer side of the corner solder joints on the component side. For the WLCSP embedded module, we believe the substrate CTE is closely matched to that of PCB. However, passives soldered on the substrate, as well as extra weight from the passives, might contribute to the unusual stress distributions and solder crack initiations. Further work addressing this is underway.

Besides board level reliability tests on the 2×3 and 7×7 daisy chain modules, functional module with a 6 pin, 0.4 mm pitch-embedded WLCSP also went through series reliability tests on 2 mm thick, two copper layer coupon PCB with multiple read points (Table 5.4). Pass/fail criteria was based on visual inspection, electrical test, and CSAM. Based on prior experience, coupon PCB presents higher stress level to SMT devices than typical JEDEC PCB used in the board level reliability assessment. However, there is no single component fail in the coupon board reliability tests. In addition, the module passed separate ESD tests.

5.5.4 Discussion

Modules with surface mounted passives and embedded daisy chain WLCSP of different sizes were built and tested for board level reliability testing; modules with embedded active WLCSP and passives were also built and tested in component level reliability tests. Modules only failed drop test at high drop count (500–600 drops) with typical solder joint crack type of failures. On the other hand, all modules passed 1,000 TMCL cycles with no failures detected. The module/embedded WLCSP via connections, which is one of the main concerning areas, show no

Table 5.4 Reliability test summary of functional module with a 2 × 3, 0.4 mm pitch-embedded WLCSP

Test	Condition	Read points	Results
DOPL	85 °C, 4.2 V	168, 500, 1000	77 × 3 pass
HTSL	150 °C	168, 500, 1000	77 × 3 pass
THBT	85 % RH, 85 °C, 4.2 V	168, 500, 1000	77 × 3 pass
TMCL	−40 ~ +125 °C	100, 500, 1000	77 × 3 pass
HAST	85 % RH, 110 °C	264	77 × 3 pass

signs of failure, either in drop or TMCL. The passive/module solder interconnections also show no signs of issues in all inspected/cross-sectioned units. All indicated robust reliability performance of modules with embedded WLCSP.

The exceptional reliability performance did not come in as a complete surprise, though extra weight from top mounted passives did generate some discussions initially. Two factors likely played important roles here. The first is the thinness of the embedded silicon. From the WLCSP study, we knew that silicon thickness determines the compliance of silicon that is soldered on the test PCB in typical board level reliability tests. Thinner silicon improves the drop and TMCL life of WLCSPs, regardless the device sizes and pitches. However, due to process and handling challenges, the practical limits of silicon thickness of a solder bumped WLCSP is around 200 μm (8 mils). In the presented study, the solder bump-less WLCSP is grinded to 50 μm, which greatly improves the compliance of silicon in drop and TMCL. The other factor is the separation of the silicon from the test PCB. For a WLCSP mounted on the PCB, this is the standoff height, and better reliability life can be expected with greater standoff height. In the studied embedded modules, the silicon is separated about 200 μm from the test PCB through dielectrics, blind vias, and module solder joints. It is effectively decoupled from the high strain locations, which are the solder joints that linked the module to the test PCB. Thus, good board level reliability life is resulted.

5.6 Summary

This chapter introduces the wafer-level discrete power chip-scale package with starting the trends of development. The standard discrete power VDMosfet WLCSP design construction and the LD Mosfet WLCSP are presented. New design concepts and constructions for direct contact VDMosfet WLCSP with both TSVs drain connects and drain cavities are presented as well. The direct contact VDMosfet WLCSP TSV based provides the excellent electrical performance, while the direct contact VDMosfet with cavities shows the thinnest wafer-level discrete chip-scale package. Then a lower cost Cu stud bumping WLCSP design construction and process technology are studied and discussed. Cu stud bumping manufacture process is simulated to study the impact of free air ball diameters and the BPSG profiles on discrete power Mosfet WLCSP. Design for experimental

study of reliability is conducted to confirm the simulation results. In the last section of the chapter, an embedded WLCSP module which integrates the active wafer-level chip-scale package and the power passives is introduced to show a new concept of 3D fan-out system in package technology. This technology adopts the mature PCB manufacturing technology and offers flexible routing for a true three-dimensional system in package solution to satisfy the customers' demands. Robust interconnect reliability at all levels can be expected from the embedding technology. With competitive cost structure to other 3D options, such as traditional fan-out 3D packages, the embedded WLCSP module packages are expected to be found in more applications in system level in the coming years.

References

1. Liu, Y.: Power Electronic Packaging: Design, Assembly Process, Reliability and Modeling. Springer, New York (2012)
2. Harman, G.: Wire Bonding in Microelectronics. McGraw Hill, New York (2010)
3. Liu, Y., Liu, Y., Luk, T., et al.: Investigation of BPSG Profile and FAB Size on Cu Stud Bumping Process by Modeling and Experiment, Eurosime (2008)
4. Liu, Y., Irving, S., Luk, T.: Thermosonic wire bonding process simulation and bond pad over active stress analysis. IEEE Trans. Electron. Packaging Manuf. **31**, 61–71 (2008)
5. Daggubati, M., Wang, Q., Liu, Y.: Dependence of the fracture of power trench Mosfet device on its topography in cu bonding process. IEEE Trans. Compon. Packaging Technol. **32**, 73–78 (2009)
6. Qu, S., Kim, J., Marcus, G., Ring, M.: 3D Power Module with Embedded WLCSP, ECTC63, Las Vegas, NV, June, 2013
7. Manessis, D., Boettcher, L., et al.: Chip Embedding Technology Development Leading to the Emergence of Miniaturized System-in-Packages. 60th ECTC, Las Vegas, NV, June 2010
8. Braun, T., Becker, K.F., et al.: Potential of Large Area Mold Embedded Packages with PCB Based Redistribution. IWLPC, San Jose, CA (2010)
9. Ryu, J., Park, S., Kim D., et al.: A Mobile TV/GPS Module by Embedding a GPS IC in Printed-Circuit-Board. 62nd ECTC, San Diego, CA (2012)
10. Syed, A., et al.: Advanced analysis on board trace reliability of WLCSP under drop impact. Microelectron. Reliab. **50**, 928–936 (2010)
11. Tee, T.Y., et al.: Advanced Analysis of WLCSP Copper Interconnect Reliability under Board Level Drop Test. 58th ECTC, Orlando, FL, June 2008

Wafer-Level Packaging TSV/Stack Die for Integration of Analog and Power Solution

6

The development of the analog and power IC package is a dynamic technology. Both analog and power IC WLCSP applications that were unattainable only a few years ago are today commonplace thanks in part to advances in WLCSP electronic package design. From portable applications such as in mobile telecommunications to consumer electronics, each imposes its own individual demands on the development of WLCSP. To meet such a diverse range of application requirements, WLCSP range encompasses from pure analog application, power application, and system-on-chip (SOC) integration of analog, logic, mixed signal, and power device to system in package which includes individual analog, logic, and power devices, most of which are subdivided into a number of outline versions of wafer-level packaging. The analog and power WLCSP offers a high thermal dissipation enabling analog and power IC usage in some of the most demanding application areas which integrate analog, logic, and power mosfets with through-silicon via (TSV), stack die technology [1, 2]. This chapter will introduce the development of the advanced wafer-level packaging with TSV and stack die concepts for integration of analog and power solutions.

6.1 Design Concept of Integration of Analog and Power Solution

In the early and middle 1980s, the power IC technology was developed from vertical diffusion metal–oxide–semiconductor (VDMOS) discrete technology and the bipolar IC technology to the mixed power technology which integrated bipolar, complementary metal–oxide–semiconductor (COMS) and double-diffused metal–oxide–semiconductor (DMOS) with bipolar-based technology (BCD). At that time, the CMOS technology used 2.5 μm, electrically erasable programmable read-only memory (EEPROM) technology used 1.2 μm and the FLASH technology used 0.6 μm. Since 2000 to now, the mixed power technologies have been evolved with CMOS (implantation)-based technology. There are three major functions:

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_6

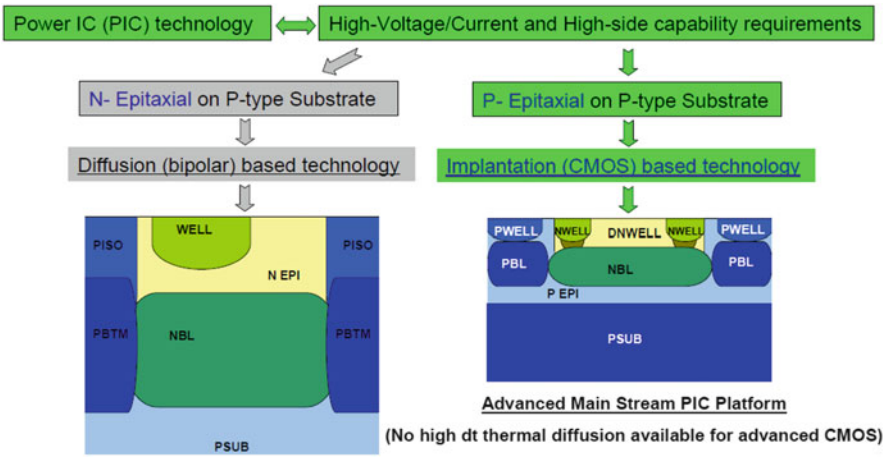


Fig. 6.1 Power IC technology development [4]

(1) mainstream CMOS compatible with 0.35 μm, 0.18 μm, 0.13 μm, and 90 nm power IC below; (2) module approach for integrating more functions; and (3) capability for many small systems integrated on a single chip (SOC). Figure 6.1 shows the evolution of the power IC technology. Figure 6.2 shows the power IC main strain CMOS implantation-based technology and the process module flexibility [3, 4], in which the technology has been designed to be highly module so that mask steps can be added or omitted according to the components required for the ICs design. About 96 power analog components may be integrated into the module. A module route creation system allows more cost-effective process flow for various application needs. The process can be either simplified or differentiated or specialized according to the application requirements. Therefore, it is possible for the modularity power IC technology to generate a large variety of very different power and analog ICs for target applications.

One of the major applications of power IC technology is the portable devices such as the mobile phones, PDAs, digital cameras, MP3 players, and portable bar code readers. Fairchild's product IntelliMAX™ family [5] of integrated load switches supports the latest generations of mobile and consumer electronic devices, which combines conventional MOSFET performance with a unique combination of protection, control, and fault monitoring features to enhance power management design. This level of integration helps designers achieve efficiency and reliability while minimizing board space requirements. Wafer-level chip-scale package is particularly suitable for the portable device due to its small size and the excellent electrical performance. Figure 6.3 shows two typical wafer-level chip-scale packages (WLCSP) for such applications.

Figure 6.4 shows the circuit diagram that integrates the control logic and mosfet with a 6-pin WLCSP for the power IC device in which one pin is not used. This product has three major functions: (1) overcurrent limit protection (OCP),

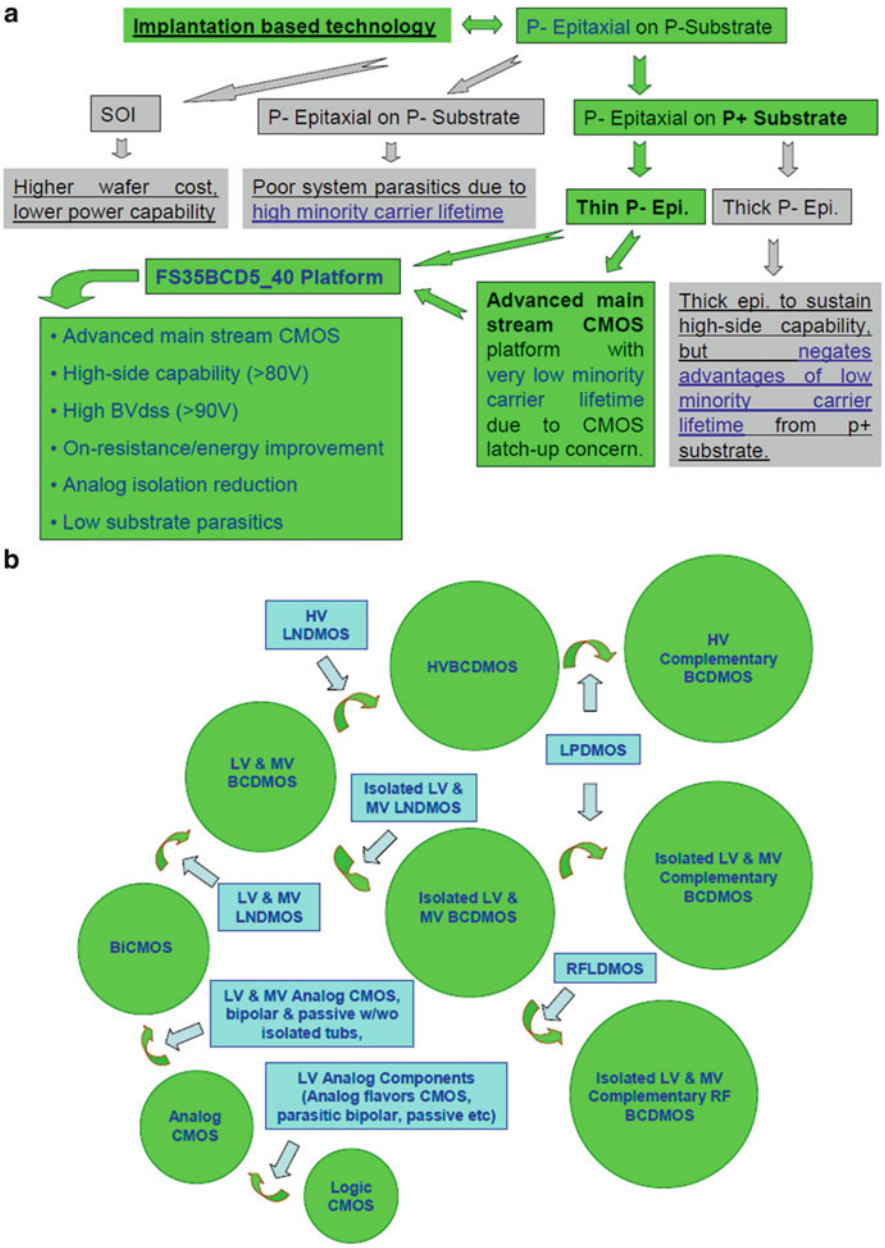


Fig. 6.2 Power IC main strain CMOS implantation-based technology [3, 4]: (**a**) implantation-based technology (**b**) process module flexibility

Fig. 6.3 Wafer-level chip-scale package for integration of Mosfet and protection/control IC

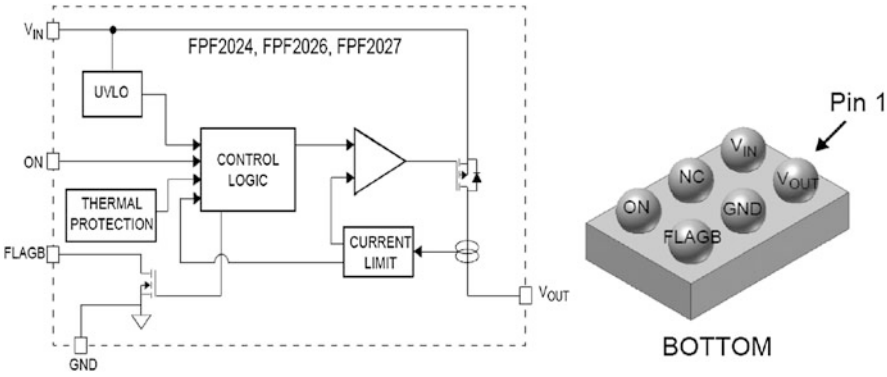


Fig. 6.4 The circuit that integrates the control logic and mosfet with thermal protection and current limit

(2) thermal shutdown protection (TSP), and (3) undervoltage lockout (UVLO). The OCP prevents excessive current and triggers one of the three fault conditions: (1) Automatic restart: the part will automatically shut down and attempt to restart at the defined "auto restart time" interval until the fault is cleared. (2) Shutdown: the part will automatically shut down and require a power cycle on the "ON" pin to clear the fault. (3) Constant current: the part will limit the current to the fixed or user-defined value. The TSP protects the part from damage due to thermal events in which the threshold is140 °C, with 10 °C hysteresis. The UVLO turns the switch off if the input voltage drops below a threshold which ensures stable operation of the device. Figure **6.5** shows another circuit diagram that integrates the control logic and mosfet with a 4-pin WLCSP for the power IC device. This 4-pin power IC WLCSP has two major functions: (1) electrostatic discharge (ESD) protection and (2) turn-on slew rate control which turns the switch on over a defined period of time, which limits the current through the device and into the load. When balanced with the load capacitance, this feature helps to prevent current spikes on the load and minimize voltage sags on the input. This WLCSP also has the output discharge optional function which turns on when the main switch is turned off, offering quick and safe discharge of the load capacitance.

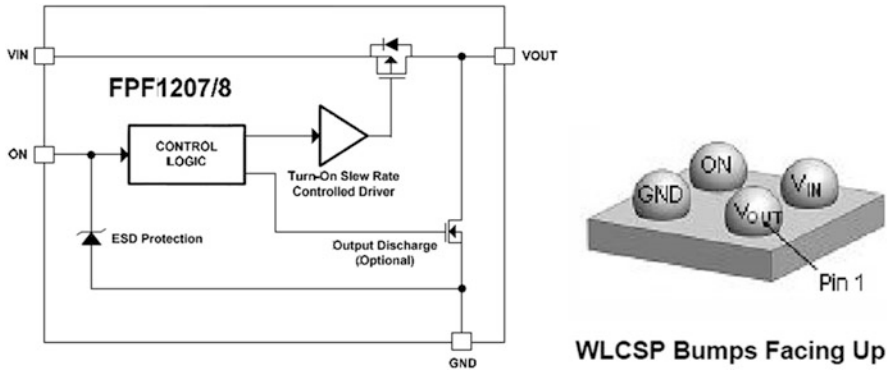


Fig. 6.5 The circuit that integrates the control logic and Mosfet with ESD protection and output discharge functions

Fig. 6.6 The WLCSP for the power IC integration [4]

For the application of the power IC in a multiple function smart module platform with power, analog, and logic integration, different voltage rating isolated tubs can be added to high-value resistors, capacitors, diodes, scalable HV/LV CMOS, bipolar, and matched mirror devices that allow the integration of isolated analog pockets to realize high-voltage level shift functions, precision voltage reference, current sensor accuracy, noise isolation, and eliminated substrate carrier injection. Trimmed components are also available in the PIC platform to make power analog product design highly accurate to provide a competitive advantage. The metal system in the platform can support multiple thin Al/Cu metal layers (4LM) for high-density interconnections plus an extra thick power metal with bond pad overlap active (BPOA) for additional high current routing layer and enhanced energy capability. The thick power metal with BPOA is easily compatible with WLCSP in which the high-density features of this modular power IC platform can offer power analog technology with a great contribution in die size reduction (Fig. 6.6).

Over the last two decades, analog and power semiconductor technology has made impressive progress with increasing the power density of application [1, 2, 4]. For relative larger power application, the technology of wafer-level system on a single chip (SOC) is not enough to meet the high-power density need. Therefore, the wafer-level system in package that utilizes the TSV and stack die technology is becoming necessary for higher-power density application. The primary driving force is from the applications like point-of-load buck converter, in which the wafer-level power system in package (SIP) with TSV and stack die technologies will evolve from single analog function WLCSP [6–8] to the heterogeneous functional integration of analog and power [9, 10]. There are several ways to build the point-of-load buck converter. The typical approach is to design the product with die side-by-side placement in a single-level module using a QFN or other standard package, but this method cannot get the benefit of smaller package size. Another way is to build the wafer-level point-of-load buck converter with TSV and stack die concept [9, 10], which may significantly reduce the parasitic electrical impact and the package size to satisfy the power management for portable application.

6.2 Analog and Power SOC WLCSP

6.2.1 Analog and Power SOC WLCSP Design Layout

Figure 6.7 shows the SOC WLCSP bumping design layout with typical industry standard metal stack for pad Al0.5Cu and under bumping metal (UBM) Au/Ni with Cu and Ti plating under the Ni layer. Between the UBM and the metal pad, there is a layer of polyimide. Figure 6.8 shows the WLCSP design layout with micro bumping in 80 μm height. The design uses the metal stack with Al0.5Cu pad and Ni/Cu UBM with To plating under the Cu layer. The polyimide layer is placed between the UBM and metal pad. The two designs may be used for both the 6-pin WLCSP and 4-pin WLCSP for integration of analog and power IC technology like in Figs. 6.4 and **6.5**.

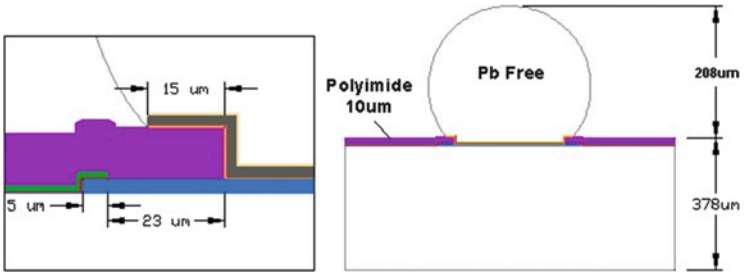


Fig. 6.7 The WLCSP bumping design with typical industry standard

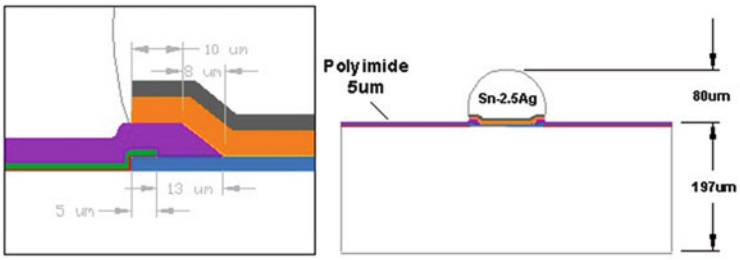


Fig. 6.8 The WLCSP bumping design with an 80 μm height

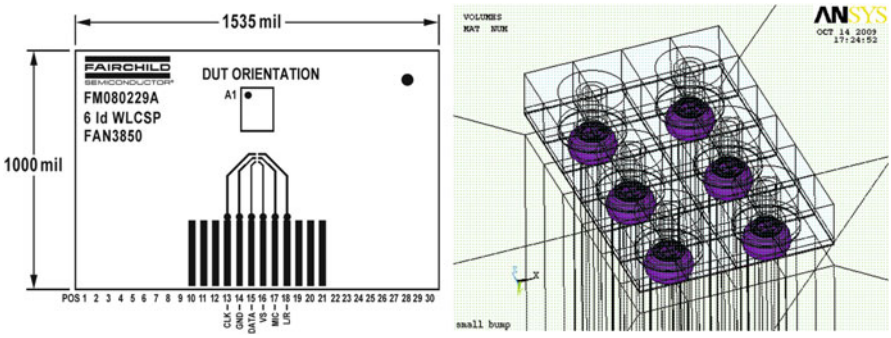


Fig. 6.9 Thermal cycling board and the 6-pin WLCSP model

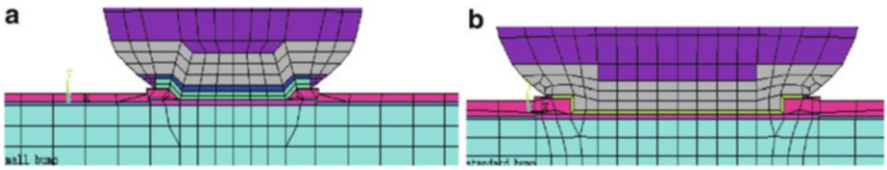


Fig. 6.10 The two models for the micro bumping and standard bumping configurations: (**a**) micro bumping model and (**b**) standard bumping model

6.2.2 Solder Joint Stress and Reliability Analysis

From the design view point, solder joint plays a key role in the product reliability. This section compares the stress and the solder joint life for the micro bumping and the standard bumping from the design concepts with 6-pin WLCSP in the thermal cycling (TMCL). The TMCL range in a cycle is −40 °C ~ 125 °C per 30 min. Figure 6.9 gives the test board layout and the model of the 6-pin WLCSP. Figure 6.10 shows the two models for the configurations of micro bumping and the standard bumping; both bumpings use the Pb-free SAC405 solder materials.

Figure 6.11 shows the von Mises stress comparison of the micro bumping and the standard bumping at −40 °C in the TMCL. The max stress appears at the corner

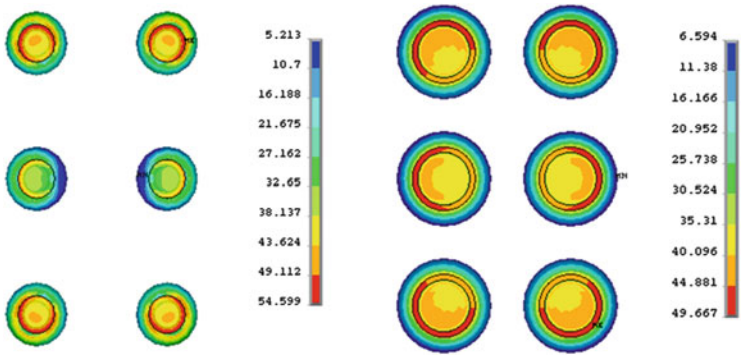


Fig. 6.11 The von Mises stresses of micro bump and standard bump at −40 °C in TMCL: (**a**) micro bump (max 54.6 MPa) and (**b**) standard bump (max 49.7 MPa)

Table 6.1 The stress comparison of the micro bumping and standard bumping in a 6-pin WLCSP

	PAD (Al–0.5Cu)		PI		Solder (SAC405)	
Item	Micro bump	Standard bump	Micro bump	Standard bump	Micro bump	Standard bump
Von Mises stress (MPa)	199.4	199.5	103.6	101.2	54.6	49.7
Von Mises plastic strain	0.77 %	0.45 %	–	–	2.1 %	1 %
Max shear stress (MPa)	100.3	74.1	34.1	33.8	30	28.3
Max shear plastic strain	1.1 %	0.29 %	–	–	3.4 %	1.8 %
Plastic energy density (MPa)	2.6	1.1	–	–	3.6	1.4

joints of die side, and the von Mises stress of micro bumping is greater than the standard bumping about 10 %. Table 6.1 lists the maximum stress comparison of the metal pad Al0.5Cu, polyimide PI, and the solder bumpings. In the metal pad, the maximum von Mises stresses of both the micro bump and standard bump are equivalent, but the maximum von Mises plastic strain, plastic energy density, maximum shear stress, and strain of metal pad in the standard bump are significantly smallerss than the metal pad in the micro bump case. All the stresses and strains and the plastic energy density in standard solder bump are small than the micro bumping. The von Mises stress and shear stress in polyimide in standard bump are slightly less than the micro bumping.

Table 6.2 shows the solder joint life comparison of the micro bump and the standard bump. It can be seen from Table 6.2 that the standard bump has much longer solder joint life than the micro bump in the first failure (52.6 %) and the character life (68.2 %). This is due to the plastic energy density in micro bump that is much greater than the standard bump. Therefore, from the design viewpoint

Table 6.2 The solder joint life comparison of micro bump and the standard bump

TMCL life	Micro bump	Standard bump
First failure	612	1,924 (52.6 % longer)
Character cycle	995	3,129 (68.2 % longer)

based on the solder joint reliability in thermal cycling analysis, the standard bumping WLCSP is better than the micro bumping. However, as the die shrinks, the micro bumping technology for the power IC technology is approaching the requirement for next-generation product; more studies and investigations for the power IC with micro bumping to get the robust design will definitely come out soon.

6.3 Wafer-Level Power Stack Die 3D Package with TSV

This section relates to the wafer-level power stack die concept with through-silicon via (TSV) and more particularly to wafer-level stack die synchronous buck converters.

Synchronous buck converters, primarily used in step-down power supply circuits, typically include two switching field effect transistors (FETs) and a series inductor to permit digital, rather than analog, control of the FETs which either supply current into the inductor or draw current back from the inductor, as shown in Fig. 6.12. Compared to analog power supplies, the synchronous buck converter with FET switching transistors is small and uses very little overhead current. Thus, they are often used for mobile electronic devices. Since space is an important consideration in such devices, the size of synchronous buck converters is important in the marketplace. The wafer-level stack die 3D package with TSV technology is an effective approach to realize the smaller package size and at the same time to keep the smaller thickness [4].

Due to large mismatch in coefficients of thermal expansion between the copper via and the silicon with TSVs, significant thermal stresses will be induced at the interfaces of copper/dielectric layer (usually SiO_2) and dielectric layer/silicon when TSV structure is subjected to subsequent assembly temperature loadings, which would influence the reliability and the electrical performance of interconnects. Thus, section discusses the design concept of the wafer-level power stack die package with TSV, thermal loading, and the impact of thermal–mechanical stress on the design variable of the package during the assembly process.

6.3.1 The Design Concept of the Wafer-Level Power Stack Die Package

The design concept comprises, in one form thereof, a wafer-level buck converter with stack die 3D package including a high side (HS) Mosfet die having source,

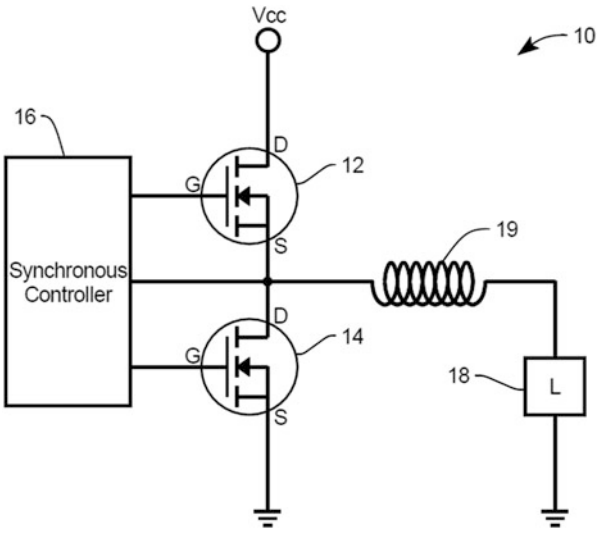


Fig. 6.12 A circuitry of synchronous buck converter

drain, and gate bonding pads on a front side of the HS die; a low side (LS) Mosfet die with a plurality of through-silicon vias (TSVs) extending from a backside to a front side of the LS die; the LS die having source, drain, and gate bonding pads located on the front side; and the drain bonding pad electrically connected to the backside of the LS die. The HS die and the LS die are bonded together such that the source bonding pad with Cu stud bumps of the HS die is electrically connected to the drain of the backside LS die through anisotropic conductive film (ACF), and each of the drain and gate bonding pads of HS die is electrically connected to separate TSVs in the LS die. Figure 6.13a shows such concept in a quarter model, in which the TSV is filled with conductive polymer, and Fig. 6.13b gives the copper TSV structure layout. Figure 6.14 shows the cross section of the wafer-level buck converters with high side wafer (wafer 1) stacked on low side wafer (wafer 2) before the singulation.

6.3.2 Thermal Analysis

The thermal analysis for the wafer-level power stack die package through simulation is introduced in this section. The package is mounted on a JEDEC 1s0p board with 76.2 × 114.3 × 1.6 mm^3 and with PCB vias. Assume natural convection is applied in the simulation. Figure 6.15 gives the quarter model of the stack die package and the system with PCB. The package size is 1.5 × 1.5 × 0.12 mm^2. Figure 6.16a gives the temperature distribution of the system including both stacked die and PCB with an input power 0.1 W on high side Mosfet die. Figure 6.16b shows the temperature distribution of the stacked die package. Table 6.3 lists the thermal resistance $R\theta_{JA}$ junction to the ambient varies with the different die sizes, in which the table includes the coupling thermal effects. $R\theta_{JA11}$ is the thermal

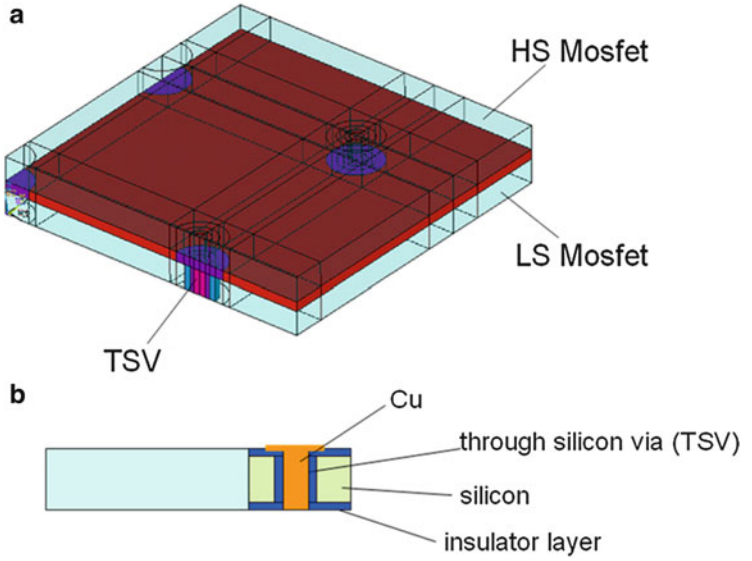


Fig. 6.13 The concept of wafer-level buck converter with stack die 3D TSV technology: (**a**) the concept of wafer-level buck converter in a quarter model and (**b**) TSV at low side die

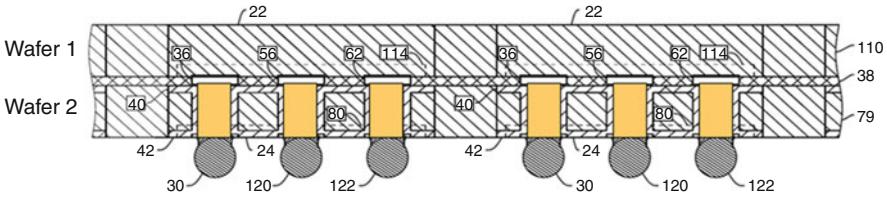


Fig. 6.14 The cross section of the wafer-level buck converter with two stacked wafers before singulation (wafer 1 for high side and wafer 2 for low side)

resistance of high side die due to the power applied on the high side die; $R\theta_{JA12}$ is the thermal resistance of low side die due to the power applied on the high side die; $R\theta_{JA21}$ is the thermal resistance of high side die due to the power applied on the low side die; and $R\theta_{JA22}$ is the thermal resistance of low side die due to the power applied on the low side die. From Table 6.3, it can be seen that the thermal resistance of the stack die power package decreases slightly as the die size increases. Table 6.4 lists the thermal resistance of the wafer-level power package under the impact of different diameter of TSV. The result does not show significantly changes. From both Tables 6.3 and 6.4, it can be seen that the wafer-level stack die power package has excellent thermal performance with low thermal resistance.

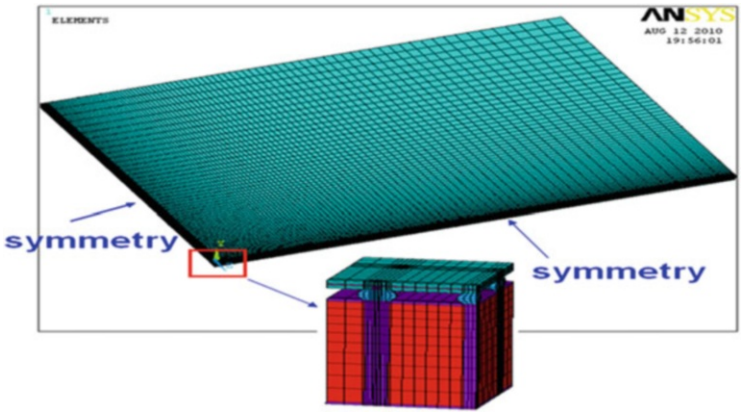


Fig. 6.15 The quarter model of the wafer-level buck converter on JEDEC 1s0p PCB with vias

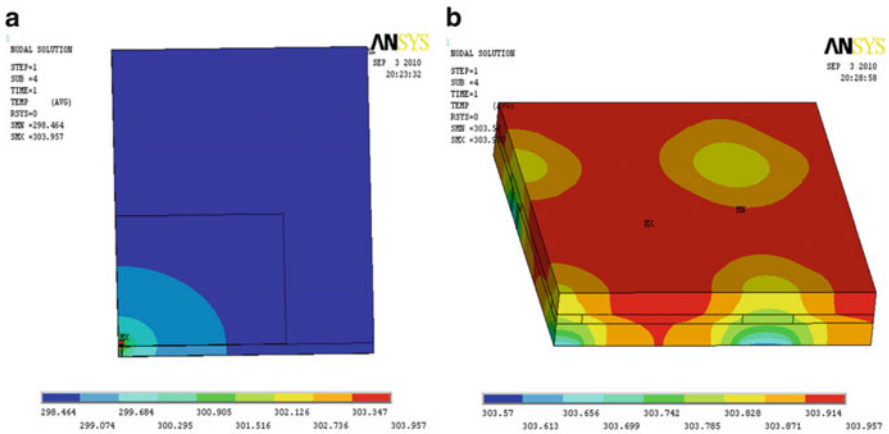


Fig. 6.16 The temperature distribution when the power (0.1 W) applied to the high side die (max temperature, 304 K): (**a**) the temperature of both PCB and die. (**b**) The temperature distribution on the stack die package

Table 6.3 Thermal resistance of junction to ambient versus the die size under natural convection and the input power is 0.1 W

Die size (mm^2)	$R\theta_{JA11}$ (°C/W)	$R\theta_{JA12}$ (°C/W)	$R\theta_{JA21}$ (°C/W)	$R\theta_{JA22}$ (°C/W)
1.2 × 1.2	60.85	60.5	60.28	60.57
1.3 × 1.3	60.29	59.99	59.79	60.05
1.4 × 1.4	59.89	59.64	59.46	59.69
1.5 × 1.5	59.57	59.35	59.2	59.4

Table 6.4 Thermal resistance of junction to ambient versus the TSV diameter under natural convection and the input power is 0.1 W

TSV diameter (μm)	$R\theta_{JA11}$ (°C/W)	$R\theta_{JA12}$ (°C/W)	$R\theta_{JA21}$ (°C/W)	$R\theta_{JA22}$ (°C/W)
20	59.49	59.26	59.11	59.31
40	59.56	59.34	59.19	59.39
50	59.57	59.35	59.2	59.40
60	59.58	59.36	59.21	59.41

6.3.3 Stress Analysis in Assembly Process

There are two major assembly processes in the stack die power package. One is the cooling stress after thermally compressing and stacking high side with metal (copper or gold) stud bumping on the lower side through ACF. The other is the reflow process to mount the wafer-level stack die power package on the PCB. This section gives the stress analysis in the two assembly processes with key design parameter variables.

6.3.3.1 Residual Stress After Stacking High Side Die on Low Side Die

Stacking the high side die to low side die is completed through thermally compressing and curing of ACF. Assume that the curing temperature of ACF is at 175 °C, in which the stress is free. After stacking high side die on low side die, the system will cool down from 175 degree to room temperature as shown in Fig. 6.17.

Finite element method has been used for insight into the state of stresses and reliability of TSV in Fig. 6.17. Table 6.5 gives the material mechanical properties. It can be seen that there is larger CTE mismatch between SiO_2 and the copper of the TSV. The conventional dielectric layer SiO_2 will induce larger stress mismatch at the interfaces of copper/SiO_2 as well as the interface of silicon/SiO_2. As a solution, a modified through-silicon via was proposed. The thin SiO_2 dielectric layer is replaced by a thick polymer isolation layer Parylene. In this section, the material Parylene is introduced as the isolation layer of TSV in the stack die power package. Conformal copper plating is used to realize the connection and the remaining hole in the copper via is filled with epoxy polymer material.

Figure 6.18 shows the tensile stress (the first principal stress S1) and the compressive stress (the third principal stress S3) of the wafer-level power packaging after the high side Mosfet die thermally stacking on the low side Mosfet die. The maximum tensile stress appears at the interface between copper TSV and its isolation. The maximum compressive stress appears at the copper stud bumping which is right above the TSV copper. Figure 6.19 shows the compressive stresses (the third principal stress S3) for the high side die and low side die. Both maximum stresses happen at the TSV areas. Those compressive stresses are much lower than silicon compressive strength. Figure 6.20 shows the stress distributions for copper stud bumpings, TSV copper, TSV isolation layer, and the ACF layer. It can be seen from Fig. 6.20 that the copper stud bumping withstands the highest von Mises stress at the TSV location after the HS die stacking on the LS die. In the TSV copper, there

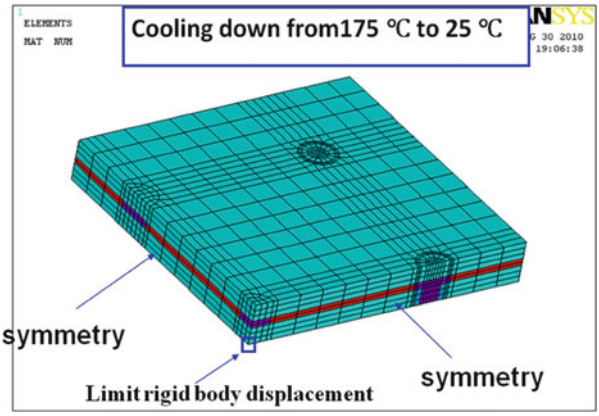


Fig. 6.17 The quarter model of the stack die power package

Table 6.5 Material properties

Material	Copper	Silicon	SiO_2	Parylene	Epoxy	ACF
Young's modulus (GPa)	127.7	131.0	60.1	3.2	3.0	3.56 at 223 K 2.76 at 298 K 1.52 at 423 K 1.44 at 523 K
Poisson ratio	0.34	0.28	0.16	0.4	0.4	0.35
CTE (ppm/K)	17.1	2.8	0.6	35	65	74 at 223 K 75 at 268 K 100 at 278 K 109 at 283 K 119 at 288 K 143 at 298 K 144 at 473 K

also appears the larger stress at the junction with copper stud bumping. The maximum tensile stress of the TSV isolation layer appears at the interface with TSV copper near the junction with copper stud bumping. The maximum von Mises stress of ACF layer appears at the corner TSV location at the interface with the copper stud. Therefore, the maximum stress of the wafer-level stack die power package is related to the TSV design, location, and its materials. Figure 6.21 has shown that the compressive stress S3 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, ACF layer, and the tensile stress S1 of the isolation layer are as varied as the design parameter of TSV diameter after stacking the high side die on the low side die. The TSV diameter has significant impact on stress of the HS and LS die, copper stud bumping, and TSV copper. As the TSV diameter increases, it can significantly reduce the von Mises stress inside TSV copper, while there are no significant changes in ACF layer and the TSV isolation layer. However, increasing the TSV diameter may induce the higher von Mises stress in copper stud bumping and compressive stress S3 in HS and LS die.

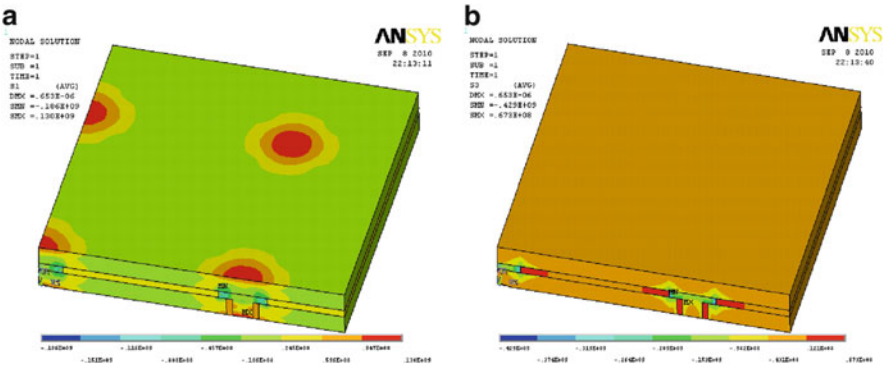


Fig. 6.18 The principal stress of the stack die power package at 25 C after the stacking process: (**a**) the first principal stress S1 (tensile) (max 130 MPa) and (**b**) the third principal stress S3 (compressive) (max 429 MPa)

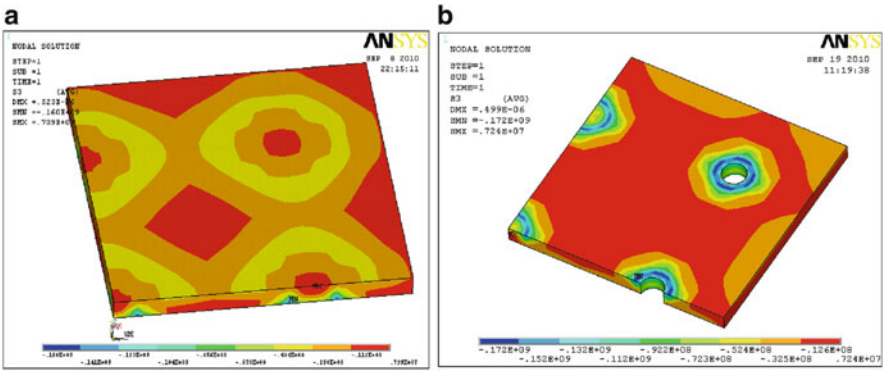


Fig. 6.19 The compressive stress at HS and LS die after the stacking process: (**a**) compressive stress S3 on HS die (max 160 MPa) and (**b**) compressive stress S3 on LS die (max 172 MPa)

Figure 6.22 shows the compressive stress S3 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, the tensile stress S1 of the isolation, and the von Mises stress of ACF layer versus the thickness of ACF layer after stacking the high side Mosfet die on the low side Mosfet die. As the ACF layer thickness increases, all the stresses increase, which include the von Mises stress of ACF layer. Especially, it can significantly increase the third principal stresses in HS and LS die, as well as the von Mises stress of copper stud and TSV copper. This indicates that for the thick ACF with copper stud bumping, it will induce greater stress in the stacking process.

Figure 6.23 shows the compressive stress S3 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, the tensile stress S1 of the isolation, and the von Mises stress of ACF layer versus the thickness of HS die. As the HS die thickness increases, the compressive stress S3 of HS die decreases and

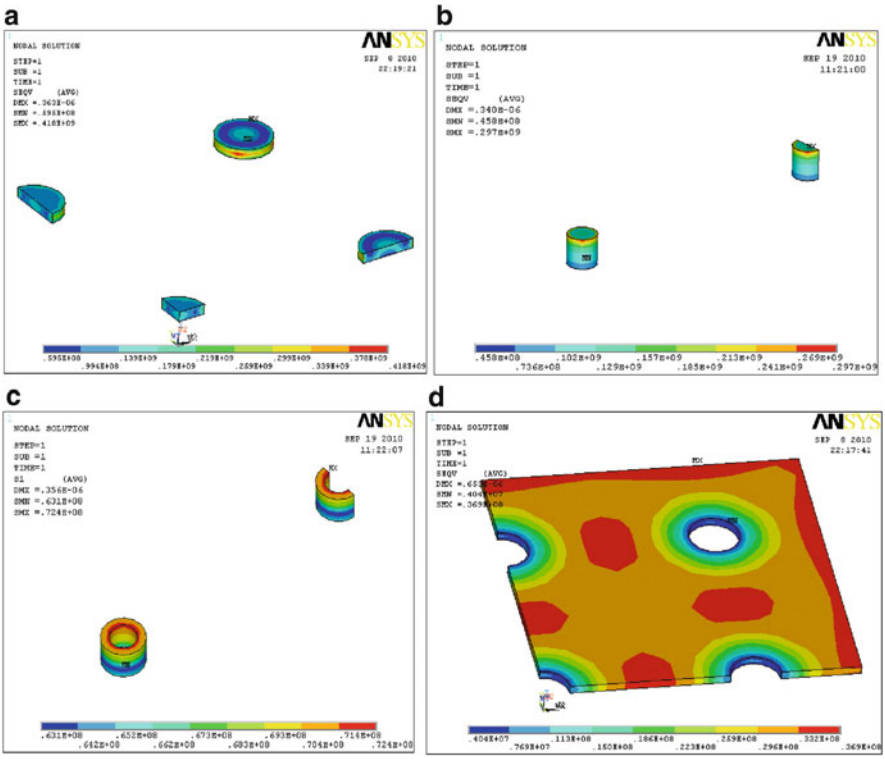


Fig. 6.20 The stress distribution of the copper stud bump, TSV/isolation, and the ACF layer after the stacking die process: (**a**) von Mises stress of the copper stud bump (max 418 MPa), (**b**) von Mises stress of the copper TSV (max 297 MPa), (**c**) the tensile stress of the isolation layer (max 72.4 MPa), (**d**) von Mises stress of the ACF layer (max 36.9 MPa)

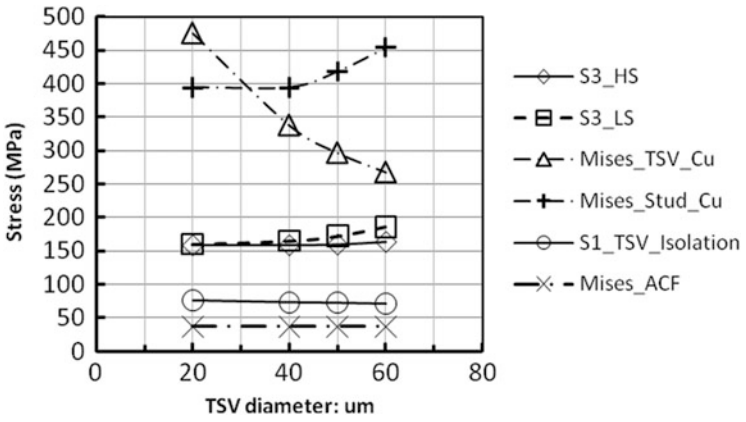


Fig. 6.21 The stresses of HS, LS, Cu stud, TSV Cu, ACF, and isolation layer versus TSV diameter

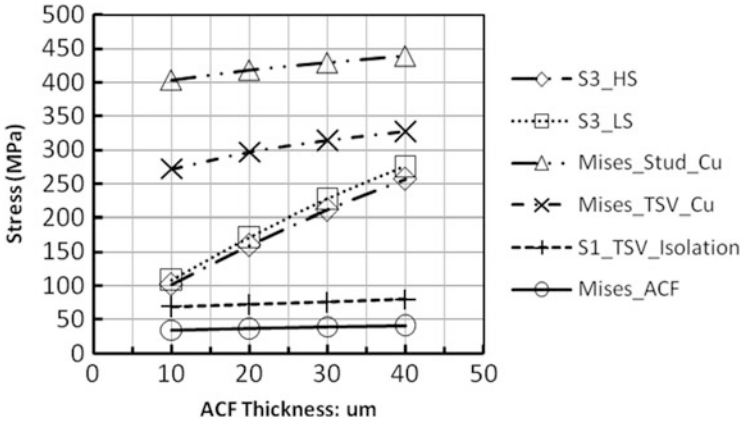


Fig. 6.22 The stresses of HS, LS, Cu stud, TSV Cu, isolation layer, and ACF layer versus ACF thickness

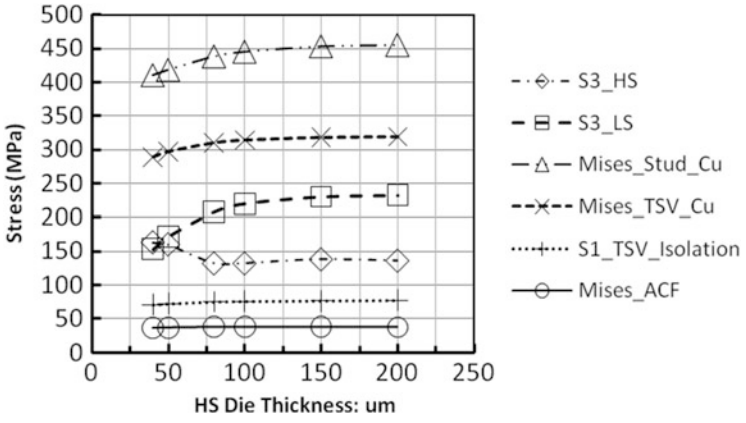


Fig. 6.23 The stresses of HS, LS, Cu stud, TSV Cu, isolation layer, and ACF layer versus HS die thickness

becomes stable after the HS die thickness exceeds 80 μm. All the rest stresses in LS die and TSV increase. The stresses in TSV isolation and ACF layer do not show significant changes, while the stress in TSV isolation layer seems to have slight increment. Figure 6.24 shows the compressive stress S3 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, the tensile stress S1 of the isolation, and the von Mises stress of ACF layer versus the thickness of LS die. As the LS die thickness increases, the compressive stress S3 of both HS and LS die and von Mises stress of copper stud increase and become stable after the LS die thickness is larger than 150 μm. The von Mises stress of TSV copper sharply increases at the beginning and then immediately decreases as the LS die thickness increases. The tensile stress S1 of the TSV isolation decreases slightly and becomes

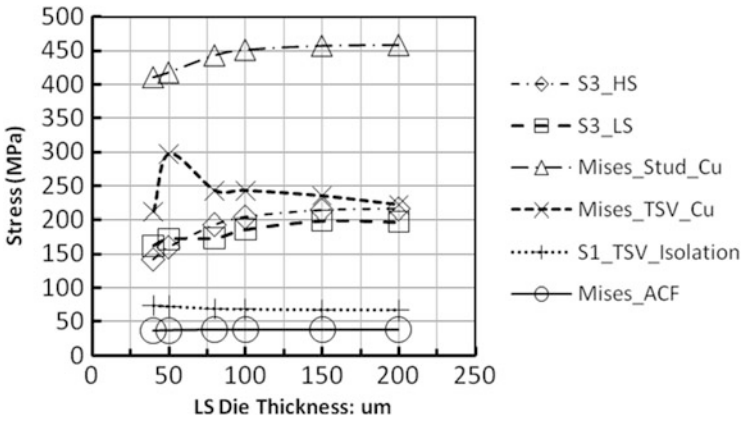


Fig. 6.24 The stresses of HS, LS, Cu stud, TSV Cu, isolation layer, and ACF layer versus LS die thickness

stable after the LS die thickness exceeds 100 μm. The von Mises stress of ACF layer has almost no change.

6.3.3.2 Reflow Stress Analysis

Reflow process is used for mounting the wafer-level stack die power package on a PCB. At this process, the solder joints in the wafer-level stacking die power package are connected to the PCB through 260 °C, which will induce high stresses inside the package as well as the solder joints. At the same time, the changes of material properties due to such high temperature are also the challenges in the reflow process. Figure 6.25 shows a quarter model of the stack die power package. Tables 6.6 and 6.7 list the solder SAC385 and PCB material properties. In Table 6.7, the viscoplastic behavior of the solder at high temperature is described by ANAND material model. All the properties of other materials are listed in Table 6.5.

In reflow process, most of the systems are subjected to tensile stress. Therefore, the tensile stresses of the HS die and LS die are checked. Figure 6.26 shows the first principal stresses (tensile) of high side die and low side die. The max stress appears at the TSV area. The tensile stress of LS die is larger than the HS die. Both tensile stresses of HS and LS die are within the range of the tensile strength of silicon.

Figure 6.27 shows the von Mises stress of the copper stud bumping and the ACF layer at reflow. The max stress (939 MPa) of copper stud bumping appears at the corner bump of which the TSV is underneath. The copper stud bumping stress is the highest inside the wafer-level stacking die power package, while the von Mises stress of the ACF layer is the very low at reflow. Figure 6.28 shows the von Mises stress of the copper TSV and the compressive stress S3 of isolation layer at reflow. The max stress of copper TSV appears at the corner of the die and at the interface that connects to the solder bumping, while the maximum compressive stress (S3) of the isolation layer appears at the interface of the copper stud bumping. Figure 6.29

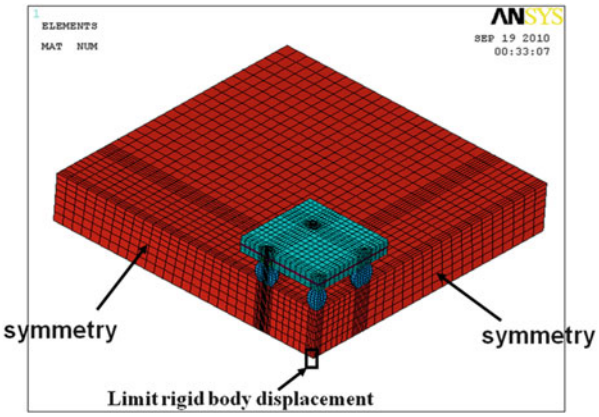


Fig. 6.25 The quarter model of the wafer-level power stack die package mounted on a PCB

Table 6.6 The material properties of solder and PCB

Material	Solder ball	PCB
Young's modulus (GPa)	75.8 − 0.152*T	EX: 25.4 EY: 11 EZ:25.4
Poisson ratio	0.35	XY: 0.39 YZ: 0.39 XZ: 0.11
CTE (ppm/K)	24.5	XZ: 16 Y: 84

Table 6.7 The viscoplastic property of solder at reflow

Description	Symbol	Units	Sn–Ag–Cu385
Initial value of s	s_0	MPa	16.31
Activation energy	Q/R	K	13,982
Pre-exponential factor	A	1/s	49,601
Stress multiplier	ξ	–	13
Strain rate sensitivity of stress	m	–	0.36
Hardening coefficient	h_0	MPa	8.0E5
Coefficient for deformation resistance saturation value	$\hat{s}$	MPa	34.71
Strain rate sensitivity of saturation value	n	–	0.02
Strain rate sensitivity of hardening coefficient	a	–	2.18

shows that the von Mises stress and the plastic energy density of the solder joints at the reflow. Both the max von Mises stress and max plastic energy density happen at the corner joint. Figure 6.30 lists the stress distribution of TSV filled with epoxy. The design concept of TSV filled with epoxy is to reduce the stress in the TSV and the power stacking die package. Figure 6.30a shows the von Mises stress of the TSV epoxy core; Fig. 6.30b shows the von Mises stress of TSV copper; and Fig. 6.30c shows the compressive stress of the TSV isolation layer. As compared to the full copper TSV case (Fig. 6.28) and the TSV cooper with filled epoxy core, the von

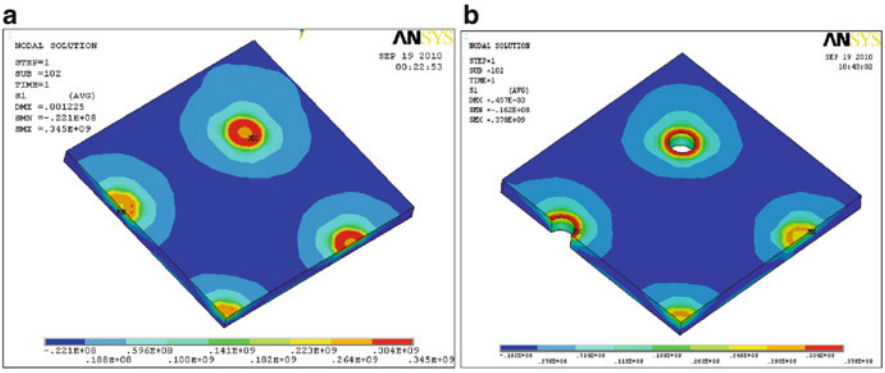


Fig. 6.26 The tensile stress S1 of HS and LS die at reflow: (**a**) the tensile stress S1 of the HS die at reflow (max 345 Mpa) and (**b**) the tensile stress S1 of the LS die at reflow (max 378 MPa)

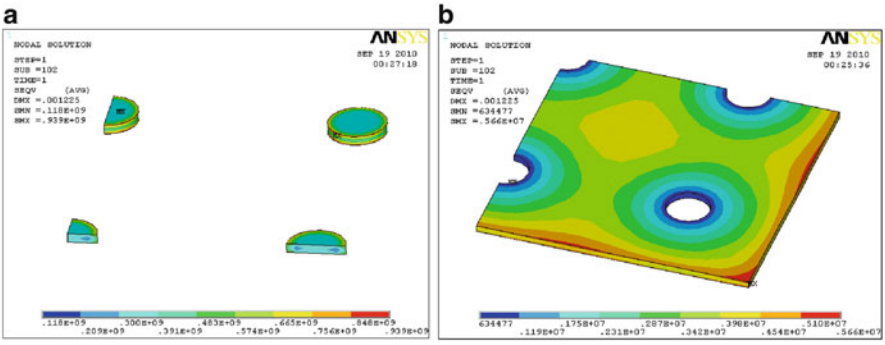


Fig. 6.27 The von Mises stress of Cu stud and the ACF layer at reflow: (**a**) the von Mises stress of Cu stud (max 939 MPa) and (**b**) the von Mises stress of ACF layer (max 5.66 MPa)

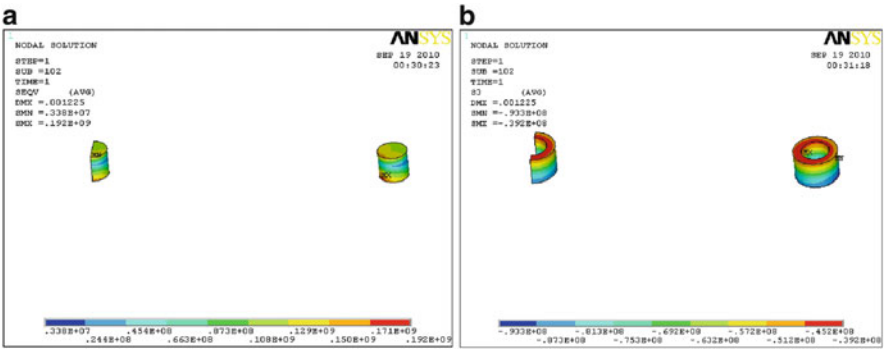


Fig. 6.28 The reflow stresses of TSV in copper and isolation layer: (**a**) the von Mises stress of TSV Cu (max 192 MPa) and (**b**) the compressive stress S3 of isolation layer (max 93.3 MPa)

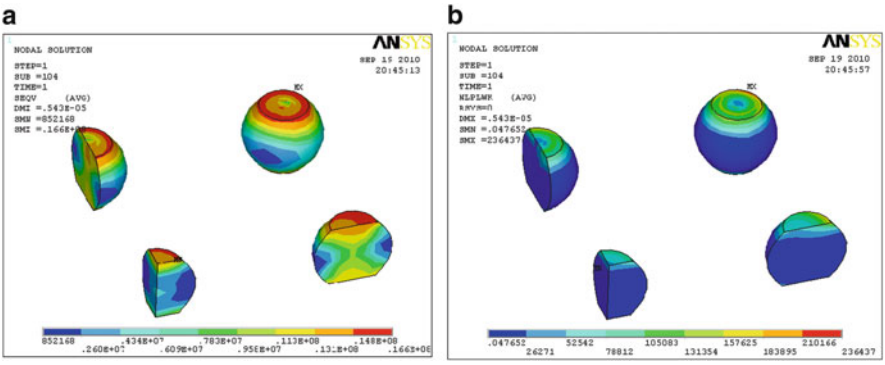


Fig. 6.29 The von Mises stress and plastic energy of solder balls at reflow: (**a**) the von Mises stress of solder bump (max 16.6 MPa) and (**b**) the plastic energy density of solder bump (max 0.236 MPa)

Mises stress of TSV copper with filled epoxy core slightly increases and the compressive stress S3 of the isolation layer slightly decreases.

Figure 6.31 has shown that the tensile stress S1 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, ACF layer, the compressive stress S3 of the isolation layer, and the von Mises stress of solder ball are as varied as the design parameter of TSV diameter at reflow. The TSV diameter has significant impact on von Mises stress of the copper stud bumping and TSV copper and tensile stress S1 of HS and LS die. As the TSV diameter increases, the von Mises stress of the copper stud bumping has significantly increased, while the von Mises stress of TSV copper increases slightly at first; after the TSV diameter is larger than 50 um, it shows the decrement. The compressive stress S3 of the TSV isolation increases as well. However, there are no significant stress changes in ACF layer and the solder ball. Figure 6.32 shows the tensile stress S1 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, ACF layer, the compressive stress S3 of the isolation layer, and the von Mises stress of solder ball versus the ACF thickness at reflow. As the ACF thickness increases, the tensile stresses S1 of both HS and LS die increase significantly (LS die stress increases fast), while the von Mises stress of copper stud bump has significantly reduced. The von Mises stress of TSV copper decreases in a wavelike function. The compressive stress S3 of the TSV isolation layer increases slightly. There are no significant changes in the von Mises stresses of ACF layer and the solder ball.

Figure 6.33 shows the tensile stress S1 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, ACF layer, the compressive stress S3 of the isolation layer, and the von Mises stress of solder ball versus the HS die thickness at reflow. As the HS die thickness increases, the tensile stress S1 of LS die and the von Mises stress of TSV copper increase slightly and become stable after HS die thickness exceeds 100 μm, while the tensile stress S1 of the HS die decreases and becomes stable after the HS die thickness is larger than 100 μm. All the other stresses do not have significant changes. Figure 6.34 shows the tensile

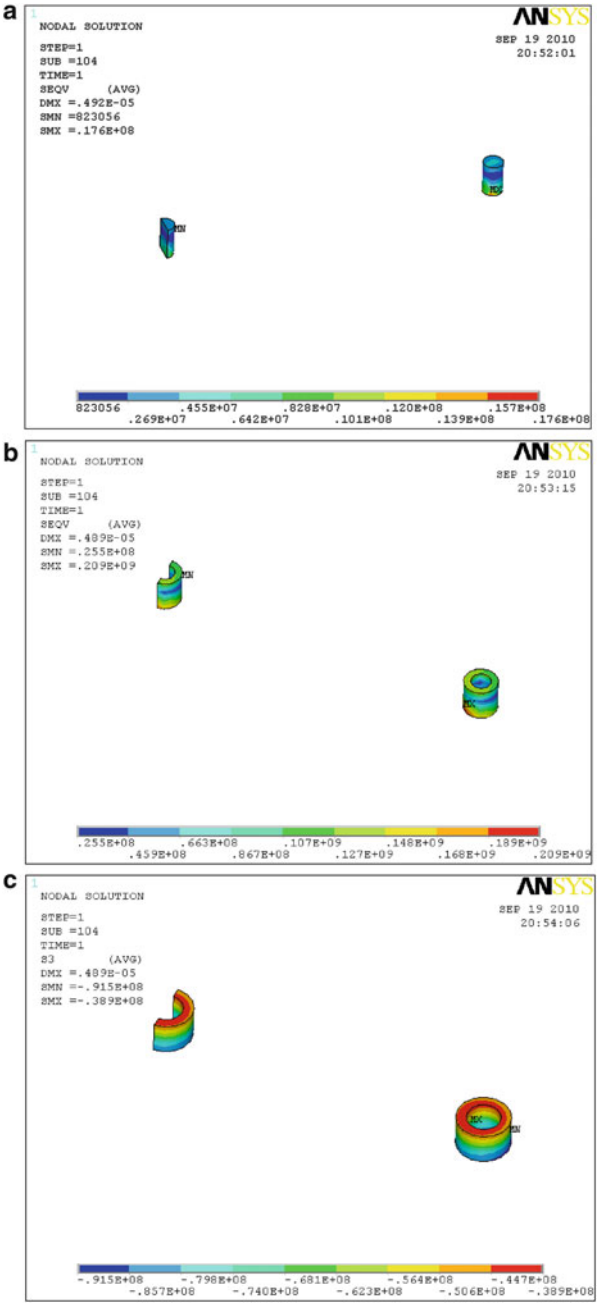


Fig. 6.30 The stress distribution of TSV filled with epoxy at reflow: (**a**) the von Mises stress of epoxy core in TSV (max 17.6 MPa), (**b**) the von Mises stress of TSV copper (max 209 MPa), and (**c**) the compressive stress S3 of isolation layer (max 91.5 MPa)

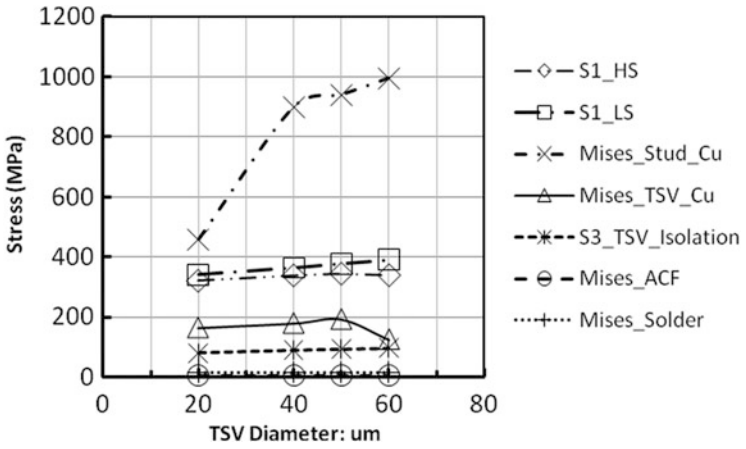


Fig. 6.31 The stresses of HS, LS, Cu stud, TSV Cu, ACF, isolation layer, and solder versus TSV diameter at reflow

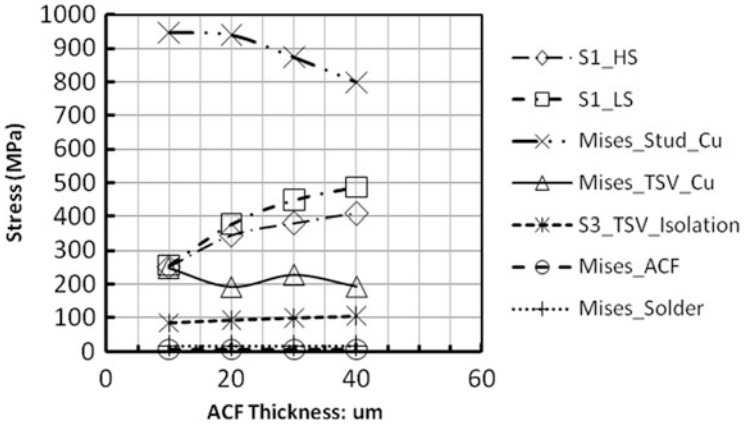


Fig. 6.32 The stresses of HS, LS, Cu stud, TSV Cu, ACF, isolation layer, and solder versus ACF thickness at reflow

stress S1 of high side die and low side die, von Mises stresses of TSV copper, copper stud bumping, ACF layer, the compressive stress S3 of the isolation layer, and the von Mises stress of solder ball versus the LS die thickness at reflow. As the LS die thickness increases, the von Mises stresses of copper stud bump and TSV copper increase. The tensile stresses S1 of HS and LS die have slightly increased, while the rest of the stresses of TSV isolation layer, ACF layer, and solder ball are almost kept the same. Therefore, through the stress analysis at reflow process, we

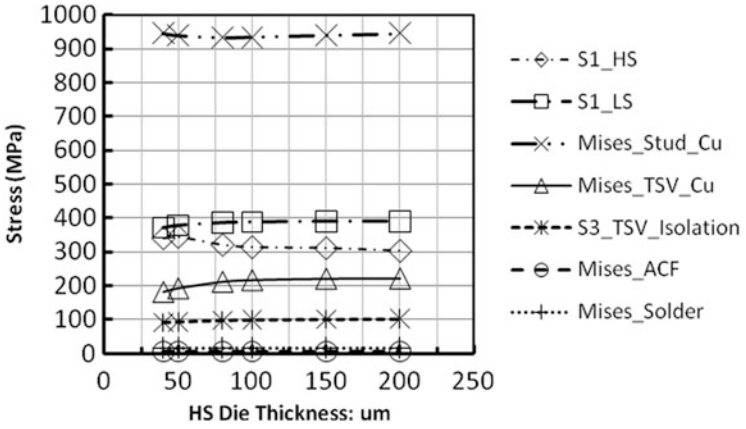


Fig. 6.33 The stresses of HS, LS, Cu stud, TSV Cu, ACF, isolation layer, and solder versus HS die thickness at reflow

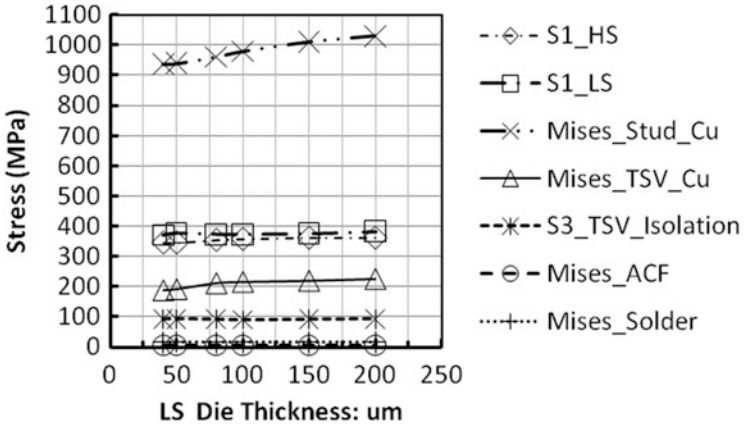


Fig. 6.34 The stresses of HS, LS, Cu stud, TSV Cu, ACF, isolation layer, and solder versus LS die thickness at reflow

understand that the stresses of ACF layer, solder ball, and the TSV isolation layer are not sensitive to the design variable of the wafer-level stack die power package. Table 6.8 lists the comparison of the stresses in copper TSV and in copper TSV filled with epoxy. The result shows there is no significant difference for both cases, although the von Mises stress of copper TSV filled with epoxy has a little higher stress (8 % higher) than the copper TSV.

Table 6.8 The comparison of stresses of TSV copper and TSV filled with epoxy

Compared items	Copper TSV (unit: MPa)	Copper TSV filled with epoxy core (unit: MPa)
Tensile stress S1 of HS	345	344
Tensile stress S1 of LS	378	377
Von Mises stress of ACF	5.66	5.62
Von Mises stress of Cu stud	939	940
Von Mises stress of solder ball	16.5	16.6
Von Mises stress of TSV (copper)	192	209
Compressive stress S3 of the TSV isolation	93.3	91.5

6.4 Wafer-Level TSV/Stack Die Concept for Analog and Power Integration

In Sect. 6.3, we introduce the stacked wafers with TSV technology for a buck converter in Fig. 6.12. In that design concept, there are stacked lower side Mosfet and high side Mosfet. However, the analog IC controller is not included in the structure of Sect. 6.3. In this section, we will introduce a new concept, which builds both the lower side and high side Mosfets in one wafer with TSV technology, then stack the analog IC die on the Mosfet wafer to complete the integration of analog and power device for a point-of-load buck converter. Figure 6.35 [11] shows schematically the cross section of such concept of integration of analog IC and power Mosfet devices, in which the analog IC die is stacked on the wafer with the first Mosfet die and the second Mosfet die.

This design concept introduces an integration package of analog and power device package which includes an analog IC die stacked on two power Mosfets. The power mosfets include a high side Mosfet and a low side Mosfet fabricated in a semiconductor substrate adjacent to one another. The analog IC die is mounted to a backside of the semiconductor substrate and coupled to the two Mosfets with a plurality of TSVs. The wafer-level stack die package is designed to have the land pattern pads on the front active side of the package. The high side source connects to the low side drain through the metal land. There is an isolation gap between the low side and high side mosfets in the semiconductor substrate. The TSVs in the semiconductor substrate connect the high side source and low side drain to the analog IC controller die through redistribution layer (RDL) metals. The analog IC is designed to be the WLCSP or flip chip that could be able to mount on the semiconductor substrate with two Mosfets. The whole stack die package is molded with wafer-level molding technology. The basic procedure of the manufacturing process includes: (1) make the power wafer with low side mosfet and high side mosfet with TSVs, RDLs, and land pattern; (2) etch the isolation grooves between high side and low side mosfet device, and then fill the grooves with isolation

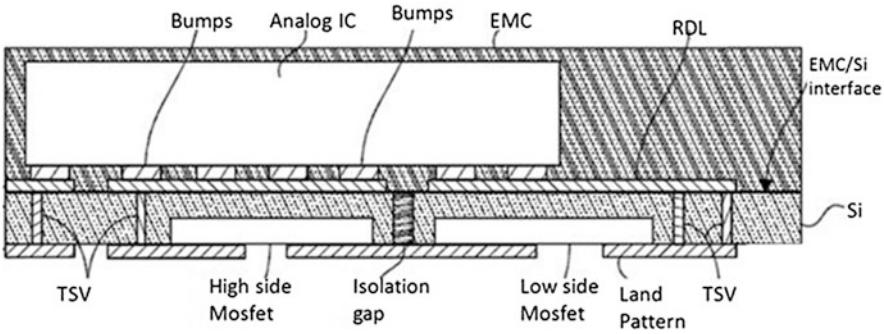


Fig. 6.35 A schematic concept of stack analog IC die with two Mosfet power device wafer

material; (3) stack WLCSP analog IC die on the RDL of the power wafer; (4) overmold the whole wafer with stacked analog IC die; and (5) singulate the package by wafer sawing.

This concept shrinks the total solution to the smallest possible footprint given a particular power semiconductor technology. It can also be very thin. The inductance and resistance of the gate drive loop can be extremely low, enabling nanosecond switching and highly efficient operation at frequencies well into the Mhz range. The power MOSFET junctions are directly coupled to the PCB, so the thermal resistance is minimized. There is no leadframe, ceramic, or organic substrate, so the solution could be very cost-effective and excellent electrical performance. It is useful for smart power stages, buck converters, or any combination of integrated controller/drivers and two or more MOSFETs.

6.5 Integrating Power Packaging with Active and Passive Chips

Section 6.3 introduces the design concept of stacked lower side Mosfet wafer and high side Mosfet wafer for a buck converter. Section 6.4 presents an idea of wafer-level stacked analog IC die on the low side mosfet and high side mosfet wafer, then mold the wafer with stacked IC die with epoxy mold compound (EMC). In the industry, however, the circuitry with both active devices (Mosfets, IC controller) and passive device-like resistor, capacitor, and inductor become more and more popular to reduce the parasitic effects. Figure 6.36 shows such circuitry with two Mosfets, one input capacitor, one output capacitor, and one output inductor.

This design concept is a wafer-level embedded active power IC die and stacked passive die of power system in package, as shown in Fig. 6.37 [12]. The active power IC die is embedded in a glass or silicon substrate with TSVs and RDL layers. The passives (inductor, in capacitor, and out capacitor) are then stacked on the RDLs. Generally, the method of manufacturing process includes making a glass or semiconductor substrate with array through vias and array through cavities. An

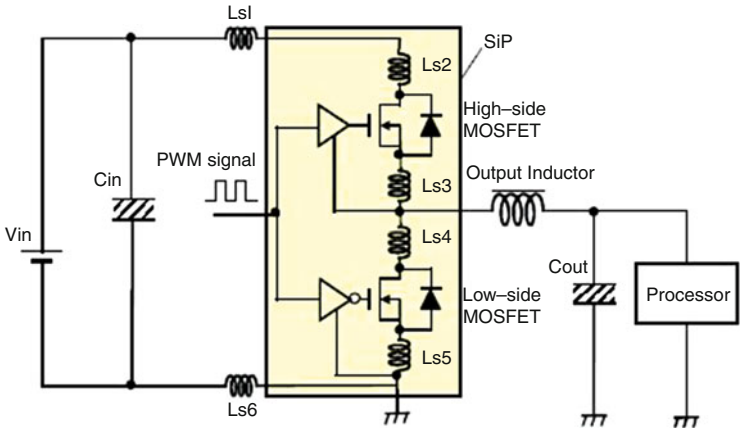


Fig. 6.36 The circuitry with both active and passive devices

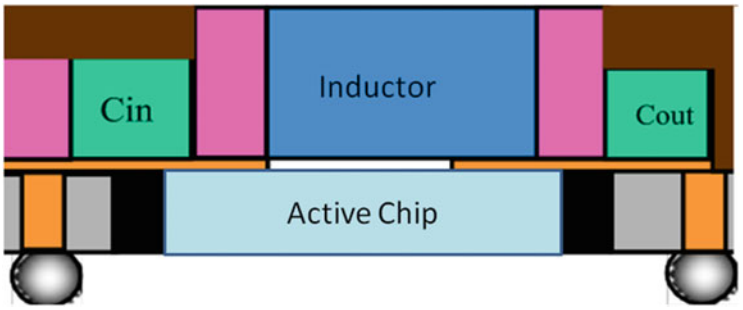


Fig. 6.37 The concept of wafer-level stack active device and passive chips

active power IC device can be placed down on the cavities of the substrate frame with gel or epoxy. The backside of the substrate frame can be removed by etching to form a through via that extends from the front side of the substrate frame. The substrate frame and the attached component can then be removed from the carrier so that the RDL metallization can be distributed on the backside of the wafer. Then the passive component inductor and capacitors are attached on the RDL of the substrate with active die by pick-and-place method. Then the wafer-level molding is applied to the wafer with passives stacked on the active die. Finally the wafer-level power system in packaging can be singulated by wafer sawing or laser cutting method.

This concept integrates both active power IC die and the passive components may extremely improve the parasitic resistance, inductance, and capacitance due to the stacking die technology and the short the distance between the active and passives with wafer-level RDL. It may also benefit the high switch frequency due to very low parasitic effect.

6.6 Summary

This chapter introduces the design concept of integration of analog and power solution, in which there are two wafer-level integrations: one is the system on chip (SOC) which integrates the power mosfets with analog IC in one wafer. Some people also called it as the smart power integration. This integration technology could make SOC chip excellent in electrical performance with high efficiency and low Rds(on). It can also make SOC easy to assemble through regular WLCSP technology. The challenges of such technology would be hard to apply to relative large power, such as in point-of-load portable application. Another wafer-level integration is the stack die technology, in which we introduce the low side Mosfet wafer to stack on high side Mosfet wafer with TSVs, analog IC die stack on Mosfets wafer with TSV and wafer molding technologies, and the wafer-level embedded technology with stacked active die and passive die. The concept of low side Mosfet wafer stacking on high side Mosfet wafer in this chapter does not include the analog IC die as well as the passive devices. While the wafer-level TSV/stack die package concept is to stack the individual analog IC die to the wafer with low side/high side Mosfet die together with the wafer-level molding technology, it does not include the passive device such as the input and output capacitors and the inductor. To integrate both the active die such as the Mosfet and IC die and the passive die such as the capacitors and inductor, the concept of wafer-level embedded die is introduced for stacking both the active and passive devices, which allows the full point-of-load application in one wafer-level packaging.

References

1. Liu, Y.: Trends of power wafer level packaging. Microelectron. Reliab. **50**, 514–521 (2010)
2. Liu, S., Liu, Y.: Modeling and Simulation for Packaging Assembly: Manufacturing, Reliability and Testing. Wiley (2011)
3. Cai, J., Szendrei, L., Caron, D., Park, S.: A novel modular smart power IC technology platform for functional diversification. 21st International Symposium on Power Semiconductor device & IC's, Barcelona ISPSD 2009 (2009)
4. Liu, Y.: Power electronic packaging: design, assembly process, reliability and modeling. Springer, New York (2012)
5. Fairchild application report, IntelliMAX^{TM} advanced load switches (2009)
6. Liu, Y., Qian, R.: Reliability analysis of next generation WLCSP, EuroSimE (2013)
7. Liu, Y.M., Liu, Y., Qu, S.: Prediction of board level performance of WLCSP, ECTC63, Las Vegas (2013)
8. Liu, Y.M., Liu, Y., Qu, S.: Bump geometric deviation on the reliability of BOR WLCSP, ECTC64, Orlando (2014)
9. Liu, Y., Kinzer, D.: (Keynote) Challenges of power electronic packaging and modeling, EuroSimE (2011)
10. Liu, Y.: (Keynote) Trends of analog and power packaging, ICEPT (2012)
11. Kinzer, D., Liu, Y., Martin, S.: Wafer level stack die package, US Patent 8,115,260, 14 Feb 2012
12. Liu, Y.: Wafer level embedded and stacked die power system-in-package packages, US patent 8,247,269, 21 Aug 2012

Thermal Management, Design, and Analysis for WLCSP

7

The operation of a wafer-level semiconductor device is sensitive to junction temperature. When the junction temperature exceeds the functional limit, the device does not operate in a normal way. It is also well known that the failure rates of semiconductor devices increase exponentially as the junction temperature rises. Figure 7.1 shows a FLIR camera image of temperature distribution inside a smart phone, which gives the sources of heat dissipation from WLCSPs mounted on the board. It is very crucial that the WLCSP designer and application engineer understand the definition, characteristics, and application of the thermal resistance of the WLCSP for proper device operation [1–6]. Power dissipation during the operation of the semiconductor device induces an increase in the junction temperature. This depends on the amount of power dissipation and the thermal resistance between the junction and the WLCSP bumps, an ambient, and some other specified reference point. This chapter introduces the thermal management, design, analysis, and cooling methods for WLCSP.

7.1 Thermal Resistance and Measurement Methods

7.1.1 Thermal Resistance Concept

The relation amongst these thermal properties such as thermal resistance, power dissipation, and the junction temperature $R\theta_{jx}$ is defined in the following Eq (6.1) [7, 11]:

$$R\theta_{jx} = \Delta T/P = \left(T_j - T_x\right)/P \tag{7.1}$$

where

P: Power dissipation per device
T_j: Junction temperature

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_7

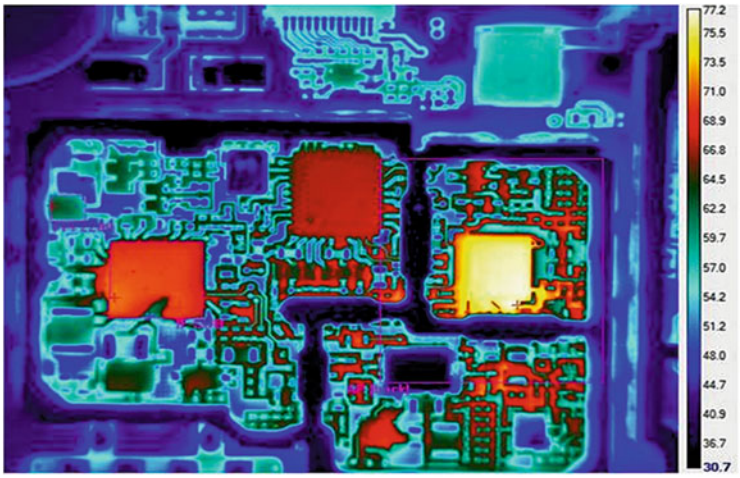


Fig. 7.1 The temperature distribution in a smart phone

T_x: Reference temperature
Unit: °C/W

The thermal resistance indicates the temperature drop between the junction and a specific reference point per unit of power consumption. It is a simplified parameter characterizing the thermal performance of a package. The selection of the reference point is arbitrary, and typical points and their abbreviations are summarized as follows:

(a) Junction-to-ambient temperature ($R\theta_{jA}$):

$$R\theta_{jA} = \frac{T_j - T_A}{P} \tag{7.2}$$

There are three major paths to air: one is the junction through package top to air; one is the junction through package bottom to either air or bottom bumps that attached to a board; and one is the junction through package side to air. In most cases, primary path is the bumps to board.

(b) Junction to the component case ($R\theta_{jc}$):

$$R\theta_{jc} = \frac{T_j - T_c}{P} \tag{7.3}$$

$R\theta_{jc}$ applies only to situations in which all or nearly all of heat is flowing out of top or bump bottom of the WLCSP through heat sink. Low $R\theta_{jc}$ means that heat will flow easily into external heat sink which connects either package top

or bottom. Here the case c could be the WLCSP top center or the bump bottom, which could be set as the room temperature 25°C.

(c) Junction to the component bump ($R\Psi_{\mathrm{jl}}$):

$$R\Psi_{\mathrm{jl}} = \frac{T_{\mathrm{j}} - T_{\mathrm{l}}}{P} \tag{7.4}$$

$R\Psi_{\mathrm{jl}}$ is not a true thermal resistance. In the formula (7.4) the total power is used because it is what is known. In standard environment, most but not all of the power flows to the board through bumps.

(d) Junction to the component top surface ($R\Psi_{\mathrm{jt}}$):

$$R\Psi_{\mathrm{jt}} = \frac{T_{\mathrm{j}} - T_{\mathrm{t}}}{P} \tag{7.5}$$

This is used to estimate the junction temperature from a measurement of top of WLCSP in actual applications. $R\Psi_{\mathrm{jt}}$ is not a thermal resistance. The total power is used in the Eq (7.5) because it is what is known. It is not the power dissipation between the junction and the top of the package. Normally only a small amount of power exits the package top.

Usually, here (c) and (d) are called thermal parameters rather than thermal resistances. From the above definitions, $R\theta_{\mathrm{jA}}$ (the junction-to-ambient thermal resistance) and $R\theta_{\mathrm{jc}}$ (the junction-to-case thermal resistance) are generally the most commonly used definitions in WLCSP packaging.

7.1.2 Temperature-Sensitive Parameter Method for Junction Calibration

The only unknown parameter is the junction temperature in the thermal resistance test, since the case or ambient temperature and power consumption are measured directly during the test. The direct measurement of the junction temperature is not possible, except in some packages, which have dies that are exposed to the air. But the P–N junction of a device reveals a specific forward voltage drop at a given temperature and a current. This forward voltage drop of a junction is called temperature-sensitive parameter (TSP), and is also known as the "diode-forward-voltage-drop" method from the original applications using power diodes or bipolar power transistors. This test method is called electrical test method (ETM), since the junction temperature is measured indirectly by an electrical relationship. Currently, ETM is the most popular technique for junction temperature measurement.

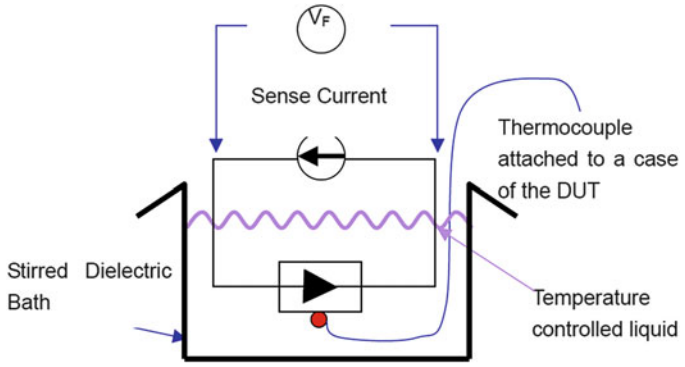


Fig. 7.2 Schematic diagram of a calibration bath for T.S.P. measurement

The relationship between the forward voltage drop and the junction temperature is an intrinsic electrothermal property of the semiconductor junction. The relationship is characterized by a nearly linear relationship between the forward-biased voltage drop and the junction temperature when a constant forward-biased current (also called Sense Current, hereafter) is applied. Figure 7.2 is a schematic illustration describing a measurement test setup of this voltage drop versus the junction temperature relationship for a diode junction.

In this test, the device under test (DUT) is heated up to a thermally equilibrated temperature in a hot bath, and a sense current is applied to the device to measure the forward-biased voltage drop at this temperature. The amount of the sense current is small enough to not heat the DUT, such as 1 mA and 10 mA, depending on the DUT's operating characteristics. By repeating the same tests at various temperatures, the following Fig. 7.3 can be obtained.

The relationship shown in Fig. 7.3 can be expressed in various mathematical equations, and a typical linear equation formula is given by Eq. 7.3 through a curve fitting.

$$T_j = m \times Vf + To \tag{7.6}$$

Here the slope "m" (° C/V) and the temperature ordinate-intercept "To" are used to quantify this straight line relationship. The reciprocal of the slope is often referred to as "K factor" in unit of mV/° C. In this case, Vf is the "Temperature-Sensitive Parameter (TSP)" for the diode junction, and the slope of the temperature–voltage calibration line is always negative, i.e., the forward conduction voltage decreases with increasing junction temperature.

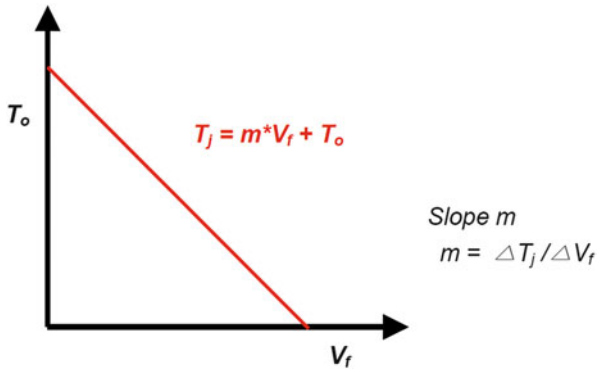


Fig. 7.3 Typical example of a T.S.P. plot

7.1.3 Thermal Resistance Measurement

Once the device is calibrated, the thermal resistance measurement test is conducted. Figure 7.4 describes a schematic of the thermal resistance measurement test circuit. This is composed of two sub-circuits, i.e., the heating and the sensing circuit. The heating circuit is operated to heat up the DUT up to T_{jmax}, given in the datasheet, by adjusting the power, while the sensing circuit is designed to measure the TSP with the sense current used in the device calibration. During the thermal resistance measurement test, an electrical switch is changed automatically for the operation of the heating circuit or the sensing circuit.

When the reference temperature T_x is measured at the package case, then T_x is referred to as T_c. The thermal resistance is written as $R\theta_{jc}$, and called junction-to-case thermal resistance. This indicates a package's power dissipation capability from the junction to the case. This is usually used for packages mounted on infinite or temperature-controlled heat sinks.

When the reference temperature T_x is an ambient temperature, then T_x is referred to as T_a. The thermal resistance is written as $R\theta_{ja}$, and called junction-to-ambient thermal resistance. This indicates a package's power dissipation capability from the junction to the ambient. This is usually used for packages mounted on PCBs without a heat sink. The detailed test environments for $R\theta_{ja}$ and $R\theta_{jc}$ are described in the following section.

7.1.4 Thermal Resistance Measurement Environments: Junction-to-Ambient Thermal Resistance

7.1.4.1 Natural Convection Environments

Figure 7.5a shows a schematic of the junction-to-ambient thermal resistance test, also known as the θ_{ja} measurement test under natural convection conditions. The major components are a still air chamber, PCB for the package mounting, and a

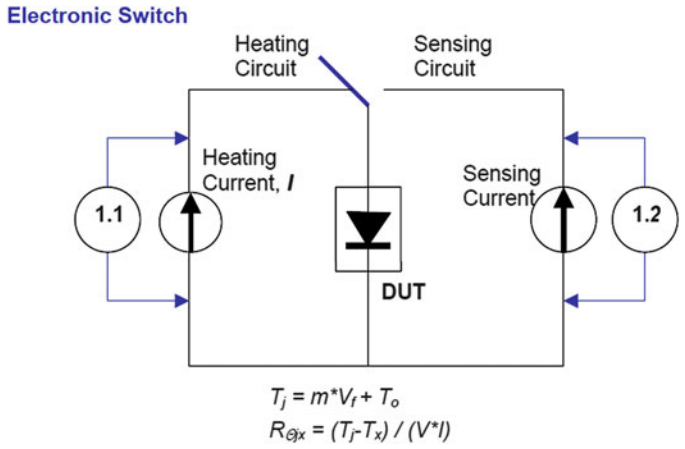


Fig. 7.4 Schematic of the thermal resistance measurement circuit

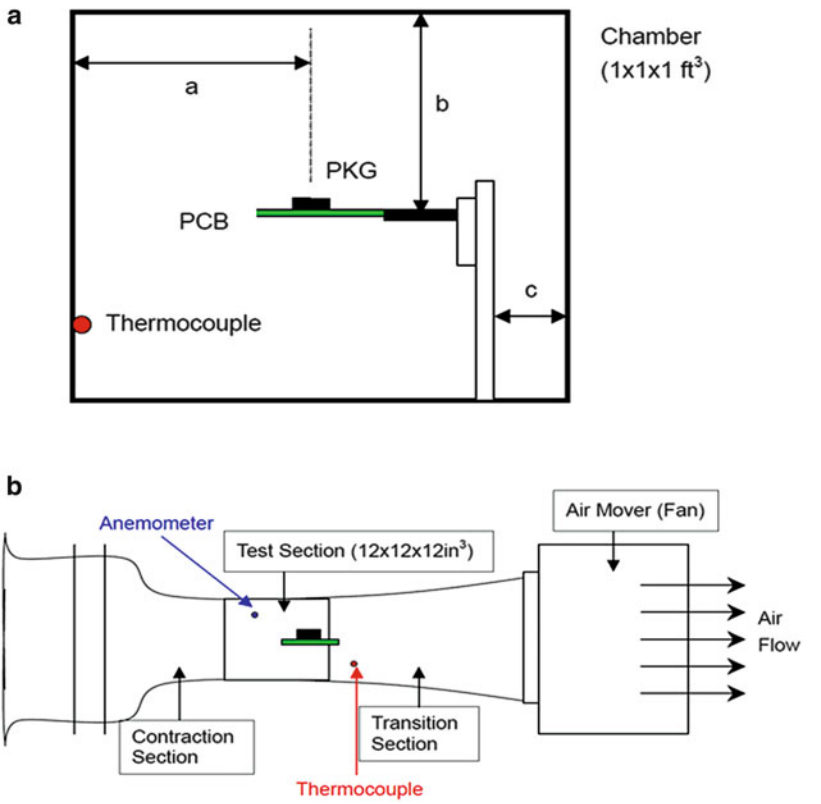


Fig. 7.5 The schematic diagram of thermal resistance measurement system with natural and forced convection. (**a**) Schematic of a still air chamber for the junction-to-ambient thermal resistance measurement test (a: 6.0 in., b: 6.5 in., c: 3.0 in., Volume $12 \times 12 \times 12$ in^3). (**b**) Schematic of a forced air wind tunnel for the junction-to-ambient thermal resistance measurement test

thermocouple. The chamber encloses one cubic-foot volume of still air, and follows the JEDEC standard recommendations. The PCB is mounted horizontally (or vertically, if requested) in the chamber, and the reference temperatures for both the inside and the outside of the chamber are measured.

7.1.4.2 Forced Convection Environment

Figure 7.5b shows a wind tunnel for the junction-to-ambient thermal resistance test, also known as the $R\theta_{ja}$ measurement test under forced convection conditions. The test setup is similar to the natural convection environment except for the still air chamber. The wind tunnel dimension is $12 \times 12 \times 74$ (in^3) with a test duct of 6 in. width. The temperature and air velocity are measured at the center of the tunnel and 6 in. ahead of the test board and package, respectively. The board and package are placed along the direction of the airflow. The air velocity is measured with a hot-wire anemometer probe, and the temperature is measured with the thermocouple.

7.2 Thermal Test Board for WLCSP

For the junction-to-ambient thermal resistance measurement test, the selection of a thermal test board or PCB for package mounting is crucial in thermal performance characterization. In industry, the specific dimension and material are recommended in the JEDEC standards JESD51-3 [8] and JESD51-7 [9]. These standards provide consistency in the thermal resistance presentation of the packages offered by the various semiconductor companies. The basic material is FR-4 with a total thickness of 1.6 mm, and the basic dimensions of the test boards are given in Figs. 7.6 and 7.7 (refer to JESD51-3).

7.2.1 Low-Effective Thermal Test Board

Low-effective thermal conductivity test boards [8] are designed to simulate a worst board mounting environment from a thermal performance point of view based on JEDEC standard. These test boards have no internal copper planes, and are named 1s0p test boards or two-layer boards. These are two-layer boards with minimal copper traces electrically connected from each package to one of the edge connectors. Two 1-oz copper layers are covered on both sides of the test board.

7.2.2 High-Effective Thermal Test Board

High-effective thermal conductivity test boards [9] are fabricated to have two evenly spaced internal planes. These boards more closely reflect applications in which ground or power planes are used in the PCB. Figure 7.8 shows the trace layers and layer thickness of a cross section in a high-effective thermal test board.

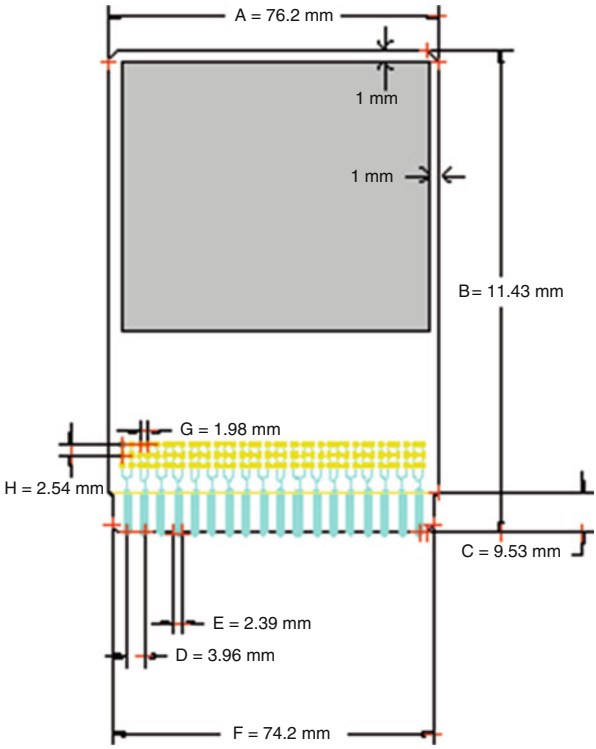


Fig. 7.6 PCB for packages <27.0 mm long 74.20 × 74.20 (mm^2) buried planes

This test board is called a 2s2p (2 signal plane and 2 power and ground planes) test board or 4-layer or multilayer board.

7.2.3 A Typical JEDEC Board for WLCSP

The application of JEDEC standard can apply to the WLCSP; a typical top trace layout of a JEDEC board is shown in Fig. 7.9 [10]. Of course, the customer may design the special trace layout that fits their own product application, which might be different from the JEDEC standard.

7.3 Thermal Analysis and Management for WLCSP

Analog and power WL-CSP becomes more and more popular in semiconductor industry due to its small profile and excellent electrical performance. However, the shrinkage of profile also results in a higher heat concentration, which brings widely concern throughout the industry. As a result of this, to make thermal simulation and analysis of its trend in the early design phase is extremely important for a good heat management and robust design of WLCSP.

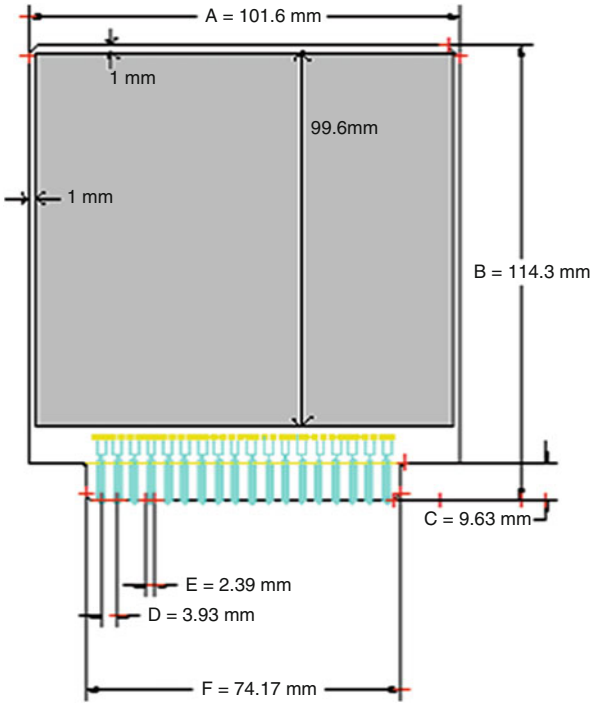


Fig. 7.7 PCB for packages >27.0 mm long 99.60 × 96.60 (mm^2) buried planes

Fig. 7.8 Cross section of multilayer PCB in JEDEC standard

The parametric design and modeling is an effective way to predict the thermal resistance, the temperature distribution, and the trend of the WLCSP design variation. Following the JEDEC standards [7–10], the thermal resistance θ_{JA} and thermal parameters including ψJB and ψJC are taken as objective concerns in this section to characterize thermal performance of WL-CSP. For traditional thermal analysis, when package modeling engineers perform thermal simulation, solid model of thermal board has to be built, and this work takes most of project time, therefore delaying the product design cycle time. In addition, although every engineer

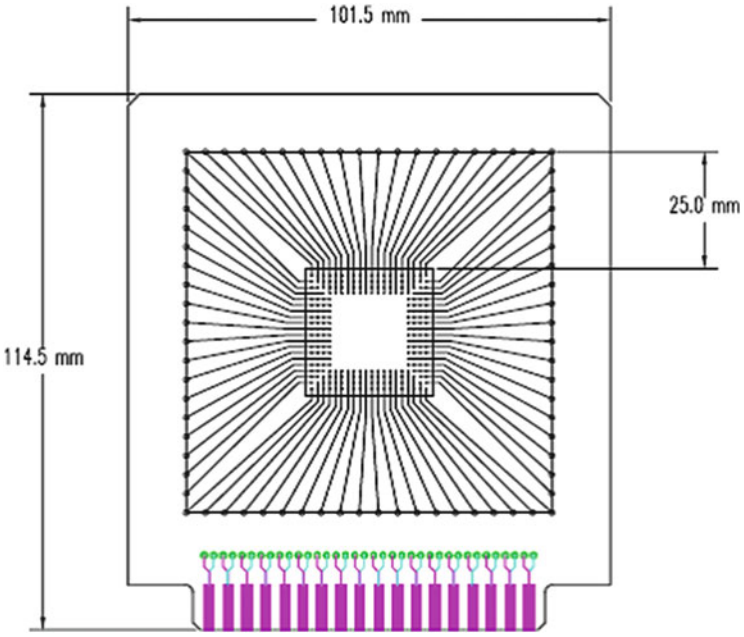


Fig. 7.9 The trace layout of a typical JEDEC board for WLCSP

follows JEDEC standards, different individuals will generate different thermal boards due to the tolerance of specified parameters which JEDEC provides (this seems unavoidable for actual manufacturing of thermal board, but can be avoidable for parameter design and simulation) and personal understanding of general specification. Comparing with traditional thermal design and analysis, the thermal parametric design and modeling have two outstanding merits.

1. It improves the efficiency of modeling engineers significantly. By employment of the parametric model, modeling engineers don't have to spend much time on understanding of JEDEC standards, and also they don't have to take pains to build model and do meshing; what they need to do is just to set parameters on demand, and then the entire thermal simulation, including meshing, loading/boundary condition, solving, and post-processing, will be dealt with and automated by ANSYS APDL coding.
2. The parametric model can avoid any variation due to unclear or un-detailed declaration for JEDEC thermal test board especially for trace layout. This will eliminate the need for the investigator to track the variations and allow focusing on the factors which are of primary concern.

This section introduces the construction of full parametric model for WL-CSP design, and the worst and the best layouts for internal trace are included as extreme cases to evaluate the influence of internal trace layout to thermal resistances or

thermal parameters. Both low- (1S0P) and high- (2S2P, 2S2P with thermal vias) effective thermal conductivity JEDEC boards are included in the model. Then experiment-validated empirical heat convection coefficients are applied to the parametric model, extensive modeling tasks are done to study the impact of solder ball number, die size, and terminal pitch on thermal resistances or parameters, and related results are systemically investigated as well. As verification, a WL-CSP with six balls will be actually tested finally.

7.3.1 Construction of the Parametric Model

As mentioned above, the parametric model consists of package and thermal test board. For a certain package, there are three types of thermal boards existing according to JEDEC standards including 1S0P, 2S2P, and 2S2P with vias. Since 1S0P is a simple case, so we will begin with that, and take 49-ball WL-CSP as an example to introduce how to construct the parametric model [11].

Figure 7.10 shows the structure of 49-ball WL-CSP which is mounted on JEDEC 1S0P thermal test board. As shown in Fig. 7.10, the model consists of silicon die (gray color), solder balls (purple color), copper pads and traces (red color), and FR4 board (green color). The silicon die in the right bottom is set transparent so that the ball array can be shown clearly. All the parameters for the 1S0P parametric model are listed in Table 7.1.

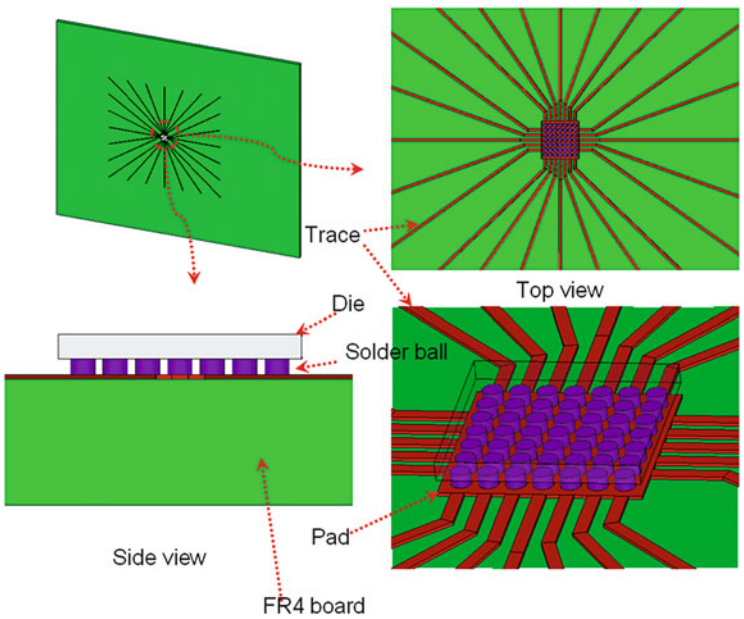


Fig. 7.10 WLCSP (49 balls) mounted on JEDEC 1S0P thermal test board

Table 7.1 Specified parameters and its description

Parameter	Description	Note
l_si	Length of silicon	
w_si	Width of silicon	
h_si	Height of silicon	
h_ball	Height of solder ball	
d_ball	Diameter of solder ball	
n1	Lead number in length of WLCSP	Need input (n1 $\geq$ 2)
n2	Lead number of in width of WLCSP	need input (n1 $\geq$ 2)
p	Pitch	
l_board	Length of board	114.5 mm [PKG $\leq$ 40 mm] 139.5 mm [40 < PKG $\leq$ 65 mm] 165.0 mm [65 < PKG $\leq$ 90 mm]
w_board	Width of board	101.5 mm [PKG $\leq$ 40 mm] 127.0 mm [40 < PKG $\leq$ 65 mm] 152.5 mm [65 < PKG $\leq$ 90 mm]
h_board	Height of board	1.6-h_trace
w_trace	Width of trace	40 % of p for $p > 0.5$ mm 50 % of p for $p \leq 0.5$ mm
h_trace	Height of trace	70 μm for $p > 0.5$ mm 50 μm for $p \leq 0.5$ mm
lt	Minimum length of trace	25 mm
s_inc	Strep increase of trace	Can be adjusted according to ball number and pitch
l_i	Initial shrinkage length of trace	Can be adjusted according to ball number and pitch

Most parameters in Table 7.1 are easily understood, so they will not be explained further. But some trace-layout-related parameters, which may bring confusion to readers, are illustrated in Fig. 7.11.

According to JESD51-9, traces to outer ball row should be flared to perimeter 25 mm from package body, as shown in Fig. 7.9. The distance from package body to perimeter is parameterized as "lt."

Parameter "l_i" indicates the length of maximum step of traces (see Fig. 7.11): in the case of more balls with finer pitch, the parameter "l_i" should be specified as a bigger value so that the traces will not touch each other to have a better spacing. Parameter "s_inc" stands for the length difference between neighbor trace steps; for the same reason, "s_inc" also should be defined properly.

One point should be noted that both "l_i" and "s_inc" are not specified in JEDEC standards. One advantage of the parametric model herein is: we can define values or rules for specifying above two parameters by ourselves. This leads us to eliminate the variation of models due to the uncontrolled factors.

For the parametric model, the major challenge herein is how to parameterize copper trace of thermal test board. For whichever side of board, the lead number

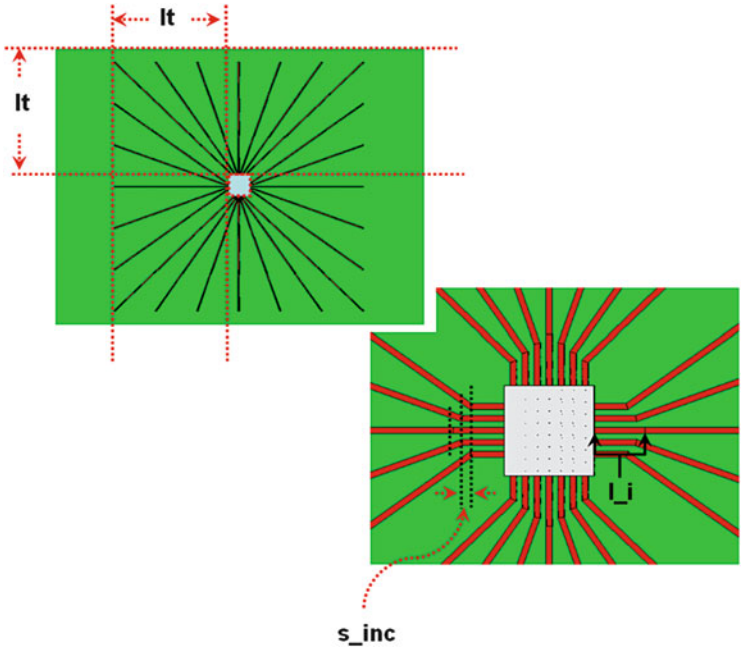


Fig. 7.11 Illustration of trace-layout-related parameters

should be either odd or even. For odd case, use a "*do. . ." comment to generate half of right side of traces except central one firstly, and then reflect along a horizontal symmetry axis, and then followed by central trace generation; by this time, all the right side of traces have already been generated. Finally reflect all the right side of traces along a vertical symmetry axis, as the results of this, all traces along long side orientation, are prepared. While for even case, it is similar to the odd case; the difference is that the central trace should not be a special concern (see Fig. 7.12).

After generation of the traces along long side orientation, hide all current traces and use the same rule to build traces along short edge orientation. One point should be noted that the boundary traces (in Fig. 7.12, the trace number is 4, 7, 11, 14 for odd case; and 4, 8, 12, 16 for even case) belong to both neighbor sides, so when building short side orientation traces, use "*do,i,1,(n2-1)/2-1" instead of "*do,i,1,(n2-1)/2" for odd case, and use "*do,i,1,n2/2-1" instead of "*do,i,1,n2/2" for even case to avoid trace repeat in the same boundary (see Fig. 7.12).

According to JEDEC, internal pads should be connected to external traces, in order to have a simplified model to improve efficiency; two extreme cases for internal trace layout are designed, which are also used to give the evaluation of internal trace layout to thermal resistances of packages. One is for the best case, which uses a thin block (same thickness as trace, same area as die) to connect all pads; refer to Fig. 7.13a). The other is for the worst case, which only connects outer pads to traces, and leaves internal pads isolated; see Fig. 7.13b).

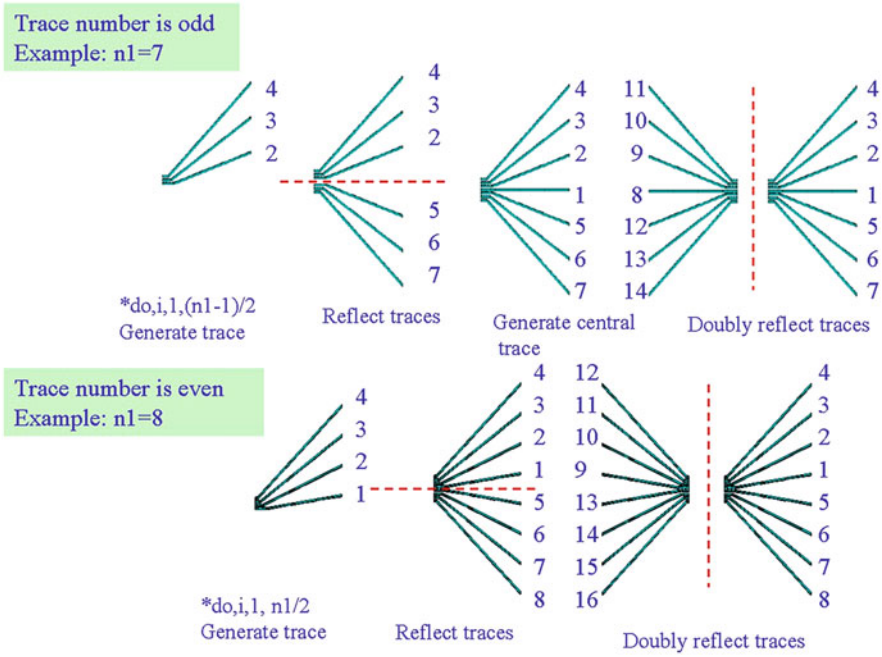


Fig. 7.12 Trace generation flowchart

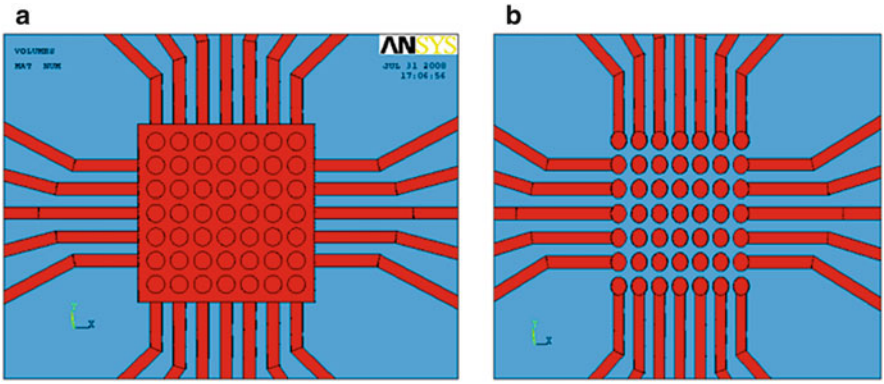


Fig. 7.13 Extreme internal trace layouts. (**a**) The best one and (**b**) the worst one

Other parts including silicon die, solder ball, and FR4 board are easily parameterized due to its simple geometry. So this part of work will not be given in detail.

JEDEC 2S2P thermal board adds bottom signal layer and two buried layers to 1S0P thermal board, while for 2S2P with vias (see Fig. 7.14), one thermal via

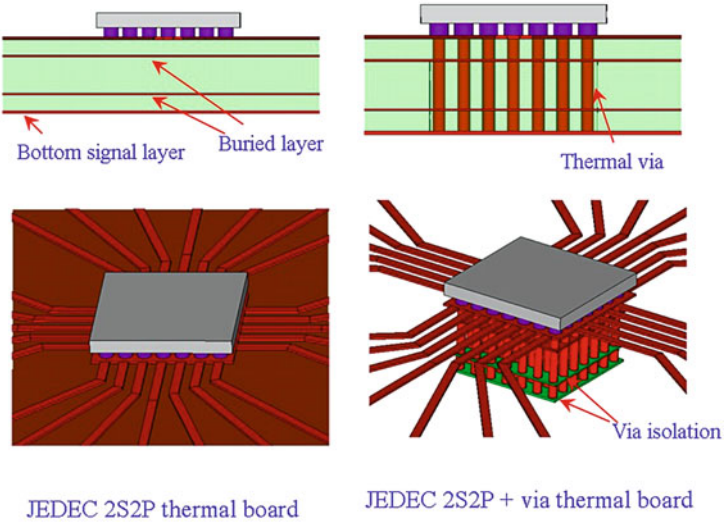


Fig. 7.14 JEDEC high-effective board

Table 7.2 Material properties

Material	Thermal conductivity (W/m · C)
Silicon die	145
Solder ball	33
Copper trace and solder	386
FR4	0.4

beneath each ball's pad is designed. These models employ same trace generation methodology and same internal trace design rule. So far all three types of thermal boards have been constructed and set up.

7.3.2 Application of the Parametric Model

Thermal performance of WL-CSP under natural convection condition is simulated by use of the parametric model developed here. The thermal conductivities for all materials in WLCSP are listed in Table 7.2.

For thermal resistance θ_{JA} evaluation, all three types of thermal boards are used, while for thermal parameters ψ_{JB}, 1S0P and 2S2P thermal boards are employed separately. The impact of ball number, die size, and pitch on these thermal resistances and parameters is studied systemically. Figure 7.15 shows the definition of different temperatures, including ambient temperature (T_A), case temperature (T_C), and trace temperature (T_B).

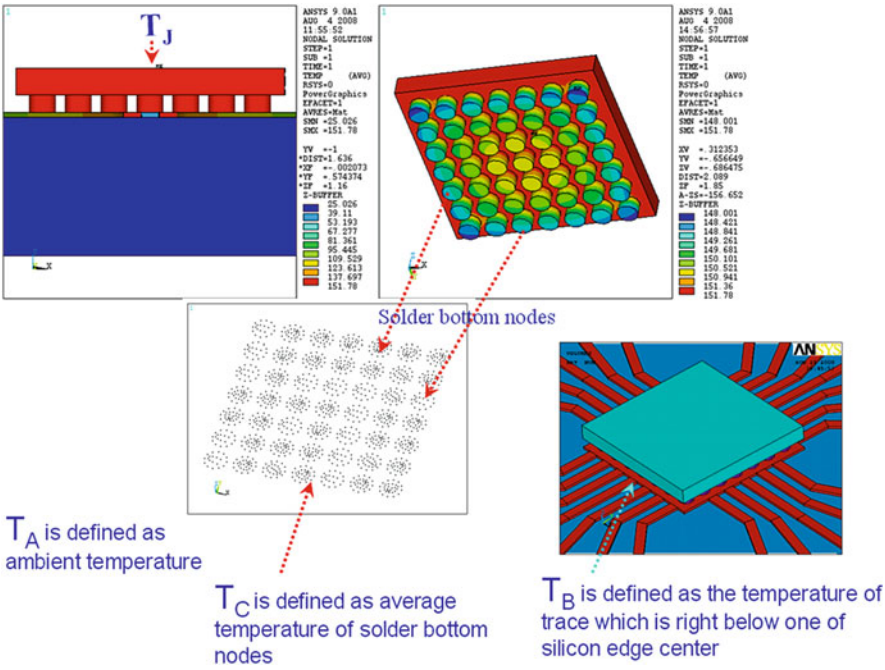


Fig. 7.15 Definition of different temperatures

7.3.3 Thermal Simulation Analysis

7.3.3.1 Impact of Solder Ball Number

In order to investigate the impact of solder ball number on thermal performance of WL-CSP, 3 × 3 mm die size and 0.4 mm pitch are selected, and power is set as 1 W, which is applied on the silicon die. And then design ball arrays as 2 × 2, 3 × 3, 4 × 4, 5 × 5, 6 × 6, 7 × 7 separately, and input these values to the parametric models including 1S0P, 2S2P, and 2S2P + via thermal test boards. Figure 7.16 gives the temperature distribution of all the WL-CSP packages with the best internal trace design for 1S0P thermal board.

All the results with two types of extreme cases for 1S0P, 2S2P, and 2S2P + via are also summarized in Fig. 7.17. It is clearly shown that:

1. Solder ball number is critical for thermal dissipation. Packages will dissipate heat with higher efficiency through more balls for all three types of boards; therefore more balls will result in a smaller θ_{JA}. So if viewed from thermal aspect, it is very important for a certain package to design enough balls to dissipate heat and secure die to work under safe temperature.
2. θ_{JA} curves for 1S0P is far above the ones for 2S2P and 2S2P + via, which means θ_{JA} is really board dependent; obviously high-effective boards dissipate heat from packages to ambient environment with higher efficiency.

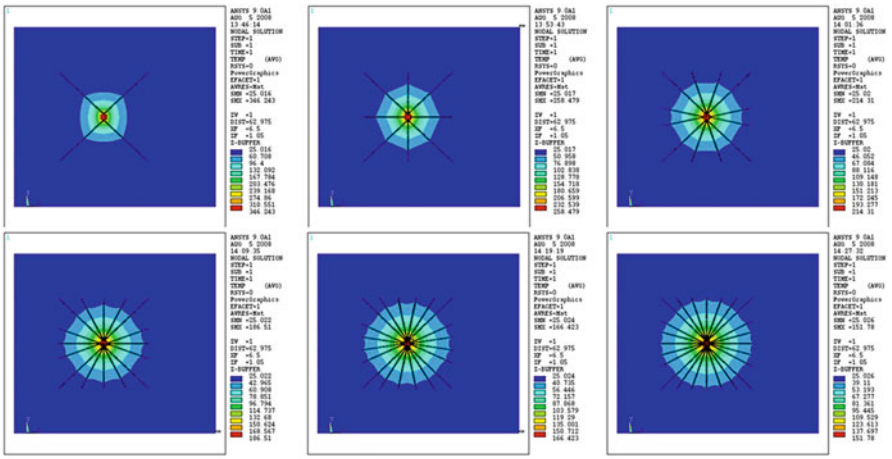


Fig. 7.16 Temperature distribution with different ball array (Best internal trace design, 1S0P)

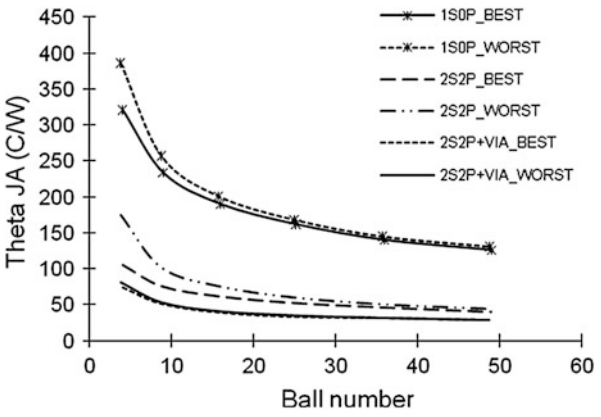


Fig. 7.17 Thermal resistance θ_{JA} versus ball number

3. The gap between two curves of extreme cases for a certain board reflects the role of internal trace layout design as the function of dissipating heat. Comparison between cases with and without vias tells whether internal trace layout design is invalid or valid for improving heat dissipation; it depends on the case if there is design for vertical heat transfer for thermal test board. In the case of thermal vias existing, heat transfers vertically from package to thermal board; for this case, internal trace layout will not work effectively due to its horizontal orientation. So for customer's application, we should suggest them to concern about internal trace layout for the PCB without effective vertical heat transfer design, especially for the packages with lower ball density.

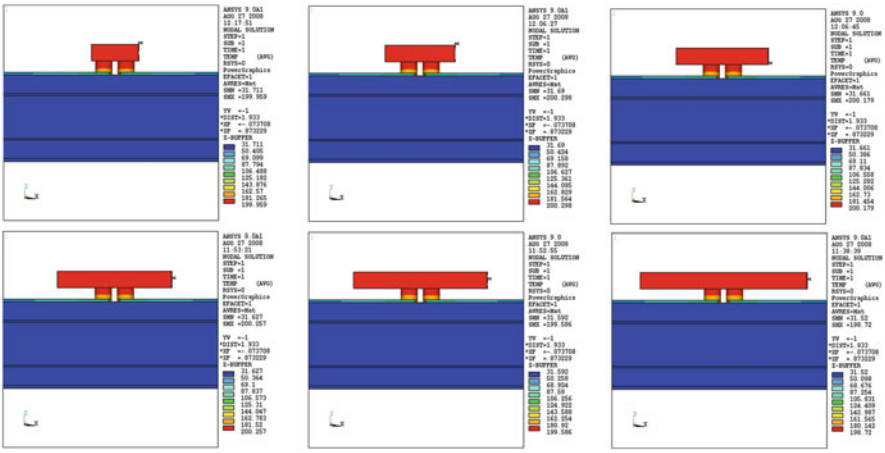


Fig. 7.18 Temperature distribution with different die size (Best internal trace design with 2S2P PCB). (**a**) Die size 0.85×0.85 mm^2. (**b**) Die size 1.25×1.25 mm^2. (**c**) Die size 1.65×1.65 mm^2. (**d**) Die size 2.05×2.05 mm^2. (**e**) Die size 2.4×2.4 mm^2. (**f**) Die size 3×3 mm^2

7.3.3.2 Impact of Die Size

For obtaining impact of die size on thermal performance of WL-CSP, two groups of ball arrays are selected including 2×2 and 4×4 cases. For 2×2 (4 balls) case, designed die size includes 0.85×0.85, 1.25×1.25, 1.65×1.65, 2.05×2.05, 2.4×2.4, and 3×3, while for 4×4 (16 balls) case, designed die size includes 1.65×1.65, 2.05×2.05, 2.4×2.4, and 3×3. For all these cases, pitch and power keep unchanged, 0.4 mm for pitch and 1 W for power individually. Figure 7.18 gives the temperature distribution of all the cases with the best internal trace design for 2S2P thermal board.

The results for all the cases above are summarized in Fig. 7.18. The curves with six points for 4-ball cases and the curves with four points for 16-ball cases are shown in Fig. 7.19. The conclusion drawn from Fig. 7.19 is listed below:

1. For all the worst internal trace layout designs, the only change is die size; from the curve, we can see that these serials of curves appear flat; this means that increased surface due to the increased die size contributes little to dissipate heat, so θ_{JA} is not sensitive to die size.
2. For the best internal trace layout designs except the cases with vias, θ_{JA} will decrease with increasing die size. Why does this happen? Why is it not the same for the cases with the worst internal trace design? If we look back to the construction of the model, we will find that with die size increasing, the thermal test board also changed. Since the copper block right below silicon die has the same area as that of the die, so with increasing die size, the copper block will also increase accordingly which enhances heat dissipation; this explains why for this case, it seems that θ_{JA} is sensitive to die size.
3. The rules for the role of internal trace layout as the function of dissipating heat are consolidated by Fig. 7.19: internal trace layout has very limited impact on

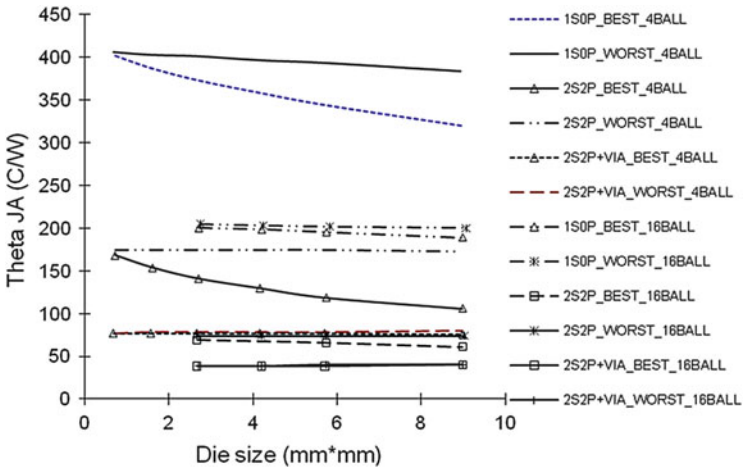


Fig. 7.19 θ_{JA} versus die size

θ_{JA} for the thermal test board with thermal vias, and it should be a concern for WL-CSP with bigger size and lower ball density when they are mounted on the board without effective vertical heat transfer path.

7.3.3.3 Impact of Pitch

To evaluate impact of pitch on thermal performance of WL-CSP, five pitch designs including 0.35 mm, 0.4 mm, 0.5 mm, 0.6 mm, and 0.7 mm, with three types of thermal boards, are prepared. For each experiment, fix die size as 3 mm × 3 mm, ball number as 16 (4 × 4), and power as 1 W. θ_{JA} according to pitch is illustrated in Fig. 7.20.

Results show that larger pitch has a smaller θ_{JA}; this may be due to two reasons:

1. According to JEDEC standard, larger pitch will also need a wider trace, so this makes PCB dissipate heat with higher efficiency.
2. For all five designed pitches, larger pitch makes solder ball space more equally, and therefore it avoids heat crowds and results in a better heat dissipation.

7.3.3.4 Actual Measurement for a 6-Ball WL-CSP

To correlate simulation result to the actual measurement, actual test is performed on a 6-ball WL-CSP, detailed package, thermal test board, and test apparatus; see Fig. 7.21. This is a 6-ball WL-CSP with 0.65 mm pitch and 1.92 mm × 1.44 mm die size. Three samples are prepared and tested, according to the results. θ_{JA} falls into the scope from 289 °C/W to 309 °C/W (see Table 7.3).

Based on the actual construction and dimension of thermal test board, simulations are run to calculate θ_{JA}. The result shows the simulated θ_{JA} is 290 °C/W for the worst internal trace design and 283 °C/W for the best case. Figure 7.22 gives the temperature contour distribution for the worst case. Table 7.4

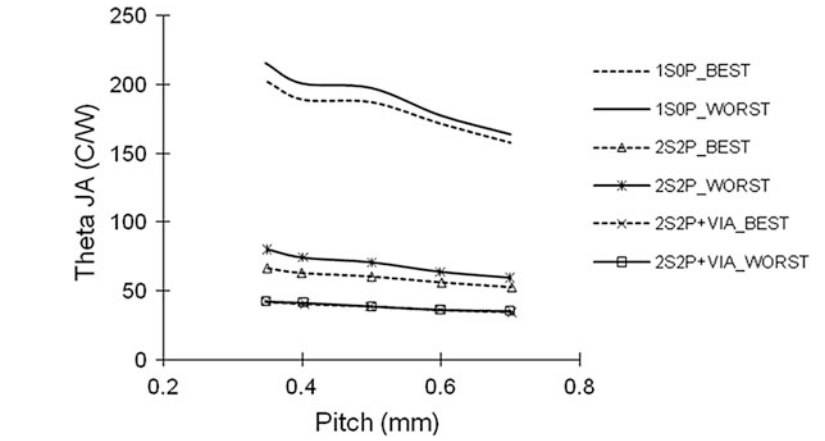


Fig. 7.20 θ_{JA} versus pitch

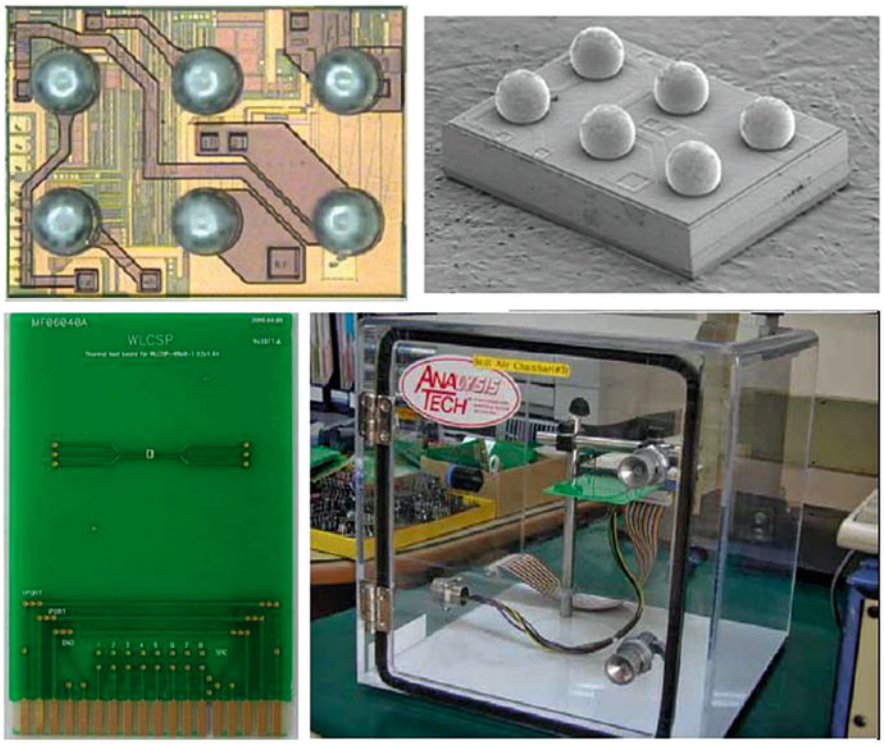


Fig. 7.21 Actual θ_{JA} measurement setup

Table 7.3 Actual measurement of 6-ball WLCSP

Sample 1	Sample 2	Sample 3	Average. θ_{JA}
291.18	309.48	289.02	296.56

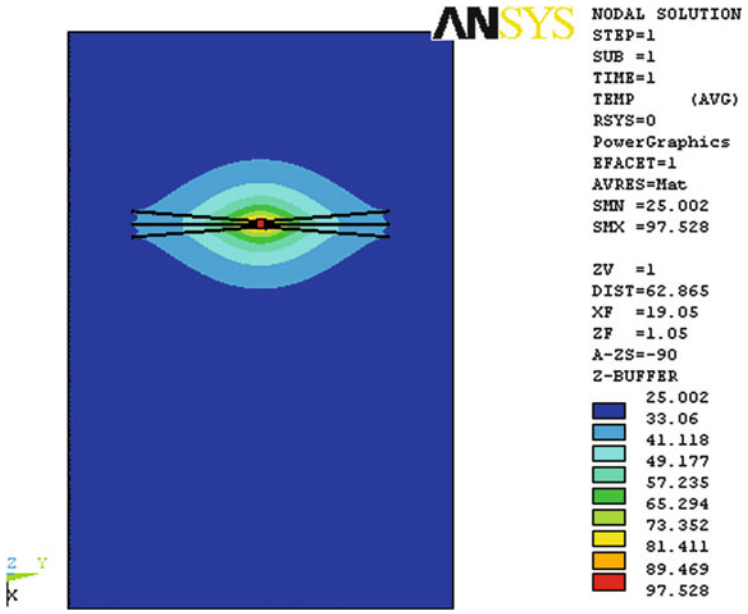


Fig. 7.22 Simulated temperature distribution of 6-ball WLCSP under 0.25 W power input (1S0P, the worst internal trace layout)

Table 7.4 Simulated θ_{JA} versus measured θ_{JA} for 6-ball WLCSP (Unit: °C/W)

	Simulation	Measurement
Min.	283	289
Max.	290	309
Median or average	286.5 [M]	296.6 [A]
Variation	3.41 % off from measurement	

gives the comparison of the simulated values and the actual measurement. As shown in Table 7.4, variation from simulation to the measurement is about 3.41 %, which proves it fits measurement quite well.

$$\theta_{JA} = 290^{\circ}\mathrm{C/W}$$

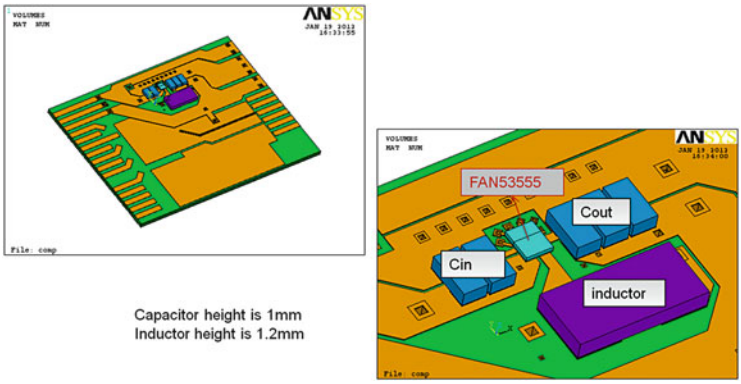


Fig. 7.23 A 4 × 5 WLCSP mounted on a non-JEDEC PCB

7.4 Transient Thermal Analysis for WLCSP

A 4 × 5 WLCSP (Fairchild FAN53555) chip is mounted on a non-JEDEC standard PCB with input capacitor, output capacitor, and inductor, as shown in Fig. 7.23. How its transient thermal performance impacts the passives on the PCB? And how its temperature propagates? This session will give the transient thermal analysis for a 4 × 5 WLCSP.

7.4.1 The Outline of 4 × 5 WLCSP and the Transient Material Properties

Figure 7.24 shows the outline of the 4 × 5 WLCSP layout. The solder material is SAC405; the package size is 1.6 × 2.0; the silicon thickness is 0.387 mm, solder pitch is 0.4 mm, and solder height is 0.15 mm. The transient thermal material properties are listed in Table 7.5

Figure 7.25 shows the finite element mesh of the 4 × 5WLCSP mounted on the non-JEDEC PCB with metal traces. Natural convection is selected in the simulation. For the heat transfer coefficients in the system, the convection heat transfer coefficient is selected as 1e-5 W/(mm^2°C) in the package top surface and solder bump surfaces, 7.5e-6 W/(mm^2°C) in the package side, bottom surfaces, and the PCB top traces, and 7.5e-6 W/(mm^2°C) in all the remaining PCB surfaces. Ambient temperature is selected to be 25 °C.

Figure 7.26 gives the temperature distribution at time 1 s, which shows the information of heat dissipation.

Figure 7.27 shows the transient heat dissipation and the dynamic temperature distribution at different time with a power 1.5 W applied on die. Figure 7.28 gives the temperature curves versus time at different locations of the system, from which the designer could estimate the speed of heat dissipation in the system and the temperature distribution in the interested points. This would be useful to understand

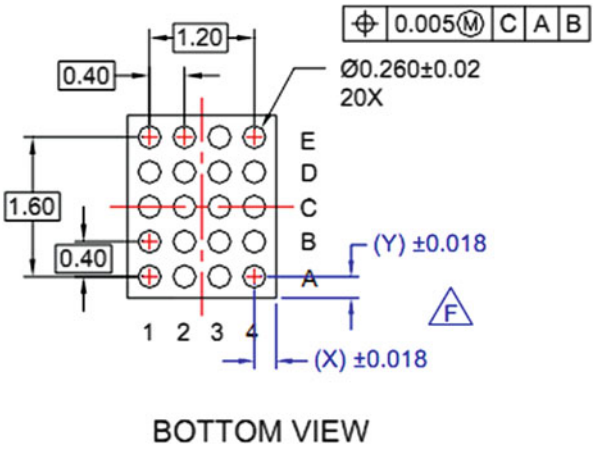


Fig. 7.24 4 × 5 WLCSP Layout

Table 7.5 Transient material properties for the 4 × 5 WLCSP mounted on non-JEDEC board

Material	Type	Thermal conductivity (w/m-k)	Density (kg/m^3)	Specific heat (j/kg-k)	Resource
si		146	2,330	708	From MatWeb
Inductor		17.6	5,400	767	Mixture method
Capacitor		59	7,210	481	Mixture method
Solder	SAC405	33	7,400	236	
FR4		0.4	1,910	600	
Cu trace		386	8,940	385	From MatWeb

how the heat dissipates from a WLCSP and how fast the heat transfers to the interested points in a PCB. Of course, changing the design and layout of PCB will impact the heat dissipation and the temperature distribution [8].

7.5 Summary

This chapter discusses the thermal management, design, and analysis for WLCSP. Section 7.1 gives the definition of the WLCSP thermal resistance and measurement methods; Sect. 7.2 introduces the thermal test board for WLCSP and JEDEC standard; Sect. 7.3 develops a thermal parametric model for WL-CSP packages, which includes parametric WL-CSP and its adaptive parametric JEDEC thermal test board. By employment of the parametric model, package geometry parameters can be set easily, and also trace layout for PCB will accordingly change to meet the requirement of WLCSP, so that the influence of all geometry parameters to thermal performance can be easily investigated for whole series of WL-CSP packages.

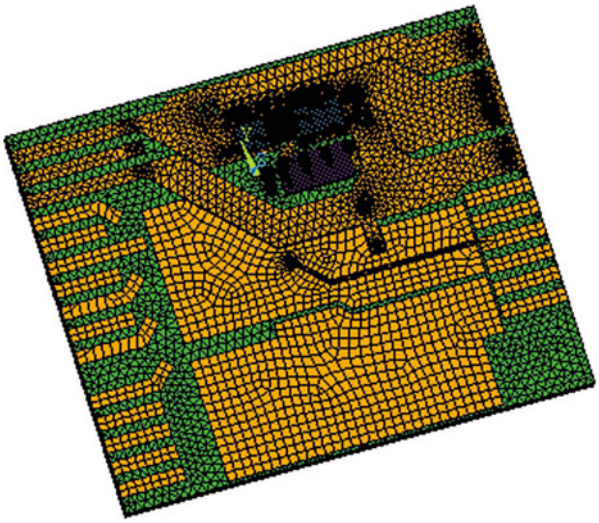


Fig. 7.25 The finite element mesh of the 4 × 5 WLCSP and PCB

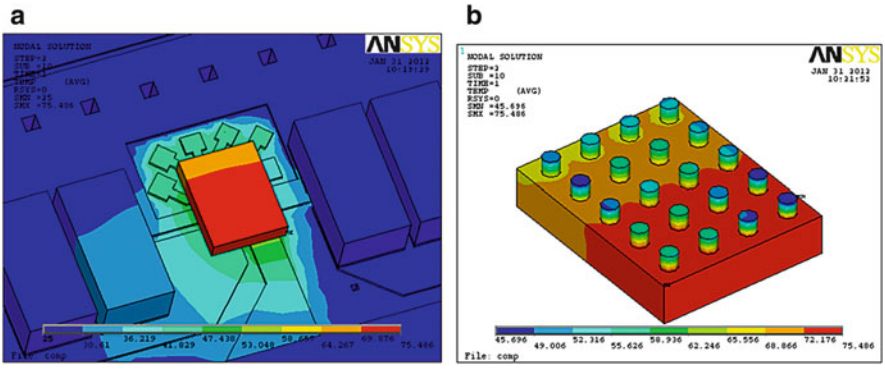


Fig. 7.26 Temperature distribution of the WLCSP on board at 1 s (power on die: 1.5 W) (**a**) The system temperature (**b**) The temperature of WLCSP (max: 75.5 °C)

Investigation of impact of ball number, die size, and pitch on thermal performance of WL-CSP shows: (1) Solder ball number is very critical to thermal performance of WL-CSP; more balls result in a smaller thermal resistance; therefore it is important to design proper balls and to secure die working under safe temperature. (2) Internal trace layout has significant impact on thermal resistance for boards without effective vertical heat transfer design, so it should be a concern for designing a proper internal trace layout to achieve a better thermal performance, and fully makes of area right below the silicon die, especially for the WL-CSP with bigger die size and lower ball density. (3) Equal distribution of balls on silicon die

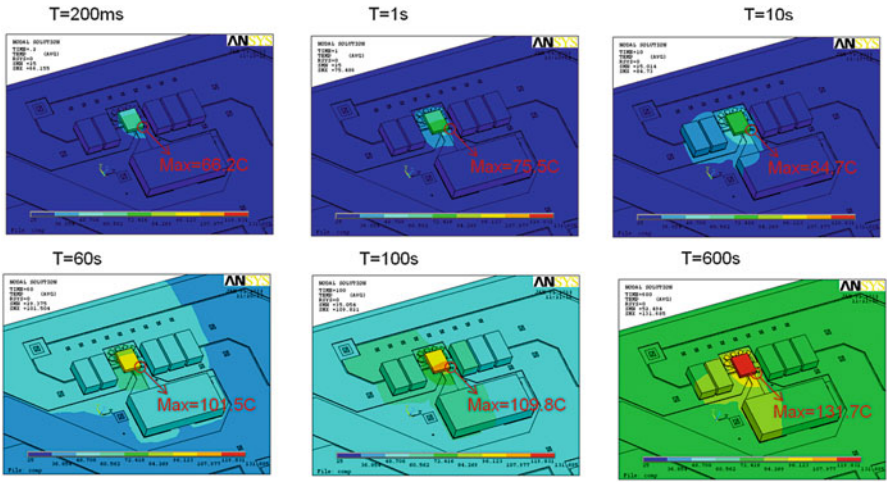


Fig. 7.27 The transient temperature growth in the system with a 1.5 W on die

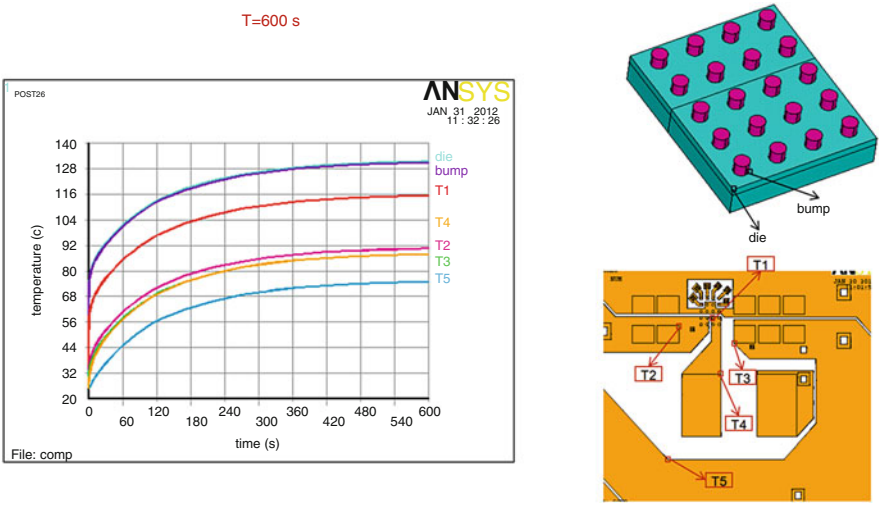


Fig. 7.28 The temperature curves at different locations with a power of 1.5 W on die

can avoid heat crowds and achieve a better performance for WL-CSP. Correlation of a 6-ball WL-CSP from simulation to measurement also proves the validation of the full thermal parametric model to some extent. Section 7.4 gives the transient thermal analysis for a WLCSP mounted on a non-JEDEC board. It shows the rate of heat dissipation and propagation along the board. The temperature distributes at different time and different locations. This is helpful for the product engineer to lay out the components and design the robust products on board.

References

1. Liu, Y.: Reliability of power electronic device and packaging. International Workshop on Wide-Band-Gap Power Electronics 2013 (ITRI), Taiwan, April, (2013)
2. Liu, Y.: Trends of power wafer level packaging. Microelectron. Reliab. **50**, 514–521 (2010)
3. Liu, Y., Liang, L., Qu, J.: Modeling and simulation of microelectronic device and packaging (Chinese). Science Publisher. (2010)
4. Hao, J., Liu, Y., et al.: Demand for wafer level chip scale package accelerates. 3D Packaging. 22. (2012)
5. Fan, X. J., Aung, K. T., Li, X.: Investigation of thermal performance of various power device packages. EuroSimE, (2008)
6. Liu, Y., Kinzer, D.: Challenges of power electronic packaging and modeling. EuroSimE, (2011)
7. JEDEC standard-JESD51-2, Integrated Circuits thermal Test Method Environment Conditions-Natual Convection (Still Air), (1995)
8. JEDEC standard-JESD51-3, Low Effective Thermal Conductivity Test Board for Leaded Surface Mount Packages, (1996)
9. JEDEC standard-JESD51-7, High Effective Thermal Conductivity Test Board for Leaded Surface Mount Packages, (1999)
10. JEDEC standard-JESD51-9 Test Boards for Area Array Surface Mount Package Thermal Measurements, (2000)
11. Liu, Y.: Power electronic packaging: Design, assembly process, reliability and modeling. Springer, New York (2012)

Electrical and Multiple Physics Simulation for Analog and Power WLCSP

8

The electrical performance (such as electrical resistance, inductance, and capacitance) is a key factor for a WLCSP product. Many studies, such as the electrical performance of different devices, effect of assembly reflow process on electrical properties, and the resistance of a solder joint, have been done to improve a product's electrical performance [1–3]. In recent years, the investigation for the electrical performance of a WLCSP has been paid more attention due to the wide applications of the WLCSP. The parasitic resistance, inductance, and capacitance (RLC) will impact the efficiency and switch speed of the WLCSP circuit. The electromigration issue of WLCSP, which is a multi-physics problem, becomes more critical due to the high current density in analog and power electronics. This chapter will introduce the electrical parasitic RLC simulation and electromigration simulation methods for WLCSP and wafer level interconnects.

8.1 Methods for Electrical Simulations: Extracting the Resistance, Inductance, and Capacitance

This section introduces the methods for extracting the self and mutual inductance, resistance, and capacitor from a WLCSP using ANSYS® Multiphysics. More general methodology can be found in the book of Power Electronic Packaging [1]. The results are to be used to generate an electrical model of the WLCSP for use in SPICE simulation.

8.1.1 Extracting the Inductance and Resistance

8.1.1.1 Theory Background for Resistance and Inductance

At alternating current (ac), the characteristic impedance can be represented by the real component "resistance" and imaginary component "reactance" [1]

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_8

$$Z_0 = R + jX \tag{8.1}$$

where R and X can be described using ac voltages and currents

$$R = \frac{V_{\text{real}} \times I_{\text{real}} + V_{\text{imag}} \times I_{\text{imag}}}{I_{\text{real}}^2 + I_{\text{imag}}^2} \tag{8.2}$$

$$X = \frac{V_{\text{imag}} \times I_{\text{real}} - V_{\text{real}} \times I_{\text{imag}}}{I_{\text{real}}^2 + I_{\text{imag}}^2} \tag{8.3}$$

Inductance can be derived from reactance as long as the frequency is known

$$L = \frac{X_L}{2\pi f} \tag{8.4}$$

Mutual reactance can be solved by

$$X_{ab} = \frac{V_{b_{\text{imag}}} \times I_{a_{\text{real}}} - V_{b_{\text{real}}} \times I_{a_{\text{imag}}}}{I_{a_{\text{real}}}^2 + I_{a_{\text{imag}}}^2} \tag{8.5}$$

Mutual inductance can be obtained from mutual reactance

$$L_{ab} = \frac{X_{ab}}{2\pi f} \tag{8.6}$$

Coupling factors can be derived based on the self and mutual inductance

$$K_{\text{ij}} = \frac{L_{\text{ij}}}{\sqrt{L_{\text{ii}} \times L_{\text{jj}}}} \tag{8.7}$$

Note: if $K_{\text{ij}} < 10\,\%$, the mutual inductance L_{ij} could be neglected.

8.1.1.2 Simulation Procedure

The unit of electrical simulation is based: MKS system. The CAD geometry structural dimensions must be accurate in every detail. Check the drawing to insure there are no tiny dimensional errors from the CAD import. All the components in the package system must be in perfect connection without initial defects. For the material data requirements, it includes resistivity and relative permeability. The 3D Element is selected with Solid97: 3-D 8-node magnetic solid and infin111: 3-D infinite boundary element. In the following simulation procedure, we will use a WLCSP as the simulation example.

1. *Define the element type*
2. *Define the material property data*
3. *Define the harmonic analysis: ANTYPE, 3*

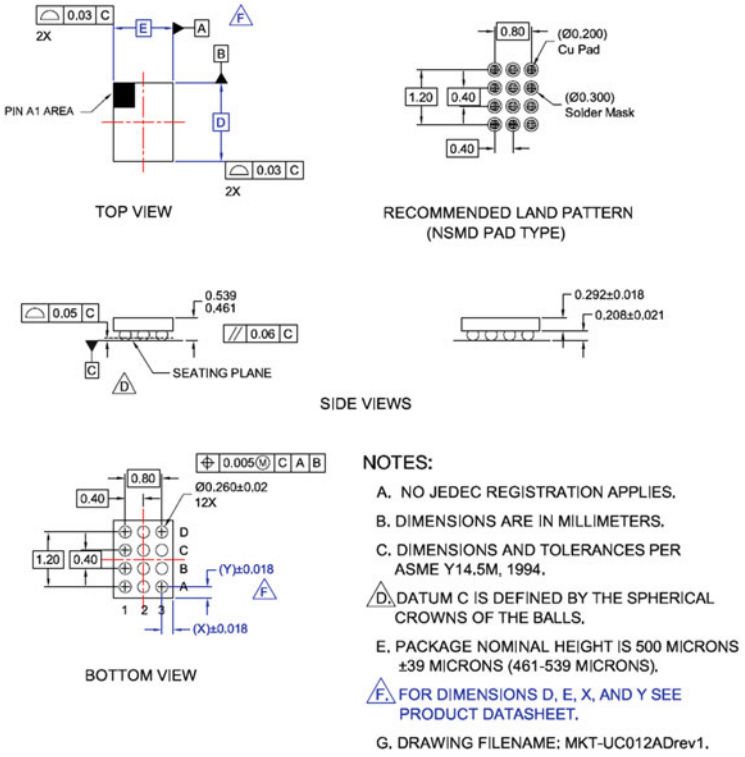
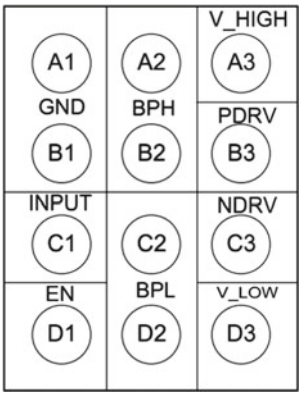


Fig. 8.1 A typical 3 × 4 WLCSP layout

4. *Generate/Import the solid model of the package*
   Figure 8.1 shows the layout of a typical 3 × 4 WLCSP. The solder bumps are the pins that would induce the parasitic electrical performance (RLC). For extraction of L/R of pins, the solid model only consists of the solder pins and no RDL metal lines/traces in the front of die. Figure 8.2 shows the pin map of 3 × 4 WLCSP and the number list.
5. *Add air volume around conductors and insulators*
   The "air" volume refers to any nonconductive material having a relative permeability of "1." The length, width, and height of the air volume should be 3–5 times greater than the length, width, and height of the package as shown in Fig. 8.3. This will allow room for the mesh to transition from swept brick to tetrahedral.
6. *Add infinite boundary to external surfaces of air using mitered volumes*
   Around the air volume, there is an external dimension of the infinite boundary which should be two times the air volume. See Fig. 8.4.
7. *Create a mesh*
   Try to mesh the conductors with hex/wedge shapes using the sweep operation. If the geometry of the conductors will not allow for this, a free tetrahedral mesh

a



b

Pin	Name	Ball
1	GND	A1
2	BPH	A2
3	V_HIGH	A3
4	PDRV	B3
5	BPH	B2
6	GND	B1
7	INPUT	C1
8	BPL	C2
9	NDRV	C3
10	V_LOW	D3
11	BPL	D2
12	EN	D1

Fig. 8.2 3×4 WLCSP pin map. (**a**) 3×4 WLCSP Pin map (**b**) Pin number

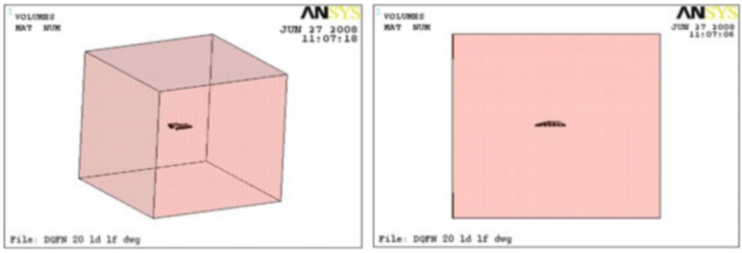


Fig. 8.3 The air volume around the conductors and insulators in package

can be used. The meshing for the conductors and air volume is shown in Fig. 8.5.

For each of the six mitered infinite boundary volumes, generate the mesh by extruding a prism mesh from the air volume. For infinite boundary, only one layer of elements is required. See the meshing result in Fig. 8.6.

8. *Apply the electrical boundary condition*

Apply the 0-V loads to one end of conductor bumps (like bump top). Couple the opposite ends of the conductor (bumps) with volt degree of freedom (DOF),

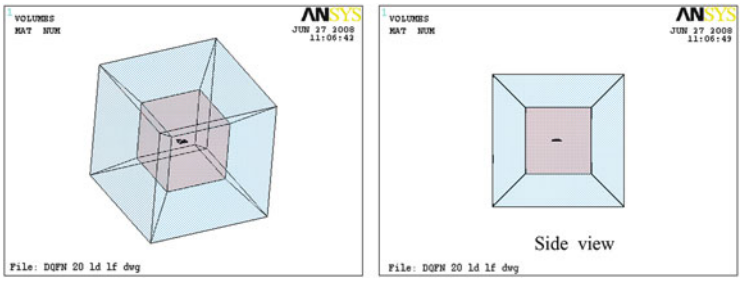


Fig. 8.4 Infinite boundary

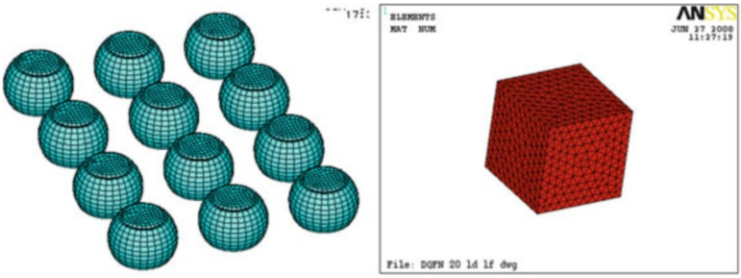


Fig. 8.5 Meshing for conductors and air volume

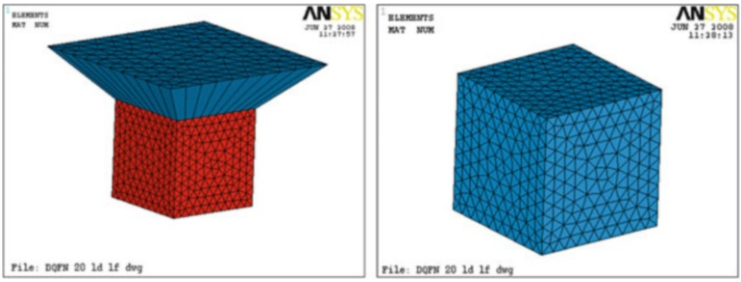


Fig. 8.6 Meshing for infinite boundary

in which the current load can be applied. The coupled set number should coincide with the package pin number; see Fig. 8.7.

9. *Apply the infinite boundary flags to the six exterior areas*
   The infinite boundary flags to the six exterior areas are set as in Fig. 8.8
10. *Set frequency for analysis*
    Extraction can be performed at various frequencies including performing a frequency "sweep" to extract R and/or L versus frequency. By default the analysis frequency has been set to 1 Hz to extract the static inductance and resistance. For high frequency analysis, it can be set to 1 MHz or higher.

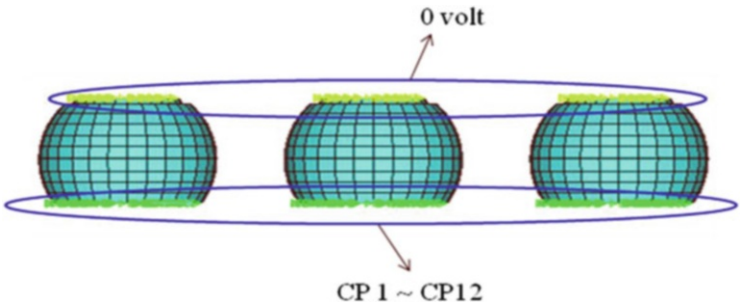


Fig. 8.7 Apply the 0 V loads on one end of bumps and couple the other end of the bumps with volt DoE

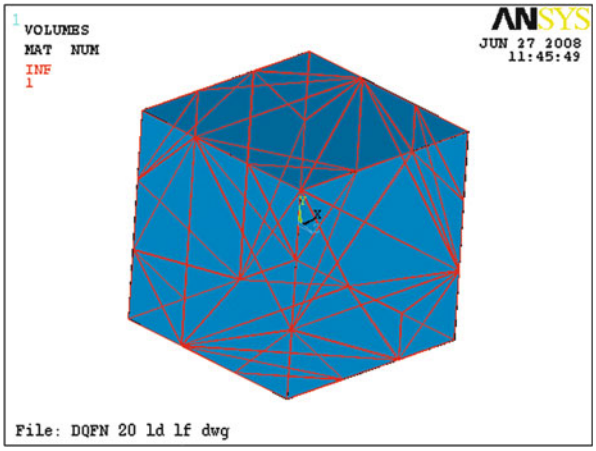


Fig. 8.8 Infinite boundary flags on six exterior surfaces

11. *Run the simulation and review the result.*
    Simulation may be run through the interface of the ANSYS software or may be run through a macro that assigns the current load, run the magnetic solution, and perform the inductance and resistance calculation. After the simulation is finished, check the result. In ANSYS, you can run *STAT to show the result.

    Sometimes we may get negative mutual inductance. The negative values are valid and just mean that the current in the path is flowing in opposite directions.

8.1.1.3 Skin Effect

The skin effect is the tendency of an alternating electric current (AC) to distribute itself within a conductor so that the current density near the surface of the conductor is greater than that at its core. That is, the electric current tends to flow at the "skin" of the conductor. The skin effect causes the effective resistance of the conductor to increase with the frequency of the current. Skin effect is due to eddy currents set up by the AC current. The skin effect has practical consequences in the design of radio

frequency and microwave circuits and to some extent in AC electrical power transmission and distribution systems. Also, it is of considerable importance when designing discharge tube circuits.

If the analysis is to be performed at high frequency, the skin depth of the conductors must be considered prior to meshing. To properly model the skin effect, the maximum element spacing should be equal to or less than one-half of the skin depth.

$$\delta = \sqrt{\frac{\rho}{\pi f \mu}} \tag{8.8}$$

δ = skin depth in meters; ρ = resistivity in Ω-meters; f = frequency in Hertz; and μ = absolute permeability in Henries/m; $\mu = \mu_0 \cdot \mu_r$, where μ_0 is the permeability of free space ($4\pi \times 10^{-7}$ N/A^2); and μ_r is the relative permeability.

8.1.1.4 Spectre Netlist Generation

The extracted inductance, resistance, and capacitance will be used to generate the Spice model including a Spectre netlist and a Cadence symbol.

When modeling a conductor there are two common structures that are used. They are referred to as LCL and CLC. We use the CLC structure because it results in fewer total components in the netlist.

Below Fig. 8.9 is a CLC circuit diagram

Below Fig. 8.10 is a LCL circuit diagram

For high frequency modeling we can divide a conductor into several CLC elements as shown in Fig. 8.11 for two CLC elements. Sometimes the elements are the same and sometimes they will change as the characteristic impedance of the conductor changes with respect to the propagation velocity of the signal.

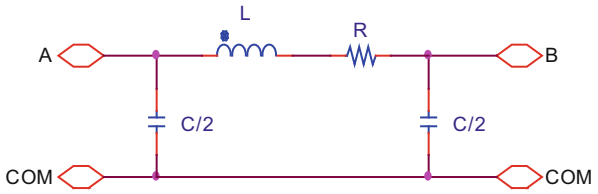


Fig. 8.9 CLC circuit element

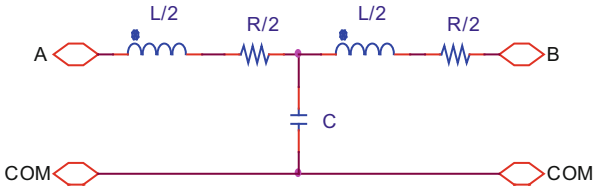


Fig. 8.10 LCL circuit element

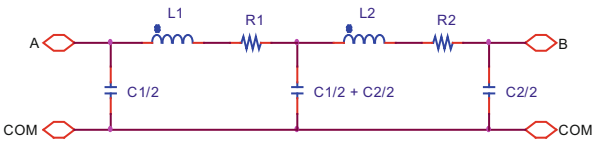


Fig. 8.11 Two CLC circuit element

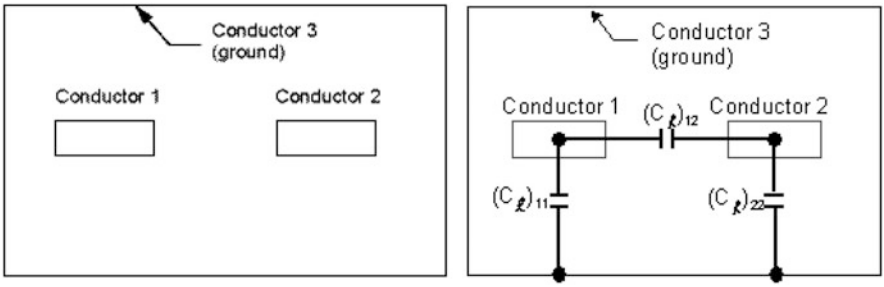


Fig. 8.12 A three conductor system

8.1.2 Methodology for Extracting Capacitance

A capacitor is a device for storing charge. It is usually made up of two plates separated by a thin insulating material known as the dielectric. The capacitance is a measure of the amount of charge a capacitor can store; this is determined by the capacitor geometry and by the kind of dielectric between the plates. For a parallel plate capacitor made up of two plates of area A and separated by a distance d, and dielectric material with dielectric constant k, the capacitance is given by $C = k\varepsilon_0 A/d$, where ε_0 is free space permittivity.

8.1.2.1 Ground Capacitance and Lumped Capacitances

Finite element simulation can readily compute and extract a "Ground" capacitance matrix of capacitance values that relate the charge on one conductor with the conductor's voltage drop (to ground). A three-conductor system (one conductor is ground) shown in the following Fig. 8.12 is used to illustrate the ground and lumped capacitance matrix. The following two equations relate charges on electrodes 1 and 2, Q_1 and Q_2, with the voltage drops for the electrodes, U_1 and U_2:

$$\begin{aligned} Q_1 &= \left(C_g\right)_{11}(U_1) + \left(C_g\right)_{12}(U_2) \\ Q_2 &= \left(C_g\right)_{12}(U_1) + \left(C_g\right)_{22}(U_2) \end{aligned} \tag{8.9}$$

where C_g represents the matrix of ground capacitances.

The **CMATRIX** command macro in ANSYS can convert the ground capacitance matrix to a lumped capacitance matrix, which is typically used in a circuit simulator such as Spice. The lumped capacitances between the conductors are

illustrated in the following figure. The following two equations then relate the charges with the voltage drops:

$$\begin{aligned} Q_1 &= (C_1)_{11}(U_1) + (C_1)_{12}(U_1 - U_2) \\ Q_2 &= (C_1)_{12}(U_1 - U_2) + (C_1)_{22}(U_2) \end{aligned} \tag{8.10}$$

where C_1 represents the matrix of lumped capacitances.

8.1.2.2 Simulation Procedure

To extract capacitance of WLCSP, an h-method electrostatic analysis will be used. The CMATRIX command will be used to solve the analysis and extract the lumped capacitance matrix. The unit used in this section is uMKS (capacitance in pF, length in uMeter). The CAD geometry structural dimensions must be accurate in every detail. Check the drawing to insure there are no tiny dimensional errors from the CAD import. All the components in the PKG system must be in perfect connection without initial defects. Material data requirements: Dielectric constant k, i.e., the relative permittivity of all the materials is needed. The 3D element is selected to be: Solid122: 3-D 20-Node Electrostatic Solid, Solid123: 3-D 10-Node Tetrahedral Electrostatic Solid, and the Infin111: 3-D infinite boundary element.

1. *Define the element type*
2. *Set the electromagnetic units to microns and picofarads:*
3. *Define the material property data*
4. *Generate/Import the solid model of the package*
   The 3D model is shown in Fig. 8.13. In this figure, the self-capacitance is Cii, and mutual capacitance is Ci,i+1 and Ci,i+2.
5. *Generate the test board*
   The WLCSP bump array and test board for a package electrical test with a thickness of 152 μm (JEDEC EIA/JEP 126) as shown in Figs. 8.14 and 8.15

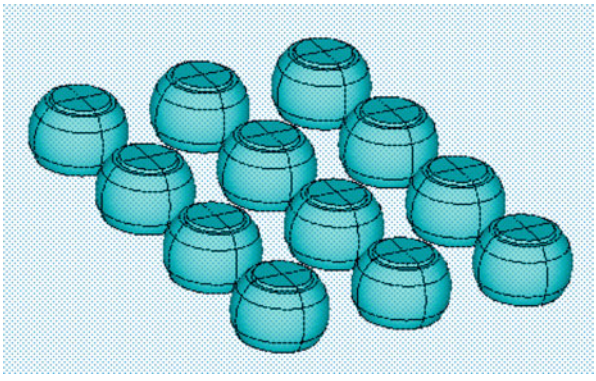


Fig. 8.13 The 3D solder bump model of WLCSP

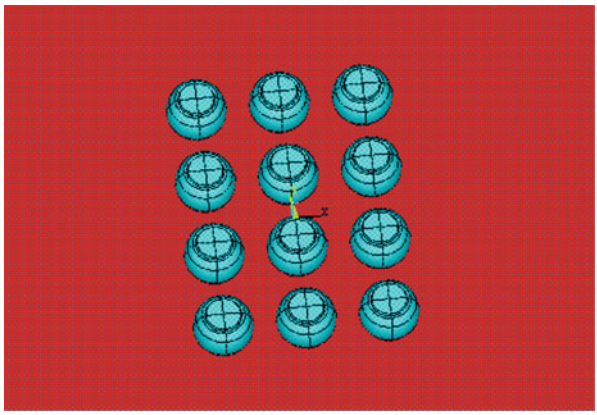


Fig. 8.14 3×4 WLCSP bumps

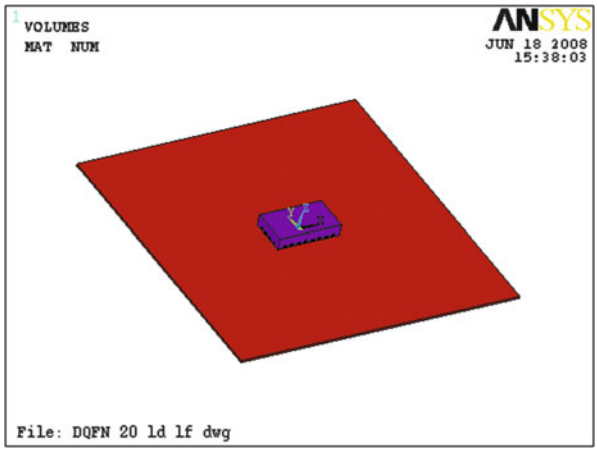


Fig. 8.15 The test board with a thickness of 152 μm (JEDEC EIA/JEP 126)

6. *Add air volume around conductors and insulators*
   The length, width, and height of the air volume should be three to five times greater than the length, width, and height of the package (see the below Fig. 8.16).
7. *Add infinite boundary to external surfaces of air using mitered volumes*
   External dimensions of the infinite boundary should be two times greater than the air volume; see Fig. 8.17
8. *Mesh generation*
   Use free tet mesh for all volume except infinite boundary (Fig. 8.18) and a sweep mesh with hex/wedge for the infinite boundary (Fig. 8.19). There is no need to mesh the conductors. But we usually do mesh them because this helps when defining the components. For infinite boundary, only one layer of elements is required.

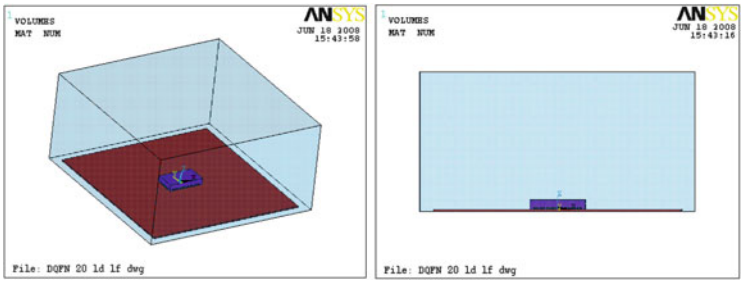


Fig. 8.16 Add air volume around the conductors and insulators of the package

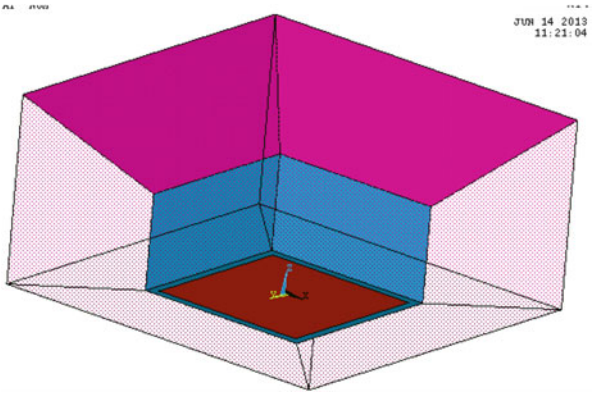


Fig. 8.17 Infinite boundary around the external surfaces of air volume

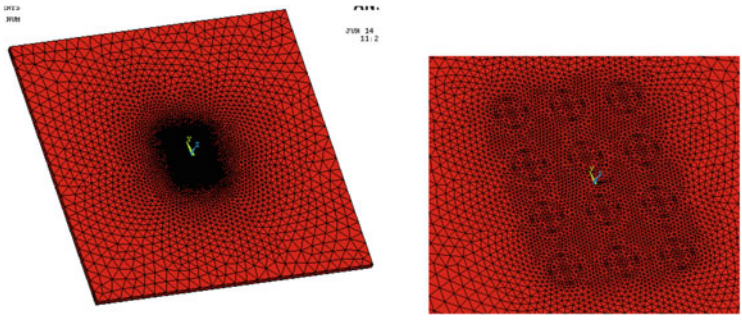


Fig. 8.18 Test board and bump mesh for capacitor model

9. *Create components from the conductors*
   Select all of the elements associated with that volume and select the nodes associated with the elements using the EXT (external) option in ANSYS. Create a component using the NODE option with the name "cond1." Repeat

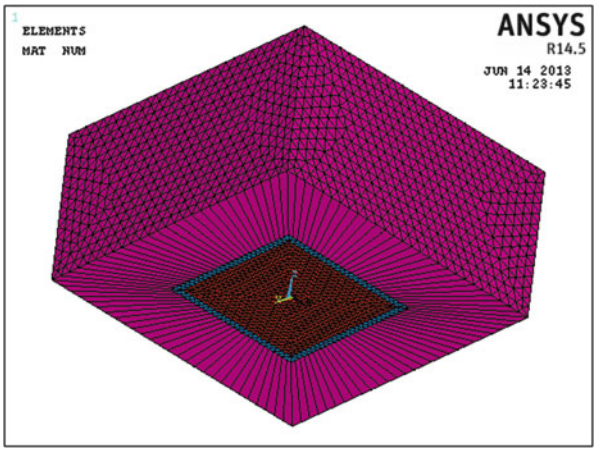


Fig. 8.19 Infinite boundary mesh for capacitor

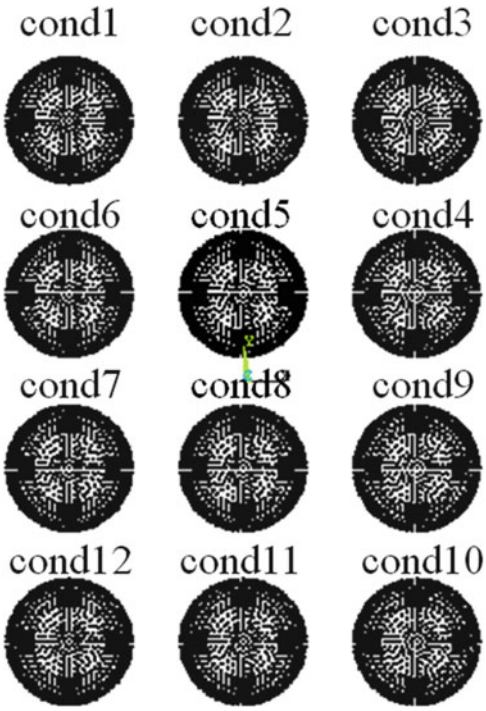


Fig. 8.20 Create the components

for each bump and use the package pin # with the "cond" prefix as the component name in each case, as shown in Fig. 8.20.

For the ground plane, select the nodes attached to the bottom surface area of the PCB and those external nodes of bumps create a component using those nodes. This component must be assigned the greatest value (an example is

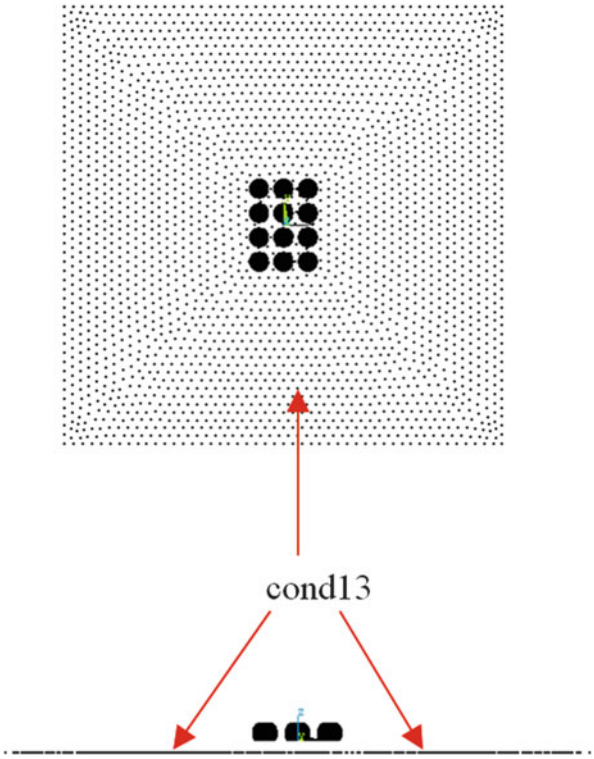


Fig. 8.21 Ground plane setting

"cond13" if there are 12 pins defined as components) as shown in Fig. 8.21. *In other cases, if the bumps are floating, the external nodes of bumps cannot be defined as ground.*

10. *Apply the infinite surface boundary condition to the five exterior areas (the sixth exterior area is the ground plane).*
    Figure 8.22 shows the boundary layout at the five outside surfaces.
11. *Run the analysis and check the results.*
    For ANSYS, use the following commends to run it.

cmatrix,symfac,'condname',numcond,grndkey

symfac: geometric symmetry factor. Defaults to 1.0

Condname: conductor name

Grndkey = 0 if ground is one of the components or 1 if ground is at infinity

Numcond: total number of components.

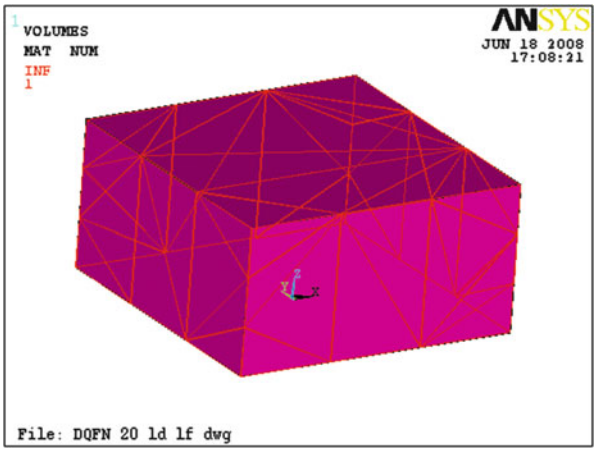


Fig. 8.22 The infinite surface boundary condition

Table 8.1 The geometry of 3 × 4 WLCSP

WLCSP	Pitch (mm)	Chip size (mm)	PI via diameter	AI Pad diameter	UBM diameter	SiN diameter	PCB pad diameter
3 × 4	0.4	1.2 × 1.6	170 μm	225 μm	205 μm	215 μm	200 μm

WLCSP	UBM thickness	SiN passivation thickness	Al pad thickness	PI thickness	PI via side wall angle	Solder joint diameter	Solder joint height (H1 + H2)
3×4	2.75 μm	1.45 μm	2.7 μm	10 μm	60°	289.55 μm	206.93 μm

The example is:

```
/SOLU
CMATRIX,1,'cond',9,0
```

Results of the analysis will be written to a file named cmatrix.txt in the ANSYS working directory as a capacitance matrix in two formats, "Ground" and "Lumped."

12. *RLC extraction for a fan-in 3 × 4 WLCSP*
    The geometry size of the fan-in 3 × 4WLCSP is shown in Table 8.1 and Fig. 8.23. The material properties of WLCSP are listed in Table 8.2. The resistance, inductance, and capacitance of the bumps under 10 MHz are listed in Table 8.3. The mutual inductance and capacitance are listed in Tables 8.4 and 8.5

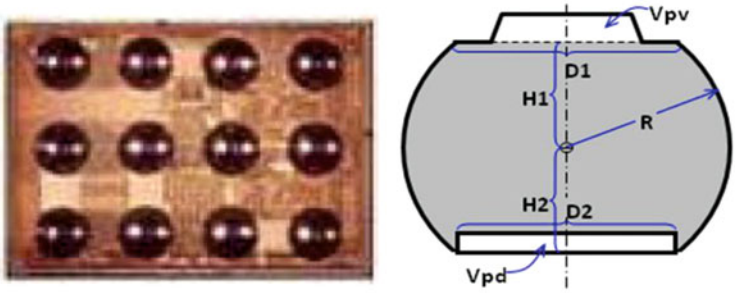


Fig. 8.23 The bump geometry definition of a fan in 3 × 4 WLCSP

Table 8.2 The material properties of 3 × 4 WLCSP

Material	Solder ball	air	Infinite boundary	PCB
Material type	SAC396	–	–	FR4
Resistivity (Ohm-m)	12.5e-8	–	–	–
Permeability	I	I	I	4.2

Table 8.3 The resistance, inductance, and capacitance of the bumps under 10 MHz

pin#	R (mOhm)	Lii (nH)	Cii (pF)
1	0.6	0.032	0.037
2	0.6	0.032	0.034
3	0.6	0.032	0.037
4	0.6	0.032	0.034
5	0.6	0.032	0.030
6	0.6	0.032	0.034
7	0.6	0.032	0.034
8	0.6	0.032	0.030
9	0.6	0.032	0.034
10	0.6	0.032	0.037
11	0.6	0.032	0.034
12	0.6	0.032	0.037

8.2 Electrical Simulation for a Fan-out Molded Chip Scale Package

8.2.1 Introduction of the MCSP

Figure 8.24 shows a construction of a molded flip chip package (MCSP) which includes a fine pitch flip chip die which is bonded to the substrate with gold-to-gold interconnection (GGI), a PCB substrate with copper trace as the redistribution layer (RDL) and vias, the mold cap with epoxy mold compound to encapsulate the flip chip, and the PCB substrate and the solder bumps for connecting the MCSP to the

Table 8.4 The mutual inductance of the bumps under 10 MHz

pin#	pin#	Lij	Kij
i	j	nH	--
1	2	0.009	0.29
2	3	0.009	0.29
3	4	0.009	0.29
4	5	0.009	0.29
5	6	0.009	0.29
6	7	0.009	0.29
7	8	0.009	0.29
8	9	0.009	0.29
9	10	0.009	0.29
10	11	0.009	0.29
11	12	0.009	0.29

pin#	pin#	Lij	Kij
i	j	nH	--
1	3	0.005	0.15
1	5	0.006	0.20
1	6	0.009	0.29
1	7	0.005	0.15
2	4	0.006	0.20
2	5	0.009	0.29
2	6	0.006	0.20
2	8	0.005	0.15
3	5	0.006	0.20
3	9	0.005	0.15
4	6	0.005	0.15
4	8	0.006	0.20
4	9	0.009	0.29
4	10	0.005	0.15

pin#	pin#	Lij	Kij
i	j	nH	--
5	7	0.006	0.20
5	8	0.009	0.29
5	9	0.006	0.20
5	11	0.005	0.15
6	8	0.006	0.20
6	12	0.005	0.15
7	9	0.005	0.15
7	11	0.006	0.20
7	12	0.009	0.29
8	10	0.006	0.20
8	11	0.009	0.29
8	12	0.006	0.20
9	11	0.006	0.20
10	12	0.005	0.15

Table 8.5 The mutual capacitance of the bumps

pin#	pin#	Cij
i	j	pF
1	2	0.004
1	6	0.004
2	3	0.004
2	5	0.004
3	4	0.004
4	5	0.004
4	9	0.004
5	6	0.004

pin#	pin#	Cij
i	j	pF
5	8	0.004
6	7	0.004
7	8	0.004
7	12	0.004
8	9	0.004
8	11	0.004
9	10	0.004
10	11	0.004
11	12	0.004

outside world. For a wafer level MCSP, the substrate can be made by silicon wafer with TSV and metal RDL. After the die attaches through the GGI process, the wafer level molding is applied. The final wafer MCSP can be obtained by the wafer

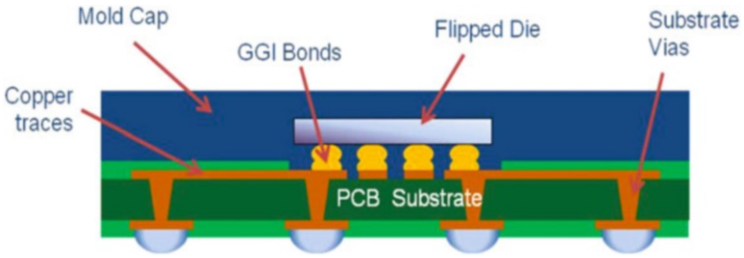


Fig. 8.24 A typical MCSP structure

singulation process, such as the wafer sawing. MCSP is a fan-out chip scale package; the bump pitch and size of the MCSP are much larger than that of flipped attached die with small gold bumps. This section studies the two types of MCSP packages: one is based on the flipped die GGI process on the substrate with RDL; another is based on the wire bonding process and the wire connects the fine pitch die to the RDL substrate. Through the comparison, we can find out how the significant difference would be with the two technologies: GGI versus wire bonding.

8.2.2 RLC Simulation for a 40 Pin MCSP with GGI Process

Figure 8.25 gives the outline of the 40 pin MCSP, which shows the top, side, and bottom views and the geometry outline of the packaging. Figure 8.26 and Table 8.6 give the pin map, RDL, and the signal pins that are corresponding to the MCSP pins for the GGI connection.

Figure 8.27 shows the model for resistance and inductance, in which Fig. 8.27a includes the air volume that covers the 40 bumps of MCSP; Fig. 8.27b shows the RDL which connects the small pitch WLCSP die to the 40 fan-out pins. Figure 8.27c shows part of mesh of the air volume and the infinite boundary. Table 8.7 gives the 40 pin MCSP electrical material properties, which lists the resistivity and the permeability.

Figure 8.28a, b shows the electrical boundary conditions for the simulation of the resistance and inductance for the 40 pin MCSP with GGI, in which the 0 V is applied at the top surface of the GGI. The bottom nodes of pin 1–pin 40 are coupled as CP1–CP40 and the current load is applied to the coupled sets, respectively. Table 8.8 lists the parasitic resistance and the inductance of the whole 40 pins. Table 8.9 gives the mutual inductance of the whole 40 pins MCSP with GGI, in which the coupling factor K_{ij} is listed as well.

Figure 8.29 shows the FEA models for the capacitance simulation of the 40 pin MCSP with a board and the material electrical permittivity property. Figure 8.30 lists the components number cond1–cond41, in which the components cond1–cond40 consist of the nodes in the exterior surface of Pin1–Pin40 and the component cond41 consists of the nodes of the bottom surface of the test board (Table 8.10).

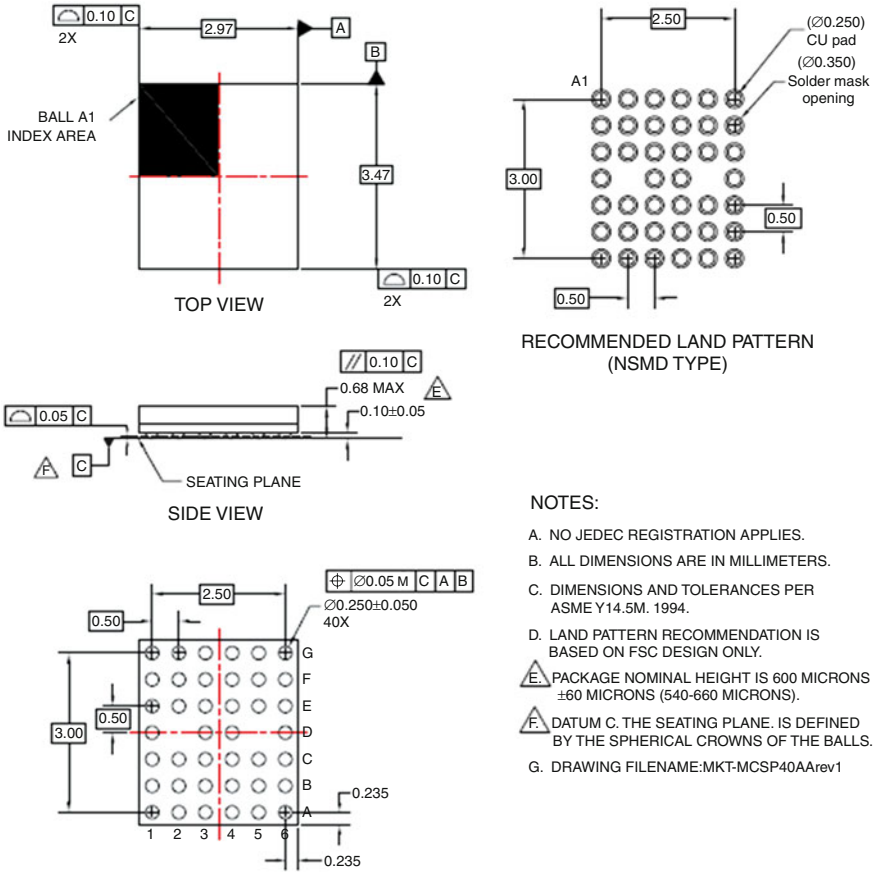


Fig. 8.25 The outline of the 40 pin MCSP with GGI process

Table 8.11 lists the parasitic self-capacitance of the whole 40 pins of the MCSP with GGI. Table 8.12 gives the mutual capacitance of the 40 pins MCSP with GGI.

8.2.3 Wire Bonded MCSP and the Electrical Performance Comparison with GGI Type of MCSP

Figure 8.31 shows the pin map and wire bonding diagram of the 40 pin MCSP with wire bond. Instead of GGI, wire bonding process is applied to connect the signal die to the RDL and the package pins. Figure 8.32 shows the interconnect structure of the wire bonding type of MCSP. The benefit of the MCSP with wire bonding process is to reduce the cost of manufacturing the MCSP. However, the parasitic electrical performance of the wire bond type of MCSP is definitely not good as the

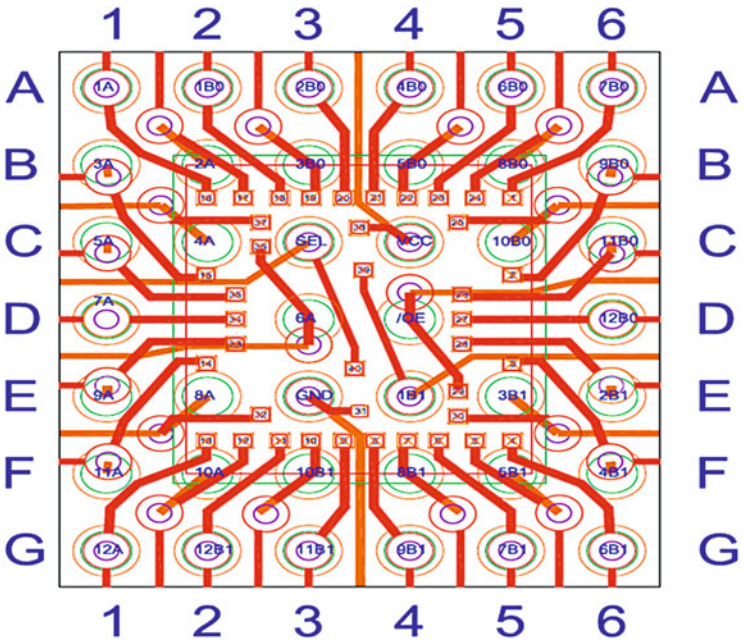


Fig. 8.26 The pin map distribution and RDL of the MCSP with GGI

Table 8.6 The pin map list of the 40 pin MCSP between die and package

Die Pin	Signal	Package Pin
1	7B0	A6
2	9B0	B6
3	4B1	F6
4	6B1	G6
5	5B1	F5
6	7B1	G5
7	8B1	F4
8	9B1	G4
9	11B1	G3
10	10B1	F3
11	12B1	G2
12	10A	F2
13	12A	G1
14	11A	F1
15	3A	B1

Die Pin	Signal	Package Pin
16	1A	A1
17	2A	B2
18	1B0	A2
19	3B0	B3
20	2B0	A3
21	4B0	A4
22	5B0	B4
23	6B0	A5
24	8B0	B5
25	10B0	C5
26	11B0	C6
27	12B0	D6
28	2B1	E6
29	/OE	D4
30	3B1	E5

Die Pin	Signal	Package Pin
31	GND	E3
32	8A	E2
33	9A	E1
34	7A	D1
35	5A	C1
36	6A	D3
37	4A	C2
38	VCC	C4
39	1B1	E4
40	SEL	C3

GGI type of MCSP. Figure 8.33 shows the comparison of the parasitic electronic resistance between GGI and wire bond type of MCSPs. Figure 8.34 shows the comparison of the parasitic electronic inductance between GGI and wire bond type of MCSPs. From both Figs. 8.33 and 8.34, it can be seen that the parasitic electrical

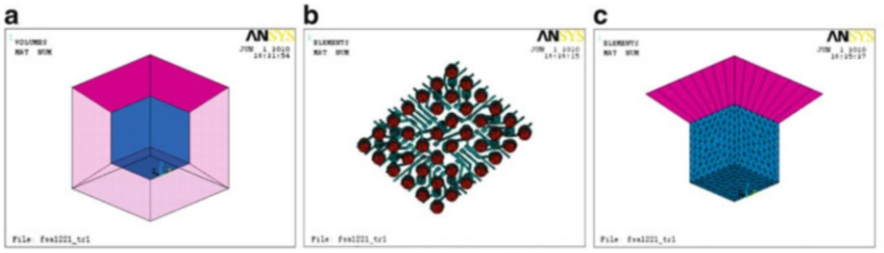


Fig. 8.27 FEA models of 40 pins MCSP for resistance and inductance: (**a**) Air and infinite volume (**b**) 40 bumps with RDL (**c**) partial mesh

Table 8.7 Material electrical properties

Material	Copper	Au	Solder ball	Air	Infinite b.c.
Material type	–	Gold bump	SAC305	–	–
Resistivity (Ohm-m)	1.73e-8	3.02e-8	1.3e-7	–	–
Permeability	1	1	1	1	1

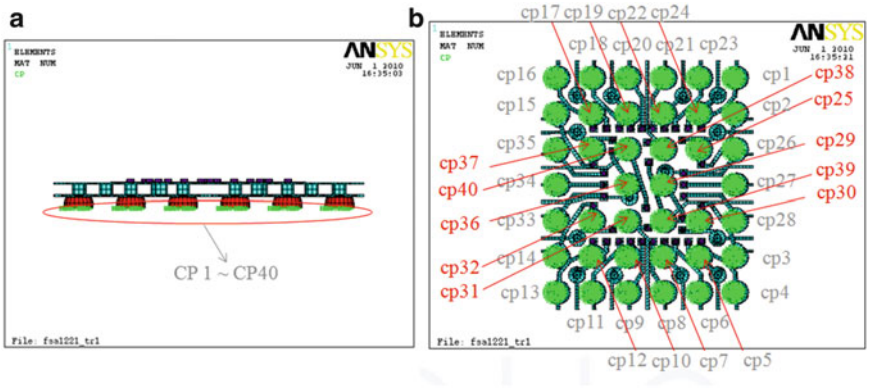


Fig. 8.28 The electrical boundary conditions of the 40 pin MCSP for resistance and inductance. (**a**) Coupling pin 1-pin 40 (**b**) Bottom view of coupled 40 pins

resistance and inductance of GGI type MCSP are much lower than that of the wire bond type of MCSP.

8.3 Electromigration Prediction and Test for 0.18 μm Wafer Level Power Technology

This section investigates the electromigration prediction and test for a 0.18 μm power technology in a wafer level reliability interconnect structure. The driving force for electromigration induced failure considered here includes the electron wind force, stress gradients, temperature gradients, as well as the atomic density

Table 8.8 The parasitic resistance and inductance of the 40 pins MCSP with GGI (1 MHz)

pin#	R	Lii
	mOhm	nH
1	17.5	0.623
2	16.9	0.595
3	16.9	0.594
4	17.8	0.630
5	13.5	0.351
6	15.0	0.551
7	11.5	0.293
8	13.6	0.501
9	13.6	0.500
10	11.5	0.294
11	15.0	0.552
12	13.5	0.354
13	17.8	0.629
14	16.9	0.594
15	16.9	0.595

pin#	R	Lii
	mOhm	nH
16	17.5	0.623
17	13.5	0.351
18	15.0	0.551
19	12.2	0.308
20	13.6	0.501
21	13.6	0.500
22	12.2	0.307
23	15.0	0.551
24	13.5	0.355
25	9.8	0.301
26	14.4	0.588
27	12.1	0.500
28	14.4	0.589
29	9.8	0.408
30	9.9	0.298

pin#	R	Lii
	mOhm	nH
31	3.7	0.170
32	10.3	0.299
33	13.1	0.509
34	10.7	0.422
35	13.0	0.509
36	9.8	0.408
37	9.8	0.298
38	3.6	0.164
39	11.9	0.579
40	11.8	0.577

gradient. Both the electromigration prediction and test for chemical–mechanical planarization (CMP) and non-CMP power devices are investigated. Parameters of different barrier metal thicknesses are studied. The results showed that the predicted electromigration mean time to failure (MTTF) is well correlated with the experimental test data of the 0.18 μm power technology.

8.3.1 Introduction

Electromigration (EM) is a phenomenon of mass transport in metallization structures when a high electrical current density is applied. It can cause progressive damage to metal interconnects in a power integrated circuit (IC). Usually, void nucleation near cathode side and hillock development near the anode side during current stressing indicates a biased mass diffusion from cathode to anode.

A lot of efforts have been made to develop the computational EM models [1–3]. The migration driving forces include the electron-wind force induced migration (EWM), the temperature gradient induced migration (TM), and the stress gradient induced migration (SM). Tan et al. [4, 5] indicated that the traditional atomic flux divergence (AFD) formulation is not accurate in the predicting void nucleation in a

Table 8.9 The mutual inductance of the 40 pin MCSP with GGI

pin#	pin#	Lij	Kij
i	j	nH	--
1	2	0.126	0.21
1	24	0.086	0.18
1	25	0.025	0.06
2	24	0.043	0.09
2	25	0.023	0.05
2	26	0.185	0.31
2	29	0.017	0.03
3	4	0.128	0.21
3	5	0.043	0.09
3	28	0.186	0.31
3	30	0.023	0.05
3	39	0.046	0.08
4	5	0.084	0.18
4	30	0.026	0.06
5	6	0.128	0.29

pin#	pin#	Lij	Kij
i	j	nH	--
5	7	0.030	0.09
5	30	0.026	0.08
6	7	0.064	0.16
7	8	0.079	0.21
7	31	0.007	0.03
8	9	0.163	0.32
8	31	-0.002	-0.01
9	10	0.079	0.21
9	31	0.005	0.02
10	11	0.064	0.16
10	12	0.030	0.09
10	31	0.010	0.04
11	12	0.128	0.29
12	13	0.083	0.18
12	14	0.043	0.09

pin#	pin#	Lij	Kij
i	j	nH	--
12	32	0.026	0.08
13	14	0.128	0.21
13	32	0.026	0.06
14	32	0.024	0.06
14	33	0.177	0.32
15	16	0.126	0.21
15	17	0.043	0.09
15	35	0.177	0.32
15	37	0.022	0.05
15	40	0.045	0.08
16	17	0.086	0.18
16	37	0.025	0.06
17	18	0.128	0.29
17	19	0.031	0.09

pin#	pin#	Lij	Kij
i	j	nH	--
17	37	0.026	0.08
18	19	0.066	0.16
19	20	0.078	0.20
19	38	0.007	0.03
20	21	0.163	0.32
20	38	-0.002	-0.01
21	22	0.078	0.20
21	38	0.005	0.02
22	23	0.066	0.16
22	24	0.031	0.09
22	38	0.010	0.04
23	24	0.128	0.29
24	25	0.026	0.08
25	26	0.034	0.08
25	38	0.025	0.11

pin#	pin#	Lij	Kij
i	j	nH	--
26	27	0.200	0.37
26	29	-0.003	-0.01
27	28	0.200	0.37
27	29	-0.018	-0.04
27	39	0.024	0.04
28	29	-0.035	-0.07
28	30	0.034	0.08
28	39	0.039	0.07
29	30	0.007	0.02
29	38	0.007	0.03
29	39	-0.106	-0.22
30	39	0.014	0.03
31	32	0.025	0.11
31	36	0.007	0.03
31	39	-0.003	-0.01

pin#	pin#	Lij	Kij
i	j	nH	--
32	33	0.031	0.08
33	34	0.159	0.34
33	36	0.002	0.01
34	35	0.159	0.34
34	36	-0.011	-0.03
34	40	0.020	0.04
35	37	0.031	0.08
35	40	0.035	0.06
36	37	0.007	0.02
36	40	-0.109	-0.22
37	40	0.014	0.03
38	40	-0.003	-0.01
39	40	-0.179	-0.31

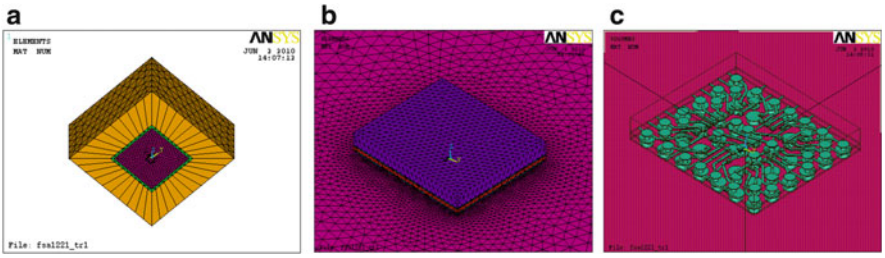


Fig. 8.29 FEA models of 40 pin MCSP for capacitance simulation. (**a**) Mesh of air and infinite volume (**b**) 40 pin MCSP (**c**) RDL and bumps

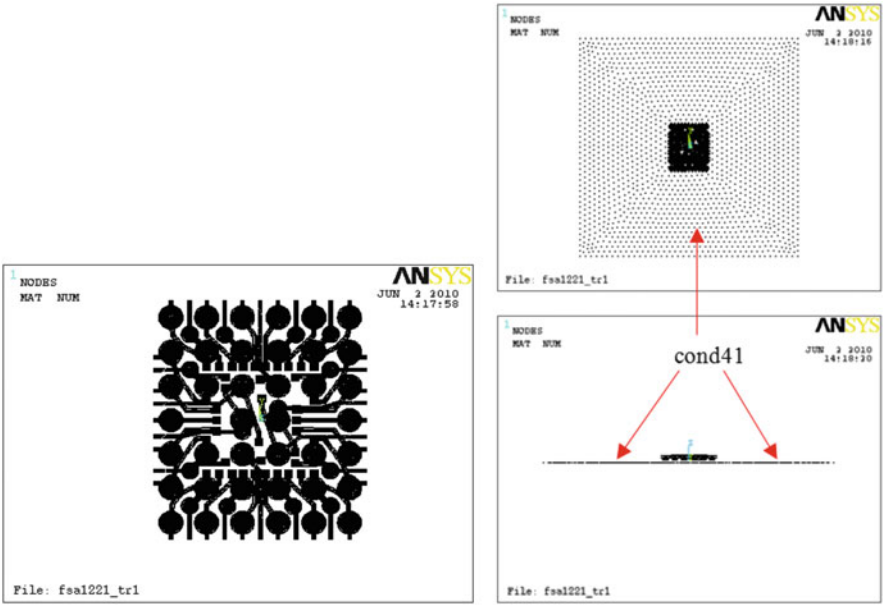


Fig. 8.30 The components cond1-cond40 (pin 1-pin 40) and cond41

Table 8.10 The material properties for capacitance simulation

Material	conductor	EMC	BT core	Solder mask	FR4	air	Infinite boundary
Permittivity	1	3.5	4.7	4.3	4.2	1	1
Type	–	CEL9700ZHF 10(V80,c30)	MCL-E-679FG	AUS320	–	–	–

Table 8.11 The parasitic self-capacitance of 40 bumps of MCSP with GGI

pin#	Cii
	pF
1	0.054
2	0.047
3	0.047
4	0.053
5	0.047
6	0.046
7	0.045
8	0.044
9	0.044
10	0.046
11	0.047
12	0.046
13	0.053
14	0.048
15	0.046

pin#	Cii
	pF
16	0.054
17	0.046
18	0.044
19	0.047
20	0.044
21	0.044
22	0.045
23	0.046
24	0.046
25	0.048
26	0.046
27	0.045
28	0.046
29	0.051
30	0.045

pin#	Cii
	pF
31	0.047
32	0.049
33	0.048
34	0.047
35	0.047
36	0.051
37	0.046
38	0.047
39	0.051
40	0.050

very thin film structure. Therefore, they proposed a modified atomic flux divergence formulation. In reality, the atomic mass transport is caused by a combination of interacting driving forces, which can generate voids at different locations. These driving forces are induced from different physical phenomena such as momentum exchange with current carriers (electron wind), temperature gradients, mechanical stress gradients, and atomic density gradient (or more general, of the chemical potential) [6]. However, the traditional AFD method that neglects the effect of the atomic density gradient (ADG) has induced larger error in modeling [7]. In this section, first we will investigate the electromigration prediction method with consideration of the ADG. Then, the wafer level electromigration test with different SWEAT (Standard Wafer-Level Electromigration Acceleration Test) structure layouts for 0.18 μm power technology is studied for different failure modes and time to failures. The detail EM test is arranged as follows: An ALSICu line with different thickness of TiN/Ti barrier metals on different topography of oxide/TEOX is prepared to enable the investigation of the effect of mechanical stress and barrier metal thickness and coefficient of thermal expansion (CTE) mismatch for 0.18 μm power IC. The EM tests are carried out on SWEAT structures by using JEDEC

Table 8.12 The parasitic mutual capacitance of 40 bumps of MCSP with GGI

pin#	pin#	Cij
i	j	pF
1	2	0.028
1	24	0.074
1	25	0.022
2	24	0.011
2	25	0.064
2	26	0.045
2	29	0.010
3	4	0.031
3	5	0.011
3	28	0.045
3	30	0.062
3	39	0.013
4	5	0.070
4	30	0.023
5	6	0.067

pin#	pin#	Cij
i	j	pF
5	7	0.012
5	30	0.023
6	7	0.057
7	8	0.062
7	31	0.013
8	9	0.026
8	31	0.036
9	10	0.063
9	31	0.037
10	11	0.058
10	12	0.012
10	31	0.021
11	12	0.068
12	13	0.071
12	14	0.011

pin#	pin#	Cij
i	j	pF
12	32	0.022
13	14	0.031
13	32	0.023
14	32	0.064
14	33	0.046
15	16	0.028
15	17	0.011
15	35	0.045
15	37	0.061
15	40	0.013
16	17	0.074
16	37	0.022
17	18	0.066
17	19	0.015

pin#	pin#	Cij
i	j	pF
17	37	0.023
18	19	0.066
19	20	0.058
19	38	0.013
20	21	0.026
20	38	0.036
21	22	0.058
21	38	0.037
22	23	0.066
22	24	0.015
22	38	0.021
23	24	0.066
24	25	0.023
25	26	0.026
25	38	0.014

pin#	pin#	Cij
i	j	pF
26	27	0.037
26	29	0.054
27	28	0.038
27	29	0.029
27	39	0.019
28	29	0.015
28	30	0.024
28	39	0.038
29	30	0.013
29	38	0.020
29	39	0.061
30	39	0.021
31	32	0.014
31	36	0.019
31	39	0.016

pin#	pin#	Cij
i	j	pF
32	33	0.026
33	34	0.034
33	36	0.044
34	35	0.035
34	36	0.028
34	40	0.019
35	37	0.025
35	40	0.037
36	37	0.014
36	40	0.063
37	40	0.021
38	40	0.015
39	40	0.029

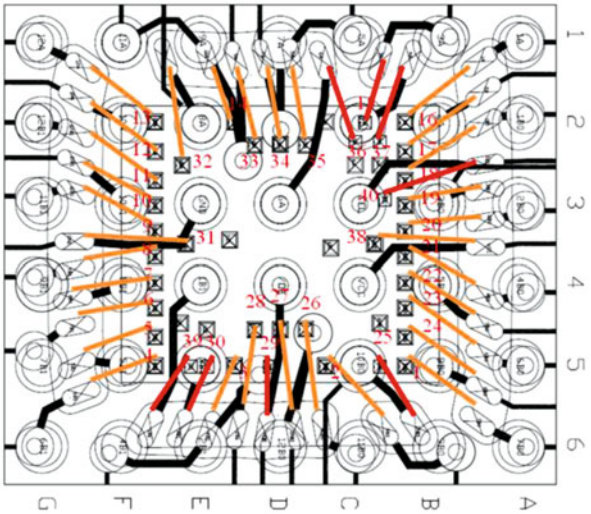


Fig. 8.31 The pin map distribution and RDL of the 40 pin MCSP with wire bond

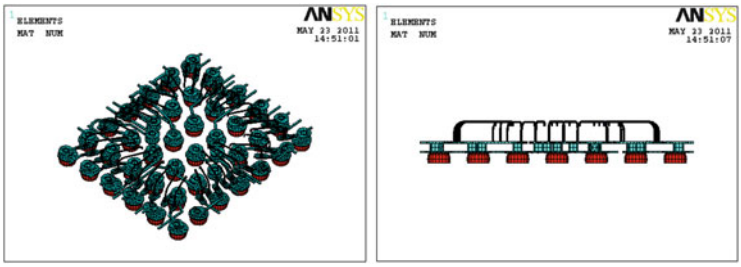


Fig. 8.32 The interconnects of 40 pin MCSP with wire bonding

standard stress methodology [12]. Finally, the EM test data are correlated with the predicted MTTF. Through this study, we will have a better understanding of electromigration on 0.18 μm wafer level power interconnects.

8.3.2 Electromigration Model Formulation

Electromigration is a diffusion controlled mass transport process in a interconnect structure. The time-dependent evolution equation of the local atomic density caused by an applied current is the mass balance (continuity) equation:

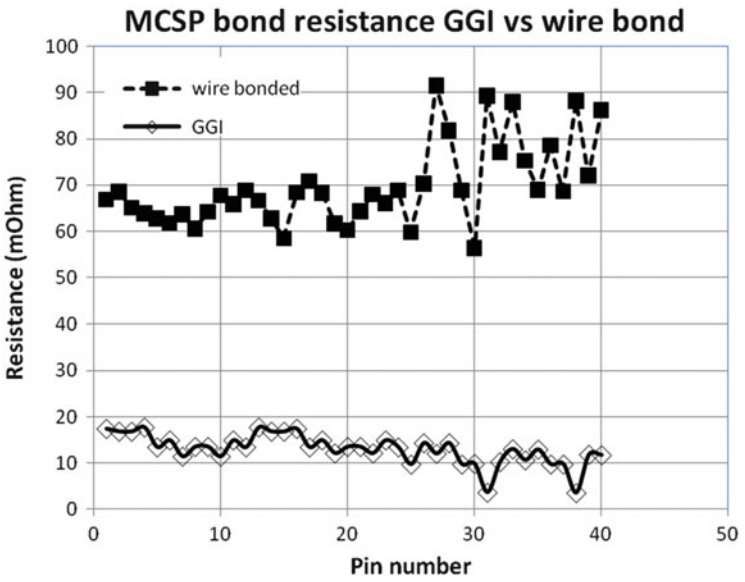


Fig. 8.33 The parasitic resistance of GGI versus wire bonding for 40 pin MCSP (1 MHz)

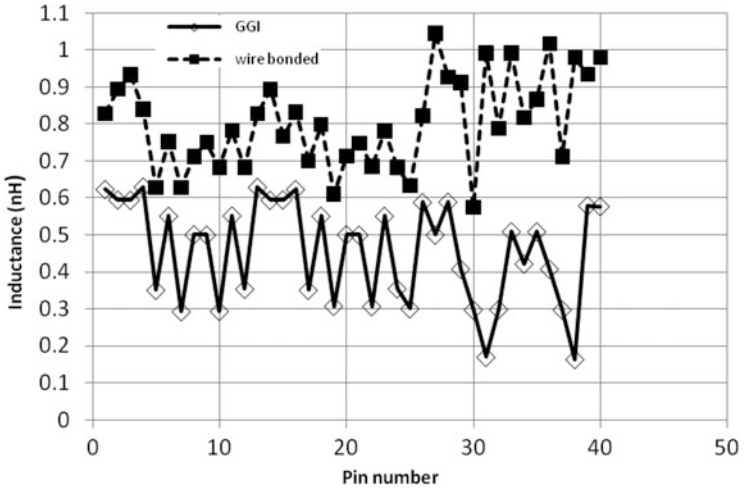


Fig. 8.34 The parasitic inductance of GGI vs. wire bonding for 40 pin MCSP (1 MHz)

$$\nabla \cdot q + \frac{\partial c}{\partial t} = 0 \tag{8.11}$$

where c is the normalized atomic density, $c = N/N_0$, N is the actual atomic density and N_0 is the initial (equilibrium state) atomic density in the absence of a stress

field, and t is the time; q is the total normalized atomic flux.

Considering that the driving forces of atomic flux include the electron wind force, the thermal gradient driving force, the hydrostatic stress gradient driving force, and the atomic density gradient driving force, respectively, the normalized atomic flux can be written as [7, 8]

$$\begin{aligned}\vec{q} &= \vec{q}_{\mathrm{ew}} + \vec{q}_{\mathrm{Th}} + \vec{q}_{\mathrm{S}} + \vec{q}_{\mathrm{c}} = \frac{cD}{kT} Z^* e\rho \vec{j} - \frac{cD}{kT} Q^* \frac{\nabla T}{T} \\ &- \frac{cD}{kT}\Omega \nabla \sigma_{\mathrm{m}} - D\nabla c = c \cdot F\left(T, \sigma_{\mathrm{m}}, \vec{j}, \cdots\right) - D\nabla c\end{aligned} \tag{8.12}$$

where

$$F\left(T, \sigma_{\mathrm{m}}, \vec{j}, \cdots\right) = \frac{D}{kT} Z^* e\rho \vec{j} - \frac{D}{kT} Q^* \frac{\nabla T}{T} - \frac{D}{kT}\Omega \nabla \sigma_{\mathrm{m}} \tag{8.13}$$

where k is Boltzmann's constant; e is the electronic charge; Z^* is the effective charge which is determined experimentally; T is the absolute temperature; ρ is the resistivity which is calculated as $\rho = \rho_0(1 + \alpha(T - T_0))$, where α is the temperature coefficient of the metallic material and ρ_0 is the resistivity at T_0; $\vec{j}$ is the current density vector; Q^* is the heat of transport; Ω is the atomic volume; $\sigma_m = (\sigma_1 + \sigma_2 + \sigma_3)/3$ is the local hydrostatic stress, where σ_1, σ_2, σ_3 are the components of principal stress; D is the effective atom diffusivity, $D = D_0 \exp\left(-\frac{E_a}{kT}\right)$, where E_a is the activation energy and D_0 is the effective thermally activated diffusion coefficient.

For the EM evolution equation (8.11) on any enclosed domain V with the corresponding boundary Γ, the atomic flux boundary conditions of a metal interconnect can be expressed as

$$q \cdot n = q_0 \quad \text{on } \Gamma \tag{8.14}$$

For blocking boundary condition,

$$q \cdot n = 0 \quad \text{on } \Gamma \tag{8.15}$$

At the initial time, the normalized atomic density for all nodes is assumed to be

$$c_0 = 1 \tag{8.16}$$

All abovementioned driving forces for atomic transport are simultaneously and self-consistently taken into account in the EM model in order to adequately describe the continuous atom redistribution and to capture the realistic kinetics of void nucleation and growth as a function of the interconnect architecture, segment geometry, material properties, and stress conditions.

The electromigration phenomenon is a complex physics coupling problem which contains thermoelectric coupling, thermo-mechanical coupling, and mass diffusion. In order to predict electromigration failure, the indirect coupled analysis of thermal–electric–structural fields based on ANSYS® is studied. The EM void evolution simulation consists of the simulation part for an incubation period and a void growth period. In the simulation for the incubation period, at first, the initial distributions of current density and temperature in the interconnect structure are obtained by the 3D finite element method analysis based on the ANSYS platform. Then, atomic density redistribution in the interconnect structure is solved based on with a user-defined FORTRAN code. The atomic density redistribution algorithm and its computation procedure can be found in our previous works [7].

8.3.3 Wafer Level Experimental Test for Electromigration

A wafer level electromigration test with different SWEAT layouts for 0.18 μm power technology is performed. The detailed EM test in SWEAT layout is arranged as follows: The metal line consisting of AlSiCu with different thicknesses of TiN/Ti barrier metals on differing topographies of oxide/TEOX is prepared to enable the investigation of the effect of electrical performance for a 0.18 μm power IC. The EM tests are carried out in two structures (CMP and non-CMP) by using the SWEAT stress methodology of the JEDEC standard [12]. This methodology involves forcing a fixed electric current into a metal line and measuring the change in the line resistance so as to get feedback on the temperature of the line through temperature coefficient of resistance (TCR) and the time to failure (TTF). The temperature measured by the change in resistance of the metal line is the average temperature of the line. The test uses this information to force a constant stress condition with stress being calculated using the current density and temperature terms from the Black equation to meet the acceleration factor. The acceleration factor is defined by

$$(t_{50\text{use}} / t_{50\text{stress}}) = (j_{\text{stress}}/j_{\text{use}})^n \, \mathrm{e}^{Ea/k\,(1/T_{\text{use}} - 1/\,T_{\text{stress}})} \tag{8.17}$$

where

$t_{50\text{use}}$ is MTTF at use condition,
$t_{50\text{stress}}$ is MTTF at stress condition,
j_{stress} and j_{use} are the current density in test,
k is Boltzmann's constant,
E_a is the activation energy,
n is the current density exponent,
T_{use} and T_{stress} are the absolute temperature at use condition and stress condition, respectively

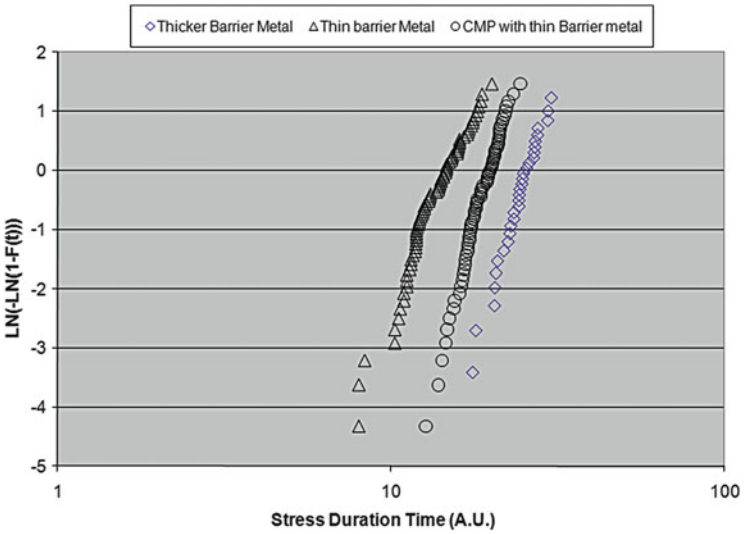


Fig. 8.35 The wafer level EM test TTF data comparison with CMP, non-CMP, and different barrier metals

The typical range for the acceleration factor is 10^5–10^9. It is very important to determine the acceleration factor. Establishing a suitable acceleration level usually requires experiments to sacrifice a few test structures. A failure condition is specified by a percent change in resistance. The test continues until the failure condition is met or the test time exceeds some specified maximum stress time.

Figure 8.35 shows the TTF of wafer level tests for CMP and non-CMP metal lines with different barrier metals. Figure 8.36 shows the void observed in the EM test for a metal line of the SWEAT structure.

8.3.4 Finite Element Simulation

8.3.4.1 Finite Element Model

The three-dimensional finite element model for a SWEAT structure after CMP process is shown in Fig. 8.37. The global thermal-electric coupled field model uses Solid-69 element and the global stress model uses Solid-45 element. The linear regular hexahedral elements were map-meshed, which can also save computing time and improve the accuracy. Due to symmetry, only a half of the structure is modeled. The related thermal, mechanical, electrical, and electromigration parameters of SWEAT structure are listed in Tables 8.13 and 8.14.

Three FE models with different TiN/Ti barrier thickness spanning from 0 to 600 A are constructed. These models are subjected to a current density of 16 MA/cm^2. In addition, the structures are considered to be stress-free at 400 °C

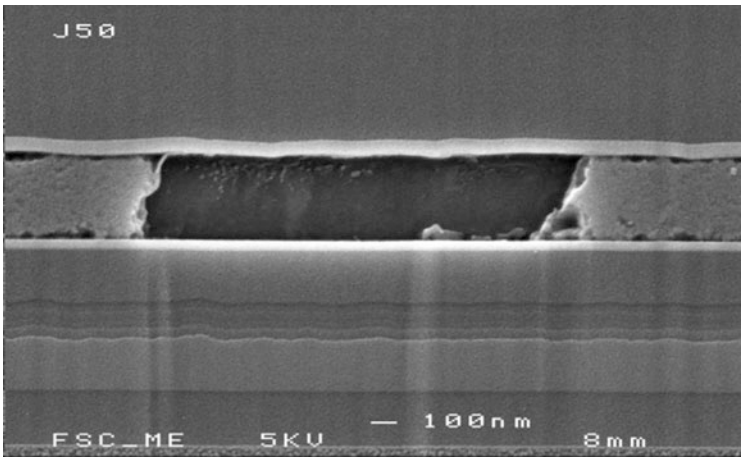


Fig. 8.36 The void observed in EM test for the metal line at cathode side

(fabrication process temperature) for the simulations. The initial atomic density for all nodes is set at $c_0 = 1$.

For the model with 300 A TiN/Ti, Fig. 8.38 shows the temperature distribution and current density distribution at initial time. Due to Joule heat, the maximum temperature occurs in the middle segment of the AlSiCu line. Therefore, the atom diffuses rapidly in the middle segment where it easily induces void. Figure 8.39 shows the hydrostatic stress distribution at room temperature and initial time of stressing current load. The hydrostatic stress releases due to the raising of temperature from room temperature to initial time of stressing current load.

Assume no EM void produces in the analysis (Static analysis). Figure 8.40 shows the normalized atomic density distribution of the AlSiCu line at different times. From Fig. 8.40a, at 10 s, we can observe that the minimum normalized atomic density locates the zone of minimum stresses where the stress gradient is very large (see Fig. 8.39b). From Fig. 8.40c, at 40 s, the zone with minimum normalized atomic density has expanded the whole quarter segment of Al metallization from left. These may be due to the hydrostatic stress that dominates the EM diffusion at the beginning, while with the increment of time, current density, and atomic density gradient dominate the EM diffusion gradually. Furthermore, according to the normalized atomic density redistribution, the void formation can be simulated, as seen in Fig. 8.41. The result is consistent with the picture observed in the experiment, as shown in Fig. 8.36 at cathode side.

To study the impact of EM with and without considering the atomic density gradient, we examined the normalized atomic density redistribution in a node (number 1512) of metal line. Figure 8.42 shows the comparison of normalized atomic density distribution with and without considering $\vec{q}_c$ in AlSiCu line, where

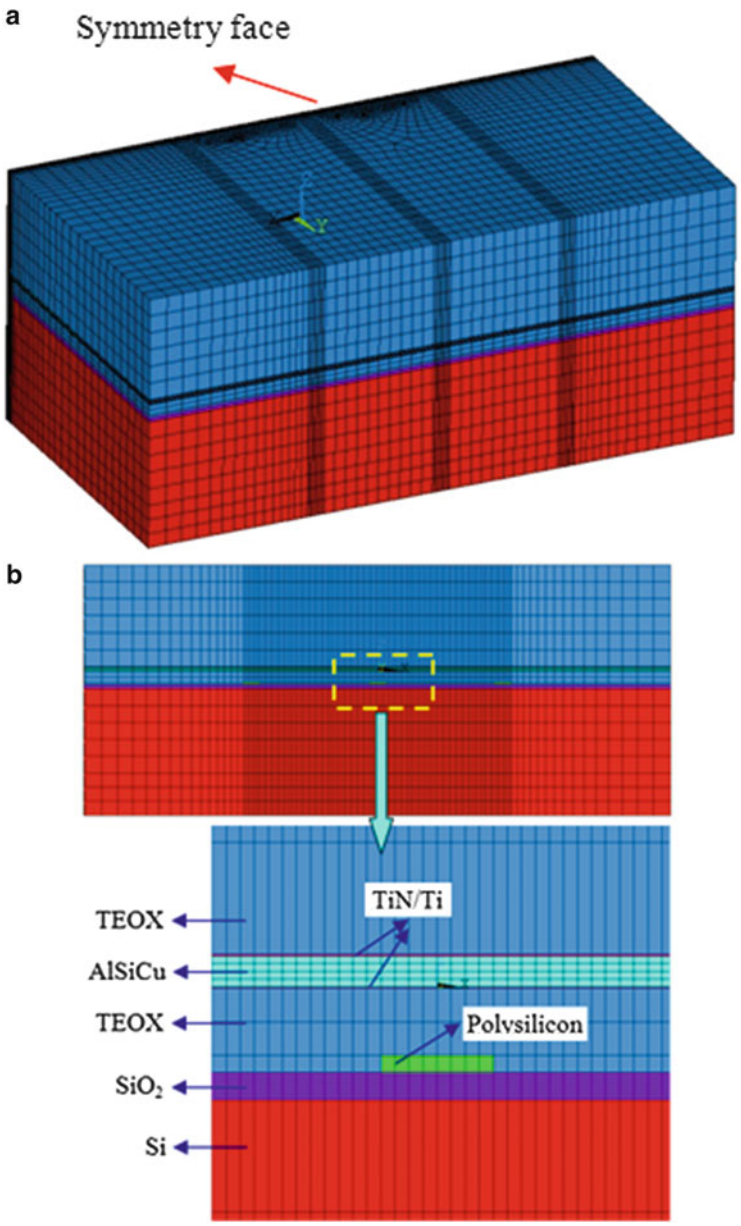


Fig. 8.37 The SWEAT structure after CMP process (**a**) Whole model (**b**) The cross section

Table 8.13 Material properties for SWEAT structure [7–11]

Material		AlSiCu	SiO_2	Si
Elastic modulus (GPa)		50	71	130
Poisson's ratio		0.30	0.16	0.28
Thermal conductivity (W/m · K)		100	1.75	80
Electrical resistivity (Ω · m)	200 K	2.139E-8	1E10	4.4
	800 K	9.194E-8		
CTE ($\times 10^{-6}$ 1/K)	200 K	20.3	0.348	2.24
	300 K	23.23	0.498	2.64
	400 K	25.1	0.61	3.2
	500 K	26.4	0.63	3.5
	600 K	28.4	0.59	3.7
	700 K	30.9	0.53	3.9
	800 K	34	0.47	4.1
Yield stress (MPa)	293 K	190	/	/
	800 K	12.55		
Material		TiN/Ti	TEOX	Polysilicon
Elastic modulus (GPa)		80.6	59	170
Poisson's ratio		0.208	0.24	0.22
Thermal conductivity (W/m · K)		26.1	2	12.5
Electrical resistivity (Ω · m)		110E-6	1E10	1.75E-5
CTE ($\times 10^{-6}$ 1/K)		9.35	1	9.4E-6
Yield stress (MPa)		/	/	/

Table 8.14 Electromigration parameters of AlSiCu [1]

Parameter	Symbol	Value
Activation energy (eV)	E_a	0.9
Effective charge number	Z*	−14
Effective self-diffusion coefficient (m^2/s)	D_0	5E-8
Heat of transport (eV)	Q^*	−0.08
Atom volume (m^3/atom)	Ω	0.166055E-28
Electrical resistivity (Ω · m)	ρ	See Table 8.1

the examined node 1512 is listed in Fig. 8.42a. It can be seen that, without considering $\vec{q}_c$, the normalized atomic density decreases linearly and rapidly with time. When $\vec{q}_c$ is considered, the normalized atomic density varies slowly with time. It means that the atomic density will be retarded due to the effect of atomic density gradient in the time-dependent EM evolution. This will delay the void generation and growth, and increase TTF. Therefore, using the time-dependent EM

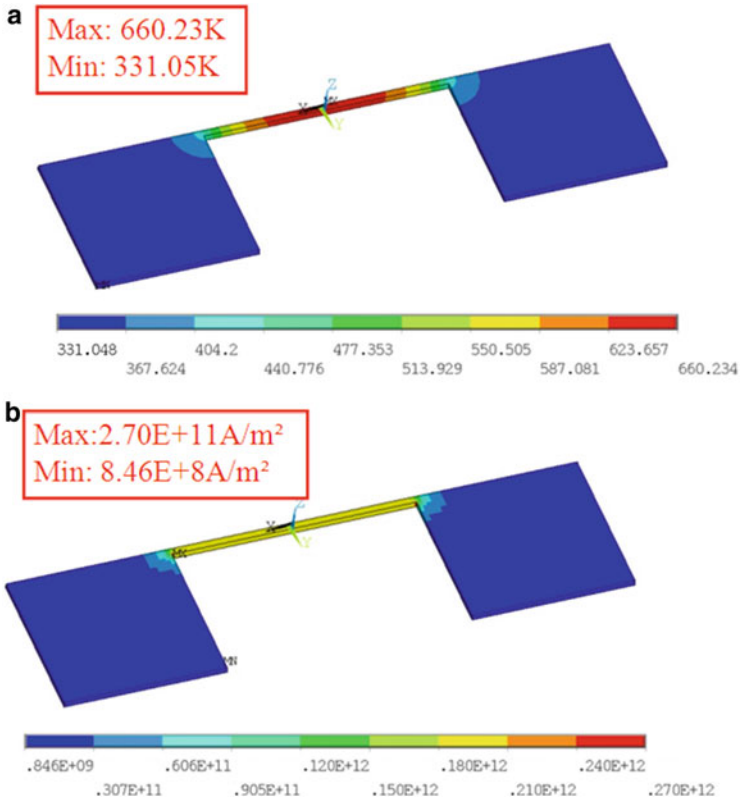


Fig. 8.38 Temperature and current density distributions at initial time (**a**) Temperature distribution (**b**) Current density distribution

evolution equation without considering $\vec{q}_c$ will underestimate the EM failure of the AlSiCu line.

8.3.4.2 The Effect of CMP and Non-CMP Process on TTF

To study the effect of CMP and non-CMP process, we developed the non-CMP SWEAT model as shown in Fig. 8.43. Figure 8.44 shows void formation of a non-CMP SWEAT structure. Figure 8.45 gives both the test and modeling comparisons of the MTTFs of the SWEAT structures with CMP and non-CMP processes. The simulation results agree with the test data quite well. The wafer level EM test data showed that the MTTF has improved by 35 % CMP process-metal line on a flat oxide/TEOX layer.

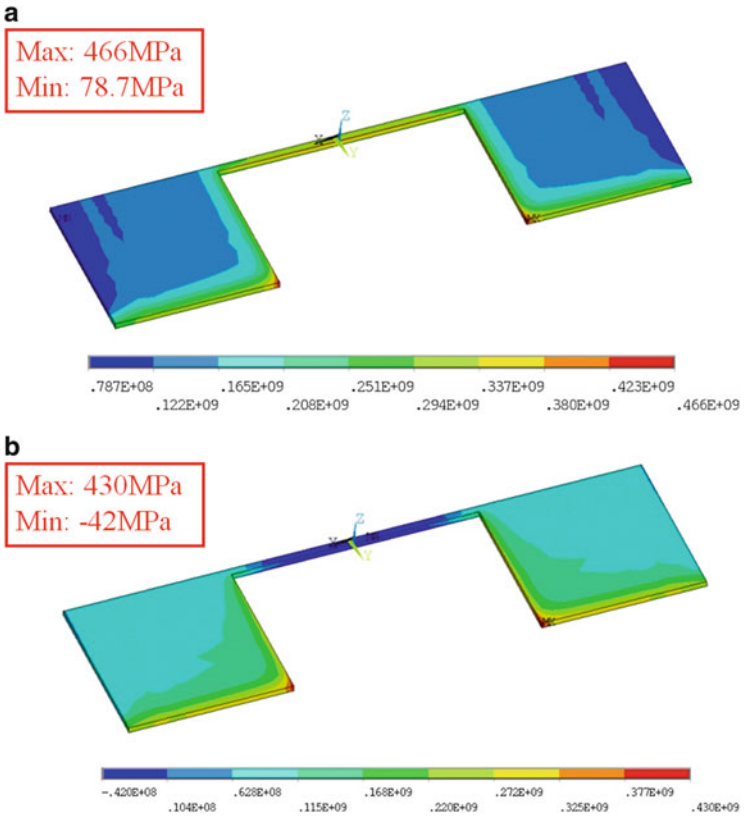


Fig. 8.39 Hydrostatical stress distribution at room temperature and initial time of stressing current load (**a**) At room temperature (**b**) At initial time of stressing current load

8.3.4.3 The Effect of TiN/Ti Thickness on TTF

A comparison of the simulation results for MTTF with different thickness of TiN/Ti for a SWEAT structure with CMP process and the experimental test results is presented in Fig. 8.46. In Fig. 8.46, the test data shows that the MTTF has been increased by 30 % on metal line with twice TiN/Ti thicker under the same current density and temperature conditions. The longer MTTF in thicker barrier metal is due to the shunting of the stress current through the TiN layer when weak points develop in the AlSiCu metal line. The thicker the barrier metal, the more current should be shunted. In addition, the interface metal between the barrier metal and AlSiCu metal may also play an important role as the barrier metal thickness.

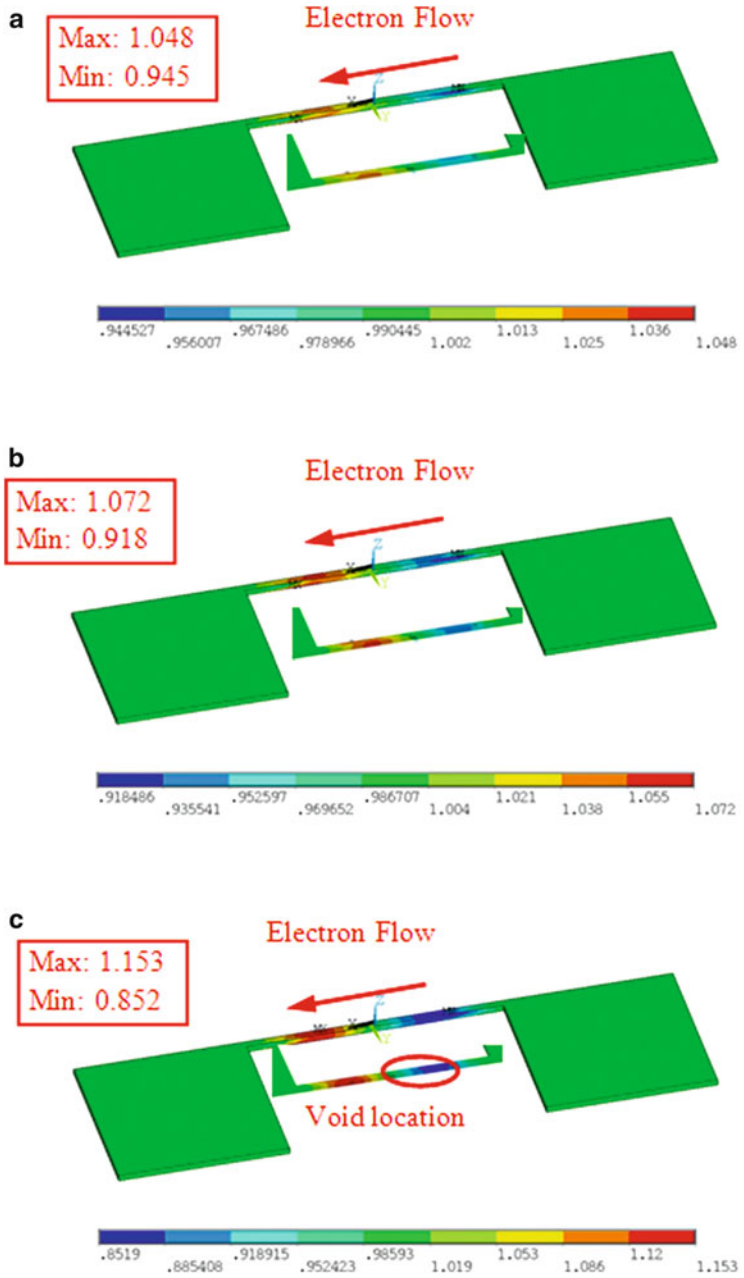


Fig. 8.40 Normalized atomic density distribution of the AlSiCu line at different time. (**a**) 10 s (**b**) 20 s (**c**) 50 s

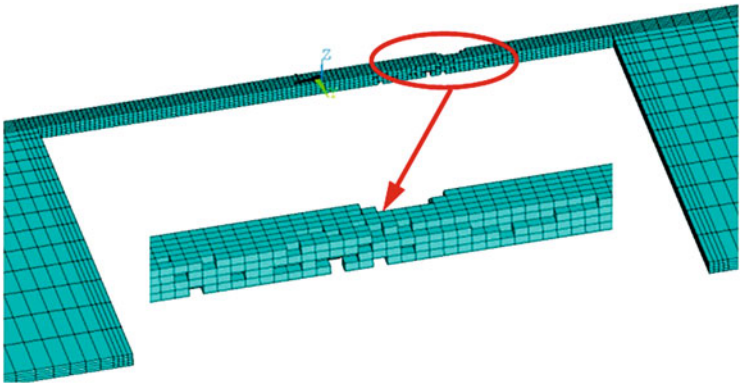


Fig. 8.41 Void formation of AlSiCu line at 18.8 s at cathode side

8.3.5 Discussion

In this section, both the electromigration prediction and test for chemical–mechanical planarization (CMP) and non-CMP 0.18 μm power devices are investigated. Parameters of different barrier metal thicknesses are studied. The simulation lists the effect comparison with and without consideration of the atomic density gradient. The results showed that the predicted electromigration MTTF are well correlated with the experimental test data. Both the test and the modeling results have disclosed the significant influence of CMP and the barrier metal thickness on the electromigration MTTF. The CMP process has improved MTTF by 35 % as compared to non-CMP. The thicker barrier metal layer (600 A) can improve MTTF by 30 % as compared to the thin barrier metal layer (300 A).

8.4 Modeling Microstructure Effects on Electromigration in Lead-Free Solder Joints

This section [13] studies the microstructure effects on electromigration in lead-free solder joints in wafer level chip scale package (WL-CSP). It is an extension of the original isotropic model [14]. The three-dimensional finite element model for solder joints with different grain structures is developed and analyzed in ANSYS. The sub-modeling technique is utilized to obtain more accurate simulation results in solder bumps. The indirect coupled analysis of electrical, thermal, and stress field is carried out. Four common microstructures of the solder bumps are modeled and anisotropic elastic, thermal, and diffusion property data are used. The results obtained from the four different microstructures are compared with each other. The microstructure effects on electromigration are drawn from the plots of the atomic flux divergence (AFD) and the time to failure (TTF) with respect to microstructure parameters.

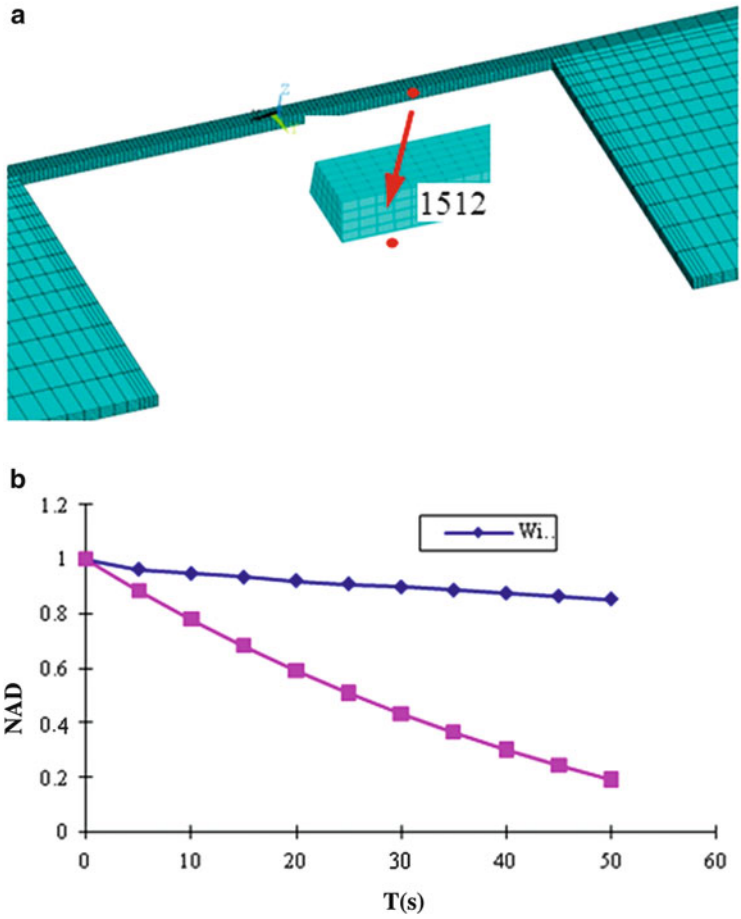


Fig. 8.42 Comparison of normalized atomic density distribution with and without $\vec{q}_c$ at an examined node (**a**) Nodal location (**b**) Comparison of normalized atomic density distribution with and without $\vec{q}_c$

8.4.1 Introduction

The microelectronics industry continues to push for higher performance and down-scaling of device dimensions, leading to an increase in current density that must be carried out by metallic interconnects and solder bumps. With increased current density, electromigration induced failure becomes a major reliability concern. Electromigration is a mass diffusion process attributed to momentum transfer from conducting electrons to atoms. This gradually leads to voids forming at the cathode, which causes an increase in resistance and may ultimately result in failure due to loss of connection. Many experiments and simulations have been done to study electromigration in the interconnects and solder bumps. Due to

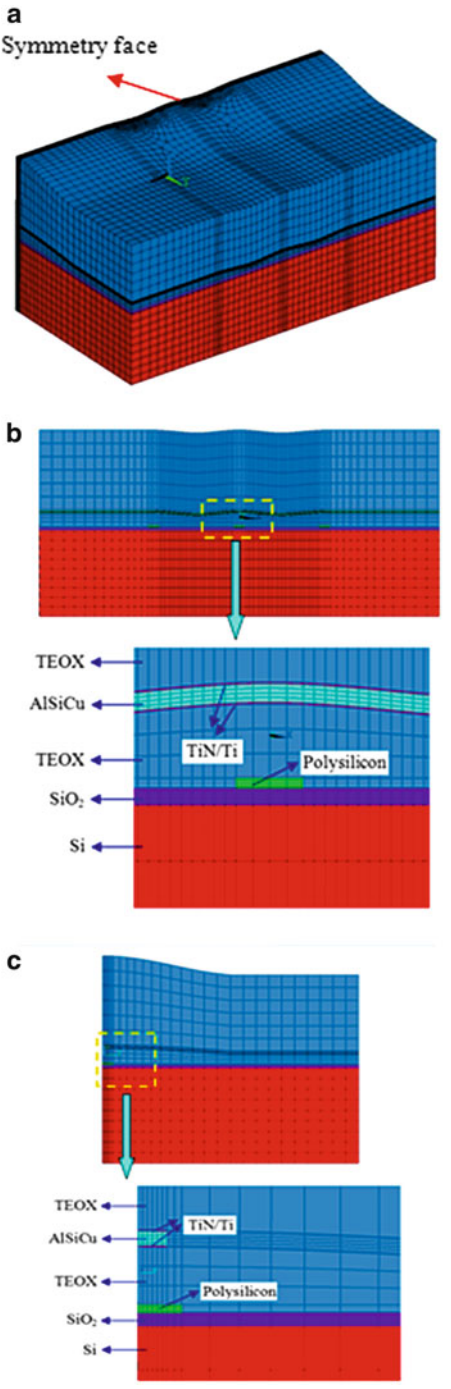


Fig. 8.43 SWEAT structure with non-CMP process (**a**) Whole model (**b**) Cross section long the metal line length (**c**) Cross section along metal line width (*Center*)

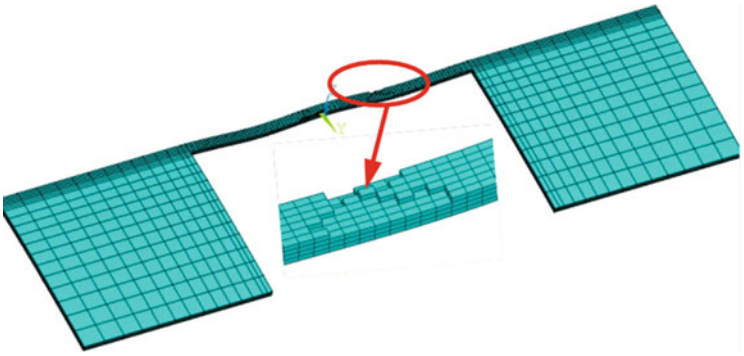


Fig. 8.44 Void formation of AlSiCu line in the SWEAT with non-CMP process at 13.9 s

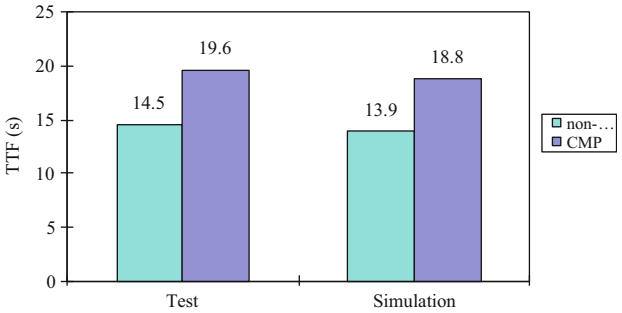


Fig. 8.45 The effect of non-CMP process on MTTF

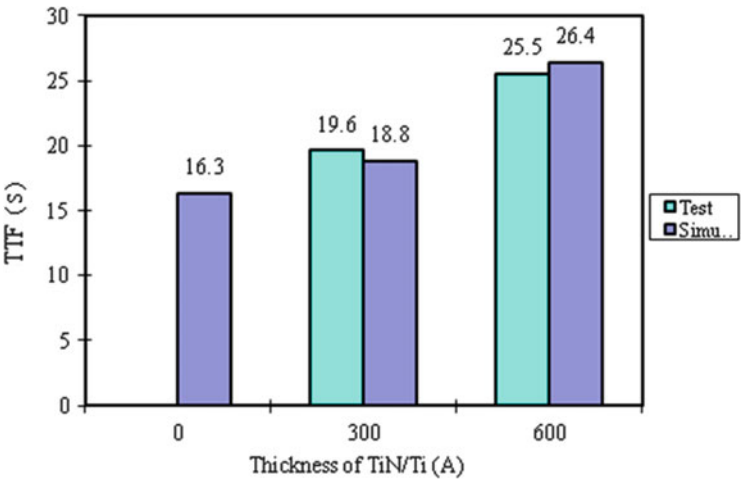


Fig. 8.46 The effect of TiN/Ti thickness on MTTF

environmental concerns, tin-based *solders have now replaced lead-based solders in microelectronic devices. SnAgCu is one of the most promising candidates because of its competitive price and good mechanical properties. The solder bumps in a wafer level chip scale package (WL-CSP) has a UBM on chip side and bump pad on substrate side and the complex geometry of the bumps leads to current crowding. The bumps are under the loads of nonuniform temperature and thermal stress which induce the thermal and mechanical forces that interact with the electrical force [15].

The near eutectic Sn–Ag–Cu solder joints are more than 95 at.% Sn and comprised of very limited β-Sn grains [16]. The small number of grains in lead-free solder joints is caused by the difficulty in β-Sn nucleation and the associated large undercooling during solidification. β-Sn has a body-centered-tetragonal (BCT) lattice structure, with parameters $a = b = 583$ pm and $c = 318$ pm. Direction [001] (side c) is almost half the length of the basal plane sides ([100] and [010]) which leads to anisotropic behaviors in mechanical, thermal, electrical, and diffusion properties. Therefore, the microstructure of Sn grains of the Sn–Ag–Cu solder joints has a significant effect on the electromigration process, and it has been reported that the anisotropy of the Pb-free Sn-based solders can lead to anomalous early failure [17]. However, very limited work has been done to study the microstructural effects on electromigration in lead-free solder joints. So the microstructural effects will be addressed in this paper.

The 3D electromigration model is further developed with consideration of the common microstructures of the lead-free solder joints in WL-CSP. The indirect electrical–thermal–structural coupled field analysis is carried out on ANSYS multiphysics simulation platform and the submodeling technique is utilized to get a more accurate simulation of the critical solder bump regions. Three mechanisms are considered in the calculation of the atomic flux divergence, i.e., the elctromigration, the thermomigration, and the stress migration. Four common microstructures are modeled and analyzed with anisotropic material properties. The ANAND viscoplastic material model is used. The distribution of current, temperature, thermal gradient, and hydrostatic stress is compared among those four microstructures. The atomic flux divergence is calculated and element birth/death function in ANSYS is used to show the void locations in the bumps. The effects of grain orientation and grain size on electromigration are investigated with the discussion of the simulation results.

8.4.2 Direct Integral Approach for Migration

If we neglect the effect of atomic density gradient and only consider the three driving forces for atomic migration which are namely electromigration, thermal gradients, and gradients of mechanical stress, the time-dependent evolution of the local atomic concentration N is given by the mass balance (continuity) Eq. (8.11) with the non-normalized atomic flux J:

$$\operatorname{div}\left(\vec{J}_{\mathrm{Tol}}\right)+\frac{\partial N}{\partial t}=0 \tag{8.18}$$

The divergences of atomic flux for electromigration, thermal migration, and stress migration can be expressed as

$$\operatorname{div}\left(\vec{J}_{\mathrm{Em}}\right)=\left(\frac{E_a}{kT^2}-\frac{1}{T}+\alpha\frac{\rho_0}{\rho}\right)\vec{J}_{\mathrm{Em}}\cdot\nabla T \tag{8.19}$$

$$\operatorname{div}\left(\vec{J}_{\mathrm{Th}}\right)=\left(\frac{E_a}{kT^2}-\frac{3}{T}+\alpha\frac{\rho_0}{\rho}\right)\vec{J}_{\mathrm{Th}}\cdot\nabla T+\frac{NQ^*D_0}{3k^3T^3}j^2\rho^2e^2\exp\left(-\frac{E_{\mathrm{a}}}{kT}\right) \tag{8.20}$$

$$\begin{aligned}\operatorname{div}\left(\vec{J}_{\mathrm{S}}\right)&=\left(\frac{E_{\mathrm{a}}}{kT^2}-\frac{1}{T}\right)\vec{J}_{\mathrm{S}}\cdot\nabla T\\&+\frac{2EN\Omega D_0\alpha_l}{3(1-\upsilon)kT}\exp\left(-\frac{E_{\mathrm{a}}}{kT}\right)\left(\frac{1}{T}-\alpha\frac{\rho_0}{\rho}\right)\nabla T\cdot\nabla T\\&+\frac{2EN\Omega D_0\alpha_l}{3(1-\upsilon)kT}\exp\left(-\frac{E_{\mathrm{a}}}{kT}\right)\frac{j^2\rho^2e^2}{3k^2T}\end{aligned} \tag{8.21}$$

where E is Young's Modulus, υ is the Poisson's ratio, and α_l is the coefficient of thermal expansion.

Then, the three divergence values are added, and the divergence of the total atomic flux can be expressed as:

$$\operatorname{div}\left(\vec{J}_{\mathrm{Tol}}\right)=N\cdot F\left(,\vec{j},T,\sigma_{\mathrm{m}},E_{\mathrm{a}},D_0,E,\cdots\right) \tag{8.22}$$

The above equation reveals that the divergence of atomic flux is proportional to the atomic concentration, and to a function F in which different physical parameters are included. In this way, Eq. (8.22) is equivalent to

$$NF+\frac{\partial N}{\partial t}=0 \tag{8.23}$$

The theoretical evolution of the atomic concentration can be obtained by

$$N=N_0e^{-F\Delta t} \tag{8.24}$$

From Eq. (8.24), we can get the expression for Δt as

$$\Delta t = -\frac{1}{F}\ln\left(\frac{N}{N_0}\right) \tag{8.25}$$

We assume that the element becomes a void when it reaches the criterion that the atomic concentration is 10 % of the initial concentration.

8.4.3 FEA Modeling of the Solder Bump Microstructure in a WLCSP

The WLCSP package in reference [18] is modeled in the commercial FEA software ANSYS. The whole structure has 36 solder bumps with 500 μm pitch. The exterior 20 solder bumps are assumed to connect with each other in a daisy chain. Due to the symmetry of the structure, a quarter of it is actually modeled. The submodeling technique in ANSYS is utilized to yield more accurate results in the critical solder bump regions. At first, a quarter global structure is modeled and analyzed with relatively coarse mesh as in Fig. 8.47. Then a refined submodel of solder bumps with UBM (Al/Ni/Cu) layer is modeled as in Fig. 8.48. The thermal electrical coupled simulation is carried out on the submodel to get the current density and temperature field followed by a structural submodel simulation to get the stress distribution. The viscoplastic ANAND constitutive material model is used here. Then the distribution of the atomic flux divergence can be calculated by adding Eqs. (8.19), (8.20), and (8.21) together. Since the highest divergence values correspond to the nucleation locations of voids, the 30 elements with the largest atomic flux divergence will be deleted using the live/death function in ANSYS. And the time by which the element will be deleted can be calculated from Eq. (8.25) by setting N/N_0 equal to 10 %. Then, the structure is automatically modified, analyzed, and calculated again, until the failure condition is reached. Figure 8.49 shows the flowchart of the analysis procedure.

8.4.3.1 Material Parameters

The anisotropic material property data of pure β-Sn is used here for the lead-free solder bump. The elastic behavior of β-Sn single crystal can be described by the engineering constants with $E_x = E_y = 76.20$ GPa, $E_z = 93.33$ GPa, $G_{xy} = 26.75$ GPa, $G_{yz} = G_{xz} = 2.56$ GPa, $\nu_{xy} = 0.473$, $\nu_{xz} = 0.170$, and $\nu_{yz} = 0.208$ [19]. Table 8.15 lists the viscoplastic ANAND model parameters for SnAgCu. The anisotropic thermal, electrical, and diffusion properties are listed in the Table 8.16.

8.4.3.2 Microstructure of the Solder Bump

Four common microstructures of the lead-free solders are modeled as in Fig. 8.50. These microstructures are based on observed microstructures in [24–26]. Figure 8.50a shows a bump of two grains with the grain boundary perpendicular to the substrate. The c direction of the left grain is almost parallel to the current direction, while the c direction of the right grain is perpendicular to that. Figure 8.50b shows a tri-grain bump and the disorientation angle between the grains is almost 60°. Figure 8.50c shows a bump of two grains with the grain boundary

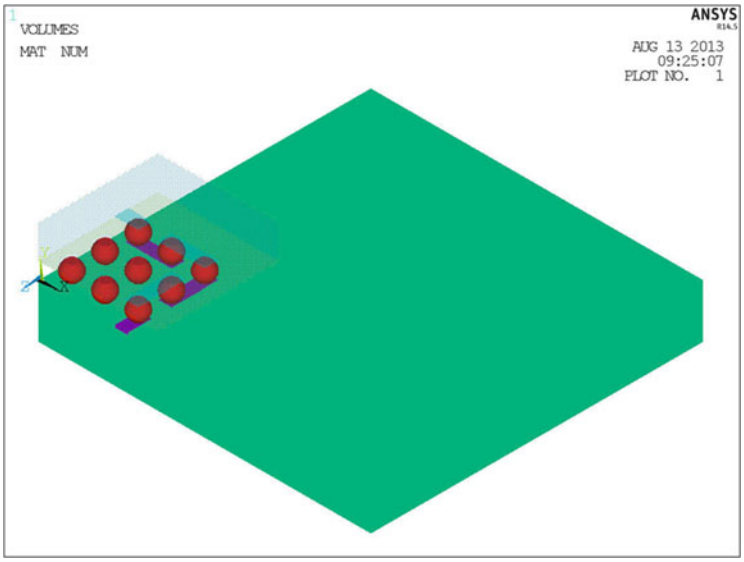


Fig. 8.47 Quarter global model

inclined at 45° to the substrate. Figure 8.50d shows a twining structure of 60° rotations with three dominant cyclic twin orientations cyclically repeating around the nucleus which is known as a Kara's beach ball structure.

8.4.4 Simulation Results and Discussion

Figure 8.51 shows the distribution of the temperature of the four different microstructures. The difference in temperature of the solder bumps is small, about 2 °C. And we couldn't see too much difference in the temperature distribution among the four microstructures. Figure 8.52 shows the current distribution of the four microstructures. There's current crowding where the current enters the bump. And the tri-grain model has the highest current density among those four microstructures. Figure 8.53 shows the distribution of thermal gradient of the four microstructures. We can see that thermal gradient is going in the opposite direction of the current and the distribution is different among the four different microstructures. Figure 8.54 shows the distribution of the hydrostatic stress of the four microstructures. The hydrostatic stress seems to be quite different among the four microstructures. The maximum hydrostatic stress of the 45° model is greater more than twice of that of the bycrystal model.

Figure 8.55 shows the dynamic distribution of the atomic flux divergence of the four representative microstructures and the nucleation of voids. We can see that the beach ball model has the minimum nucleation void volume and the bi-grain model has the maximum voids among the four microstructures. Since elements with

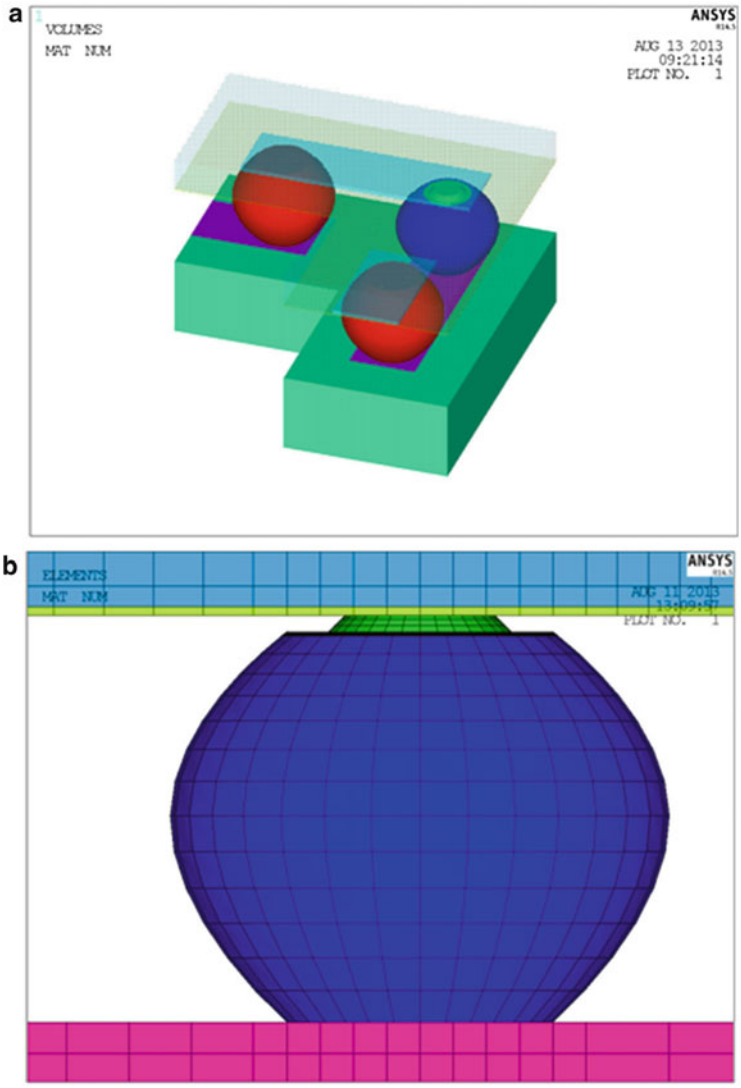


Fig. 8.48 The Submodel and mesh of the corner solder joint (**a**) Submodel of the local corner joint (**b**) Front view of the fine meshed solder bump

higher atomic flux divergence will void faster, this distribution gives us information about the nucleation of voids and time to failure.

Table 8.17 gives the time to failure for the four representative microstructures. The bi-grain model has the minimum time to failure followed by the tri-grain model and the 45°, and the maximum TTF is from the beach-ball model. This has something to do with the orientation of the grains. For the bi-grain model to the beach ball model, the grain boundary layers increase, which might induce the slow

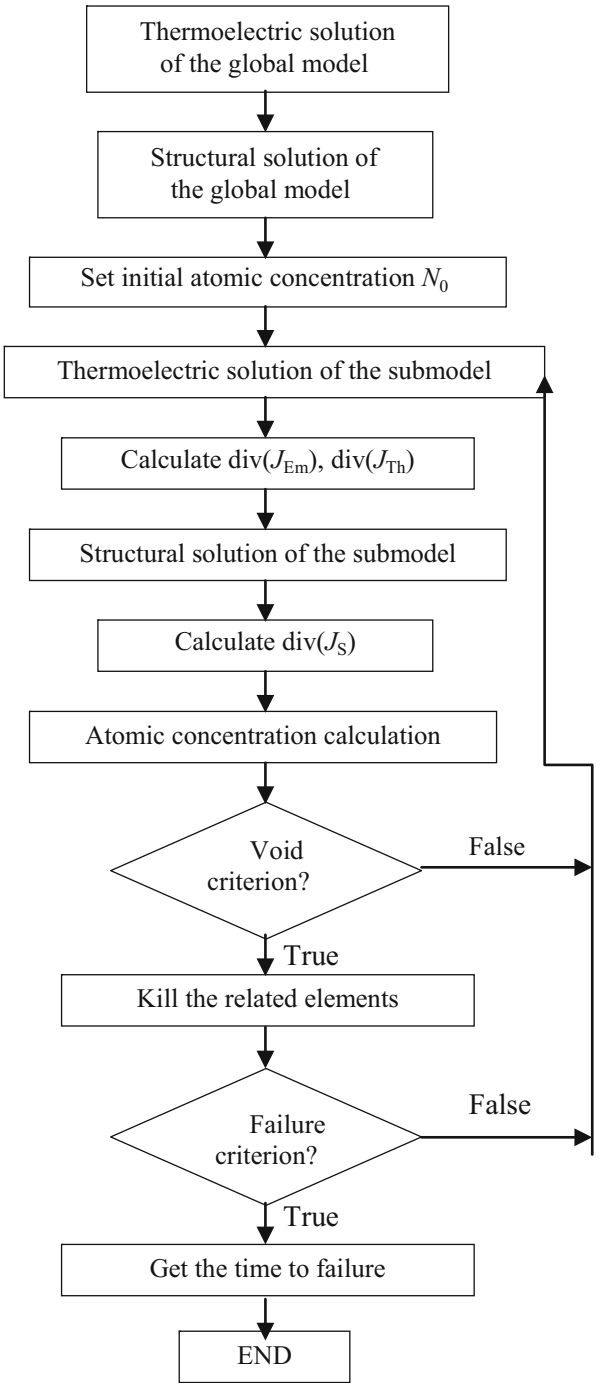


Fig. 8.49 Flowchart of the analysis

Table 8.15 ANAND model parameters for SnAgCu [19]

Description	Symbol	95.5Sn4.0Ag0.5Cu
Pre-exponential factor	A(1/s)	325
Activation energy	Q/R(K)	10561
Stress multiplier	ξ	10
Strain rate sensitivity of stress	m	0.32
Coeff. for deformation resistance saturation value	$\hat{S}$ (MPa)	42.1
Strain rate sensitivity of saturation value	n	0.02
Hardening coefficient	h_0(MPa)	800000
Strain rate sensitivity of hardening coeff.	a	2.57
Initial value of s	s_0(MPa)	20

Table 8.16 Anisotropic thermal, electrical, diffusion properties

	⊥c	‖c
Coefficient of thermal expansion (/°C)	15.8×10^{-6}	28.4×10^{-6} [20]
Electrical resistivity at RT (Ω·m)	9.9×10^{-8}	14.3×10^{-8} [21]
Temperature coefficient of resistivity (/°C)	0.00469	0.00447 [21]
Effective charge number	−16	−10 [22]
Self diffusion coefficient (m^2/s)	0.0021	0.00128 [23]
Activation energy (J/molecule)	$25.9 \times 6.95 \times 10^{-21}$	$26 \times 6.95 \times 10^{-21}$ [23]

mass diffusivity in the current flow direction. Therefore, it results in longer time to failure.

In order to better understand the effects of the grain orientation on electromigration, we run the single crystal case with different grain orientations. Figure 8.56 shows the life of time to failure versus the angle of crystal orientation. It shows when c-direction is aligned with the current flow direction, the TTF is the largest. As the c-direction gradually deflects from the current direction, the TTF decreases until the c-direction becomes perpendicular to the current direction. When the c-direction is perpendicular to the current flow direction, the TTF reaches the minimum, which makes sense because the self-diffusivity of tin along a-direction is faster than that along c-direction at the working temperature. One interesting phenomenon is at the angle of c-direction 60°, the TTF raises a little longer as compared to the TTF at the c-direction 30°. This possibly might be the coupling effect of thermal mechanical stress and the pure electro-wind induced migration at the 60°, which induces a slight smaller mass migration.

8.4.5 Discussion

The three-dimensional indirect electrical, thermal, and structural coupled analysis for the solder bumps in WL-CSP is carried out with consideration of representative microstructures. The distributions of the current, thermal gradient, hydrostatic

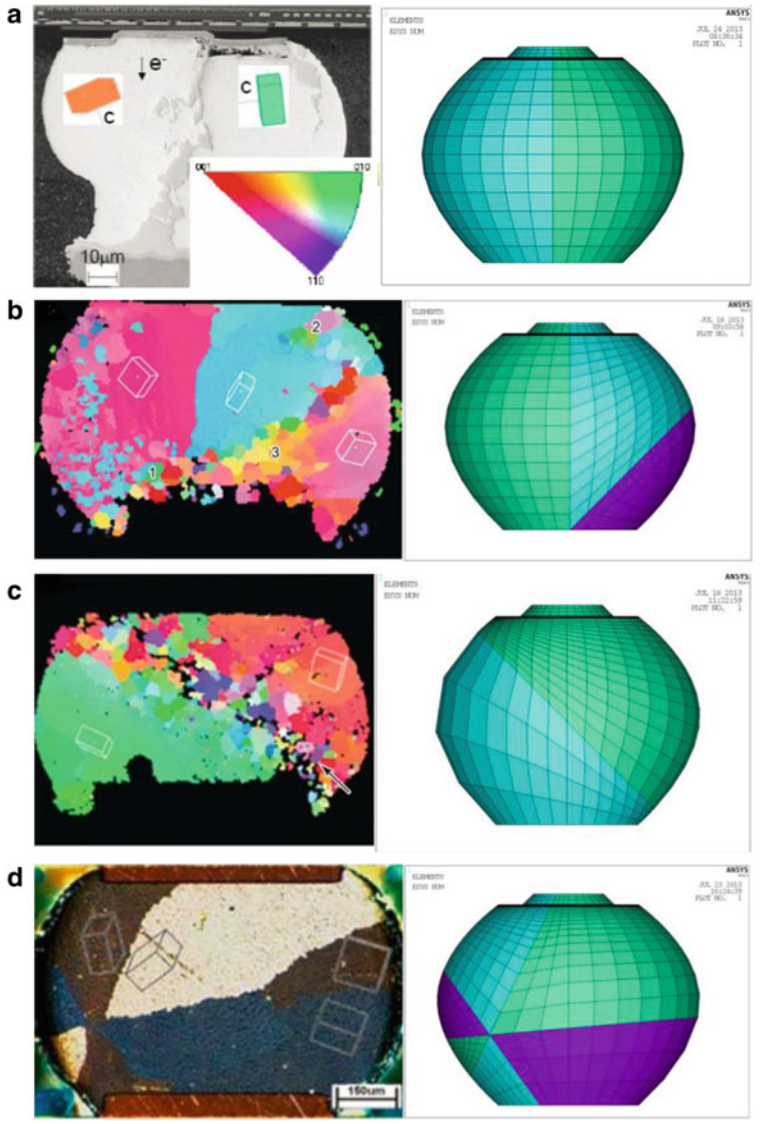


Fig. 8.50 Four microstructures of the solder bump in WLCSP (**a**) SEM image of a bicrystal bump [24], 3D Ansys finite element model (**b**) EBSD orientation map of a tricrystal bump [25], 3D Ansys finite element model (**c**) EBSD orientation map of a bump with grain boundary inclined at 45° to substrate [26], 3D FE model (**d**) Cross-polarizer optical micrograph of a beach-ball bump [26], 3D finite element model

stress, and atomic flux divergence are presented and compared among four representative microstructures of lead-free solder bumps. The effects of grain orientation and grain size on electromigration are studied and discussed. From the curve of TTF

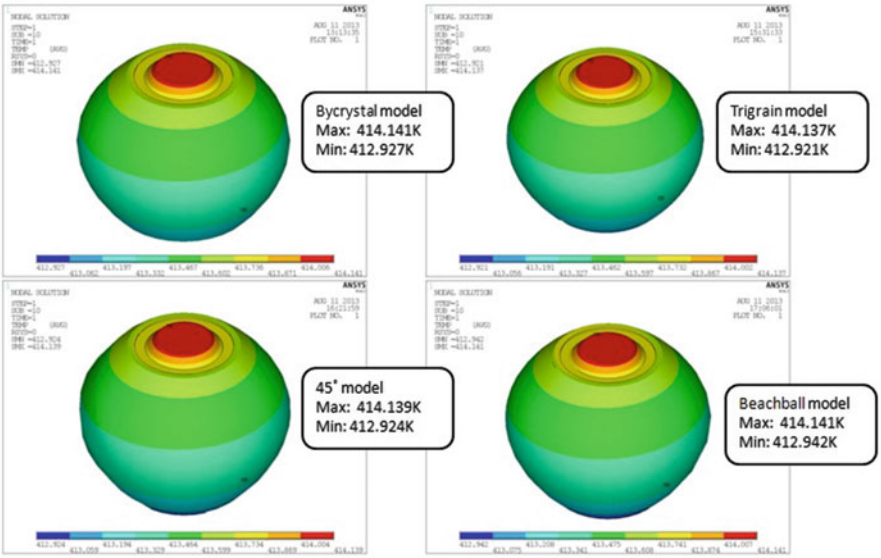


Fig. 8.51 Temperature distribution of four different microstructures

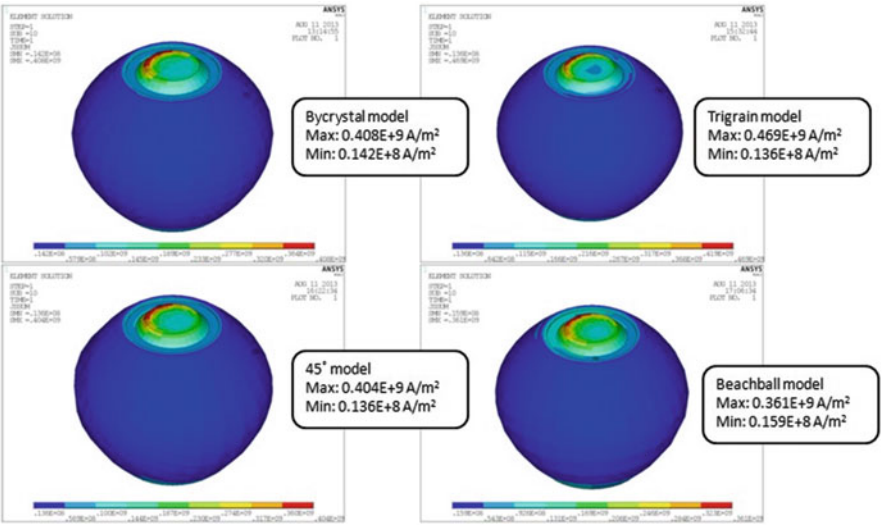


Fig. 8.52 Current distribution of four different microstructures

versus crystal orientation angles, we can see that as the c-direction gradually deflects from the current flow direction, the TTF will decrease, which means that the bump is more prone to electromigration.

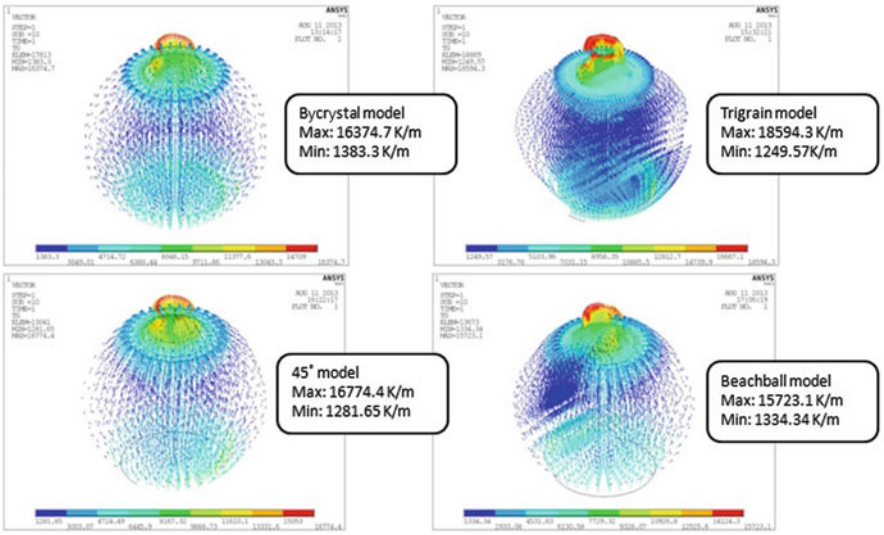


Fig. 8.53 Thermal gradient of four different microstructures

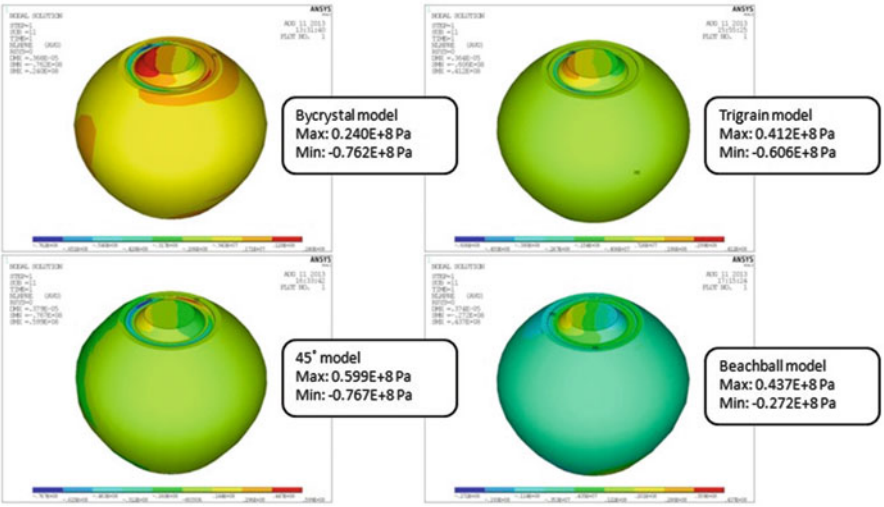


Fig. 8.54 Hydrostatic stress of four different microstructures

8.5 Summary

This chapter discusses the simulation methodologies of electrical parasitic resistance, inductance, and capacitance for wafer level chip scale package. The fan-in WLCSP and fan-out MCSP are studied through the electrical simulation. By

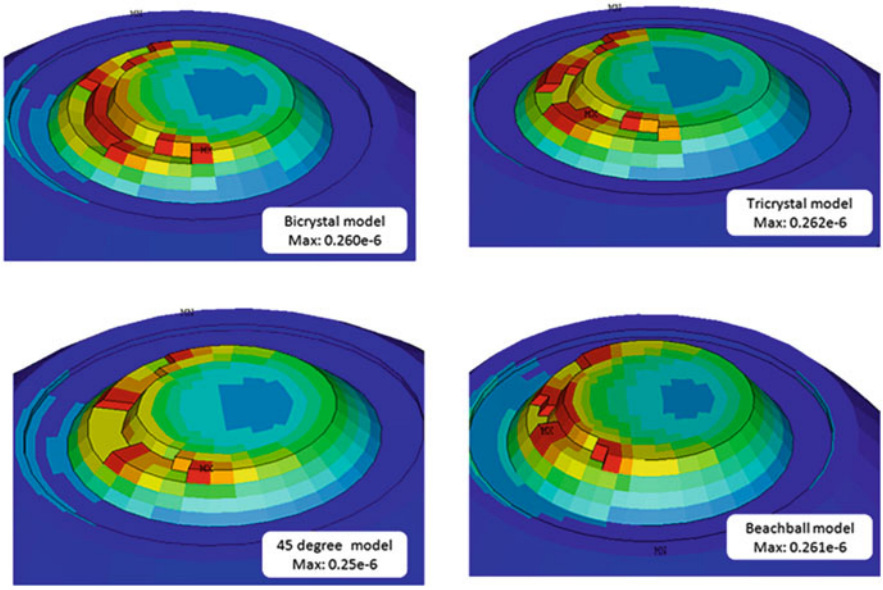


Fig. 8.55 Atomic flux divergence of four different microstructures

Table 8.17 TTF for four different microstructures

Model	Bycrystal	Trigrain	45°	Beachball
TTF (h)	1137	1296	1370	1471

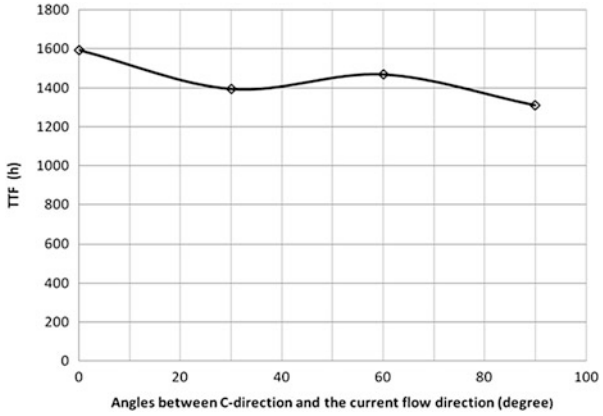


Fig. 8.56 TTF versus angles between c-direction and current flow direction in a single crystal

comparison between the MCSP with GGI RDL and wire bonding interconnect, the MCSP with GGI has shown much better electrical performance (much lower electrical resistance and inductance) than the wire bond MCSP. Then the

multiphysics simulation for 0.18 μm wafer level power technology and the WLCSP bumps with different microstructures and grain orientations are introduced. In the 0.18 μm wafer level power technology, the modeling methodology for CMP and non-CMP process with barrier metal are developed; the simulation results are correlated very well with the electromigration test data of the 0.18 μm wafer level power technology. In the WLCSP electromigration study, the microstructure of the bump which includes the different grain structures and different orientations are investigated with FEA simulation. The results have shown the signicant impact of the grain structure and orientation to the bump stress and the automatic flux divergence while it seems no much difference in temperature distribution and temperature gradient.

References

1. Liu, Y.: Power electronic packaging: Design, assembly process, reliability and modeling. Springer, Heidelberg (2012)
2. Sasagawa, K., Hasegawa, M., Saka, M., Abe, H.: Prediction of electromigration failure in passivated polycrystalline line. J. Appl. Phys. **91**(11), 9005–9014 (2002)
3. Sukharev, V., Zschech, E.: A model for electromigration-induced degradation mechanisms in dual-inlaid copper interconnects: Effect of interface bonding strength. J. Appl. Phys. **96**(11), 6337–6343 (2004)
4. Tan, C.M., Hou, Y.J., Li, W.: Revisit to the finite element modeling of electromigration for narrow interconnects. J. Appl. Phys. **102**(3), 1–7 (2007)
5. Tan, C.M., Roy, A.: Investigation of the effect of temperature and stress gradients on accelerated EM test for Cu narrow interconnects. Thin Solid Films **504**(1–2), 288–293 (2006)
6. Tu, K.N.: Recent advances on electromigration in very-large-scale-integration of interconnects. J. Appl. Phys. **94**(9), 5451–5473 (2003)
7. Liu, Y., Zhang, Y.X., Liang, L.H.: Prediction of electromigration induced voids and time to failure for solder joint of a wafer level chip scale package. IEEE Trans. Component Packag. Technol. **33**(3), 544–552 (2010)
8. Dalleau, D., Weide-Zaage, K., Danto, Y.: Simulation of time depending void formation in copper, aluminum and tungsten plugged via structures. Microelectron. Reliab. **43**(9–11), 1821–1826 (2003)
9. Wilson, S.R., et al.: Handbook of multilevel metallization for integrated circuits: Materials, technology, and applications, p. 116. Noyes, Park Redge, NJ (1993)
10. Zhao, J.H., Ryan, T., et al.: Measurement of elastic modulus, Poisson ratio, and coefficient of thermal expansion of on-wafer submicron films. J. Appl. Phys. **85**(9), 6421–6424 (1999)
11. Sharpe, William N., Yuan, Jr. B., Vaidyanathan, R.: Measurement of Young's modulus, Poisson's ratio, and tensile strength of polysilicon. In: Proc 8th IEEE International Workshop on Microelectromechanical Systems, Nagoya, Japan. (1997)
12. JEDEC, JEP119A: A Procedure for performing SWEAT. (2003)
13. Ni, J., Liu, Y., Hao, J., Maniatty, A., OConnell, B.: Modeling microstructure effects on electromigration in lead-free solder joints. ECTC64, Orlando, FL (2014)
14. Liu, Y., Irving, S., Luk, T., et al.: 3D modeling of electromigration combined with thermal-mechanical effect for IC device and package. EuroSime 2007
15. Liu, Y.: Finite element modeling of electromigration in solder bumps of a package system. Professional Development Course, EPTC, Singapore (2008)
16. Yang, S., Tian, Y., Wang, C., Huang,T.: Modeling thermal fatigue in anisotropic Sn-Ag-Cu/Cu solder joints. International Conference on Electronic Packaging Technology and High Density Packaging (ICEPT-HDP), (2009)

17. Subramanian, K.N., Lee, J.G.: Effect of anisotropy of tin on thermomechanical behavior of solder joints. J. Mater. Sci. Mater. Electron. **15**(4), 235–240 (2004)
18. Gee, S., Kelkar, N., Huang,J., Tu, K.: Lead-free and PbSn Bump Electrmigration Testing. In: Proceedings of IPACK2005, IPACK2005-73417, July 17–22
19. Wang, Q., et al.: Experimental determination and modification of the Anand model constants for 95.5Sn4.0Ag0.5Cu. Eurosime 2007, London, UK, April, (2007)
20. Zhao, J., Su, P., Ding, M., Chopin, S., Ho, P.S.: Microstructure-based stress modeling of tin whisker growth. IEEE Trans. Electron. Packag. Manuf. **29**(4), 265–273 (2006)
21. Puttlitz, K. J., Stalter, K. A: Handbook of Lead-free solder technology for microelectronic assemblies. pp. 920–926. (2004)
22. Huntington, H.B.: Effect of driving forces on atom motion". Thin Solid Films **25**(2), 265–280 (1975)
23. Huang, F.H., Huntington, H.B.: Diffusion of Sb124, Cd109, Sn113, and Zn65 in tin. Phys. Rev. B **9**(4), 1479–1488 (1974)
24. Lu, M., Shih, D., Lauro, P., Goldsmith, C., Henderson, D.W.: Effect of Sn grain orientation on electromigration degradation in high Sn-based Ph-free solders". Appl. Phys. Lett. **92**, 211909 (2008)
25. Xu, J.: Study on lead-free solder joint reliability based on grain orientation. Acta Metallurgica Sinica **48**(9), 1042–1048 (2012)
26. Park, S., Dhakal, R., Gao, J.: Three-dimensional finite element analysis of multiple-grained lead-free solder interconnects. J. Electron Mater. **37**(8), 1139–1147 (2008)

WLCSP Assembly

9

9.1 Introduction

Assembly of WLCSP components involves surface mounting technology (SMT), which includes pickup WLCSP components from tape and place them onto a printed circuit board (PCB), solder reflow, and optional underfill. Figure 9.1 provides a schematic diagram of a typical assembly line setup involving WLCSP, in which solder paste or flux is first printed or dispensed on the PCB respectively before WLCSP pick and placement. Reflow is followed to finish the solder joint formation between the WLCSP component(s) and PCB. Optical or X-ray inspections are arranged at various stages of assembly line to ensure correct solder paste printing (height, area, and volume of the solder paste bricks deposited on the PCB soldering pads), accurate component placement (XY offset and skewness), and proper solder joint formation. In-circuit test (ICT) is an electrical probe test on the assembled PCBs, checking for shorts, opens, resistance, capacitance, and other basic quantities which will show whether the assembly was correctly fabricated. Post solder reflow, underfill could be applied after flux clean to provide the needed protections to WLCSPs and solder joints going through subsequent assembly processing steps and to ensure robust reliability of the WLCSP devices in the everyday usage. Underfill is especially desired when die size is big or when low-K dielectrics are present in the WLCSP.

WLCSP is a bare die package benefiting from the smallest possible form factors and unprecedented performance and low cost. Fine pitch, such as 0.4 mm and 0.35 mm bump pitch or even 0.3 mm bump pitch, is common for WLCSP, though 0.5 mm and 0.65 mm pitches are still encountered on specialty or older-generation devices. The fine pitch and bare die nature determine that WLCSP assembly requires special considerations in the electronics assembly flow.

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_9

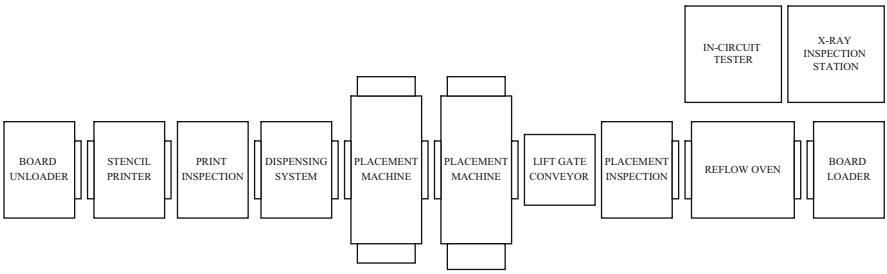


Fig. 9.1 Schematic diagram of a typical assembly line setup

9.2 PCB Design

IPC-SM-7351 is the often referenced standard designing PCB land patterns. It provides guidelines on how to design a land pattern for surface mount components; the information provided includes the size, shape, and tolerance of the land pattern, and these dimensions are based on industry-registered component specifications, board manufacturing standards, and component placement accuracy capabilities. IPC-9701 is another standard that provides information on board designs and reliability performance of the surface mount device. However, for WLCSP, especially fine pitch 0.35 mm and under, additional scrutiny is often more than necessary in specific areas, to ensure high-yielding assembly and robust reliability of WLCSP in the field use.

9.2.1 SMD and NSMD

There are two types of PCB pads for WLCSP mounting: SMD (solder-mask-defined) or NSMD (non-solder-mask-defined) (Fig. 9.2). NSMD is recommended to fine pitch WLCSP due to better overall land pattern registration accuracy and wider space for routing between adjacent pads. In combination with the right surface finish, NSMD also allows the solder to wet the pad sidewalls that effectively increase the solder joints cross section on the PCB side (Fig. 9.3). In comparison, SMD pad is defined by low accuracy solder mask process. The design also creates the notches near the solder mask edge that could be a stress concentration point, therefore a potential reliability risk in the field. Typical clearance of 0.075 mm is set between the solder mask and the Cu pads for a NSMD pad and an overlap of 0.050 mm between the solder mask and the SMD Cu pads. Tighter space is often required for fine pitch WLCSP and is acceptable so long as it satisfies PCB manufacture design rules and manufacturability targets.

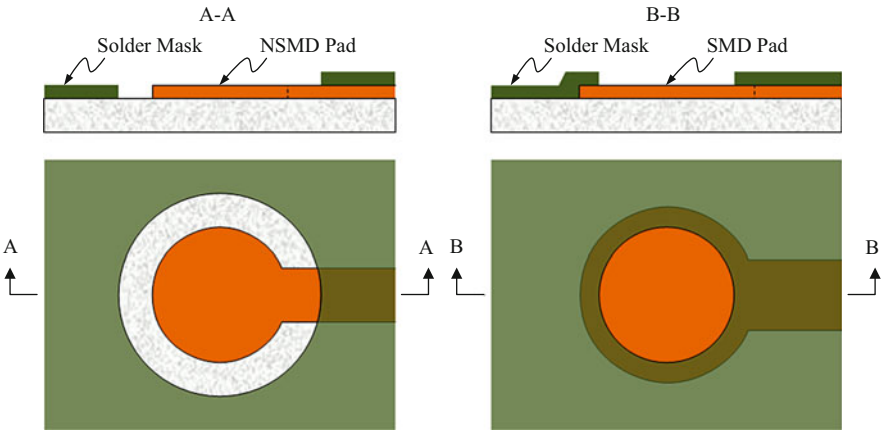


Fig. 9.2 Top view and cross-sectional view of NSMD and SMD PCB pad designs and solder wetting on sidewalls of NSMD pad and solder neck on SMD pad

Fig. 9.3 Wetting of solder on sidewalls of NSMD pad and solder necking on SMD pad

9.2.2 Land Pad Size

PCB land pad size is typically matched to the UBM diameter of the WLCSP components, because imbalanced solder joint shape with one end significantly bigger or smaller than the other end is not desired for robust board level reliability performance. This principle applies to either SMD or NSMD design. WLCSP from different suppliers tend to have different UBM sizes: some follows the long-standing 80 % bump size rule to define their UBM diameter; others opt to larger UBM size for improved drop and TMCL performance. Large UBM size and matching PCB land pad size do bring the penalties of making it harder to route traces between adjacent pads (Fig. 9.4) and reducing the standoff height of solder joints, while at the benefits of increasing the solder joint cross sections, which is the main reason for the reliability improvements (Fig. 9.5). It could also impose challenges on stencil design for fin pitch WLCSP because of the reduced webbing dimensions. At the same time, solder joint diameter increase might bring negative impact to assembly yield due to narrow gaps between joints and thus greater tendency of bridging solder joints.

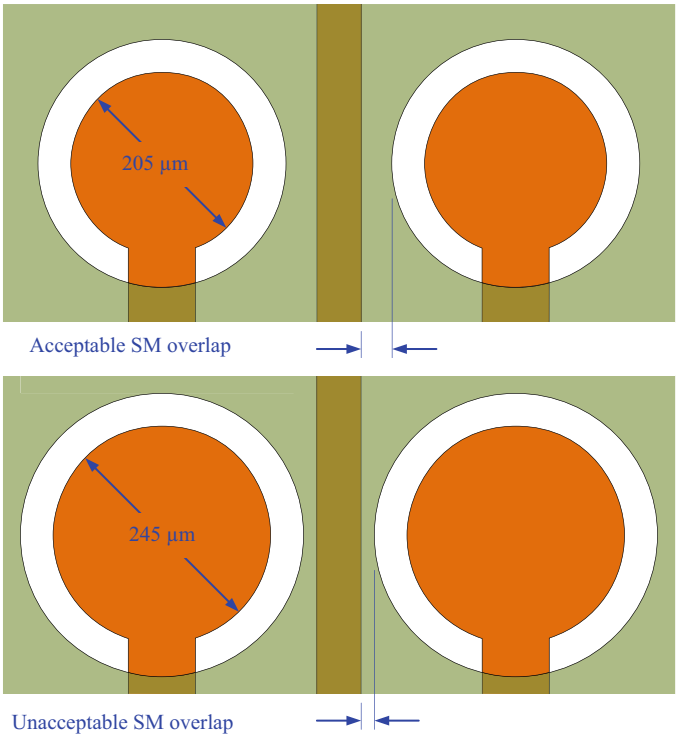
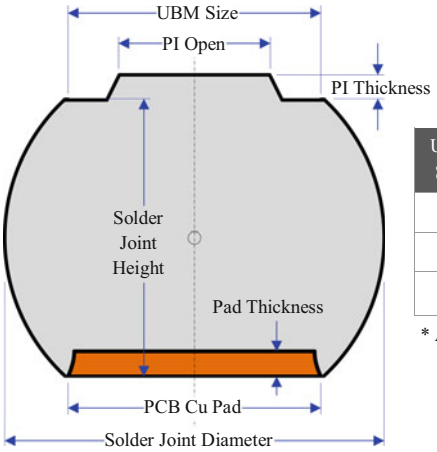


Fig. 9.4 Larger PCB pad imposes challenges to routing between adjacent landing pads

9.2.3 PCB Pad Surface Finish

ENIG, OSP (Organic Solderability Preservative), and HASL (Hot Air Surface Leveling) are commonly used PCB pad surface finishes. Immersion Ag (silver) is becoming popular as well. ENIG offers excellent surface solderability. However, solder wetting of pad is nearly impossible due to the mushroom top cross section of typical pads with ENIG finish (Fig. 9.6a). Typical Ni thickness is 2.5–5.0 µm, while for Au it is 0.08–0.23 µm. Low-cost OSP has found wide adoptions for WLCSP. Desired sidewall wetting is common for pad with OSP finish, even for additively plated pad with nearly vertical sidewall profiles (Fig. 9.6b). HASL is not preferred for WLCSP due to poor flatness. Immersion silver (Ag) is a lead-free alternative and has an excellent flatness and OSP-like sidewall wetting; however, special handling may be required.



UBM Size	PI Via Open	PCB Pad Size	Solder Joint Diameter	Solder Joint Height
210	175	205	281.11	189.61
230	195	225	289.93	179.68
250	215	245	300.19	169.82

* All dimensions are in microns.

Fig. 9.5 UBM/pad size-dependent solder joint size for typical 0.4 mm pitch WLCSP with 250 µm diameter solder balls reflowed

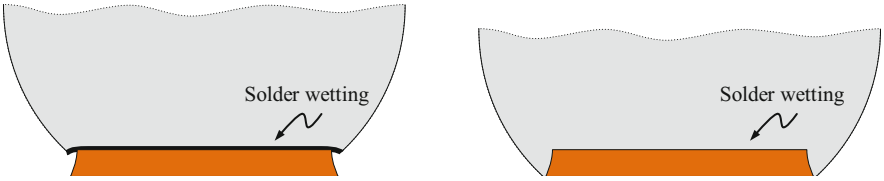


Fig. 9.6 Wetting of solder joints on ENIG finish (*left*) and OSP/immersion Ag finish (*right*)

9.2.4 Vias Under WLCSP

Improperly designed/placed through PCB vias within WLCSP land pattern induce excessive stress in solder joints during solder reflow and reliability testing and should be therefore avoided in any PCB design for WLCSP. Blind vias, either via in pad or routed, are recommended when addition routing is needed under WLCSP footprint. Solder mask over the routed trace and via is mandated for routed vias (Fig. 9.7) to prevent undesired solder wetting beyond pad area. For the via-in-pad design, bottom-up via fill plating is recommended to avoid difficulties in solder screen printing and to prevent excessive voids or irregular solder joint shapes. Via-in-pad design is recommended for improved thermal dissipation of WLCSP.

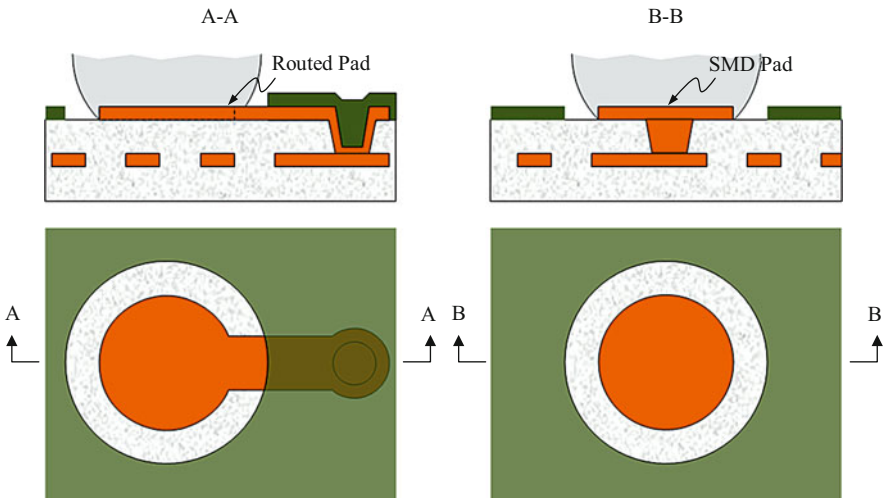


Fig. 9.7 Cross-sectional view and top view of routed PCB copper pad and via-in-pad PCB copper pad. Filled copper via is desired in the via-in-pad design

9.2.5 Local Fiducials

Local fiducials are normally placed on the board to assist the automatic placement equipment, placing WLCSP accurately onto the board. Diagonally placed fiducials outside the component land pad are most common (Fig. 9.8). Sufficient clearance between the WLCSP component edge and local fiducials is important for the placement tool vision system, though excessive clearance is not desired for high-density SMT boards, because it takes up precious surface area.

9.2.6 PCB Materials

High-temperature FR4 or BT laminate ($T_g = 170–185$ °C) is generally recommended for WLCSP assembly with lead-free solder bumps, which typically requires peak temperature of solder reflow profiles from 240 to 260 °C. More costly polyimide materials are also found in PCB with demanding requirements for temperature endurance. For high pin count WLCSP, advanced low CTE laminates should also be considered besides bumping technology, UBM size, and underfill materials for reduced thermal/mechanical stress and improved reliability. Use of the thinnest copper foil thickness that meets the electrical requirements of the circuit design is also recommended for similar reasons.

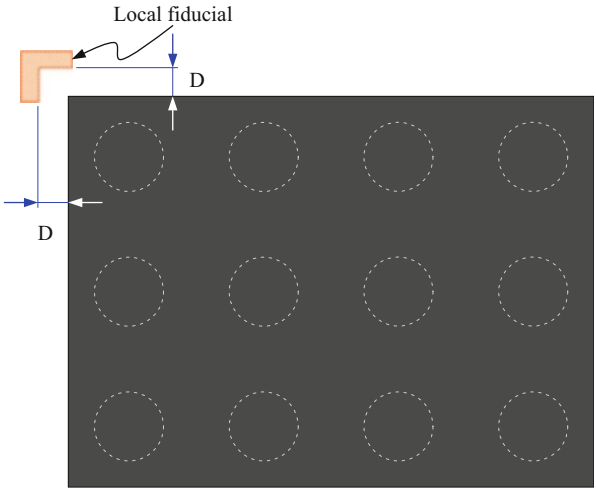


Fig. 9.8 Local fiducial for WLCSP surface mounting. Critical gap D between package and the fiducial is also highlighted

9.2.7 PCB Trace and Copper Coverage

A balanced fan-out of traces from land pads in the X and Y direction will help to avoid the unintentional component movement as a result of unbalanced solder wetting forces during reflow. Use of smaller trace routings from NSMD land pads is sometimes necessary to prevent solder migration that can result in lower standoff height of the bumps. Typically PCB trace width is less than 2/3 of the board pad diameter. For high current connections, traces could be widened under solder mask coverage. Orientation of fan-out PCB trace from peripheral land pads might require special considerations for certain WLCSP test board to avoid potential copper trace crack in the reliability tests. PCB warpage during solder reflow could be an issue for larger size fine pitch WLCSP components. Balanced copper trace and coverage in layers symmetric to the neutral plane is often necessary for improved PCB warpage control.

9.3 Stencil and Solder Paste

9.3.1 General Stencil Design Guidelines

SMT printing begins the SMT production processes and is the first most important process to achieve high quality in SMT operation. Of all the factors that affect the finishing solder paste print quality, such as the well-maintained equipment, optimized printing parameter and setup, trained operator and tooling, etc., a good stencil plays one of the most important roles in printing. It is often believed that up to 60–70 % of SMT solder defects for the whole SMT process can be prevented with well-designed, fabricated, and maintained stencils.

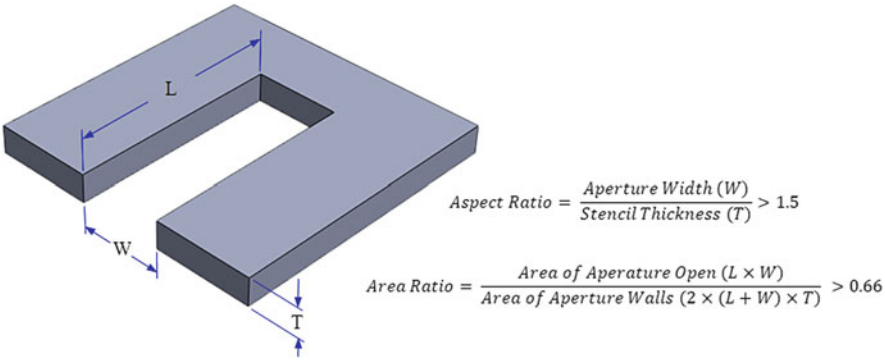


Fig. 9.9 A rectangular stencil open and definitions of aspect ratio and area ratio

For SMT stencil design, IPC-7525 outlines three major factors that affect the solder paste print/release from the stencil: (1) the area ratio and aspect ratio of the aperture, (2) the sidewall angle, and (3) the surface finish of the stencil walls.

Area ratio is defined as ratio between the aperture sidewalls and the PCB pad (Fig. 9.9). It is as a matter of fact a measure of releasing force of PCB pad to the solder paste to the holding force of aperture to solder paste. When weight is factored in, an empirical minimum area ratio of 0.66 is needed for good solder release on the PCB pad when the printing squeeze sweeps the solder paste over the aperture. Aspect ratio is a ratio between stencil thickness and short linear dimension of stencil aperture and is considered a simplified version of area ratio (Fig. 9.9). The aspect ratio is a useful guide when the aperture length (L) is much larger than aperture width (W), such as greater than 5 $\times$ the width. A ratio greater than 1.5 is recommended for good solder release. In certain cases depending on terminal size and pitch, the aspect ratio and area ratio may depart from the recommendations, and it should be tested individually.

There are two primary stencil technologies used within the industry: electroformed and laser-cut. Chemical etching can also be used to create steps for step stencils; however, final stencil apertures are generally laser-cut. Three-dimensional (3D) electroformed stencils are also available for use when unique steps are needed, providing release qualities required for fine feature printing. One distinctive advantage of electroformed stencil is the relative straight sidewalls and a smooth mirrorlike finish that promotes greater solder paste release. It is possible with this technology to print apertures with area ratios down to 0.50, which is helpful as component sizes and pitches continue to decrease.

Laser-cut stencil with complementary electropolishing is the most common one for typical SMT applications. Common stencil thicknesses for WLCSP are 0.100 and 0.125 (4 and 5 mils, respectively). Tapering on the stencil aperture, which also helps paste release, is usually between 2 and 5°, which could be achieved by making the PCB contact side 0.025 mm larger than the squeegee side. Additionally, curved corners for rectangular aperture promote better paste release and stencil

cleaning. Reducing the stencil aperture to less than that of the board land pad is desirable to enhance the process of printing, reflow, and stencil cleaning. This minimizes the board pad and stencil opening misalignment.

9.3.2 Solder Paste

Common information for solder paste selection includes solder alloy, flux materials and activity level, particle size, and metal content in the paste. Solder pastes are classified based on the particle size by JEDEC standard J-STD 005. Commonly used solder pastes in WLCSP assembly printing application are type 3 (average 36 μm particle size) or type 4 (average 31 μm particle size) solder paste with high percentage, i.e., 89.5 wt% metal loading. Flux composition can be either no clean or water soluble. No-clean fluxes may be rosin (RO) based, resin (RE) based, or free of rosins or resins, which are then classified as organic (OR). Flux activity level, which is a measure of the ability to dissolve existing oxides on the metal surface and promote wetting with solder, may be low (L), moderate (M), or high (H). Highly active fluxes are often of acidic and/or corrosive nature. Water-soluble fluxes generally have organic (OR) composition and high (H) activity level where cleaning post reflow is mandatory. Halide-free fluxes are also preferred which are designated by a 0, while 1 means the presence of halides. A flux with ROL0 classification means rosin based, low activity, and halide-free. Sn–Ag–Cu solder alloys are the most often used for the assembly of WLCSP with for lead-free solder bumps. The liquidus temperature is about 217–220 °C.

Optical solder paste inspection (SPI) before component placing can reduce the incidence of solder-related defects to statistically insignificant level and is recommended to ensure uniform solder paste coverage over PCB pads, such as height, area, and volume of the solder paste deposited by the screen printer. Either in-line system or off-line systems are available from manufacturers worldwide.

9.4 Component Placement

Automated fine pitch placement machines with vision alignment are recommended to place WLCSP components. Local fiducials are typical on the PCB board to support the vision systems and achieve placement accuracy. “Pick and place” systems using mechanical centering is prohibited due to the high potential for mechanical damage to the bare silicon WLCSP devices. Minimal pick and place force (typically <0.5 N), with all vertical compression forces controlled and monitored, is suggested to avoid damages due to the same considerations. Z-height control methods are recommended over force control during pick and place of WLCSP. It is also highly recommended to use of low-force nozzle options and compliant materials (e.g., rubber tipped) to further avoid any physical damage to the WLCSP device. Whenever manual handling is required, use only vacuum pencils with soft tip materials. Placement accuracy studies should be conducted to

ensure adequate compensation is provided for high accuracy placement. Adequate validation of the SMT pick and place process is also recommended to ensure WLCSP die integrity is not compromised.

9.4.1 Pick and Place Process

There actually exists allowable maximum component placement force that highly depends upon WLCSP bump structures and mounting board materials. However, to avoid physical damage to the back and front (active) side of WLCSP devices, Z-height control is recommended over force control for picking and placing operations. During component pickup from carrier tape, it is recommended a Z-height distance between the WLCSP and the pickup tool (Fig. 9.10). Define vacuum pressure at proper level, i.e., approximately 60–70 kPa, to lift the WLCSP from the pocket of the carrier tape. This practice prevents direct contact on the backside of silicon of the WLCSP during pickup. Similarly, when placing WLCSP onto PCB with solder paste printed on, set the Z-height to zero or to a critical distance between the PCB and WLCSP before it is dropped or placed (Fig. 9.10). Avoid setting a bond force, which can overdrive the package to the surface of the board. This is a common setup mistake, which could lead to WLCSP component damages when the resultant force drives back the solder bumps towards the active side of the WLCSP die.

The WLCSP devices used in high-volume SMT applications are generally supplied in tape in a bump-down configuration. There are special cases where production is set up such that WLCSP is supplied on film frame on a UV dicing tape with WLCSPs in a bump-up configuration. In this setup, individual WLCSP die will be picked up from the dicing tape first, flipped, and then transferred to another pickup tool before being placed on a PCB. Adequate steps should be taken to ensure proper UV exposure is performed before the die pick application.

The proper ratio between pick and place tool and package size must be observed at a minimum of 80 % to provide uniform distribution of stress on the package during placement. Optimized conveyor speed and transfer are also important to

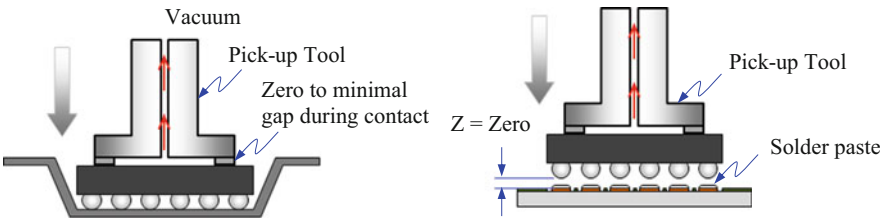


Fig. 9.10 Set Z-height distance between the WLCSP and the pickup tool to zero or with minimal gap, as shown on the *left*. The vacuum lifts out the package from the pocket of the carrier tape. Similarly, set Z-height during placement to zero or with a minimum gap height to avoid overdrive during board placement, as shown on the *right*

avoid component displacement or skewing prior reflow. If realignment of the package on the board prior to reflow is required, only soft tool such as a vacuum pen or equivalent is allowed.

9.4.2 Placement Accuracy

WLCSP with ball drop options typically has solder-bump height greater than 0.15 mm and that provides robust self-alignment with screen-printed solder paste. Ultrathin WLCSP often opts for plated solder bump for the low package profile, typically has standoff less than 0.15 mm, and is often less tolerant to placement offset due to poorer self-alignment.

Where components are to be placed, the printed circuit board normally has flat soldering pads. Solder paste is already applied in the prior printing step, either with stencil printing process or jet printing mechanism. After pasting, the boards then proceed to the pick and place machines, where they are placed on a conveyor belt. The components to be placed on the boards are usually delivered to the production line in either paper/plastic tapes wound on reels or plastic tubes. Some large integrated circuits are delivered in static-free trays. Numerical control pick and place machines remove the parts from the tapes, tubes, or trays and place them on the PCB.

Automated placement equipment with vision alignment system is used for placing WLCSP onto PCB. The allowable package offset with respect to pad should be determined. Generally, a 50 % misalignment during placement of the component on the board is tolerable; WLCSP with sufficient solder-bump size tends to self-align during reflow process. The placement machine nozzle Z-height should have enough overtravel to allow the bumps to be submerged about 50 μm (2 mils) or half of the solder paste height to the printed paste to allow the self-centering of the package. This will also prevent the package from moving during transit from the pick and place equipment to the reflow oven.

The two most popular methods for package alignment are the package silhouette (look down camera) and the ball recognition system (look up camera). In the package silhouette system, the vision system locates the package outline only, while in the ball recognition, the vision system locates the ball array pattern. It can also detect missing balls.

9.4.3 Nozzles and Feeders

No SMT machine can place accurately and run efficiently without well-functioned nozzles and feeders. Nozzles are the first and last thing to touch all parts placed. They have to hold the part during transport to the board while the machine is moving and/or rotating. Feeders move various types of components to the correct position for nozzles to pick up. Lack of proper nozzle/feeder maintenance and/or poor quality nozzles could lead to many process problems. Here are some of the most common problems:

1. Poor pickup location on part. It will cause a loss of vacuum and result in parts shifting on the nozzle during transport.
2. Short/worn nozzles result in poor pickup and can cause the part not to be imbedded into the paste. When the part is not placed into the paste correctly, there is not enough surface tension to hold the part while the PCB is moving. Parts will shift.
3. Sticking nozzles may be responsible for many issues as it drastically changes the height at which the nozzle is presented.
4. Higher than normal rejection rate at part inspection caused by:

   - Parts not being presented to the nozzle in a consistent position
   - Poor nozzle lighting
   - Incorrect nozzle height
   - Stuck nozzles from part height incorrectly set in program

In summary nozzles and feeders touch thousands of parts per hour. They are highly critical to the pick and place process. The need for proper preventative feeder and nozzle maintenance along with the use of high quality nozzles is essential to all SMT process.

9.4.4 High-Speed Surface Mount Considerations

With growing market of mobile consumer electronics, assembly at extreme speed with variety of electronics components in all forms and sizes at high mounting density is placing unprecedented demands of speed and flexibility on electronics manufacturers and ultimately the placement tools they select.

Flexibility requirement translates into the ability to deal with the complete SMD package-form range, from tiny passives down to 0.4×0.2 mm (EIA 01005 or IEC/EN 0402 MLCC chip capacitors) to fine pitch (0.3 mm) large size BGA package placement on PCB. Placement modules should also accept various component supply formats, such as matrix trays (waffle packs), embossed tape, and even wafer on saw tape. For speed, the key factors to be considered are the placement principle, component feeding, vision technique, and possibly high-speed fluxing. One critical feature, which must be considered in context with the placement rate, is the placement accuracy. However, physical laws oftentimes do not allow extreme speed and accuracy for one given technical approach. As a result, tools with fundamental placement principles need to be thoughtfully considered in practice to balance the speed and accuracy requirements simultaneously.

9.4.5 Placement Accuracy Requirement

Key determining factors for the required ball/bump placement accuracy for area array packages are bump count and package weight. One of the CSP's advantages is that the placement accuracy requirement is strongly relaxed, compared with leaded ICs (QFPs/SOs) with the same pitch.

The maximum acceptable placement error is equal to half of the PC board substrate pad diameter in the case of circular pads without a solder mask. Misplacement of the solder paste over half of the PC board pad diameter can occur, but a mechanical contact between the ball/bump and the pad will still take place. As a result, even though the solder paste is misregistered, perfect self-alignment is virtually guaranteed. The placement accuracy expectation in reality, however, is much higher. To achieve a good process capability index (cpk), users are demanding 4 sigma placement accuracies—clearly better than 100 µm.

9.4.6 Placement Principle Options

Placement accuracy depends on the quality of the positioning axes, *x*, *y*, and theta. With pick and place machines, the placement head is typically carried by an *x*–*y* gantry system and can be freely positioned within the predefined *x* and *y* range. This positioning allows specific lateral movements, such as those needed for component pickup, placing, and/or multiple measurements on the component over a stationary, upward-looking camera. Within the placement head, the most important axis is the rotational axis for correct theta or chip orientation, but the precision of the *z*-axis movement should not be neglected, especially for sensitive WLCSP pick and place. In high-performance systems, the *z*-axis movement is typically controlled by a microprocessor, utilizing sensors for determining both the length of the vertical stroke and the required placement force.

Pick and place clearly provides the best placement accuracy as long as the machine is equipped with a minimum number of placement heads. The high precision systems are specified with an *x*, *y* accuracy of ± 20 µm based on a 4 sigma quality level. The fundamental disadvantage of high accuracy pick and place systems, which normally provide only one high precision placement head, is the very limited placement rate. The rate of such systems is usually below 2,000 cph, excluding any additional process activities such as fluxing.

Today's pick and place systems typical have much flexible systems, providing a high precision pick and place head in combination with a multi-nozzle revolver head on a single gantry. Here the high precision head is responsible for the placement of large BGA or QFP odd-shaped components and very challenging fine pitch flip chips. The high-speed jobs, with smaller and less demanding components (in terms of placement accuracy), are performed by the revolver (shooter) head. These "less demanding" jobs include CSPs down to a ball pitch of 0.5 mm (20 mil). The placement principle used is "collect and pick and place," which is a deviation from the traditional chip shooter concept.

Traditional chip shooters are normally defined by a horizontally rotating turret head, which simultaneously picks the components from a moving feeder bank and places them onto a moving PC board. Very high theoretical placement rates in the range of 40,000 cph are enabled. For high-speed mounting of area array packages, classical chip shooters can only be used to a very limited degree, due to limitations in component pickup, insufficient placement accuracy for most CSP users (typical values >100 μm at 4 sigma), and inability to perform component fluxing, etc. In theory, traditional chip shooters are still capable for high-speed mounting of area array packages with ball diameters >0.3 mm in conjunction with package outline centering and standard embossed tape as a feeding format.

The most advanced version of today's collect-and-place system (based on ball centering) can perform high-speed placement of CSPs with at least 5,500 cph with a placement accuracy of 60 μm based on 4 sigma. This system also allows for high-speed flip chip mounting, assuming bump diameters of 110 μm on bump pitches in the range of 200 μm.

In an alternative dual-beam collect-and-place shooter system, two revolver heads are carried by two separate *x*, *y* gantries. Each of the two revolver heads is equipped with 12 nozzles and has random access to waffle packs or matrix trays. This system can achieve 90 μm/4 sigma overall placement accuracy (including theta deviation) with 20,000 cph for the standard SMD package spectrum. For area array packages, a dual-beam shooter system, due to time-intensive ball/bump find algorithms, can operate with a placement rate of >11,000 cph in many cases. If, instead of ball centering, outline centering is applied, the maximum placement rate will be 20,000 cph.

9.4.7 Vision System

Modern SMD placement all employs machine vision techniques that consist of a combination of component vision systems and PCBs with special features designed in. In meeting the sometimes extreme placement accuracy requirements (especially for and with flip chips) of a modern assembly, the importance of PC board fiducial and inkspot recognition should not be underestimated. Global fiducial and inkspot determination can be very difficult, due, in part, to color and contrast conflicts. Fortunately, chip-scale and other area array packages have lower placement accuracy demands, which enable the number of local fiducial readings to be reduced.

Most typically, CSP application will be on a relatively small size PCB that is a part of a larger size panel for assembly. To obtain maximum throughput within an SMD production line containing several placement modules, it is sometimes beneficial to only spend time on the first placement module for pattern recognition and transfer the individual fiducial/inkspot situation to the subsequent modules and thus saves precious machine time.

Powerful component vision systems are with today's SMT equipment to meet demands of widely varying materials and surface properties of SMDs. Component

vision system capabilities (and vision system capabilities in general) depend on both lighting (camera) technique and the algorithms applied by the evaluation unit.

Outline centering, using backlighting or laser-side illumination, is generally suitable for WLCSPs. However, WLCSP outline tolerances could bring negative impact on placement accuracy because of outline tolerances. Bump centering, on the other hand, is mandatory for most users. Vision ball centering is only possible with a front-lighted system. Maximum recognition reliability and repeatability can only be achieved with the use of a sophisticated, flexible lighting technique employing various light sources. Each light source should have a specific light emission angle. The near-perfect quality of the WLCSP image means high contrast bump/substrate images, with slightest visibility of underlying chip/package routing structures, which can be further suppressed easily by specific bump location algorithms. High-performance SMT placement systems, which have to deal with the complete package-form range, including fine pitch devices, must have two or more component cameras. The fine pitch camera must have a different lighting approach and a much higher camera resolution (magnification) than a standard camera.

Another strong argument for bump centering is the orientation check (commonly called Pin 1 recognition) for area array packages. This is the only SMD placement machine vision feature that can reliably prevent placing these packages with the wrong orientation. The orientation check is automatically included in the placement process when ball centering is carried out with non-regular (nonsymmetric) ball arrays. For package with symmetric bump array, however, orientation centering is not yet in place.

9.4.8 Algorithms

Algorithms suitable for standard SMDs cannot readily be applied for bump centering on area array packages. Complex, time-consuming, but interference-tolerant contour search methods are more advantageous.

Though WLCSP bumping manufacturers all perform 100 % bump inspection on the dedicated optical inspection systems, the capability for automated SMT vision inspection of the solder bumps is sometimes desired. With a powerful, flexible lighting approach and specific inspection algorithms, bump inspection with respect to deformation presence/absence is possible, but only to a limited degree.

One should be aware that the primary job of a component vision system is the precise and fast centering of the various packages. Fast optical centering is only possible with a single shot (no multiple measurements). Also, the large field-of-view always results in a relatively coarse resolution which is contrary to the requirements of precise ball inspection. There is a strong conflict between the demands of full array ball inspection, precision, and a good placement rate. In most applications to achieve acceptable bump find calculation times and thus high placement rates, only a few (e.g., five) balls in each corner of the package should be programmed.

9.4.9 Component Feeding and Fluxing

Feeding of BGAs and CSPs is typically done with matrix trays or standard embossed tapes. It is worth remembering that only collect-and-place shooters work with matrix trays.

When CSPs have to be used in applications with a wide mix of standard SMDs, the critical process step is solder paste printing. If the selected stencil thickness is too high, the solder paste destined for the pads at the CSP site might remain in the aperture. There are two ways to overcome this potential problem: (1) Use special stencils that enable applying different solder paste thicknesses. The various thicknesses within a stencil can be accomplished either by step etching or by additive methods. Since special stencils are more costly and create some PWB layout restrictions, their use is limited in the SMT industry. (2) For reliable soldering and good positional stability during PC board transport—to and through the reflow oven—use CSP dip fluxing. The flux carrier is typically a rotating drum on which a thin film (e.g., 75 μm) of flux is adjusted by a doctor blade. This principle is most suitable for high viscosity fluxes. The amount of flux involved in the process is very small since only the ball underside receives flux.

Additional times per placement cycle, depending on placement principle, are (1) ~0.7 s, pure pick and place, and (2) approx. 0.3 s, collect and place.

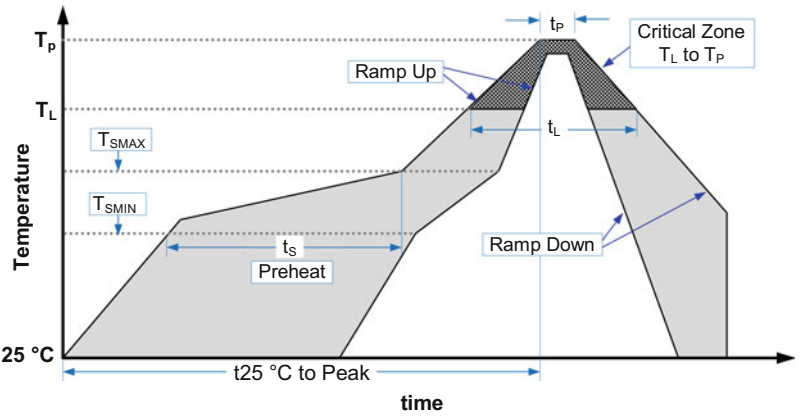
9.4.10 Summary

With respect to surface mounting of WLCSP components, placement equipment with key parameters like operating principle, flexibility, speed, accuracy, component feeding, and fluxing should be carefully examined to satisfy demanding assembly requirements of volume electronics manufacturing.

9.5 Solder Reflow

Reflow soldering is the most common method of attaching surface mount components to a circuit board. Reflow is also used on boards containing a mix of SMT and THT components. Doing through-hole reflow allows the elimination of a separate wave soldering step from the assembly process, potentially reducing the overall assembly costs.

The goal of the reflow process is to melt the solder and heat the adjoining surfaces, without overheating and damaging the electrical components. Also self-alignment of surface mounting components, i.e., WLCSP, to their corresponding PCB soldering pads is also expected due to the pulling force of surface tension of the molten solder. This self-alignment nature is important since components could be shifted from designated position either due to pick and place accuracy or due to the transportation movement from placement station to the reflow oven. In the conventional reflow soldering process, so-called reflow profile with stages,



Time Feature	
t_S	Soak time
t_L	Time above liquidus
t_P	Time at actual peak temperature
Temperature Features	
T_{SMIN}	Minimum soak temperature
T_{SMAX}	Maximum soak temperature
T_L	Liquidus temperature of solder

Fig. 9.11 Solder reflow profile and common terms without specific values. Individual reflow profile setting should be evaluated based on PCB design, component type, size and quantities, solder type and solder paste/flux types, as well as available equipment, such as heat/cooling control and number of equipment zones

sometimes called "zones," are often referred. Typical stages include preheat, soak, reflow, and cooling. Figure 9.11 showcase a typical reflow profile with unique terms used by the industry.

There are a number of techniques for reflowing solder. One is to use infrared lamps; this is called infrared reflow. Another is to use a hot gas convection. Another technology which is becoming popular again is special fluorocarbon liquids with high boiling points which use a method called vapor phase reflow. Due to environmental concerns, this method was falling out of favor until lead-free legislation was introduced which requires tighter controls on soldering. Currently, convection soldering is the most popular reflow technology using either standard air or nitrogen gas. Each method has its advantages and disadvantages. With infrared reflow, the board designer must lay the board out so that short components do not fall into the shadows of tall components. Component location is less restricted if the designer knows that vapor phase reflow or convection soldering will be used in production. Following reflow soldering, certain irregular or heat-sensitive components may be installed and soldered by hand, or in large-scale automation, by focused infrared beam (FIB) or localized convection equipment.

9.5.1 Preheat Zone

After surface mounting and inspection for placement qualities, the boards with components held by sticky solder paste/flux are then conveyed into the solder reflow oven. They first enter a preheat zone, where the temperature of the board and all the components, large or small, is gradually, uniformly ramped up. The preheat zone is often the lengthiest of all the reflow zones. The ramp-up-to-preheat rate is usually somewhere between 1.0 and 3.0 °C per second, often falling between 2.0 and 3.0 °C (4–5 °F) per second. If the rate exceeds the maximum slope, damages (cracks) could occur to components that are sensitive to thermal shock. Also undesired is the explosive vaporization of low boiling point materials that are either part of the flux or have been absorbed during use. Alcohol and other solvents and absorbed moisture can explode if heated too fast passing boiling points. Solder paste and flux spatter are the most common indicators that heating is too fast. If needed, slow down ramp-up rate to 130 °C to dry the paste/flux more gradually so it does not explode. If the rise rate (or temperature level) is too low, evaporation of volatiles could be incomplete.

9.5.2 Soak

Much of this soak concept is rooted back to the early time of the surface mount technologies, when the infrared oven is the primary means for reflow soldering. Infrared energy absorption is very inconsistent on a populated circuit board due to uneven heating of components of different surface color and finishes, and shadowing under big neighboring components. So a "soak zone" is designed to let the board and components equalize in temperature after heating up to a safe temperature below the solder reflow temperature. With infrared oven, temperature differentials from point to point of more than 40 °C were not uncommon. It takes some time for thermal energy to conduct around and achieve point-to-point temperature differential of less than 5 °C, hence the aforementioned "soak zone."

9.5.3 Reflow

The boards then enter a zone where the temperature ramps up quickly to above the melting point of the solder particles in the solder paste and solder bumps on the WLCSP and/or BGA components, joining the components or component lead to the pads on the circuit board. This is the stage where molten solder is to wet out on the soldering surface and sidewalls in a controlled fashion. Also surface tension of molten solder self-aligns the mounted components to the centering locations to minimize the overall system energies.

To facility solder wetting, flux has to be activated to prepare the surface for the molten solder to wet out. After preheat, beyond the 130 °C point, where water and most low boiling point materials have finished evaporating, at the temperature

range up to alloy solidus, the activators in the flux fulfill their function, cleaning the solder and substrate of oxides. Spend too long here and the flux activity can be mostly or completely consumed, resulting in poor wetting and what appears to be unreflowed solder despite achieving a peak well above alloy liquidus. Longer heating processes can be accommodated with higher activity flux formulations. As a goal, do not spend any more time between 130 °C and the alloy solidus than is required for acceptable point-to-point temperature differential on the product.

Time above liquidus and peak temperature should be based on process robustness. Typical recommendations of peak temperature 15–40 °C above liquidus are based on the fact that solder alloy wets better the hotter it gets. Temperatures in this range are considered necessary to ensure optimal wetting and formation of desired solder joint shape for reliability. In practice, solder joints form at just a few degrees above liquidus, but wetting may not look typical. These reflow recommendations are also based on common processing limitation for circuit board materials. If circuit board material is not as temperature sensitive, getting hotter may not have a down side.

Keep in mind the goal of a good reflow is to get the right amount of heat to the right location within the targeted cycle time. Success is measured in first pass yield and throughput, not degrees C per second and time above liquidus. An ideal heating cycle is no longer than the time it takes to make sure every joint has reflowed and wet out completely. So the time above liquidus only matters to the extent that all the solder needs to reflow and finish wetting.

The truth of the matter is that it only takes a second or three for liquid solder to finish moving and wetting to the available surfaces after it reaches reflow temperature. Time beyond that spent above liquidus is not generating any benefit with regard to joint quality. On the other side, extra time thickens the often brittle intermetallic layer and increases substrate scavenging.

Intermetallic compounds (IMC) form when two unlike metals diffuse into each other. In soldering this is tin into copper, nickel, and other solderable materials. Excessively thick intermetallic layers should be avoided if little topography exists on soldering surface, because brittle IMC fracture could happen a lot earlier than typical solder joints with controlled IMC layer.

Scavenging is most often of concern with plated parts and printed thin films. In extreme cases, the entire solderable surface can be dissolved. Again, more time above liquidus is not a benefit.

So to summarize, reflow zone should be well thought through to ensure good quality solder joints and acceptable IMC layer and controlled consumption of soldering pad metal thickness.

9.5.4 Cooling

As soon as it reaches peak temperature and holds for a time often limited by the reflow oven, cooling should start. In general, it is advantageous to cool near the maximum rate, which is typically 6.0 °C/s. Maintaining high cooling rate keeps

the solder from crystallization and formation of large grain or even single crystal solder joints, which is more prone to happen for lead-free solder with high tin concentration (>95 %) than the old eutectic solder (67 % Sn). Polygrain solder joints are important for isotropic mechanical behavior and improve the mechanical properties of the solder. Solder joints with shiny finish after cooling indicate controlled crystallization; a dull finish is the evidence of crystallization.

Cooling rates after solder solidus point can also be critical. A gradual cooling rate allows the still soft solders to creep and release the mechanical stresses from cooling down of CTE mismatched assembly components and PCB, which is proportional to the product of the CTE differences, temperature differential (~200 °C for typical lead-free solder), and distance to the neutral point of the component. Large components and components with CTEs significantly different than the PCB CTE are more susceptible to CTE mismatch cooling damages.

9.5.5 Reflow Oven

Many of the solder reflow control and fine-tuning are only possible with today's state-of-the-art equipment. Ever since the first forced convention reflow oven was introduced in 1987, this mainstream technology has evolved to accommodate challenging demand of today's electronics manufacturers. Today's reflow oven is being compared in the number of programmable precision heating and cooling zones, maximum heating/cooling rates, throughput, energy and inert gas (nitrogen) consumption, and whether there is need for frequent maintenance. Differential heating of top and bottom of PCB assembly is also possible for better substrate/PCB warpage management.

To accelerate profile set up, today's advanced reflow oven also comes with software that includes a database of over 1,000 different solder paste formulas from dozens of different solder paste manufacturers. Each database entry includes the manufacturer's recommended solder profile for that specific solder paste formula and identifies critical specifications (i.e., maximum ramp rate; maximum cooling rate; temperatures for preheat, soak, and reflow; and peak, allowable times above those temperatures, with upper and lower limits for each) that define the process window. Instant reflow profile can be set by simply entering the length, width, and weight of the PCB, an extensive profile and paste library with dynamic structure does the rest of the work for you! After selecting the solder paste, users can either use the profile as is or can modify the specifications to create a process window that meets their own unique criteria.

With addition of more heating/cooling zones for flexible yet precise reflow profile control, the most advanced reflow oven takes substantial floor space than their predecessors and takes more thoughtful planning in production setup. The example given in Fig. 9.12 is a reflow oven that boasts 13 top/bottom independently controlled heating zones; plus 3 blow-through cooling modules that could also be configured as top/bottom independently controlled. The oven features a belt speed

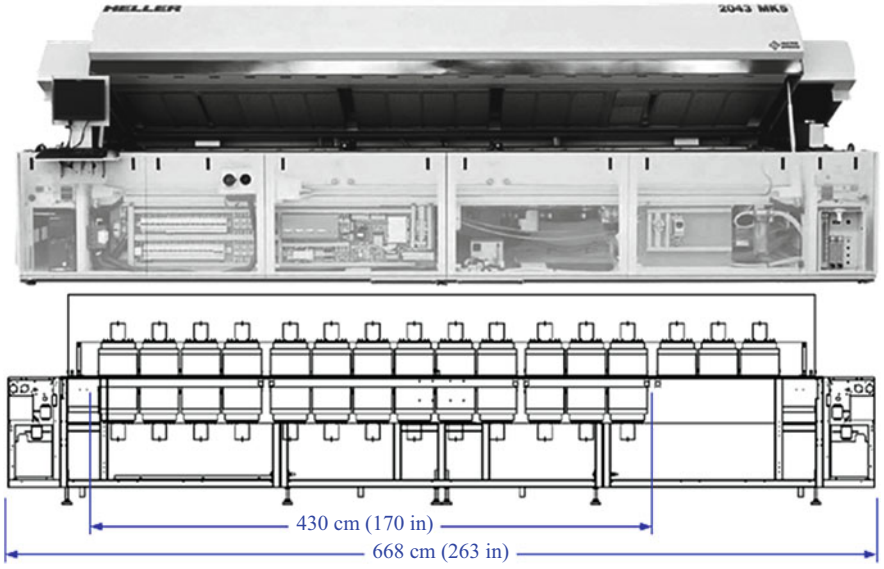


Fig. 9.12 Front view and configuration of a 13 + 3 zone forced convection solder reflow oven

up to 1.4 m/min to match with fast pick and place systems, and size is an impressive 6.68 m in length.

9.5.6 WLCSP Reflow

The Sn–Ag–Cu solder alloy melts at ~217 °C. Reflow peak temperature at joint level should be 15–20 °C higher than the melting temperature. FSC WLCSP is qualified for 260 °C reflow. A typical temperature/time profile for the lead-free (Sn–Ag–Cu) solder and the corresponding critical reflow parameters based on JEDEC JTSD020D are shown below. The actual profile for individual application depends on many factors such as the size of the package, complexity of the PCB assembly, oven type, solder paste type, temperature variation across the board, oven tolerance, and thermal couple tolerance.

9.5.7 Reflow Profile and Critical Parameters for Lead-Free (Sn–Ag–Cu) Solder

Ramp-down rate should be below 6 °C/s to prevent incurring stresses on the package and to create finer solder grain structure, which may affect reliability. On the other hand, slower than 2 °C/s ramp-down rate is prone to produce large Ag_3Sn aggregations in high silver solder and that is also not desired. Convection

type or combined convection IR reflow may be used. Nitrogen purge environment is preferable to improve solder wetting. Normally oxygen level should be less than 1,000 ppm.

WLCSPs are generally rated at a moisture sensitivity level (MSL) of 1; thus, prebaking is not required before the assembly. SMT reflow profiles suitable for solder paste/flux selection can be used as the base for optimization surface mounting WLCSP on to the selected substrates. Ranges of recommended parameters for the WLCSP reflow profiles are shown in Table 9.1 below

In all cases, a temperature gradient of 3 °C/s or less should be maintained to prevent warpage of the package and to ensure that all joints reflow properly. Additional soak time and slower preheating time may be required to improve the outgassing of solder paste. The reflow profile also depends on the PCB density and the type of solder paste used. Final adjustments of the reflow profile should be made according to the device's application requirements. Standard no-clean solder paste is generally recommended. If another type of flux is used, removal of flux residual may be necessary. If solder balling occurred during reflow, the paste aperture may be decreased to reduce the amount of paste deposited to the PCB. Nitrogen is also recommended to help counteract the tendency of the paste to form solder balls as well as provide wider solder wetting window. During the assembly process, the PCB must be supported properly to ensure board flatness. Typically, supports under the board are provided at each workstation but are not always available in a conveyor system. When thin or large boards are populated with components, the weight of the board, itself, plus the weight of the components may deform the PCB in reflow operation and cause the board to sag. This effect may shift a component from a designated location on the board after placement. Consequently, there is a large variation in solder joint height, which increases the potential for solder joint defects such as "bridging" and "opens." Generally, carriers need to be designed and used for large or thin boards to ensure proper board flatness. Although a certain amount of voids does occur in the solder joints, a 20 % voiding distributed across the solder joints as small voids should be acceptable.

Table 9.1 Typical SMT reflow profile parameters for WLCSP

Profile feature	Values
Average ramp rate (T_L to T_P)	1–3 °C/s
Preheat	
Minimum temperature (T_{SMIN}) Maximum temperature (T_{SMAX}) Time (T_{SMIN} to T_{SMAX}), t_S	130 °C 200 °C 60–75 s
Ramp-up rate (T_{SMAX} to T_L)	1.25 °C/s
Time maintained above liquidus (t_L)	60–150 s
Liquidus temperature (T_L)	217 °C
Peak temperature (T_P)	255–260 °C
Time within 5 °C of actual peak temperature (t_P)	20–30 s
Ramp-down rate	3 °C/s max
Time from 25 °C to peak temperature	480 s maximum

9.5.8 Double-Sided SMT

It is not uncommon to see mobile electronics board with SMT components on both sides. This typically requires 2X reflow with fixtures and glue dots to hold on the bottom (first reflow) SMT component during reflow and during second solder paste printing. If a wave soldering process is used, then the parts must be glued to the board prior to wave soldering to prevent them from floating off when the solder holding them in place is melted.

Underfill is probably the worry-free option for WLCSP if they have to be reflowed on the bottom side of a double-sided PCB. In theory small parts get held in place by the surface tension of the solder. So glue dots might not be necessary. Differential top/bottom heating possible on modern reflow ovens that helps keep the bottom side cooler could be another option for managing the double-sided PCB reflow.

If WLCSP is reflowed on the bottom side without glue or underfill, it is important to know if there is enough surface tension on the solder joints to hold the package. If analysis determines that the package would drop, then the application and curing methodology of an appropriate adhesive should be provided.

The need for adhesive can be estimated using the following empirical calculation:

$$\text{Actual weight of the component} \leq \text{Total solder contact surface area of WLCSP in mm}^2 \times 0.665$$

9.5.9 Post Reflow Inspection

Visual and X-ray inspection is recommended after solder reflow for WLCSP solder joint size and shape irregularities. Solder joints with uniform surface appearance, shape, and size are evidence of good wetting and reflow process. Dull or grainy solder surface is not that uncommon for lead-free solder. These solder joints are acceptable.

High-resolution automated X-ray inspection has rapidly expanded its usage in production solder joint inspection. Many material flaw and quality characteristic affecting the solder joints can be detected, such as missing solder fillets, voids and blisters, solder joint bridges, non-wetting defects, and missing ball.

Though X-ray microtomography is available and offers unique nondestructive 3D virtual models, the traditional through transmission, 2D approach still provides more cost-effective, high-throughput imaging solutions that are far capable of differentiating a good joint from a bad one. When additional joint information is needed, off axis X-ray imaging can be acquired by positioning the device assembly at an angle to the X-ray source. It is widely accepted that bridges between the solder joints are easy to detect and the challenge comes with detecting opens. Here, off axis 2D imaging is often used to tell if there is a problem with that connection. Needless to say, it requires experiences; however, great amount of information

concerning the package connections can be obtained using this technique. To summarize, for high-volume inspection, automated 2D X-ray inspection (AXI) system is a must-have capability.

9.5.10 Flux Clean

After soldering and depending on the application, soldering flux activity, and board surface finish used and whether underfill will be applied, the boards may be washed to remove flux residues and any stray solder balls that could short out closely spaced component leads. There are three typical methods in cleaning the assembly: (1) boiling liquid bath with or without ultrasonic agitation, (2) liquid bath plus vapor, and (3) aqueous spray cleaning (most commonly used in the PCBA industry when cleaning is required using water-soluble fluxes). Rosin flux is removed with fluorocarbon solvents, high flash point hydrocarbon solvents, or low flash solvents, e.g., limonene (derived from orange peels), which require extra rinsing or drying cycles. Water-soluble fluxes are removed with deionized water and detergent, followed by an air blast to quickly remove residual water. However, most electronic assemblies are made using a "no-clean" process where the flux residues are designed to be left on the circuit board [benign]. This saves the cost of cleaning, speeds up the manufacturing process, and reduces waste.

Caution must be observed when conducting ultrasonic cleaning, because it may result in weakening of the solder joints of the assembly, especially for the fine pitch WLCSP. Typical liquid cleaning solutions are alcohol at 40 °C or other water-based cleaning solutions. An appropriate drying methodology must be used to ensure no water entrapment under the package.

Certain manufacturing standards, such as those written by the IPC—Association Connecting Electronics Industries—require cleaning regardless of the solder flux type used to ensure a thoroughly clean board. Even no-clean flux leaves a residue which, under IPC standards, must be removed. Proper cleaning removes all traces of solder flux as well as dirt and other contaminants that may be invisible to the naked eye. However, while shops conforming to IPC standard are expected to adhere to the association's rules on board condition, not all manufacturing facilities apply IPC standard, nor are they required to do so. Additionally, in some applications, such as low-end electronics, such stringent manufacturing methods are excessive both in expense and time required.

IPC/EIA J-STD-001 provides acceptance criteria post soldering cleaning, particularly on the rosin flux residues, ionic residues, and other surface organic contamination. IPC-TM-650 provides the test methodology for these.

Finally, the boards are visually inspected for missing or misaligned components and solder bridging. If needed, they are sent to a rework station where a human operator repairs any errors. They are then usually sent to the testing stations (in-circuit testing and/or functional testing) to verify that they operate correctly.

9.5.11 Rework

WLCSP components removed during PCB rework should not be reused for final assemblies. A WLCSP that has been attached to a PCB and then removed has seen 2–3 solder reflows depending on whether the PCB is double sided. This is at or near the end of the tested and qualified three solder reflows' survivability for typical WLCSP. The removed WLCSP components should be properly disposed of so that they will not mix in with fresh equivalent WLCSP components.

WLCSP component removal and replacement procedures should be established and qualified. A reference rework process follows this flow:

1. The floor life of moisture-sensitive components starting from the time the moisture barrier bag is opened and mounted component(s) is exposed to ambient conditions—the floor life of the component should not have been exceeded before rework. Baking may be required in case this has been exceeded.
2. Preheating of the whole PCB assembly before localized heating of the component for rework—preheating reduces the overall heating time and prevents potential substrate warping when localized heating is applied on the region for rework. Typical preheating temperature is around 100 °C.
3. Localized heating on the region for rework—it is recommended to conduct localized heating on the component for rework to minimize the heat exposure to the surrounding components. Hot air gun equipped with a thermocouple to monitor the temperature at the component site is preferred. Once solder interconnect reaches the specified reflow temperature, a vacuum pickup tool is used to remove the component from the board.
4. Cleaning of the board land pads—residue solder is removed using a soldering iron and a braided solder wicking material. A vacuum-desoldering tool can also be used to extract the solder by continuous vacuum aspiration of the solder. After removing the residue solder, leftover flux must also be removed.
5. Solder paste or flux printing—solder paste is usually deposited using a miniature stencil and squeegee. In case where space is limited, flux is deposited onto the solder pads.
6. Component placement on the board—replacing WLCSP component can be placed onto the board using automatic placement equipment.
7. Soldering or reflow—this can either be selectively soldering the component using the same tools for removing the component or by passing the whole board to the original reflow profile.

9.5.12 Underfill

Although underfilling WLCSPs is often considered to be undesirable due to the added process complexity and cost, it has been demonstrated to be beneficial in board level reliability testing that includes thermal cycling, drop testing, and board

bending performance. Besides, it helps to hold component fall-off when reflowed upside down.

As a result, underfills have been effectively used to improve solder joint reliability. Underfills enhance WLCSP board level reliability with the impact conditions associated with mobile electronics. A number of applications demand high reliability, which include medical, automotive, industrial, and military electronics. The decision to select and use a specific underfill should be carefully considered, and the effectiveness of the desired underfill should be evaluated by the customer.

Underfill selection depends on several factors such as the combination of WLCSP size, WLCSP construction material and dimensions, circuit board structure and materials, and reliability requirements. The following are a list of recommendations when selecting the underfill:

- High CTE epoxy is not acceptable as an underfill material, because CTE mismatch causes higher stress on WLCSP products. Likewise, silicone cannot be used, because it does not provide the enhanced mechanical support needed for an underfill material. Underfill CTE must be close to the CTE of the solder joint.
- Significant differences are often observed in performance from different underfill types and suppliers. It is strongly recommended the use of optimized underfill material derived from an actual DOE.

9.5.13 WLSCP Underfill Process Requirements

9.5.13.1 Needles

Needles are very important in the manipulation of underfill flow. Many types and sizes of needles are available on the market:

- Conventional metal shafts—For additional heat application on needle for improved dispense
- Plastic tips and shafts—Prevents die chipping and scratch on PCB
- Tapered plastic tips—Reduces back pressure in the pump and ideal for fine pitch dispensing

0.25-inch needle keeps back pressure on the pump low. Needle diameter controls the line width during dispensing. It is recommended to start with 22-gauge needles, with a 410 μm inner diameter, to dispense underfills at a rate of 10–20 mg/s. Use lower-gauge needles when dispensing minimal underfill on very small packages. Twenty-gauge needles with a 610 μm inner diameter enable good control at larger flow rates. For manually dispensed underfill, use a plastic conical tip to reduce contact on the edge of the die and reduce mechanical damage.

9.5.13.2 Prebake

Substrates must be free from moisture for a good and reliable underfill. Prebake is necessary to prevent voiding and delamination during underfill cure. Plasma

cleaning improves underfill wetting, fillet height, and uniformity and promotes effective interface adhesion. As a result, plasma treatment prevents delamination and void formation that can result in a shortened lifetime for microelectronic devices.

9.5.13.3 Dispensing

It is recommended to use an auto-dispensing machine for underfill dispense to reduce mechanical damages caused by manually dispensing the underfill on the edge of the die. Underfill volume is controlled to optimize reliability and appearance. Ideal underfill should be dispensed to completely fill the solder ball area of the die and provide good fillet that covers greater than 50 % of the edges of the die but not more than 75 %. Dispense volume variations lead to undesired fillet size variations. Estimation of volume is possible with simple calculations. Determine the final volume by trial and error by processing a number of assemblies each with a different volume of underfill and reliability testing. A change of substrate supplier or substrate manufacturing process or solder ball type requires another volume evaluation.

Because the total bond area of the solder ball is always much smaller than the respective areas of the die and the substrate, the stress on an individual solder ball is relatively large. By absorbing energy during thermal cycling, the underfill reduces this stress by a factor of about 10. When no underfill is present, the solder ball absorbs the stress created by the mismatch in CTE between the package and the PCB which is typically large.

Underfill also prevents solder extrusion during thermal cycling. Successfully applied underfill functions as an isotropic compression container around each solder ball and prevents them from extruding to form shorts within each other. At the same time, the underfill prevents the initiation of cracks in solder balls by the elimination of free surfaces at grain boundaries where cracks can propagate.

To some degree, underfill also serves as a heat sink to dissipate heat from the die. However, for this to occur, all regions of the cured underfill must have the same thermal characteristics; variations can cause overheating in the die.

9.6 WLCSP Storage and Shelf Life

Floor life of the package is the allowable period after removal of the SMD components from the moisture barrier bag and before the solder reflow process in an environment not exceeding 30 °C and 60 % RH ambient conditions. Classifications of moisture-sensitive packages per J-STD-020 and the floor life per J-STD-033 are tabulated in Table 9.2 below. WLCSP with little moisture absorbing materials, such as polymer repassivation materials which typically measure less than 20 μm total thicknesses, is rated as level 1 moisture sensitive without much surprise. However, fan-out WLCSP, with substantial epoxy over mold plus additional polymer repassivation, is often MSL level 3 rated.

Table 9.2 Classification of moisture-sensitive package

Level	Floor life at 30 °C/60 % RH environment
1	Unlimited
2	1 year
2a	4 weeks
3	168 h
4	72 h
5	48 h
5a	24 h
6	Mandatory bake before use, must be reflowed within the time limit specified

9.7 Summary

Unique requirements exist for WLCSP assembly mainly due to the fact of being the only bare die package on a PCB assembly. Large bump array at fine pitch can also be a challenge. Besides typical assembly defects seen on other type of IC packages, major WLCSP assembly failure could be related to the damage on the silicon chip itself, and it could often trace back to the poor handling of WLCSP during machine pick and place and post-placement handling of assembly PCB by operators. In many occasions, underfill will not only provide stress relief to the critical solder joints but also add protection to the sensitive active side of the WLCSP.

References

1. Schiebel, G.: Criteria for reliable high-speed CSP mounting. Chip Scale Rev. September 1998
2. IPC-2221: Generic Standard on Printed Board Design
3. IPC-SM-7351: Generic Requirements for Surface Mount Design and Land Pattern Standard
4. IPC-7525: Stencil Design Guideline
5. J-STD-004: Requirements for Soldering Fluxes
6. IPC/EIA J-STD-001: Requirements for Soldered Electrical and Electronic Assemblies
7. IPC-TM-650: Test Methods
8. IPC-9701: Performance Test Methods and Qualification Requirements for Surface Mount Solder Attachments
9. IPC/JEDEC J-STD-020: Moisture/Reflow Sensitivity Classification for Non-hermetic Solid-State Surface Mount Devices
10. JEITA Std EIAJ ED-4702A: Mechanical Stress Test Methods for Semiconductor Surface Mounting Devices
11. JESD22-A113: Preconditioning Procedures of Plastic Surface Mount Devices Prior to Reliability Testing
12. IPC/JEDEC-9702: Monotonic Bend Characterization of Board-Level Interconnects
13. IPC/JEDEC J-STD-033: Handling, Packing, Shipping and Use of Moisture/Reflow Sensitive Surface Mount Devices
14. CEI IEC 61760-1: Surface mounting technology—Part 1: Standard method for the specification of surface mounting components (SMDs)
15. IEC 60068-2-21 Ed. 5: Environmental Testing—Part 2-21: Tests—Test U: Robustness of terminations and integral mounting devices

16. JESD22-B104: Mechanical Shock
17. JESD22-B110: Subassembly Mechanical Shock—Free state, mounted portable state, mounted fixed state
18. JESD22-B111: Board Level Drop Test Method of Components for Handheld Electronic Products
19. JESD22-B113: Board Level Cyclic Bend Test Method for Interconnect Reliability Characterization of Components for Handheld Electronic Products
20. IPC-7095: Design and Assembly process implementation for BGAs
21. Fan, X.J., Varia, B., Han, Q.: Design and optimization of thermo-mechanical reliability in wafer level packaging". Microelectron. Reliab. **50**, 536–54 (2010)
22. Syed, A., et al.: Advanced analysis on board trace reliability of WLCSP under drop impact. Microelectron. Reliab. **50**, 928–936 (2010)
23. Liu, Y., Qian, Q., Qu, S., et al.: Investigation of the Assembly Reflow Process and PCB Design on the Reliability of WLCSP", 62nd ECTC, San Diego, California, June 2012
24. Schiebel, G.: Criteria for Reliable High-Speed CSP Mounting. Chip Scale Rev. September 1998
25. Oresjo S.: When to Use AOI, When to Use AXI, and When to Use Both", Nepcon West, December 2002

WLCSP Typical Reliability and Test 10

WLCSP is one of the fastest growing segments in semiconductor packaging industry due to the rapid advances in integrated circuit (IC) fabrication, small form factor, and low cost. This technology results in a lower cost per die (vs. traditional wirebond) when the die count per wafer is high. As the number of I/O per die increases (and thus the die size and the distance to neutral point increases), the WLCSP may not achieve prescribed solder joint reliability requirements; the metal stack (UBM and the Al pad), passivation, or polyimide may also appear to fail, especially when the WLCSP is mounted on the PCB. The board level reliability is a big concern for both analog and power WLCSP packaging. This chapter will discuss the WLCSP typical reliability test.

10.1 WLCSP Reliability Test in General

This section introduces the general concepts of reliability life, failure rate, and typical reliability tests for power semiconductor packaging [1].

10.1.1 Reliability Life

According to the international standards the term quality is defined as totality of characteristics of a power electronic product that bear on its ability to satisfy stated and implied needs. Reliability is the property of the power semiconductor to maintain all its functions during usage. Since it is not possible to determine the long-range reliability of power semiconductor components prior to production release under realistic conditions, accelerated life tests must be applied which allow reliable results about the reliability of the components after a short test period. To achieve an acceleration effect the reliability tests are carried out under greater stress than in application. According to the familiar failure rate curve (bathtub curve) we distinguish between early failures, random failures (failures

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9_10

with constant failure rate), and failures resulting from wear-out and fatigue. While applying a so-called "burn-in" for integrated circuits to catch early failures prior to the final application this does not make sense for power semiconductors because of substantial higher costs. Provided there is no misapplication caused by the user early failures must be avoided by complete control and mastery of the manufacturing processes. Excluding short time overload during operation random failures are determined by reproducibility and safety margins of the manufacturing parameters. The early design phase of a product already decides about failures caused by wear-out and fatigue designing parts and processes and selecting the material.

10.1.2 Failure Rate

It is difficult to find a distribution function which will allow the whole bathtub curve. However, for each section of it the Weibull distribution is applicable.

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] \tag{10.1}$$

where

$F(t)$: probability that the device fails in the interval [0,t]
η: characteristic life
β: shape parameter
t: time; number of cycles

From this equation follows the failure rate (hazard function):

$$\lambda(t) = \left(\beta/\eta^{\beta}\right) \times t^{\beta-1} \tag{10.2}$$

The shape parameter means:

$\beta = 1$ constant failure rate (random failures)
$\beta < 1$ decreasing failure rate (early failures)
$\beta > 1$ increasing failure rate (wear-out, fatigue)

(a) Random failures

In terms of electrical failures usually the shape parameter $\beta = 1$ is applicable. This particular Weibull distribution then is called Exponential distribution

$$F(t) = 1 - \exp(-\lambda \times \text{ t }) \tag{10.3}$$

where $\lambda = 1/\text{MTTF}$ the constant failure rate.

Failure rate often is estimated by experiment using the formula:

$$\Lambda = r/(n \times t) \tag{10.4}$$

where

r: number of failures
n: sample size
t: test time

MTTF: Mean Time To Failures, time in which 62.3 % of devices failed

Because of statistical nature of this number an extended formula takes into account a confidence limit—Upper Confidence Limit (UCL) = 60 %, which is the common value. Moreover, it is possible to calculate reliability data by computer models. The available computer models, however, have not been developed for power semiconductors. The failure rate model according to MIL-HDBK 217 serves therefore for the time being only as rough estimate.

$$\lambda[\text{ FIT}] = [(r + \Delta r)/(n \times t)] \times 10^9 \tag{10.5}$$

Δr: depending on confidence limit and number of failures
FIT: Failures In Time.

Constant failure rates allow reliability prediction by using an acceleration factor. This acceleration factor is calculated by means of the Arrhenius equation:

$$a_\text{f} = \exp\{E_\text{a} \times [T2 - T1/(T1 \times T2)]/k\} \tag{10.6}$$

E_a: activation energy
k: Boltzmann's constant $k = 8.6 \times 10^{-5}$ eV/K
$T1$: absolute application junction Temperature [K]
$T2$: absolute test junction Temperature [K]

The abovementioned acceleration factor is applicable at constant temperature. In case of temperature differences (temperature cycle, power cycle) other formulas have to be used.

To estimate life times for different ΔT frequently the Manson–Coffin relation is used which has been established originally for metals under low cycle plastic deformation. If plastic strain dominates, then

$$N_f \approx C \times (\Delta T)^{-n} \tag{10.7}$$

N_f: number of cycles to failure
ΔT: temperature difference
C: A material-dependent constant
n: an experimentally determined constant.

An extended life prediction may be obtained through finite element analysis for the estimated life for thermal cycling and power cycling in Sect. 9.2.

(b) Early Failures

Constant failure rate allows simple computation and prediction of life. In the regime of early failures prediction of failure rate is a challenge and failure rate is strongly time dependent.

(c) Wear-Out

Power semiconductors rarely come up to fatigue or wear-out. Even accelerated tests take a long time to show this. Usually reliability tests are stopped at a time or number of cycles far from the fatigue.

10.1.3 Typical Reliability Tests for Analog and Power WLCSP

Typical reliability tests used for analog and power WLCSP include:

1. **Solder Reflow Preconditioning (PRECON)**: The preconditioning stress sequence is performed for the purpose of evaluating the capability of semiconductor devices to withstand the stresses imposed by a user's printed circuit board assembly operation. A properly designed device (i.e., die and package combination) should survive this preconditioning sequence with no measurable changes in electrical performance. Furthermore, preconditioning of properly designed devices should not produce latent defects which lead to degraded reliability during life or environmental stress tests. Changes in electrical characteristics and both observable as well as latent physical damage during this stress sequence result principally from mechanical and thermal stresses and from ingress of flux and cleaning agents. Effects include die and package cracks, fractured wire bonds, package and leadframe delamination, and corrosion of die metallization. The preconditioning stress condition is listed in Table 10.1.

   Reference Industry Standard: JESD22-A113C

Table 10.1 Preconditioning stress conditions

Step	Stress	Conditions
1	Initial electrical test	Room temperature
2	External visual inspection	40× Magnification
3	Temperature cycling	5 cycles at –40 °C (max) to +60 °C (min) (Step is optional)
4	Bake out	24 h (min) at 125 °C
5	Moisture Soak	Per MSL rating
6	Reflow	3 cycles per referenced profile
7	Flux Application	10 s immersion in water soluble flux @ room temp
8	Cleaning	Multiple DI water rinses
9	Dry	Room temperature
10	Final Electrical Test	Room temperature

2. **Power Cycle (PRCL):** The power cycle test is performed to determine the effects on solid-state devices of thousands of power-on/power-off operations such as would be encountered in an automobile. The repetitive heating/cooling effect caused by multiple on/off cycles can lead to fatigue cracks and other degrading thermal and/or electrical changes in devices which generate significant internal thermal heating under maximum load conditions (i.e., voltage regulators or high-current drivers). This test forces junction temperature excursions at the rate of ~30 cycles per hour (typical for small WLCSP packages).
3. **High Temperature Reverse Bias Test (HTRB):** The HTRB test is configured to reverse bias major power handling junctions of the device samples. The devices are characteristically operated in a static operating mode at, or near, maximum-rated breakdown voltage and/or current levels. The particular bias conditions should be determined to bias the maximum number of solid-state junctions in the device. The HTRB test is typically applied on power devices.

   Stress Conditions: 150 °C T_j, Biased
   Reference Industry Standard: JESD22-A108B

4. **High Temperature Gate Bias Test (HTGB):** The HTGB test biases gate or other oxides of the device samples. The devices are normally operated in a static mode at, or near, maximum-rated oxide breakdown voltage levels. The particular bias conditions should be determined to bias the maximum number of gates in the device. The HTGB test is typically used for power devices.

   Stress Conditions: 150 °C T_j, Biased
   Reference Industry Standard: JESD22-A108B

5. **Temperature Humidity Biased Test (THBT):** The steady-state temperature–humidity–bias life test is performed for the purpose of evaluating the reliability

of non-hermetic packaged devices operating in humid environments. It employs severe conditions of temperature, humidity, and bias which accelerate the penetration of moisture through the external protective material (encapsulant or seal) or along the interface between the external protective materials and the metallic conductors passing through it. When moisture reaches the surface of the die, the applied potential forms an electrolytic cell, which can corrode the aluminum, affecting DC parameters through its conduction, and eventually causes catastrophic failure by opening the metal. The presence of contaminants such as chlorine greatly accelerates the reaction as does excessive phosphorus in the PSG layers (passivation, dielectric, or field oxide).

Stress Conditions: 85 %RH, 85 °C
Reference Industry Standard: JESD22-A101B

6. **Highly Accelerated Stress Test (HAST):** HAST is performed for the purpose of evaluating the moisture resistance of non-hermetic packaged devices operating in high humidity environments. Bias is applied minimizing current draw using alternating potentials wherever possible. The test approximates a highly accelerated version of the THBT test. These severe conditions of pressure, humidity, and temperature, together with bias, accelerate the penetration of moisture through the external protective material (encapsulant or seal) or along the interface between the external protective material and the metallic conductors passing through it. When moisture reaches the surface of the die, the applied potential forms an electrolytic cell, which can corrode the aluminum, affecting DC parameters through its conduction, and eventually causes catastrophic failure by opening the metal. The presence of contaminants such as chlorine greatly accelerates the reaction as does excessive phosphorus in the PSG layers (passivation, dielectric, or field oxide).
Care must be taken when using HAST as a stress technique for WLCSP that has polyimide layer or mold compound with low T_g, since uncharacteristic failures may result.

Stress conditions: 130 °C, 85 %RH, 18.6 psig or 110 °C, 85%RH, 3psig
Reference Industry Standard: JESD22-A110B

7. **Autoclave (ACLV):** The autoclave (or pressure cooker) test is performed for the purpose of evaluating the moisture resistance of non-hermetic packaged devices. No bias is applied to the devices during this test. It employs severe conditions of pressure, humidity, and temperature not typical of actual operating environments that accelerate the penetration of moisture through the external protective material (encapsulant or polyimide) or along the interface between the external protective material and the metallic conductors passing through it. When moisture reaches the surface of the die, reactive agents cause leakage paths on the surface of the die and corrode the die

metallization, affecting DC parameters and eventually catastrophic failure. Other die-related failure mechanisms are activated by this method including mobile ionic contamination and various temperature- and moisture-related phenomena.

The autoclave test is destructive and produces increasing failure rates when repetitively applied. It is useful for short-term, comparative evaluations such as lot acceptance, process monitors, and robustness characterization but generates no absolute information since accelerating factors relating to the operating environment are not well established. In addition, the autoclave test can produce spurious failures not representative of device reliability, due to excessive chamber contaminants. This condition is usually evidenced by severe external package degradation, including corroded device terminals/leads or the formation of conducting matter between the terminals, or both. The autoclave test is therefore not suitable for measurements of package quality or reliability.

ACLV test is not required in the qualification of standard WLCSP products. However, during the development stages of a WLCSP, ACLV can be used to understand inherent weakness in the technology. Caution must be taken when interpreting results because failure mechanisms may be due to exceeding the capabilities of the package, producing unrealistic material failures. ACLV may be a required test for some customers or markets, such as the automotive market.

Stress Conditions: 100 %RH, 121 °C, 15psig
Reference Industry Standard: JESD22-A102C

8. **Temperature Cycle (TMCL):** The temperature cycle test is conducted for the purpose of determining the resistance of devices to alternating exposures at extremes of high and low temperatures. Permanent changes in electrical characteristics and physical damage produced during temperature cycling result principally from mechanical stress caused by thermal expansion and contraction. Effects of temperature cycling include solder cracking or cratering of die, cracking of passivation, delamination of metallization, and various other changes in the electrical characteristics resulting from thermo-mechanically induced damage.

   Stress Conditions: Various. −40 °C to +125 °C or –65 °C to +150 °C are typical.
   Reference Industry Standard: JESD22-A104D

9. **Board Level Temperature Cycle (BTMCL):** The BTMCL test is intended to provide fatigue-related wear-out information on the solder joint attachment of devices to circuit boards. Daisy chain structure test devices are mounted to circuit boards and cycled through temperature extremes typically in the range of 0 °C to +100 °C. During stress, the solder joint resistance is continuously

monitored and a unit is considered failing when five cumulative incidences of elevated resistance (>1,000 Ω) are detected. Ideally, testing should continue until a cumulative 63 % failure rate of the test sample has been observed.

Stress Conditions: 0 °C to +100 °C, 2 cycles/h
Reference Industry Standard: IPC-SM-785

10. **High Temperature Storage Life (HTSL):** The high temperature storage (also called the stabilization bake test) is employed for the purpose of determining the effects of storing devices at elevated temperatures without electrical stresses applied. It is also a useful test for determining the long-term reliability of wire bonds which are susceptible to formation of intermetallic voids (such as gold wire bonds on aluminum bond pads).

    Devices under test are subjected to continuous storage in a chamber with circulated air heated to +150 °C. At the end of the specified stress period, the devices are removed from the chamber, allowed to cool, and electrically tested. Interim measurements are made if specified in the detailed test procedure.

    Stress Conditions: 150 °C or 175 °C
    Reference Industry Standard: JESD22-A103B

11. **Solderability:** The purpose of the solderability test is to determine the solderability of all WLCSP terminations that are normally joined by a soldering operation. This determination is made on the basis of the ability of these terminations to be wetted or coated by solder. These procedures will verify that the treatment used in the manufacturing process to facilitate the soldering is satisfactory and that it has been applied to the required portion of the part which is designed to accommodate a solder connection. An accelerated aging test is included in this test method.

    The referenced standard also provides optional conditions for aging and soldering for the purpose of allowing simulation of the soldering process to be used in the device applications. It provides procedures for solderability testing of through-hole, axial, and surface mount devices and reflow-simulated use testing for surface mount packages. The WLCSP devices under test are first "aged" by exposure to steam for a period of 8 h. After aging the bumps of the device are fluxed in a solder bath heated to a temperature of 215 °C (SnPb board assembly processing) or 245 °C (Pb-free board assembly processing) for 5 s.

    Reference Industry Standard: JESD22-B102C

12. **Board level drop test**: This test is intended to evaluate and compare drop performance of WLCSPs for handheld electronic product applications in an accelerated test environment, where excessive flexure of a circuit board causes product failure. It is particularly applied to the power Mosfet and analog wafer-level chip-scale package.

Experience with different board orientation has suggested that the horizontal board orientation with components facing down results in maximum PCB flexure and, thus, the worst orientation for failures. Therefore, it requires that the board shall be horizontal in orientation with components facing in downward direction during the test. Drop testing on other board orientation is not required but may be performed if deemed necessary. However, this is an additional test option and not a replacement for testing in required orientation.

Drop test requires JEDEC Condition B (1,500 Gs, 0.5 ms duration, half-sine pulse), as listed in JESD22-B110 or in JESD22-B104-B, as the input shock pulse to the printed circuit assembly. This is the applied shock pulse to the base plate and shall be measured by accelerometer mounted at the center of base plate or close to the support posts for the board. Other shock conditions, such as Condition H (2,900 Gs, 0.3 ms duration), in addition to the required condition can also be used. In situ electrical monitoring of daisy chain nets for failure is required during each drop. The electrical continuity of all nets should either be detected by an event detector or by a high-speed data acquisition system. The event detector should be able to detect any intermittent discontinuity of resistance greater than 1,000 Ω lasting for 1 μs or longer. The high-speed data acquisition system should be able to measure resistance with a sampling rate of 50,000 samples per second or greater.

Reference Industry Standard: JESD22-B111

10.2 WLCSP Solder Ball Shear Performance and Failure Mode

The shear test under high strain rate is becoming a popular approach to investigate the fracture behavior of a thermally attached solder ball under different strain rates. In this study, the experimental results regarding effects of shear loading speed are illustrated, and then three-dimensional explicit finite element analysis is employed to study dynamic responses of solder joints under ball impact testing. Through a three-dimensional explicit element analysis incorporated with a cohesive model, fracturing and fragmentation mechanisms, transient fracturing of the solder joint subjected to high-speed impact test is investigated

10.2.1 Introduction

High-speed shear test is the evaluation method of the strength of solder ball that was attached to a WLCSP die. Although this test is simple and convenient to implement, the details of performing the test have not yet been standardized for all uses of the ball shear test.

Chai T.C. [2] also studied the BGA solder ball shear test and found the relationship between shear angle and reaction force. For convenience, the failure modes are classified under a few categories, namely bulk failure (mode 3), bulk-intermetallic

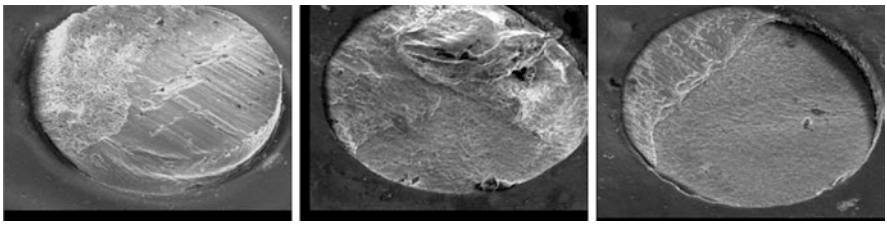


Fig. 10.1 Failure modes observed from shear test [2]

compound (IMC) partial failure (Mode 2), and IMC failure (Mode 1) illustrated in Fig. 10.1. In actual case, the most failure modes in the shear test are attributed to the bulk-IMC partial failure. This phenomenon is more complex and difficult to replicate by numerical simulation. Furthermore, fewer investigators have combined the experimental and numerical investigation for carrying out more in-depth studies.

In general, the past investigations, regardless of the test or simulation, are mostly based on BGA modules. Meanwhile, for the numerical study, the researches mainly targeted the pure bulk or IMC failure, but ignored a more important failure style: bulk-IMC partial failure. Moreover, the experimental data is scarcely used to testify their numerical results.

In this study, a WLCSP module is investigated for its mechanical response under an impact action. To allow realistic stress assessments of WLCSP modules during shear events, the dynamic mechanical behavior of SAC405 solder ball is determined by a combination of fast shear tests and FEM simulations. From the experimental result, the most failure mode is bulk-IMC partial failure. In this work a dynamic 3D finite element simulation, based on ANSYS®/LS-DYNA, is utilized in an attempt to catch such a typical failure mode. Comparing the results from experimental and numerical investigation, it can be concluded that the finite element analysis used in this study is reliable and able to approximate the experimental observation.

10.2.2 Test Procedure and Specimen

In the experiment, the sample of WLCSP mainly consists of a SAC405 solder ball (300 μm in diameter) and CuNiAu UBM (2 μm in total thickness). An intermetallic compound (IMC), about 2 μm thickness, is generally formed in the interface between the ball and the UBM, as shown in Figs. 10.2 and 10.3. Note that fracture in the IMC is frequently observed, especially in high impact speed cases. Accordingly, the IMC ineffective area is taken as an indication for comparing different failures. In general the IMC has a higher modulus and is more brittle than either the ball alloy or the UBM components. Examples of the pure metal and IMC Young's modulus are listed in Table 10.2 [3].

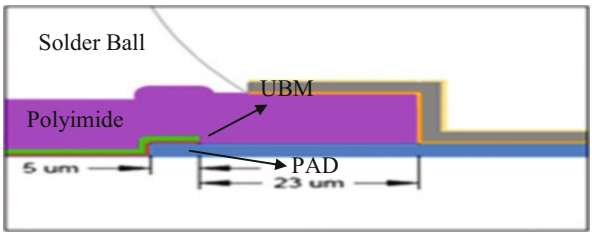


Fig. 10.2 Typical WLCSP modules

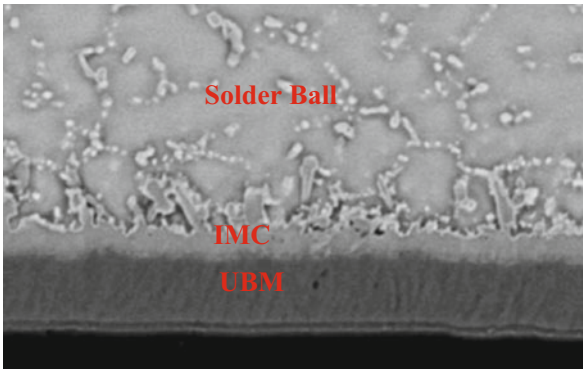


Fig. 10.3 A magnification of IMC layer in the solder interconnection

Table 10.2 Pure metal and IMC modulus values

Metal or alloy	Young's modulus	Source
Nickel	200–214 Gpa	Pure metal data table
Tin	50 Gpa	Pure metal data table
Copper	110–128 Gpa	Pure metal data table
Gold	78 Gpa	Pure metal data table
Silver	83 Gpa	Pure metal data table
Lead	16 Gpa	Pure metal data table
Intermetallic compounds		
Cu_6Sn_5	112.3 ± 5.0 Gpa	Nanoindentation
$(Cu, Ni)_6Sn_5$	157.82 ± 5.69 Gpa	Nanoindentation
Cu_3Sn	134.2 ± 6.7 Gpa	Nanoindentation
Ag_3Sn	78.9 ± 3.7 Gpa	Nanoindentation
Ni_3Sn_4	140–152 Gpa	Nanoindentation
$(Ni,Cu)_3Sn_4$	175.14 ± 4.12 Gpa	Nanoindentation

After the impact tests, the sheared samples are inspected by microscope for failure mode identification. To determine a meaningful shear speed range for the experiment, some initial shear tests have been performed to ensure the IMC ineffective area ranges from 50 % to 10 %. To induce this bulk-IMC partial failure, the tool speeds selected for this present study ranged from 400 mm/s to 800 mm/s.

Table 10.3 The statistical data of experimental failure modes

Speed (mm/s)	Residual solder on UBM after high-speed shear			Remaining solder (%)
	Mode 1 (10 %)	Mode 2 (50 %)	Mode 3 (100 %)	
400	7	2	6	51
600	7	7	1	34.7
800	12	4		20

10.2.3 Experimental Investigation of Impact Test

Through experimental tests, the sheared samples were inspected under optical microscope and SEM for their failure areas. Some experimental data is represented in Table 10.3. The designations of: Mode 1, Mode 2, and Mode 3 relate to three area percentages of solder alloy remaining on the fracture surface. The data in the last column represents the average area percentage of solder alloy remaining on the underlying UBM after shear at different speeds. From characterization of the failure modes as illustrated in Table 10.3, a conclusion can be derived: as the shear speed increases, the area of solder alloy remaining declines. Therefore more interfacial failures occurred at the IMC layer vs. within the bulk solder alloy.

In the experiment, the force–displacement curves are also recorded, as shown in Fig. 10.4 [3], with various shear speeds for SAC405 samples. For low shear speed below 400 mm/s, the peak of curve is smooth and flat. As the speed increases, the peak becomes relatively sharper. Another obvious rule regarding the peak force can also be seen from the curves: as the speed increases, the peak force rises, but the final displacement decreases.

10.2.4 Simulation and Analysis Based on FEM

Finite element simulation is well capable of visualizing the solder joint behavior and estimating the magnitude of the stress components in all parts of the WLCSP module. However, the accuracy of such virtual approaches is determined by adequacy, comprehensiveness, and reliability of the models describing the relevant behavior of the materials and the structure.

In this work, a 3D dynamic simulation based on ANSYS/LS-DYNA is conducted. In order to simulate the fracture of interface (IMC layer) subjected to an impact load, a cohesive zone model is used in combination with the general elastic-plastic constitutive model. Since the material property of IMC layer is difficult to obtain due to its thickness and morphology, an iterative calculation based on presuming different model parameters for IMC layer is performed to approximate the experimental observation. Different shear speeds are then applied to validate the above presumed parameters suitable for these different cases.

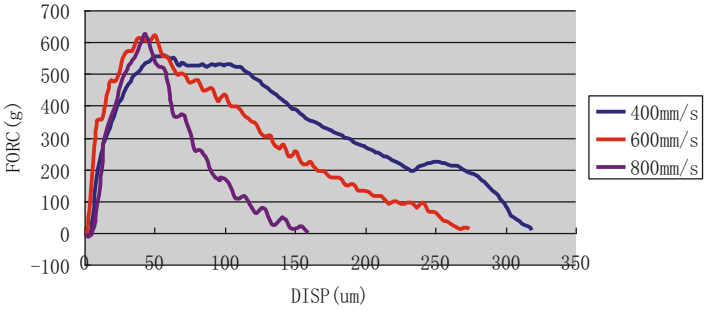


Fig. 10.4 Comparison of load–displacement response curves under different impact speeds [3]

10.2.4.1 Finite Element Model

Figure 10.5 is one-half finite element model of the shear test with a single solder ball, surrounded by a solder mask layer, the underlying structure of UBM on an aluminum pad. The IMC layer is located between UBM and solder ball. The shear tool was assumed to be a rigid body with constant velocity to impact the solder ball. Because the IMC layer is an important object of investigation, a magnification of this part is separately shown in Fig. 10.6, and the element number of complete IMC layer is noted beside.

10.2.4.2 Materials Parameters

Elastic properties for the constituent components are listed in Table 10.4. In the table, E is the Young's modulus, υ the Poisson's ratio, and ρ the mass density. The shear tool is assumed to be rigid.

The material of the solder ball is modeled by a strain rate-dependent bilinear elastic-plastic constitutive relationship. In the dynamic analysis, Cowper-Symonds model is widely applied for considering strain hardening due to strain rate dependence of material. It presents the ratio between dynamic flow stress and static flow stress as a function of strain rate for impact problems. It is well known that the solder alloy is highly strain rate dependent. Accordingly, Cowper-Symonds model is considered in this simulation. As the formula of model, two material constants must be set; in our research, initialize B as 106 and q as 2.35. The criterion for damage is met when the following condition is satisfied: the equivalent plastic strain reaches 0.6.

In determining the failure of solder ball under shear impact, the equivalent plastic strain criterion was performed. When the accumulated or current equivalent plastic strain reaches a critical value, the related elements will be deleted. As for the more complicated interface fracture (IMC layer failure), a cohesive zone model is adopted for simulating its failure.

In recent years, cohesive zone models have been widely employed to simulate fracture and delamination in solids. The approach is based on the cohesive zone concepts of Dugdale [4] and Barenblatt [5]. The cohesive zone model (CZM) originally was suggested by Needleman [6] for simulation of inclusion debonding

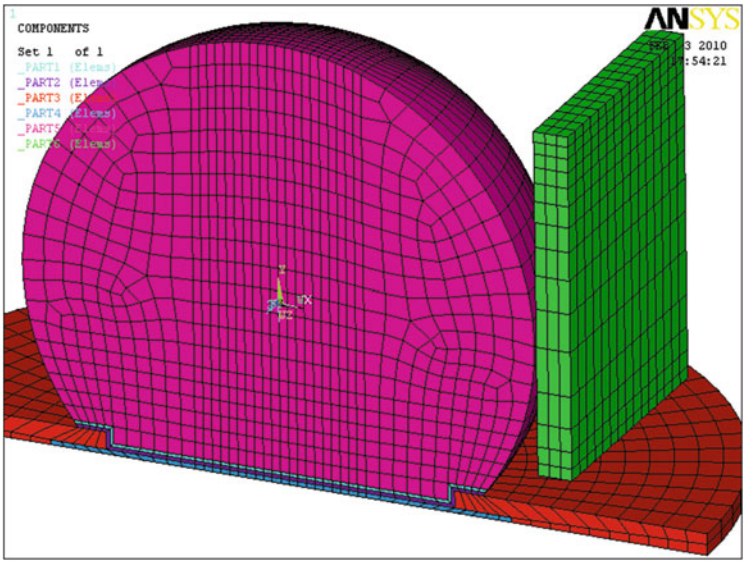


Fig. 10.5 Asymmetric finite element model of solder interconnection

from a metal matrix. In various numerical investigations like crack growth in homogeneous ductile materials, interface debonding [7], impact damage in brittle materials, and the analysis of sandwiched structures [8], the cohesive zone approach has been successfully employed. For the modeling of crack growth, cohesive zone elements have become an attractive concept, where geometric and constitutive thicknesses are defined independently.

In the cohesive zone model, a traction–separation law is imposed on the interface where the cohesive elements are embedded. The material characteristics of the cohesive elements are mainly determined by two important parameters: the peak traction and the maximum separation. In this work, a cohesive zone model as shown in Fig. 10.7 is chosen, which was firstly proposed by Tvergaard and Hutchinson [7]. Except for the above two parameters, fracture energy G is also an important parameter, which can be derived from the other parameters directly. From Fig. 10.7, we can derive that fracture energy is numerically equal to the area between the abscissa and curve. Here, σ_c represents the peak traction. When the displacement δ reaches to δ_{cn}, the material totally fails. The initial parameters are listed in Table 10.5.

10.2.4.3 Simulation Result

Through the simulation, the region of high plastic strain in the solder ball is found to occur initially at the contact point between the shear tool tip and the solder ball and expand through the solder in parallel to the pad. It implies a strong likelihood of crack initiation and growth through this region, as shown in Fig. 10.8.

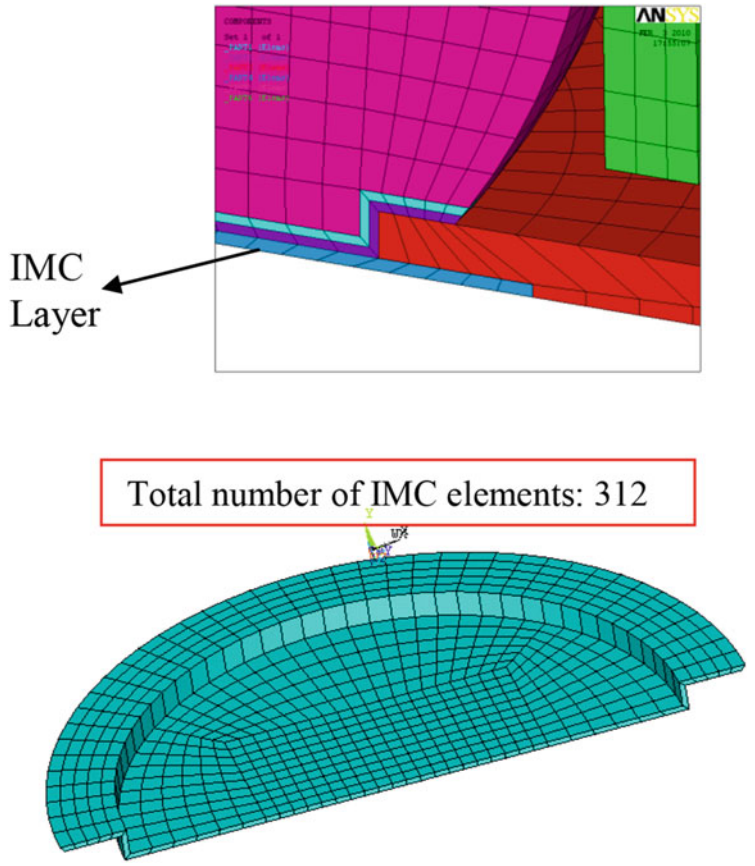


Fig. 10.6 IMC layer simulated by cohesive element

Table 10.4 Elastic properties of constituents

Material	ρ (kg/m^3)	E (GPa)	υ
SAC405	7.5×10^3	26	0.4
Al-Cu pad	2.7×10^3	69	0.33
Solder mask	1.47×10^3	3.5	0.35
UBM	9.7×10^3	200	0.3
Shear tool	7.9×10^3	rigid	

Figures 10.9 and 10.10 show the whole process of the failure at 400 mm/s and 800 mm/s impact velocity.

Figures 10.11, 10.12, and 10.13 show different bulk-IMC partial failure cases after total failure under different impact speed. In these sets of figures, Figs. 10.11, 10.12, and 10.13a show the plastic strain plots at different speeds, and also disclose the degree of solder ball failure. Figures 10.11, 10.12, and 10.13b display the remaining IMC elements at the corresponding speeds. Through calculation, the

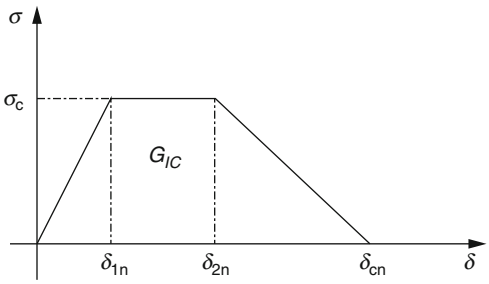


Fig. 10.7 Traction–separation response of cohesive zone model

Table 10.5 Initialization of cohesive material

600	0.05	0.95	1

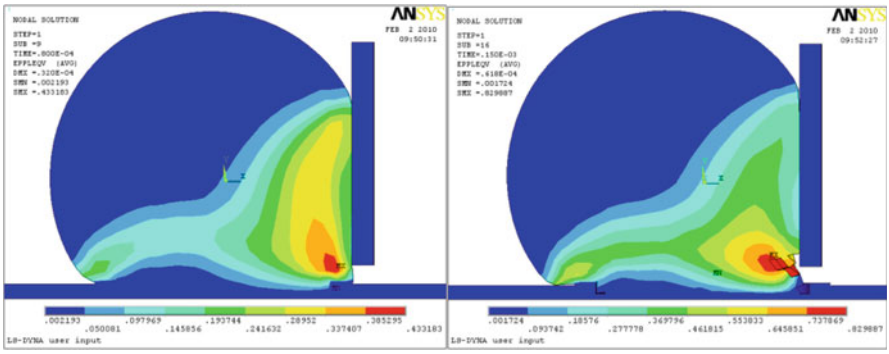


Fig. 10.8 Contour plots of equivalent plastic strain

number of the remaining IMC elements is 164 at 400 mm/s. Comparing with the intact IMC layer shown in Fig. 10.6, the remaining elements are around 50 %. According to the same calculation, 28.8 % and 7.4 % correspond to the cases of 600 mm/s and 800 mm/s, respectively. When the speed increases, the remaining elements of IMC layer or solder ball decrease. This tendency is in accordance with the experimental data listed in Table 10.3.

Figure 10.14 shows the calculated load–displacement response curves at 400 mm/s, 600 mm/s, and 800 mm/s. From this figure, the tendency coincides with the experimental results, as shown in Fig. 10.4. As the speed increases, the final failure displacement declines; however, the peak force increases. Additionally, the curve at low speed is much smoother.

From the results mentioned above, it can be concluded that the finite element simulation used in this study has the ability to approach the results obtained in the experiments. However, if this method is to be applied in the industry, it needs more work to obtain the accurate model parameters and better failure criterion.

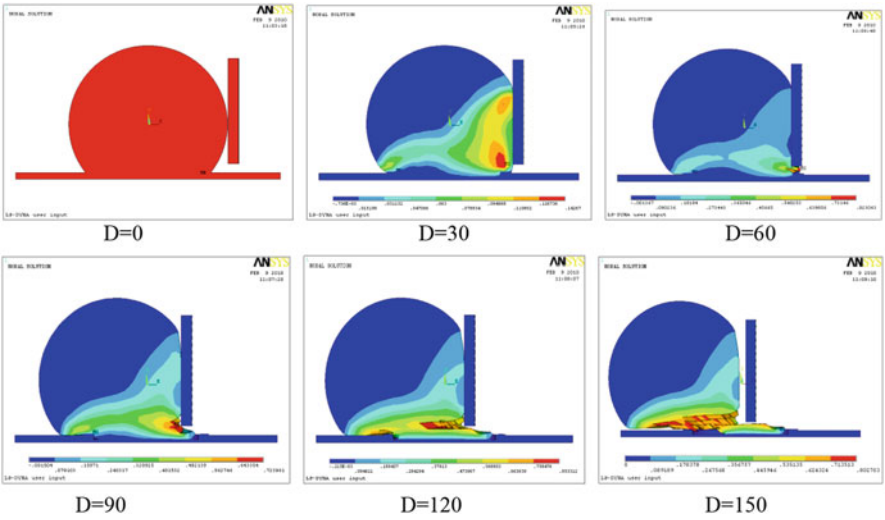


Fig. 10.9 The whole process of 400 mm/s. *D* is the shear tool displacement (μm)

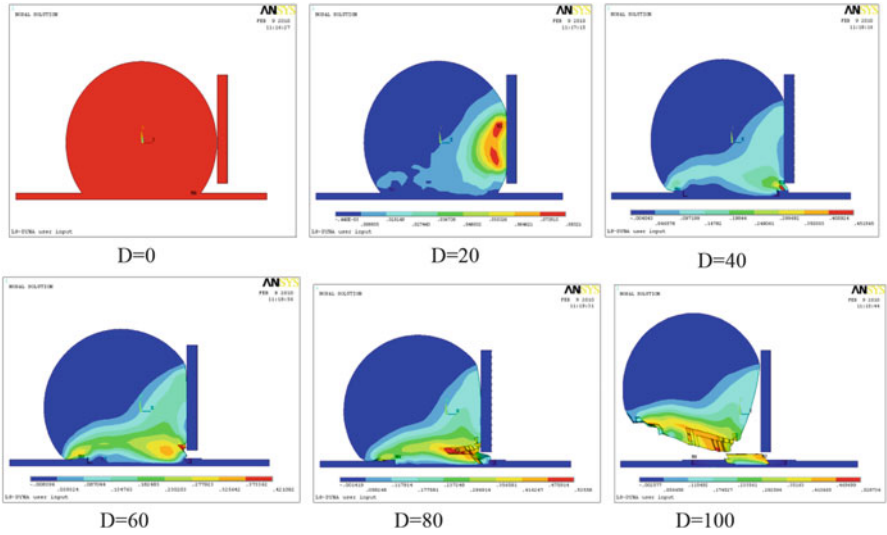


Fig. 10.10 The whole process of 800 mm/s. *D* is the shear tool displacement (μm)

10.2.5 Discussion

Bulk-IMC partial failure is a much more complicated mixed failure phenomenon, which combines the failures of solder bulk and IMC layer. At present, the numerical research regarding this issue is still in the early stages of development.

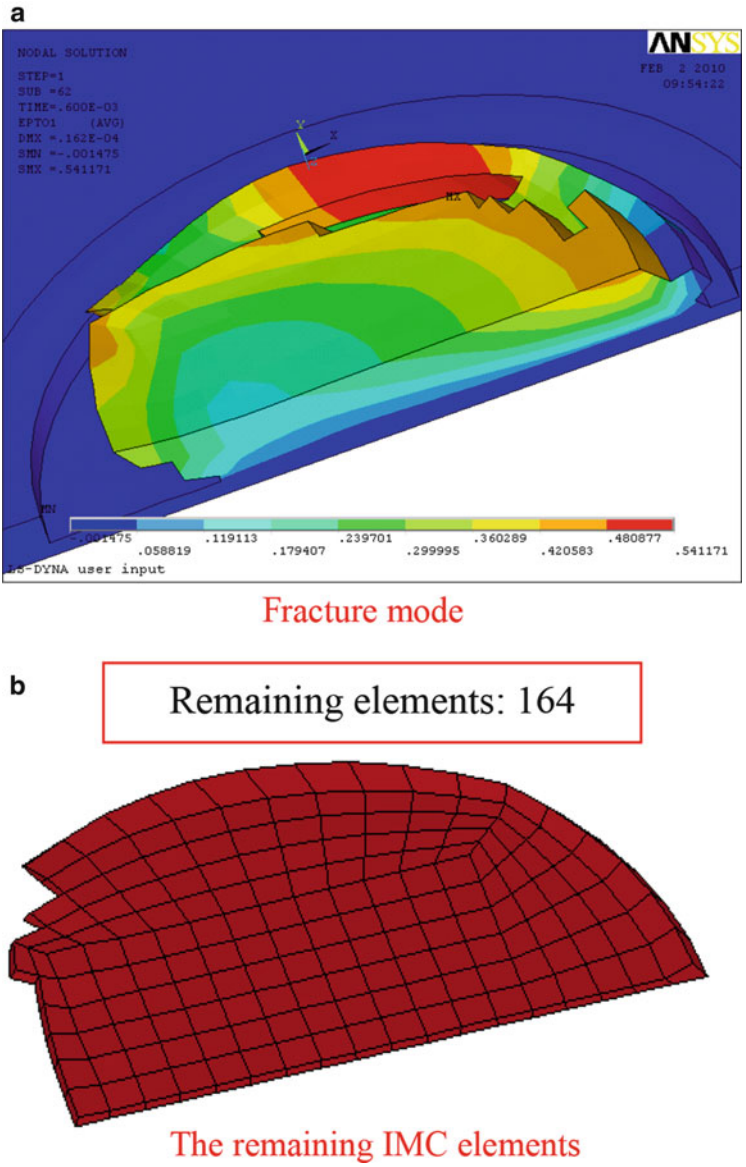


Fig. 10.11 Fracture surface (**a**) and remaining IMC (**b**) at 600 mm/s

In this section, an experimental study of shear test under different impact speeds was performed for WLCSP interconnection. Subsequently, a 3D dynamic simulation based on ANSYS/LS-DYNA FEM tool was conducted in an attempt to replicate the experimental phenomenon. Some significant results have been obtained. The main conclusions can be outlined as follows.

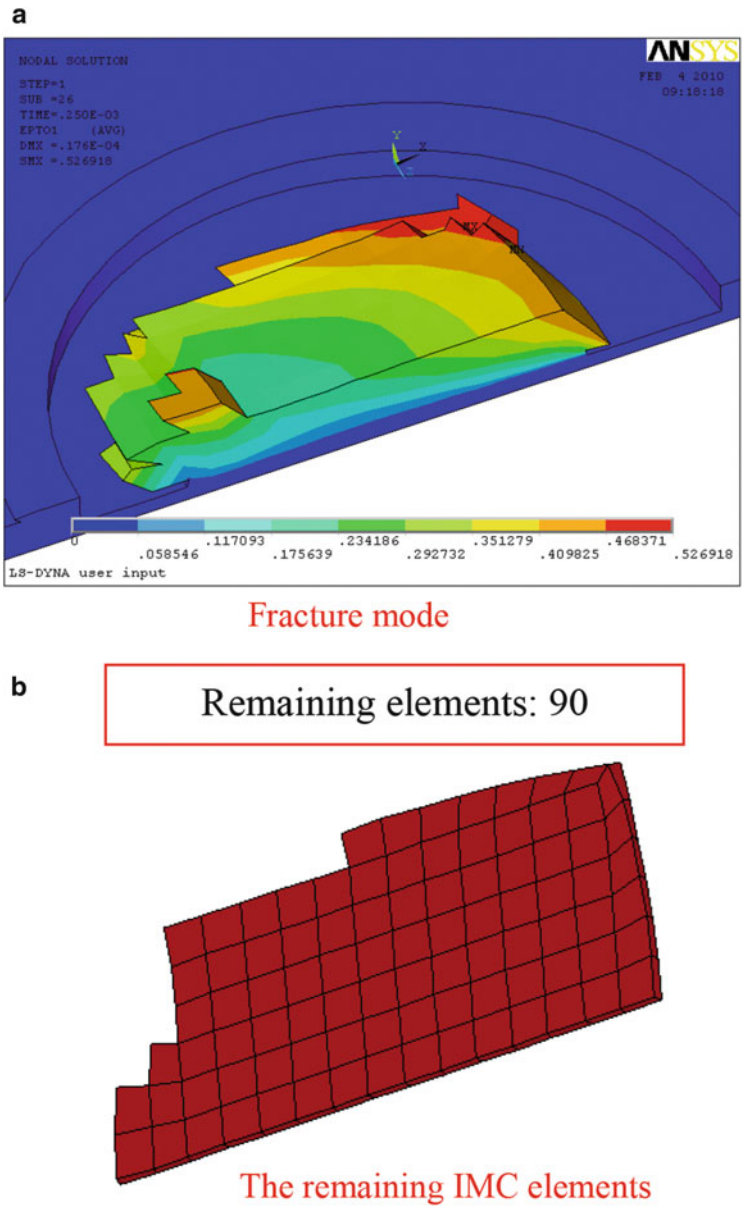


Fig. 10.12 Fracture surface (**a**) and remaining IMC (**b**) at 600 mm/s

1. Different failure modes frequently occurring in the shear test under different impact speeds can be successfully predicted by using a cohesive zone model combined with an equivalent plastic strain-based criterion and a strain rate-dependent constitutive relationship.

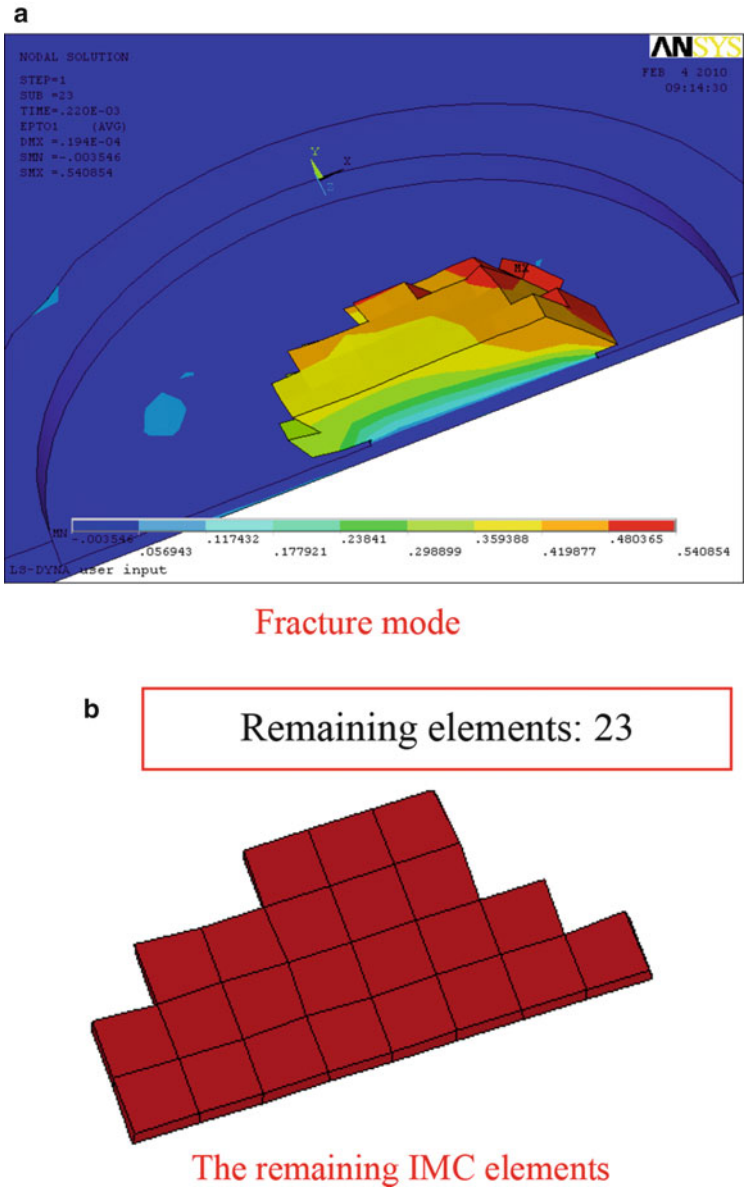


Fig. 10.13 Fracture surface (**a**) and remaining IMC (**b**) at 800 mm/s

2. Both simulation and experimental tests indicate that more brittle fracture take place in IMC layer when the impact speed increases.
3. The simulated and experimental load–displacement response curves reveal that the required fracture energy in high speed is less than that in low speed; however, the peak traction is evidently improved.

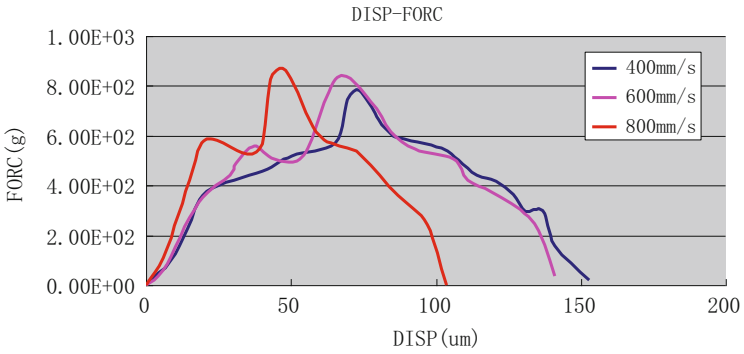


Fig. 10.14 Loading curves in simulation

In the future, the material parameters of IMC layer will be adjusted and validated further and some more suitable models or criteria will be applied to better describe the failure of material. Finally, a quantitative simulation and analysis of interconnection failure can be carried out.

10.3 Reliability of WLCSP Assembly Reflow Process and PCB Design

This section investigates the solder joint failures of WLCSP mounted on test PCBs. In particular, stress in assembly reflow process was studied [9]. The solder balls of WLCSP in the study have 5×5 ball array, which corresponds to 16 outmost solder joints and nine inner solder joints, all soldered to the matching copper pads on the test PCB. Three PCB designs were modeled to understand the impact of PCB through via arrangement on stresses in solder joints during assembly reflow process: design #1 has no PCB through vias at all; design #2 has plated through vias under nine inner PCB copper pads; and design #3 has plated through vias under all 25 PCB copper pads. The modeling results disclose that PCB design #2 with plated through vias under nine inner PCB copper pads induces the highest solder stress in all three models. Contrary to common sense of higher stress on corner solder joints due to coefficient of thermal expansion (CTE) mismatch of silicon and PCB, the maximum stresses of design #2 actually occur on the inner solder joints. The simulation results match well with experimental observations. For PCB design #1 and #3, highest solder stress is lower than stress in design #2. In addition, in both cases, the maximum stress locates on the corner solder joints. New PCB design guidelines have since been implemented based on the simulation. Due to the improvement of the design, premature solder joint failure has not been recorded.

10.3.1 Introduction

Wafer-level chip-scale package (WLCSP) with micro solder joints is widely adopted because it provides the smallest form factor in all semiconductor device packages. Small size and short signal/thermal path also come with the benefits of superior electrical/thermal performance. The trend of today's WLCSP technology is towards bigger and thinner body size at finer solder joint pitches to satisfy the fast growing market for lighter, faster mobile electronic products. This growing trend in WLCSP technology development demands better understanding of solder intermetallic compound (IMC) growth, influence of package architecture, and package/PCB interactions.

In terms of package/PCB interaction, it is well known that mismatch of thermal expansion of silicon and organic laminates induces deformations, such as warpage, in the assembly of semiconductor devices. Also well known is mismatch of CTE of WLCSP and PCB is one of the root causes of board level reliability failures.

Over the years, few researchers studied the impact of PCB design on the solder joint reliability. There is even less published work on the effect of layout of the PCB and placement of vias on the reliability in assembly reflow process. Yet, thorough understanding of stress in solder joints during assembly process is needed as WLCSP devices become bigger, thinner, and finer pitch.

In a typical WLCSP device qualification, coupon PCB is often used to mount WLCSP devices with passives for functional tests before and after environmental stresses, such as autoclave (ACLV), highly accelerated stress test (HAST), temperature cycle (TMCL), operational life (OPL), and high temperature storage life (HTSL). Due to cost and time considerations, PCB with plated through vias is still the primary choice for component testing and, at the same time, provides sufficient power/signal routing capability.

In the qualification of a 5×5, 0.4 mm pitch WLCSP, unexpected early failures were recorded in ACLV and HTSL. Electrical test and subsequent failure analysis revealed unusual failure mode on inner solder joints of the 5×5 array. Unlike typical thermal–mechanical fatigue failure that often first occurs at the corner solder joints, here, catastrophic solder joint failures all occur at the inner solder joints with no signs of failure initiation on the corners (Fig. 10.15).

Further analysis identified the root cause of the unique failure mode. Quickly realized was of the components just mounted on the test PCB (without any environment stress), cracks were already initiated on inner solder joints. Also noticed in the review was for this particular WLCSP, the test PCB was designed to have through PCB vias under only the nine center solder joints (where early failures and crack initiation were located) for routing at the specific PCB line/space requirements. Because of the uniqueness of this PCB layout and high-level correlation between solder joint failure and through PCB vias, investigations were then focused on the effect of PCB through via and stress in the assembly solder reflow process.

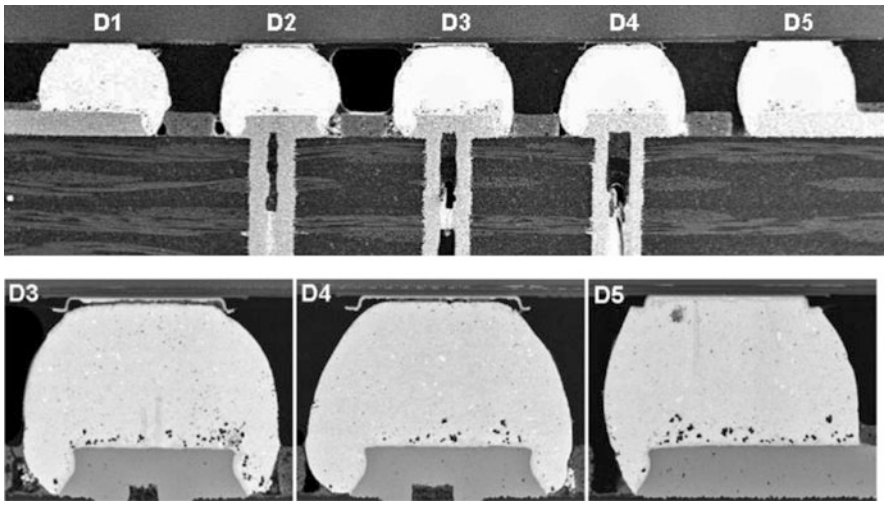


Fig. 10.15 Inner solder joint fails on a component that failed ACLV test

10.3.2 Three PCB Designs and Their FEA Models

Three PCB designs were considered with 5 × 5, 0.4 mm pitch WLCSP. In the first model, none of the 25 copper pads on top PCB surface connected to the bottom PCB surface's pads with through-board via. In the second model, the nine inner copper pads on top PCB surface connected to the corresponding pads on the bottom PCB surface with nine through-board vias. The 16 outmost PCB pads did not connect to any through-board via. In the third model, all of the 25 copper pads on top PCB surface connected to the bottom PCB surface's copper pads with 25 through-board vias. Figure 10.16 shows a finite element model for the WLCSP and the PCB board. The WLCSP die size is 2.1 mm × 2.42 mm, and thickness is ~0.38 mm. In Fig. 10.17, a cross-section view of a finite element model of WLCSP mounted on the PCB board is shown.

Figure 10.18 shows a zoomed in cross-section view of the WLCSP. The basic setting includes a 2.7 μm thick aluminum pad, a nitride passivation, and polyimide repassivation that overlaps the aluminum pad; adhesion metal layer and plating seed layer; and 2 μm thick nickel-based UBM with flash of gold on top (shown as one UBM layer). The sidewall angle of the polyimide repassivation is set at 60°. The UBM connects to the aluminum pad through the polyimide open and the solder is placed/reflowed on the UBM and connect to copper pad on the PCB.

Figure 10.19 shows a finite element model for the PCB board and a detailed view for its top trace. The actual PCB board dimension is 21.6 mm × 39 mm × 1.52 mm. The top layer has a copper trace 0.55 mm wide and the bottom layer has a copper trace 0.25 mm wide. In the model, only 6 mm × 6 mm top trace and bottom trace are included for simplification. Two buried copper metal layers were also considered in the model.

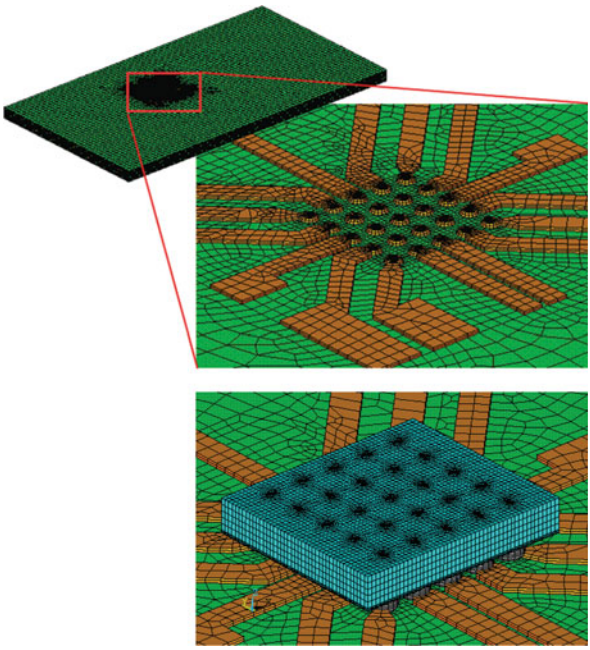


Fig. 10.16 Finite element mesh model of PCB with 5×5 WLCSP mounted on top

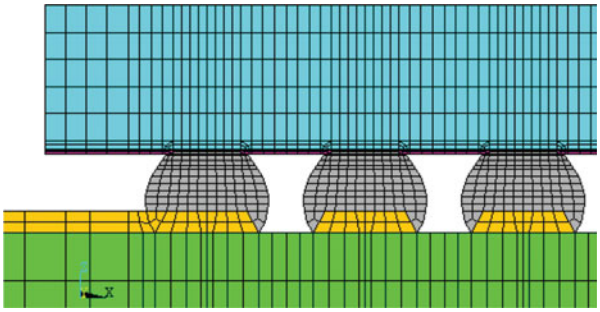


Fig. 10.17 Cross section of 5×5 WLCSP mounted on the PCB board

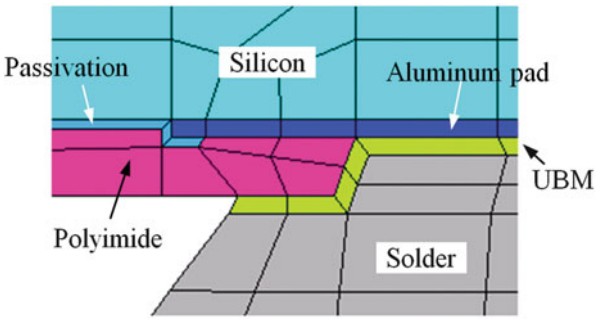


Fig. 10.18 Local (near UBM) cross-section view of WLCSP

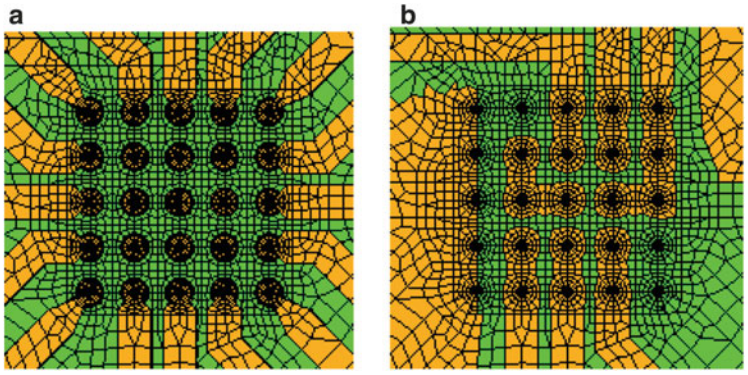


Fig. 10.19 Finite element model for top PCB board. (**a**) Top trace. (**b**) Top buried copper plane

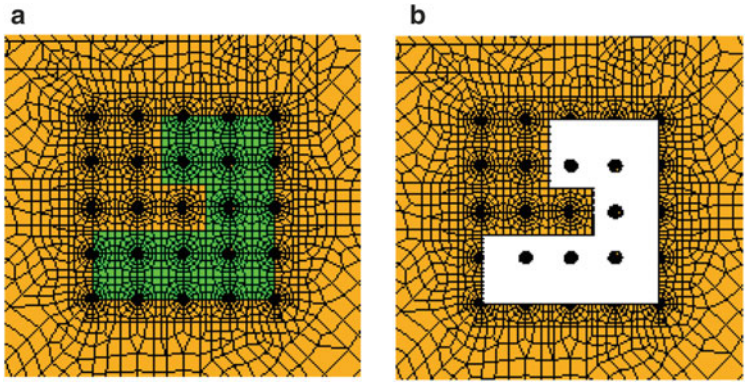


Fig. 10.20 Finite element model for bottom PCB board. (**a**) Bottom buried copper plane. (**b**) Bottom trace

Figure 10.20 shows a finite element model for the bottom copper layers and two buried copper layers. The yellow element shows the copper trace or copper pad. The green element shows the FR4.

Tables 10.6, 10.7, 10.8, and 10.9 list the material properties used in the FEA simulation. Table 10.6 defines the elastic modulus, Poisson's ratio, and CTE of each material. FR4 in PCB, silicon, nitride passivation, polyimide, and UBM are all considered as linear elastic materials. UBM is a 2 μm thick nickel-based metal with flash gold on top, a plating seed layer, and adhesion metal under. In the FEA model, the UBM is simplified as a single layer with its material properties made up of a combination of nickel, gold, plating seed metal, and adhesion metal.

The aluminum pad is considered a bilinear material. Table 10.7 shows its yield stress and tangent modulus at different temperatures. The copper-based PCB pad and via are also considered as bilinear material. Table 10.8 shows their yield stress and tangent modulus.

Table 10.6 Elastic modulus, Poisson ratio, and CTE of the simulation materials

Materials	Modulus (GPa)	Poisson ratio	CTE ($\times 10^{-6}$ ppm/°C)
Silicon	131	0.278	2.4
Solder joint	See table 4	0.4	21.9
Passivation	314	0.33	4
Polyimide	3.5	0.35	35
FR4	Ex = 25.42	Nuxy = 0.11	Alpx = 14
	Ey = 25.42	Nuxz = 0.39	Alpy = 16
	Ez = 11	Nuyz = 0.39	Alpz = 45 (<180 °C)
	Gxz = 4.97		
	Gyz = 4.97		Alpz = 220 (>180 °C)
	Gxy = 11.45		
Copper	117	0.33	16.12
Al Pad	68.9	0.33	20
UBM	124.5	0.299	15

Table 10.7 Yield stress and tangent modulus of Al PAD

Temperature (°C)	25	125
Yield stress (MPa)	200	164.7
Tangent Modulus (MPa)	300	150

Table 10.8 Yield stress and tangent modulus of copper trace, pad and vias

Yield stress (MPa)	70
Tangent Modulus (MPa)	700

Table 10.9 Elastic modulus of solder at different temperatures

Temperature (°C)	35	70	100	140
Modulus (GPa)	26.38	25.8	25.01	24.15

Table 10.9 gives temperature-dependent elastic modulus of solder materials. Table 10.10 gives a rate-dependent Anand model of the solder joint.

Figure 10.21a–c below shows cross sections of FEA models of WLCSP mounted on PCB boards with three different through-via configurations. Figure 10.21a is the first model with no through PCB vias under all 25 solder joints. Figure 10.21b is the second model with nine through PCB vias under the copper pads of center 3×3 solder joints array. Figure 10.21c is the third model, where all 25 copper pads on the top PCB surface are connected to the bottom PCB copper pads with plated through PCB vias.

Figure 10.22 shows the temperature loading applied to the model. The reference temperature was set at the peak reflow temperature, 246 °C. It is then decreased to room temperature. After dwelling at room temperature, it rises again to another high temperature point of 210 °C.

Table 10.10 Experimentally determined and fitted Anand model constants for solder alloy [8]

Description	Symbol	Constant
Initial value of s	s_o	1.3 MPa
Activation energy	Q/R	9,000 K
Pre-exponential factor	A	500/s
Stress multiplier	ζ	7.1
Strain rate sensitivity of stress	m	0.3
Hardening coefficient	h_o	5,900 MPa
Coefficient for deformation resistance saturation value	$\hat{s}$	39.4 MPa
Strain rate sensitivity of saturation value	n	0.03
Strain rate sensitivity of hardening coefficient	a	1.4

10.3.3 Simulation Results

Solder joint's Z component stress comparison for the three models at high temperature (210 °C) is shown in Fig. 10.23a–c. First, it is evident that the maximum Z component stress on a solder joint is at the interface of solder and the UBM. It can also be seen that Model 2, with nine center through PCB vias connecting the top and the bottom PCB copper pads, has the highest Z component stress on solder joint. Model 1, with no through PCB vias, has the lowest Z component stress on solder joint. Model 3, with 25 through vias under every solder joint, resulted in a Z component stress that is between Model 1 and Model 2. It is also evident that in Model 2 the maximum Z component stress occurs at the corner of the inner 3×3 solder joint array, while for the Model 1 and Model 3, the max Z component stress occurs at corner joints of the 5×5 solder joint array.

Figure 10.24a shows a cross-section position for the modeled WLCSP. Figure 10.24b shows a cross-section view of solder joint Z component stress of Model 1, which has no through PCB vias under component solder joints. The stress at the outmost solder is much higher than stress at the inner joints. Also, the stress close to the solder and UBM interface is higher than that at the solder and PCB interface. Figure 10.24c shows a cross-section view of solder joint Z component stress for Model 2, which has nine through PCB vias under the center 3×3 solder array. In this case, stresses on the inner solder joints at both solder/UBM interface and solder/PCB pad interface are higher than on outer solder joints. Figure 10.24d shows a cross-section view of solder joint Z component stress for Model 3, which has through PCB vias under all 25 component solder joints. Similar to Model 1, the maximum stress is again on the outmost joints. However, the stress at the solder/UBM interface solder/PCB pad interface is higher than Model 1.

Figure 10.25a–c shows the first principle stress S1 at the same cross-section location as of Fig. 10.24a. Figure 10.25a shows that for Model 1, the center joints are under slight compressive stress and the outmost joints are under partial tensile and partial compressive stress. The maximum first principle stress of solder joint is

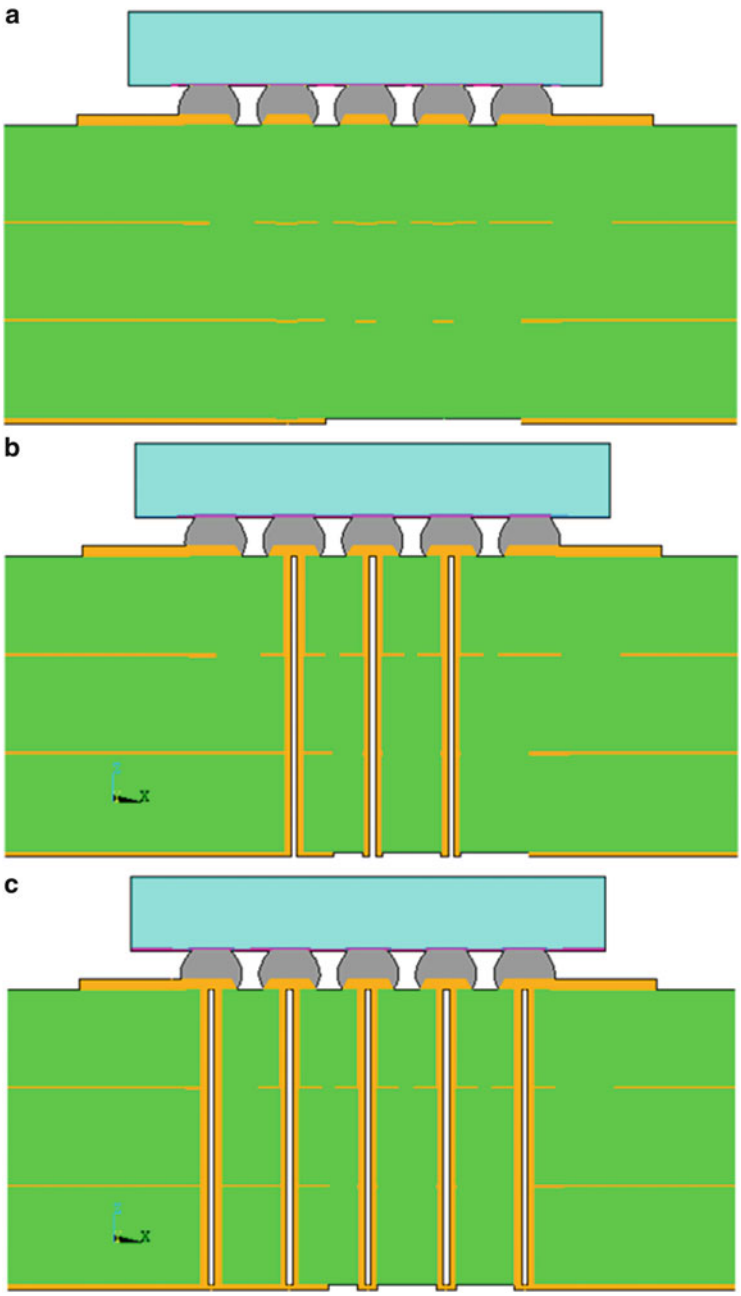


Fig. 10.21 Cross-section view of three models through PCB via configurations. (**a**) Model 1: no through board via in PCB. (**b**) Model 2: nine through board vias in PCB. (**c**) Model 3: 25 through board vias in PCB

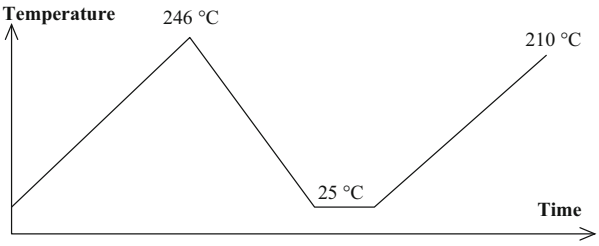


Fig. 10.22 Modeled temperature loading

18.7 MPa at the outmost joints. Figure 10.25b shows that for Model 2, the inner nine joints are under tensile stress and the outmost 16 joints are under compressive stress. The maximum first principle stress of solder joint S1 is 22 MPa, which is the highest in the three models. Figure 10.25c shows that for Model 3, all of the solder joints are under partial tensile and partial compressive stress. All of the solder joints are under tensile stress at solder/PCB pad interface. The stress at 16 outmost joints is larger than that at nine inner joints. The max stress is 20.1 MPa.

10.3.4 Discussion and Improvement Plan

The simulation results clearly indicated high stresses can be introduced by the placement of the copper plated through vias. In the case where through vias are only under some of the solder joints, excessive stress should be expected on solder joints with copper plated through vias under.

An intuitive explanation of simulation results can also be made: FR4 in PCB is anisotropic in expansion. While in XY plane, CTE is match to that of copper (~17 ppm/°C) with glass cloth, Z direction CTE is typically in the range of 40–60 ppm/°C at low temperature and much higher (220 ppm/°C) above the glass transition temperature of the FR4 resin materials. When there are copper plated through vias under specific solder joints, PCB expansion surrounding the vias is limited by the copper inside the via. If neighboring solder joints don't have plated through vias under, free Z direction expansion at high temperature under the neighboring solder joints will push the WLCSP up, exerting high tensile on solder joints with plated through vias under. This is exactly what the simulation tells us (see Fig. 10.24b) and the root cause of unexpected early failures of 5 × 5 WLCSP in HTSL.

On the other hand, moisture-induced swelling is a more applicable explanation to early ACLV failures. Similar to thermal expansion, anisotropic hygroscopic expansion due to the presence of glass cloth in FR4 makes similar stress contributions to the solder joints with copper plated through vias under and causes early and surprisingly identical failure mode as that in HTSL.

To avoid the early failures seen in 25 ball WLCSP qualification, design changes in test PCB have to be made. If routing is not an issue, PCB design without through vias is recommended. If more than one layer is needed for signal/power/ground connection, blind vias should be in favor of through vias. If through vias are

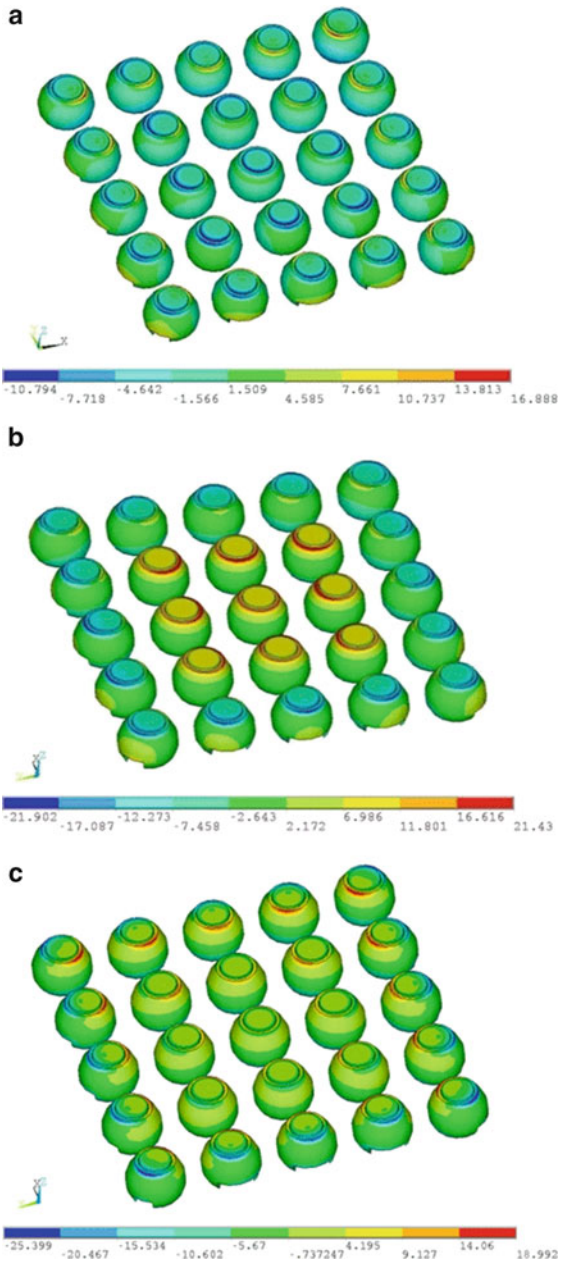


Fig. 10.23 Solder joint's Z component stress comparison. (**a**) Model 1: no through PCB vias. Maximum Sz at the corner of 5×5 array: 16.9 MPa. (**b**) Model 2: nine through PCB vias. Maximum Sz at the corner of inner 3×3 array: 21.4 MPa. (**c**) Model 3: 25 through PCB vias. Maximum Sz at the corner of 5×5 array: 19 MPa

designed due to other considerations, it is recommended to place through vias under every solder balls. For the case of 25 ball WLCSP, simulation results showed that full array through vias reduces the stress and deformation by more than 50 % compared to the partial through vias design.

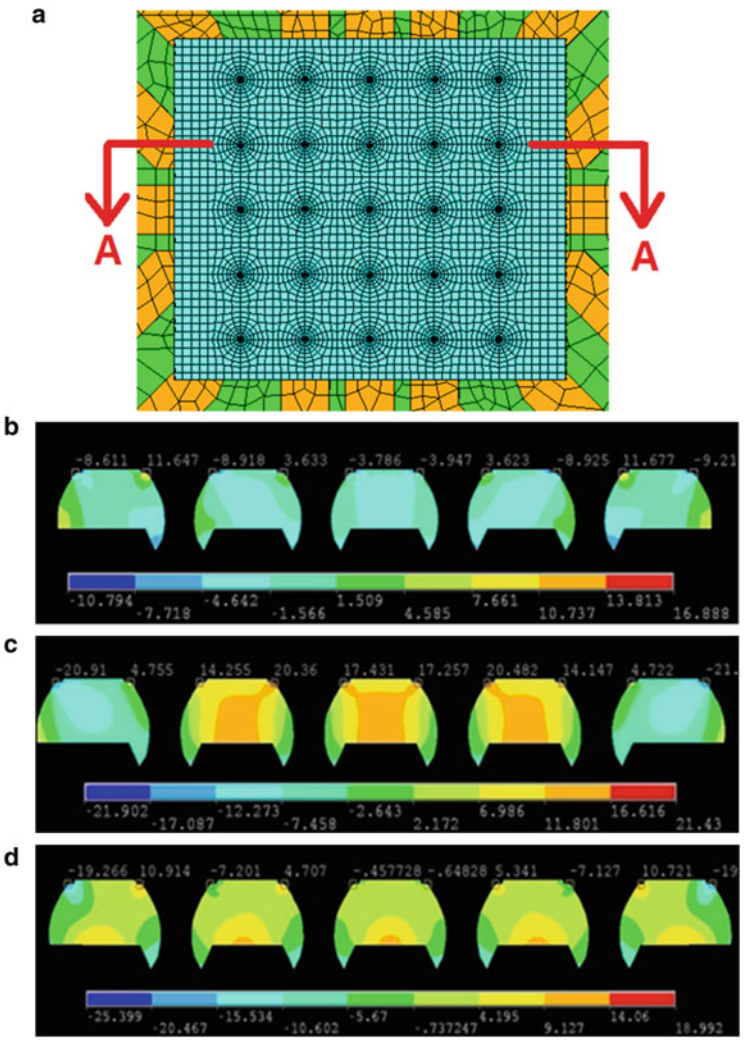


Fig. 10.24 Solder joint Z component stress. (**a**): solder joint's cross-section position. (**b**): A–A cross-section view of Sz stress for Model 1 (no via). Maximum Sz: 16.8 MPa at corner joints. (**c**): A–A cross-section view of Sz stress for Model 2 (9 vias). Maximum Sz: 21.4 MPa on inner joints. (**d**): A–A cross-section view of Sz stress for Model 3 (25 vias). Maximum Sz: 18.9 MPa at corner joint

This simple design principle has been adopted and implemented in the following qualification PCB design cycles. One latest test PCB design for 9×9 WLCSP chip placed through vias under all 81 solder joints and the component test has successfully passed predefined reliability test (Fig. 10.26).

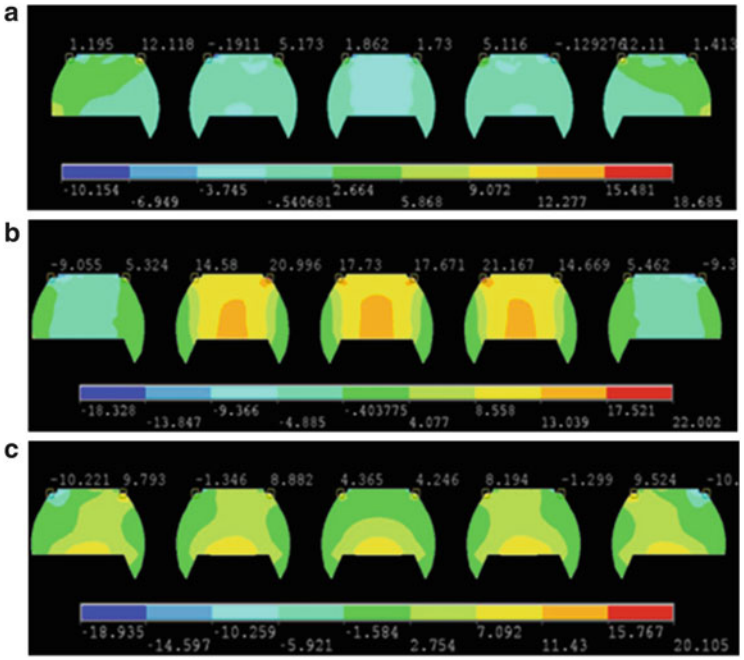


Fig. 10.25 Cross-section view of the first principle stress in solder joints (**a**): A–A cross-section view of S1 stress for Model 1 (no via). Maximum S1: 18.7 MPa. (**b**): A–A cross-section view of S1 stress for Model 2 (9 vias). Maximum S1: 22 MPa. (**c**): A–A cross-section view of S1 stress for Model 3 (25 vias). Maximum S1: 20.1 MPa

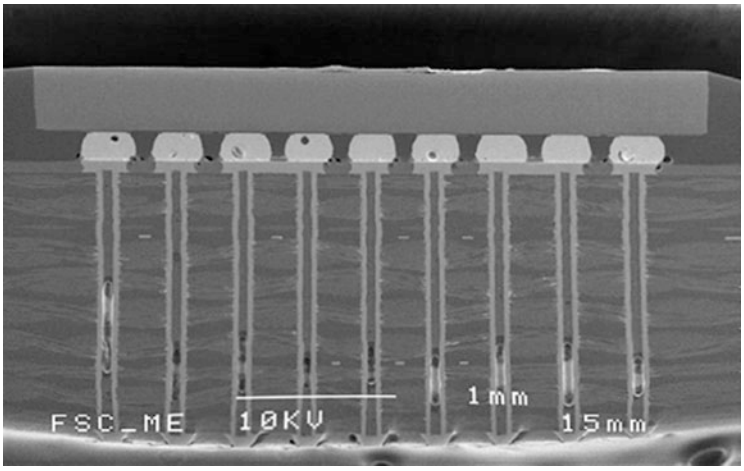


Fig. 10.26 Cross section of a 9×9 WLCSP device soldered to the test PCB. Copper plated through vias are placed under every solder joints to avoid uneven stresses

10.4 WLCSP Board Level Drop Test

Board level drop test is very critical for the reliability of WLCSP. The dynamic responses of the WLCSP parameter design with different UBM geometry, different polyimide sidewall angle and thickness, different metal stack thickness, and different solder joint height are studied in this section; The drop test under JEDEC standard is conducted. The results from both the drop test and modeling show that the corner joints of each corner located WLCSP fail first as compared to the chips at other locations. The test results agree with the simulation for the failure modes and the locations.

10.4.1 Introduction

The trends of next-generation WLCSP are towards thinner and finer pitch with micro bumps. The mechanical shock resulted from mishandling during transportation or customer usage may cause WLCSP package solder joint failure. Since the board level drop test is a key qualification test for portable electronic products, it is becoming a topic of great interest for many researchers. This section will address the board level drop test of the WLCSP.

10.4.2 WLCSP Drop Test and Model Set Up

The drop test setup is based on the JEDEC Standard JESD22-B111. The board, with a dimension of 132 mm × 77 mm × 1.0 mm, accommodates 15 components of the same type in a three-row by five-column format.

Due to the symmetry, a quarter finite element model (66 mm × 38.5 mm × 1 mm) of a JEDEC board with WLCSP chips is selected [10]. Figure 10.27a shows the test system and a finite element model of the lower left quarter part of the test board with six components U1, U2, U3, U6, U7, and U8, which are numbered according to the JEDEC standard.

Figure 10.27b shows the finite element model of the cross section of the corner joint for a WLCSP structure. The basic setting includes a 2.7 μm thick aluminum pad, 2 μm thick UBM with 0.5 μm Au and 0.2 μm Cu, and a 0.9 μm thick passivation that covers 5 μm edge of the aluminum pad. A 10 μm thick polyimide layer is above the passivation and the aluminum pad. There is a 200 μm diameter via open in the polyimide layer; its sidewall angle (between its slope and the bottom surface) is 60°. The UBM connects to the aluminum pad through the via and the solder is placed on the UBM and connects to copper post on the PCB board.

Table 10.11 defines the elastic modulus, Poisson's ratio, and density of each material. The silicon, passivation, polyimide, PCB, and UBM are considered as linear elastic material, while the solder ball, aluminum pad, and PCB copper pad are considered as the nonlinear material properties. Table 10.12 gives the nonlinear property for solder SAC405 that is considered as a rate-dependent Peirce model; see

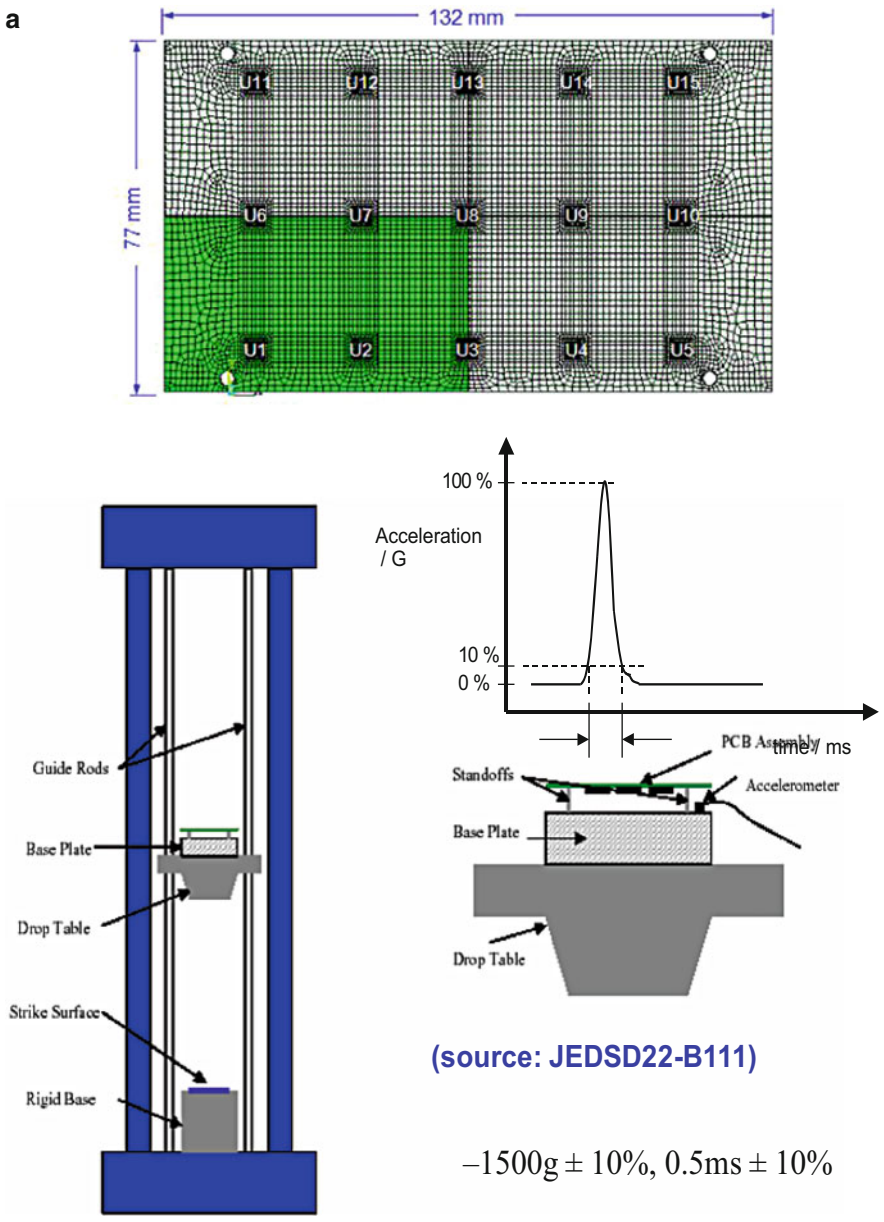


Fig. 10.27 Finite element model of the WLCSP. (**a**) Finite element model of PCB with quarter chip units (U1, U2, U3, U6, U7, U8) and drop test setup (**b**) The cross section and design variable

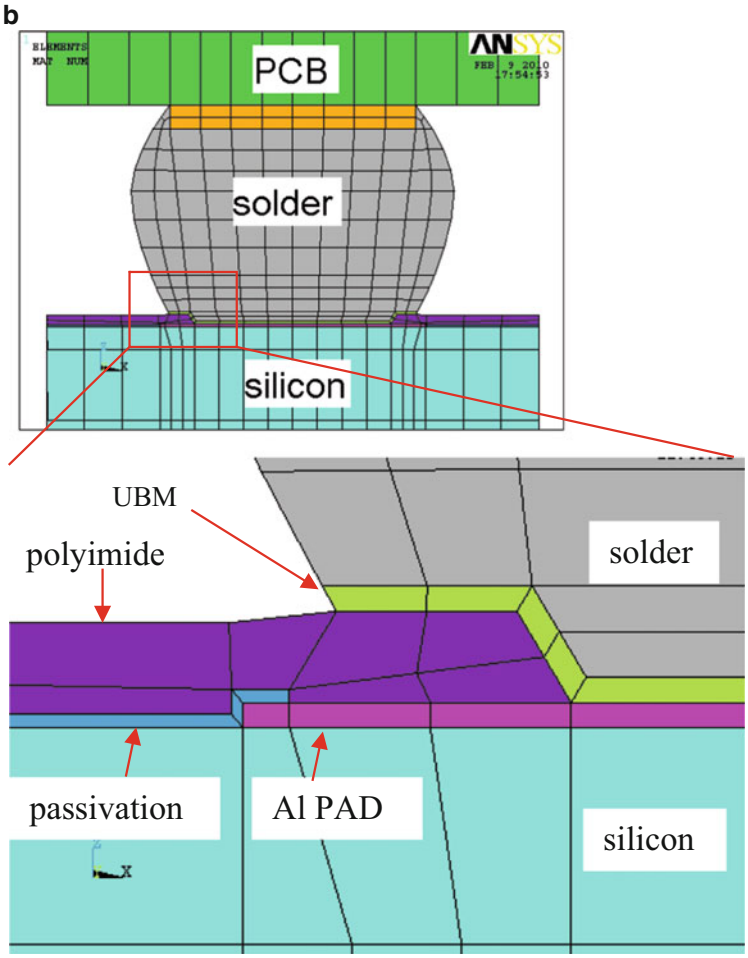


Fig. 10.27 (continued)

Eq. (10.8). The data was obtained through the Hopkinson dynamic material high-speed impact test.

$$\sigma = \left[1 + \frac{\dot{\varepsilon}^{pl}}{\gamma}\right]^{m} \sigma_0 \tag{10.8}$$

where σ is the dynamic material yield stress, $\dot{\varepsilon}^{pl}$ is the dynamic plastic strain rate, $\sigma_0$is the static yield stress, m is the strain rate hardened material, and γ is the material viscosity parameter.

The direct acceleration input (DAI) method is applied in this study [11]. In this method, an acceleration impulse is applied as an inertia that is specified with the linear acceleration of the structure at each time step. The surfaces of mounting holes

Table 10.11 Elastic modulus, Poisson ratio, and density of each material

	Modulus (Gpa)	Poisson ratio	Density (g/cm^3)
Silicon	131	0.278	2.33
Solder	26.38	0.4	7.5
Passivation	314	0.33	2.99
Polyimide	3.5	0.35	1.47
PCB	Ex = Ey = 25.42 Ez = 11 Gxz = Gyz = 4.91 Gxy = 11.45	Nuxy = 0.11 Nuxz = Nuyz = 0.39	1.92
Cu PAD	117	0.33	8.94
Al PAD	68.9	0.33	2.7
UBM	196	0.304	9.7

Table 10.12 Rate-dependent Peirce model of solder SAC405

	Static yield stress (Mpa)	γ	m
Solder (SAC405)	41.85	0.00011	0.0953

are constrained during dynamic impact. Therefore, the problem formulation becomes

$$\{M\}[\ddot{u}] + \{C\}[\dot{u}] + \{K\}[u] = \begin{Bmatrix} -\{M\}1,500g \sin\dfrac{\pi t}{t_w} & for \quad t \le t_w, \quad t_w = 0.5 \\ 0 \quad for \quad t \ge t_w \end{Bmatrix} \tag{10.9}$$

with initial conditions

$$[u]|_{t=0} = 0$$
$$[\dot{u}]|_{t=0} = \sqrt{2gh} \tag{10.10}$$

where h is the drop height, and the constraint boundary condition is

$$[u]\,|_{at\_holes} = 0 \tag{10.11}$$

10.4.3 Drop Impact Simulation/Test with Different Design Variable and Discussion

10.4.3.1 Impact of Design Variable Polyimide Sidewall Angle

The polyimide layer connects both the aluminum pad and UBM; see Fig. 10.27b.

Figure 10.28 shows the max peeling stress comparison of the solder, copper pad, and aluminum pad with different polyimide sidewall angles at location U1. The stresses in the copper pad, aluminum pad, and solder interface connected to the

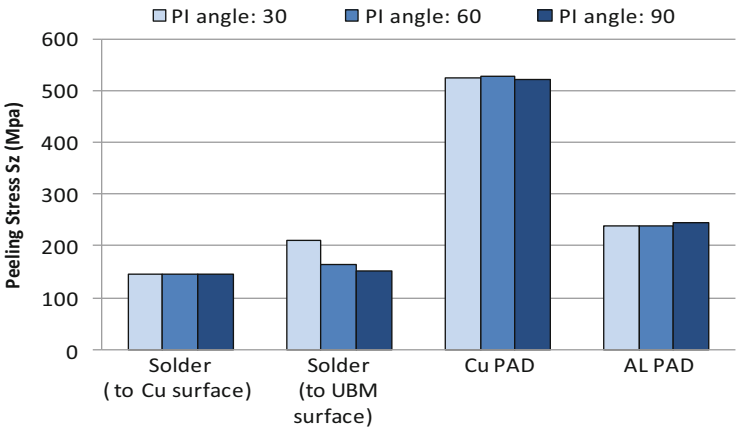


Fig. 10.28 Comparison of max peeling stress at U1 with different polyimide sidewall angles

copper pad show that there is no significant difference with different polyimide sidewall angles. However, there is impact on the solder joint interface that adhered to UBM.

10.4.3.2 Impact of Design Variable Polyimide Thickness

The polyimide thickness is selected to be 5 μm, 10 μm, and 15 μm, respectively. Figure 10.29 shows the peeling stress for the solder, copper pad, and aluminum pad with different polyimide thickness at location U1. The peeling stress on the aluminum pad increases as polyimide thickness increases from 5 to 15 μm. The solder stress at the interface with UBM decreases when the polyimide thickness increases from 5 to 10 μm; however after 10um, there is no significant difference.

10.4.3.3 Impact of Design Variable UBM Structure

A copper UBM structure is designed to compare with the existing 2 μm nickel standard UBM with 0.5 μm Au and 0.2 μm Cu. The thickness of copper UBM is 8 μm.

Figure 10.30 shows the peeling stress comparison for the solder joint, copper pad, and aluminum pad with copper UBM and with standard UBM of the package at location U1. From Fig. 10.30, it can be seen that the peeling stress on the solder joint interface attached to the standard UBM is greater than that attached to the copper UBM. However, the peeling stress on Al pad with copper UBM is greater than standard UBM.

10.4.3.4 Impact of Design Variable Aluminum Pad Thickness

Different aluminum pad thicknesses with 0.8 μm, 2 μm, 2.7 μm, and 4 μm are simulated.

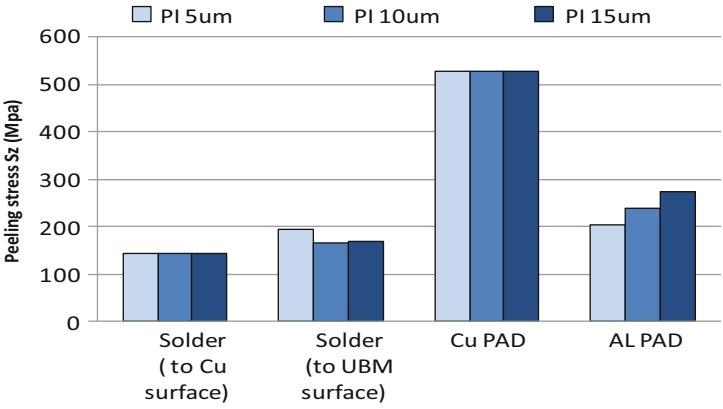


Fig. 10.29 Comparison of max peeling stress at U1 of different polyimide thicknesses

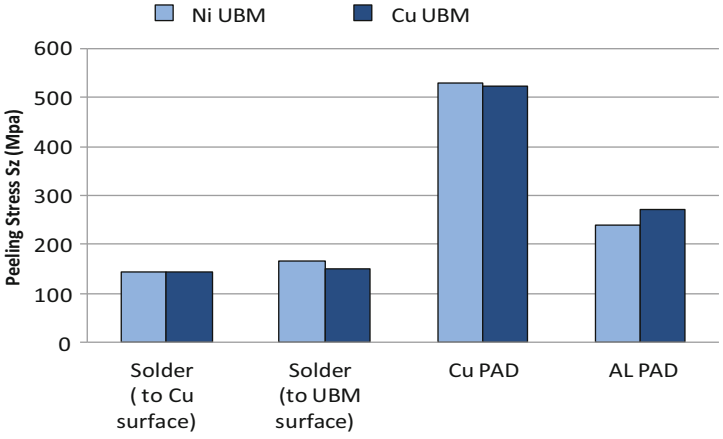


Fig. 10.30 Comparison of max peeling stress at U1 with two different UBM designs

Figure 10.31 gives the polyimide (PI) stresses versus the Al pad thickness and its correlation with the drop test life (black dot line). There is an optimized Al thickness from both modeling and drop test result.

10.4.3.5 Impact of Design Variable Solder Joint Height

Different solder joint heights with 50 μm, 100 μm, 200 μm, and 300 μm are considered. Figure 10.32 gives the trends of solder joint plastic energy density with different heights. In all three package locations, the higher solder joint, the less plastic energy density. This indicates that the higher solder joint height can help to improve the WLCSP drop test dynamic plastic energy performance.

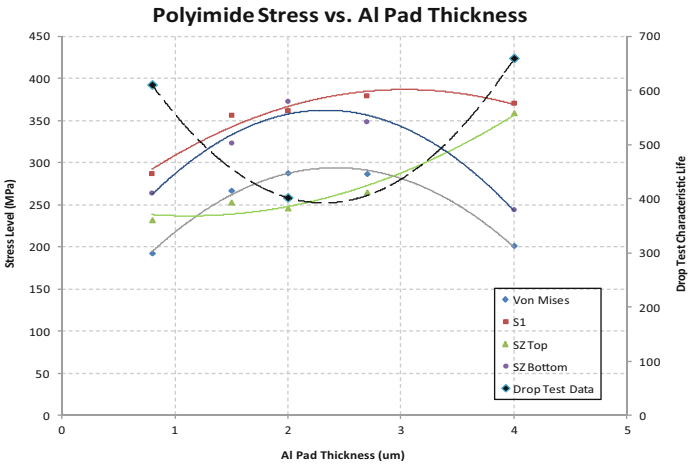


Fig. 10.31 Polyimide stress vs Al pad thickness

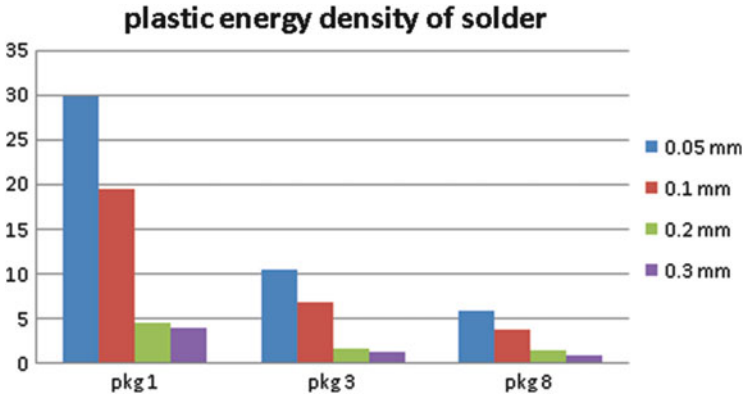


Fig. 10.32 Plastic energy density of solder joint (MPa)

10.4.4 Drop Test

The drop test was done based on JEDEC standard JESD22-B111. The test condition is 1,500 g with half sine wave in 0.5 ms. Drop count is 1,000. A total of 90 units were investigated and mounted on eight JEDEC PCBs for two groups with and without the vias under the copper pads. The drop test results are shown in Figs. 10.33 and 10.34. From Fig. 10.33 it can be seen that most of the drop failure appeared at the corner location U5, U11, and U15. Figure 10.34 gives the copper pad/crack which locates at the interface of solder and copper to the PCB. The test results have correlated the simulation result that the max first principal strain appears at the location of U1 which has the same behaviors of U5, U11, and U15

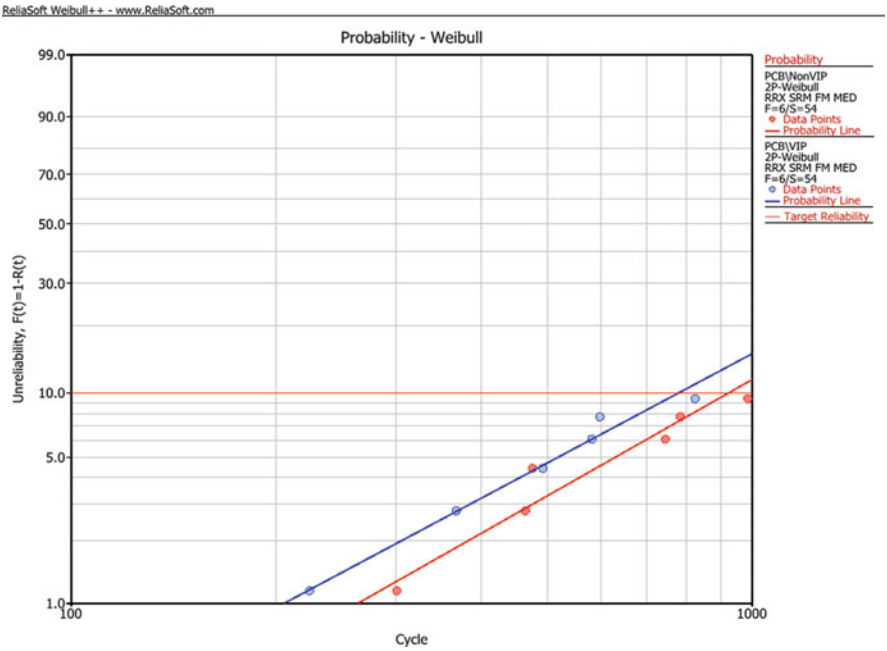


Fig. 10.33 Drop test result

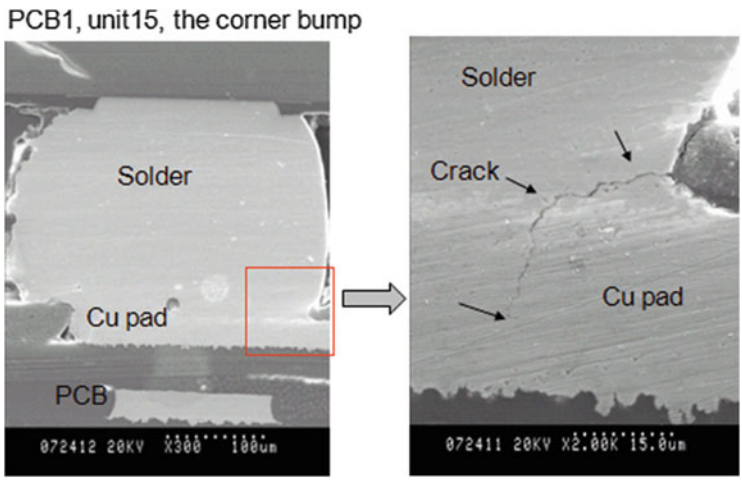


Fig. 10.34 Drop test failure mode

due to the model symmetric property. Figure 10.35 gives the first principal strain curves at the interface of the copper pad, solder, and PCB with the different package location U1, U3, and U8. The failure rank is U1 > U3 > U8. When the dynamic first

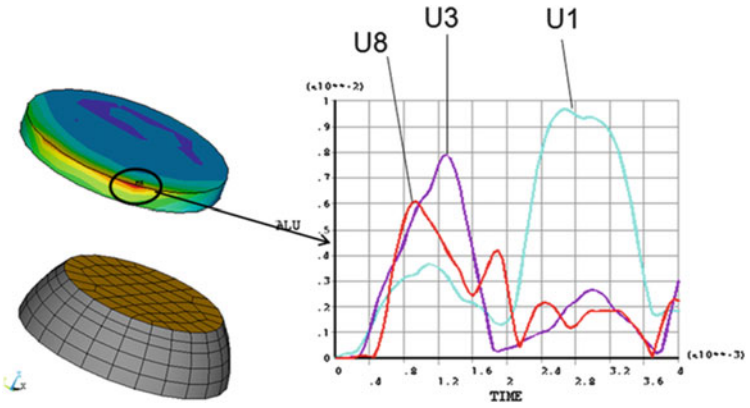


Fig. 10.35 First principal strains of copper pad at the interface of solder, copper pad, and PCB with package location U1, U3, and U8

principal strain reaches the failure strain, the copper pad/trace will break/crack, like in Fig. 10.34.

10.4.5 Discussion

In this section, the WLCSP design variable modeling and test study are carried out to investigate the dynamic behaviors of WLCSP subjected to drop impact with different design parameters of polyimide sidewall angle, thickness, UBM geometry, aluminum metal stack thickness, and solder joint height. Both the drop test and modeling results showed that the corner joints of each WLCSP chip (U1, U5, U11, U15) near the PCB screw holes fail first as compared to the chips at other locations. Next the corner solder joints of the chip location U3 and U13 fail, followed by the U8 chip (center) and the rest chips. The drop test results correlated the simulation for the failure modes and the PI stress trends with different Al pad thickness. In addition, the design variable solder joint height can significantly improve the drop test plastic energy performance of WLCSP.

10.5 WLCSP Design for Reliability

Comprehensive study of WLCSP design for reliability is presented in this session. The finite element simulation is carried out to improve the performance of designs of WLCSP. Major work will include (1) a design with one-layer redistribution layout (RDL) copper with etched pocket in the non-covered UBM area and one-layer polyimide structure (1Cu1Pi design) with different polyimide layouts, copper thicknesses, pocket parameters, and non-covered UBM diameters; (2) a stacked metal design with the sputtered copper UBM stacked on the RDL copper

layer, with one polyimide layer between them (2Cu1Pi) for the WLCSP. Parameter study of different UBM diameters with the same solder volume and different UBM diameters with the same solder joint height is conducted.

10.5.1 Introduction

Solder joint reliability of WLCSP is a topic of great interest for many researchers. Both of thermal cycling test and board level drop test are key qualification tests for portable electronic products. Most of the current works on WLCSP are focused on solder joint material, solder joint geometry, and solder joint array. Few papers studied the design for reliability, especially the critical design variable of the metal stack under the bump, such as under bump metallurgy (UBM) design and polyimide (PI). Actually, the UBM structure is the direct interface between the solder bump and the final chip metallization. It is a critical design component of a solder interconnect in a WLCSP. A lot of solder joint failures happened at the interface between solder bump and UBM layer. The polyimide layer acts not only as the solder mask, but also acts as a buffer layer to improve the reliability of solder joint. In this session, the critical designs of WLCSP include copper metal RDL, UBM, and polyimide layout. First, a design with 1Cu1Pi of WLCSP is studied. The copper UBM is covered with polyimide layer around its edge. Solder joint is then bumped to the non-covered area of the copper UBM. The copper UBM is etched with a pocket shape at the interface with solder bump. The polyimide layouts, copper UBM thickness, and pocket parameters are examined through modeling. Three models with different polyimide layouts are studied and discussed in both thermal cycling test and drop test. In the model one, polyimide contacts with solder bump and the solder bump are deformed. In the model two, polyimide just starts to contact with solder bump and the solder bump is slightly deformed. In the model three, polyimide is far away from the solder bump. There is no contact between polyimide and the solder bump. The concept of etched pocket on the UBM is studied to check the reliability performance. The models with non-etched UBM, slightly etched UBM, and deeply etched UBM are discussed. The last parameter for 1Cu1Pi WLCSP is the diameter of non-covered UBM. Two models with different UBM diameters are studied. Secondly, sputtered copper-based UBM (2Cu1Pi design) WLCSP is studied. Parameter study of different UBM diameters with the same solder volume and different UBM diameters with the same solder joint height is conducted for the drop test and thermal cycling performance. The correlation and comparison of the failure mechanism between the modeling and the test are discussed finally [12–14].

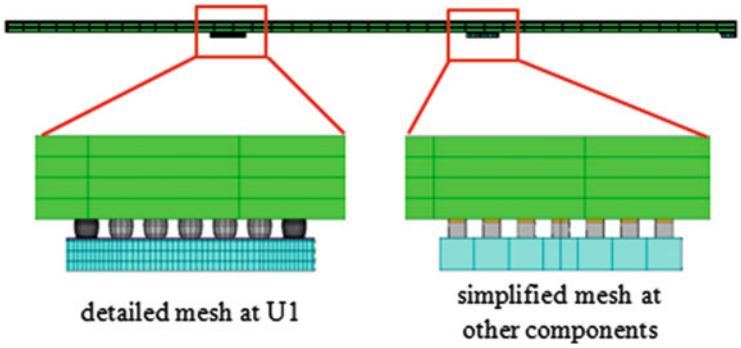


Fig. 10.36 Quarter PCB with detailed mesh at U1 and simplified at other components (U2, U3, U6, U7, and U8; see Fig. 10.27)

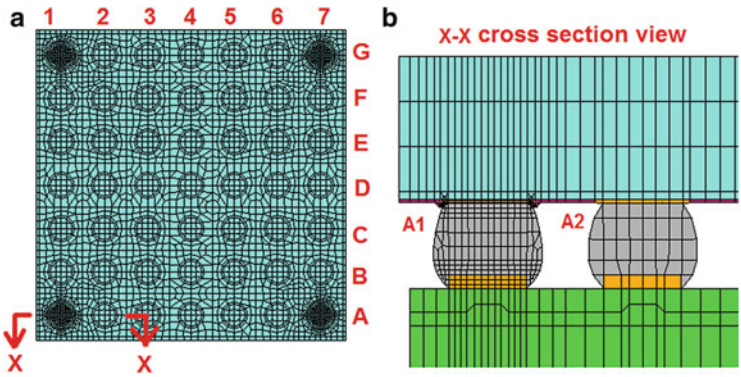


Fig. 10.37 Cross section of WLCSP. (**a**) WLCSP 7 × 7 top view. (**b**) A–A cross-section view

10.5.2 Finite Element Model Set Up

10.5.2.1 Drop Test Model

The drop test is based on the JEDEC standard: JEDEC22-B111. The board, with a dimension of 132 mm × 77 mm × 1.0 mm, accommodates 15 components of same type in a three-row by five-column format (see Fig. 10.27 in Sect. 10.4)

Figure 10.36 shows a finite element model of the detailed mesh at component U1 and simplified mesh at other components. All solder balls in component U1 are modeled with detailed structure, while, in the other simplified components, all of the solder balls are simplified as rectangular block elements.

Figure 10.37a shows a top view of the wafer-level chip package (WLCSP) with 0.4 mm solder ball pitch and 2.8 mm × 2.8 mm package size. Figure 10.37b shows cross-section view of two solder balls. Since the failure locates at the corner solder bumps, the corner bump A1 (same as A7, G1, and G7) is meshed with fine elements; the other solder bumps are meshed with coarse elements for fast simulation. The material data are listed in Tables 10.11 and 10.12.

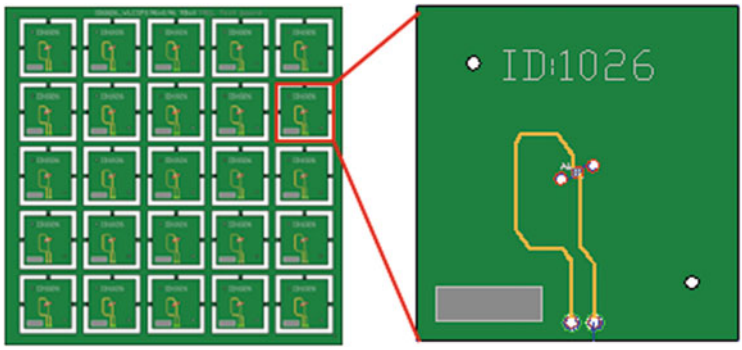


Fig. 10.38 TMCL board

10.5.2.2 Thermal Cycling Model

The thermal cycle test is completed according to the JEDEC specifications. The temperature range is −40°C–125°C. The thermal load is considered to be the uniform temperature which is applied to all elements of the model.

Figure 10.38 shows the thermal cycling test board. It is composed of 25 small Units. WLCSP package is attached on each unit. Each unit connects with the whole test board with three tie bars. A finite element model with a quarter unit is selected for the simulation due to symmetry. Figure 10.39 shows the finite element mesh for such quarter model. The PCB board includes two-layer copper trace at surfaces and six buried copper planes with 40 % copper coverage or 70 % copper coverage. The corner ball A1 (same as A7, G1, and G7) is meshed with fine elements; the other solder balls are meshed with coarse elements for simplicity.

Tables 10.13, 10.14, and 10.15 show the material properties for WLCSP and PCB board. The data of Young's modulus versus temperature are listed in Table 10.14. Table 10.13 defines the coefficient of thermal expansion (CTE) of each material. Table 10.15 gives the nonlinear property for solder ball.

10.5.3 The Results of Drop Test and Thermal Cycling Simulations

10.5.3.1 Different Polyimide (PI) Layouts, UBM Thickness, and Etched UBM Structure for 1Cu1Pi WLCSP

Three different polyimide layouts are studied in both thermal cycling modeling and drop test modeling. Figure 10.40 shows the actual SEM pictures for the different polyimide layouts. The dash line is the free edge of the final solder joint if the polyimide does not contact the solder after the reflow. Figure 10.40a shows the polyimide contacts with solder ball, which induces the deformed solder outline in reflow. Figure 10.40b shows the polyimide just starts to contact with the solder ball. The solder ball is slightly deformed. Figure 10.40c shows the polyimide does not contact with the solder. The outline of the final solder ball has no deformation.

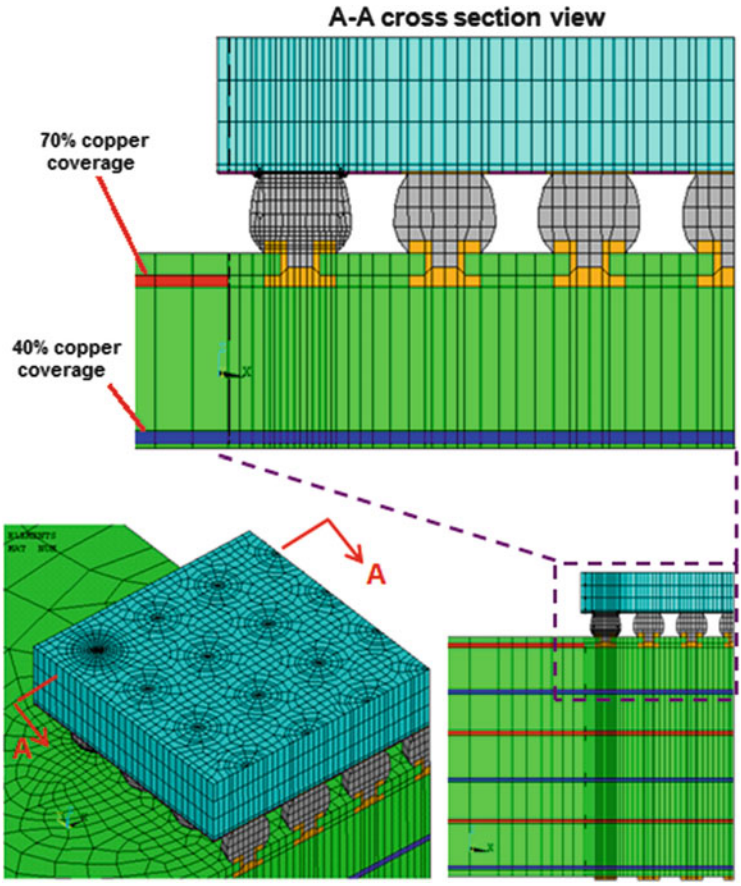


Fig. 10.39 The finite element model for TMCL

Table 10.13 Material CTE

	CTE (×1e-6)
Silicon	2.4
Solder joint	21.9
Passivation	4
Polyimide	35
FR4	Alpx = 16 Alpy = 16 Alpz = 60
Copper	16.12

Table 10.14 Young's modulus of solder at different temperature

Temperature (C°)	35	70	100	140
Modulus (Gpa)	26.38	25.8	25.01	24.15

Table 10.15 Anand model constants for solder alloy

Description	Symbol	Constant
Initial value of s	s_o	1.3 MPa
Activation energy	Q/R	9,000 K
Pre-exponential factor	A	500/s
Stress multiplier	ζ	7.1
Strain rate sensitivity of stress	m	0.3
Hardening coefficient	h_o	5,900 MPa
Coefficient for deformation resistance saturation value	$\hat{s}$	39.4 MPa
Strain rate sensitivity of saturation value	n	0.03
Strain rate sensitivity of hardening coefficient	a	1.4

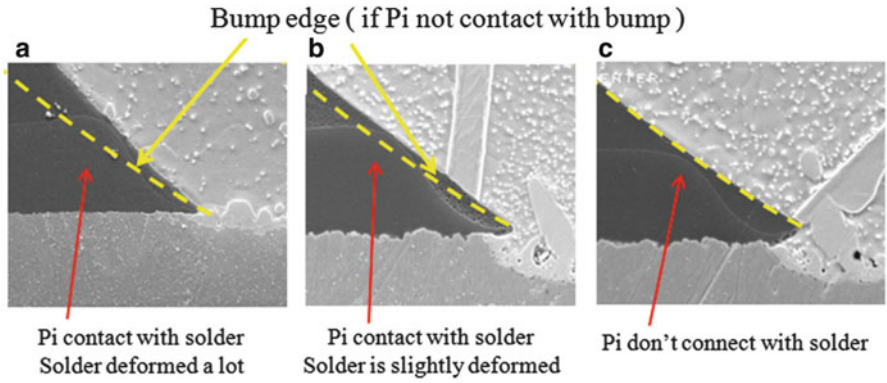


Fig. 10.40 SEM pictures for three different polyimide layouts: (**a**) PI layout 1, (**b**) PI layout 2, and (**c**) PI layout 3

Figure 10.41 shows the FEA models of the three polyimide layouts as shown in Fig. 10.40. A contact pair between solder and polyimide is set up in all three polyimide layouts.

The designs with different copper UBM thicknesses and etched pocket are studied. Figure 10.42 shows two models with 10 μm thick copper UBM and 7.5 μm thick copper UBM. Figure 10.43 shows the model of the UBM with an etched 1 μm pocket.

12 FEA design of experiment (DoE) models with different polyimide layouts, different UBM thicknesses, and different pocket designs in UBM are simulated. Table 10.16 shows the 12 models in 4 groups. Each group has 3 PI layouts with same UBM thickness and same UBM pocket design, but with different polyimide layouts.

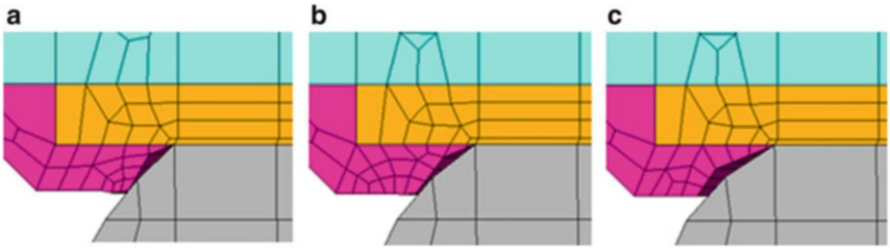


Fig. 10.41 Finite element model of the three different polyimide layouts: (**a**) PI layout 1, (**b**) PI layout 2, and (**c**) PI layout 3

Fig. 10.42 Finite element models of two different copper UBM thicknesses

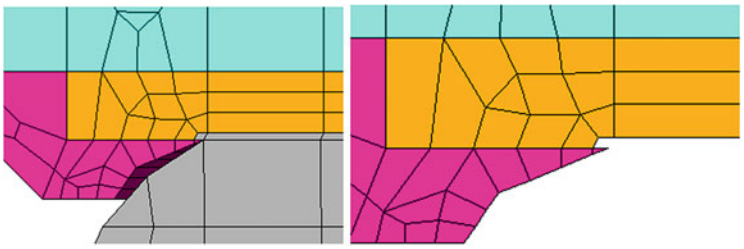


Fig. 10.43 Finite element model of copper UBM with 1 μm thick pocket

Table 10.16 Four group models with different polyimide layout, UBM thickness, and UBM pocket design

Group	Polyimide layouts	UBM thickness	UBM pocket
Group 1	3 polyimide layouts	10 μm	1 μm
Group 2	3 different polyimide layouts	10 μm	No
Group 3	3 different polyimide layouts	7.5 μm	1 μm
Group 4	3 different polyimide layouts	7.5 μm	No

Figure 10.44 shows cross-section contour view of the first principle stress S1 distribution in solder bump in drop test simulation. The UBM thickness is 10 μm with polyimide layout without contacting the solder ball. Figure 10.44a shows the

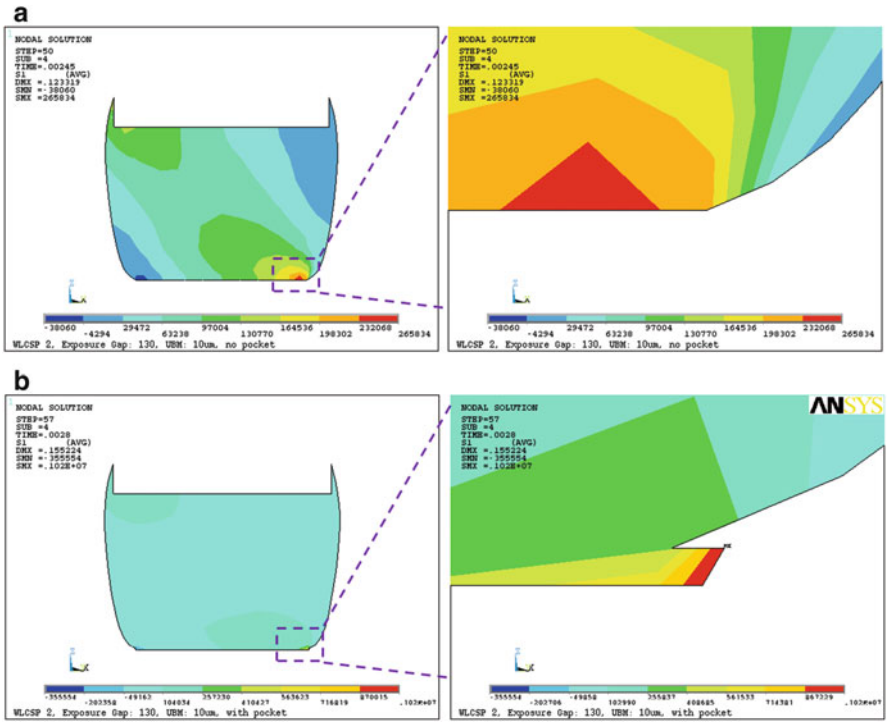


Fig. 10.44 First principal stress S1 of with pocket versus without pocket designs in drop test. (**a**) Max solder S1: 265.8 MPa (no pocket in UBM). (**b**) Max solder S1: 1,020 MPa (1 μm pocket in UBM)

stress contour of non-pocket model. The max first principle stress of solder locates at the junction of solder, polyimide, and UBM. Figure 10.44b shows solder stress contour of the pocket model. The pocket tip is filled with solder which has resulted in a sharp stress as compared to the model without pocket. Failure will start at the solder pocket tip.

Figure 10.45, 10.46, 10.47, and 10.48 show first principle stress S1, von Mises stress, Z component peeling stress Sz, and max shear stress S_{XZ} comparison of solder at the interface of UBM in drop test simulation. The dynamic stresses of the design with etched pocket are greater than the design without the pocket. In drop impact, the first principal stress and peeling stress will dominate the failure of solder joint. Therefore, for the design without the etched pocket, the polyimide layout 1 in which the polyimide contacts the solder ball with thicker UBM is the best solution. While for the design with etched pocket, the polyimide layout 2 in which the polyimide just starts to contact the solder ball is the better solution.

Figure 10.49 shows first failure cycle comparison of solder ball at the interface of solder and UBM in thermal cycling simulation. Thicker UBM with the pocket design in UBM has increased the first failure cycle. The larger air gap between

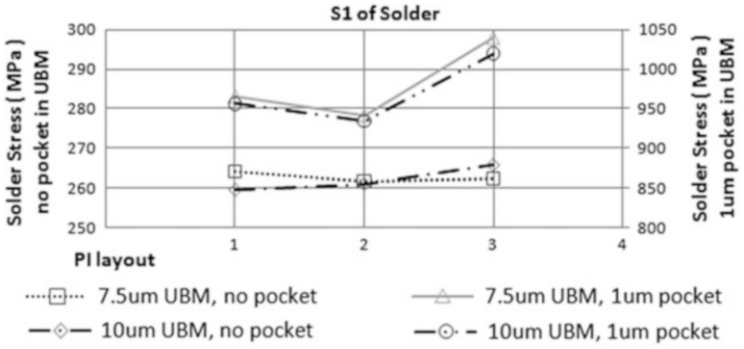


Fig. 10.45 First principle stress comparison with different design parameters in drop test

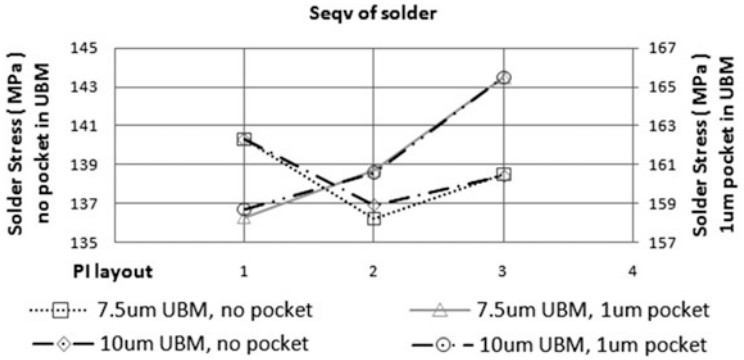


Fig. 10.46 Von Mises stress comparison with different design parameters in drop test

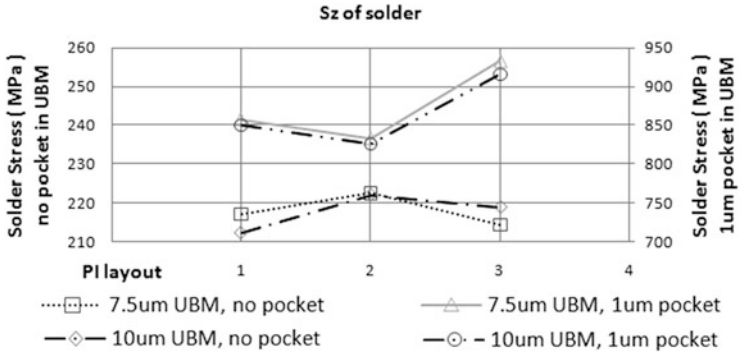


Fig. 10.47 Z component peeling stress comparison with different design parameters in drop test

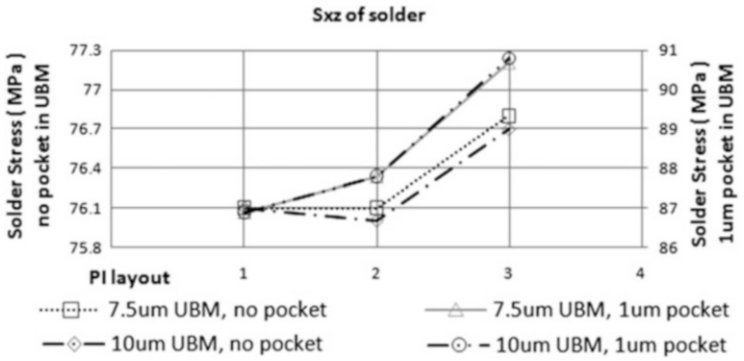


Fig. 10.48 XZ shear stress comparison with different design parameters in drop test

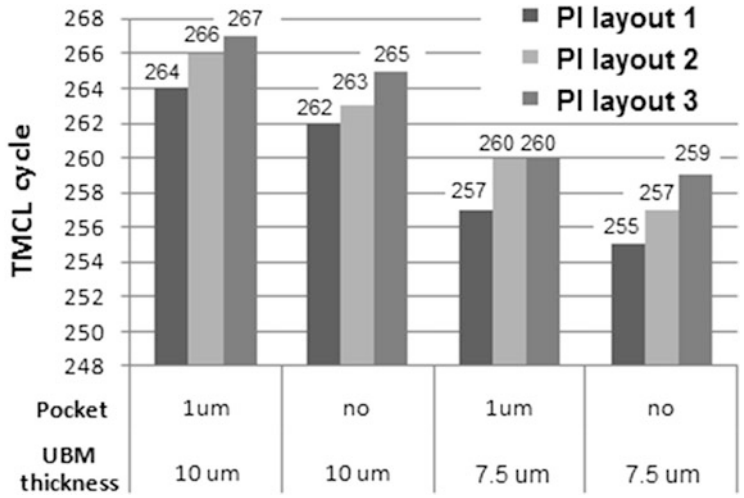


Fig. 10.49 First failure cycle comparison of solder at the interface to UBM (different polyimide layouts, UBM thicknesses, and UBM pocket designs)

polyimide and solder ball has increased the first failure cycle number as well. Among the three PI layouts, the PI layout 3 seems to have the longest first failure life. The thicker UBM with etched pocket has showed the longest thermal cycling life. Figure 10.50 shows the character life cycle comparison of solder ball at the interface of solder and UBM in thermal cycling. The trends of character cycle are similar to the trends as in the first failure cycle shown in Fig. 10.49.

Figure 10.51 shows the finite element models with different etching depths in UBM. Three design layouts with 0 μm etched (non-etched) UBM, 1 μm deep etched UBM, and 4 μm depth etched UBM are considered in thermal cycling loading. Figure 10.52 shows first failure cycle and character life cycle comparison of solder ball at the interface to UBM. It shows that deeper pocket in UBM slightly increased

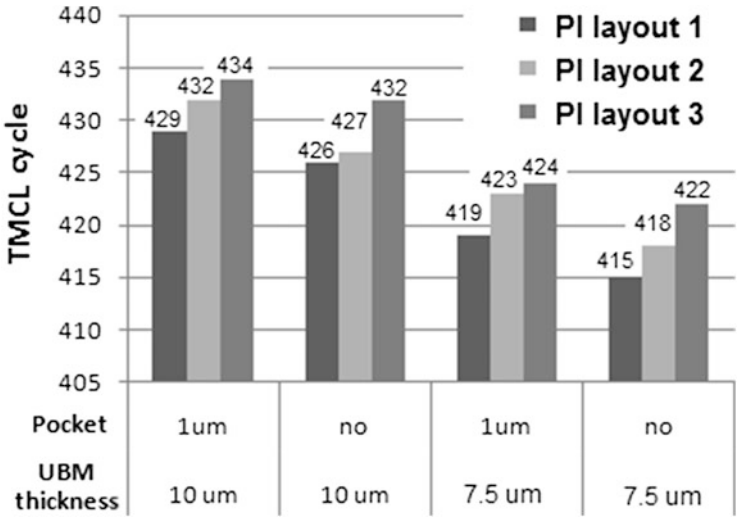


Fig. 10.50 Character life cycle comparison of solder at the interface to UBM (different polyimide layouts, UBM thicknesses, and UBM pockets)

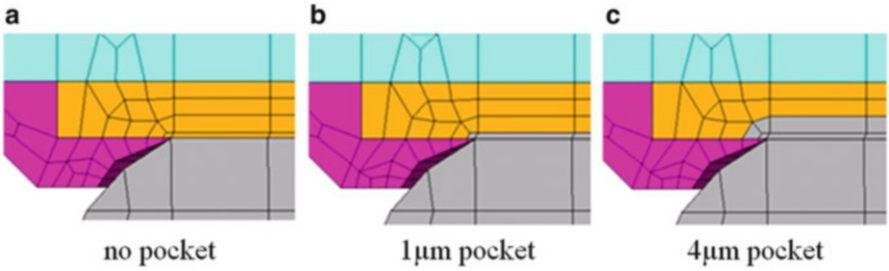


Fig. 10.51 The finite element model cross section of the WLCSP with different pocket depths in UBM: (**a**) no pocket, (**b**) 1 μm pocket, and (**c**) 4 μm pocket

the first failure cycle and character cycle in thermal cycling. However, the impact is not significant.

10.5.3.2 Different Non-covered UBM Diameter for 1Cu1Pi WLCSP

Above modeling results show the impact of the polyimide layout on the solder joint in thermal cycling and drop test modeling. A WLCSP with new simple polyimide layout is studied. Figure 10.53a shows the SEM picture of the WLCSP with a new polyimide layout. Figure 10.53b shows the finite element model of the simple polyimide layout in WLCSP. The diameter of non-covered copper UBM is 205 μm. The solder ball diameter is 280 μm. The solder ball height is 185 μm. Figure 10.53c shows a model with larger diameter of non-covered copper UBM. The solder volume is same as the previous models. As the diameter of non-covered

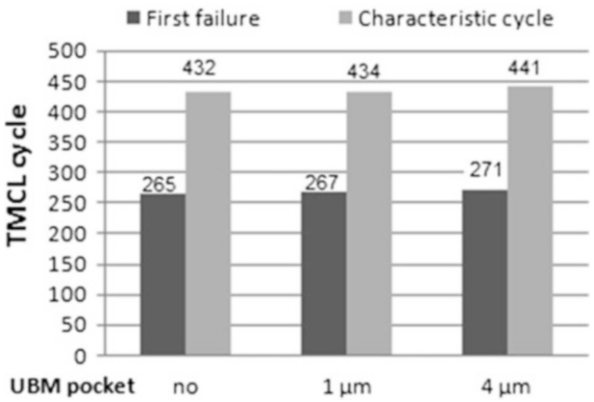


Fig. 10.52 Thermal cycling life cycle comparison of solder at the interface to UBM with different pocket depths

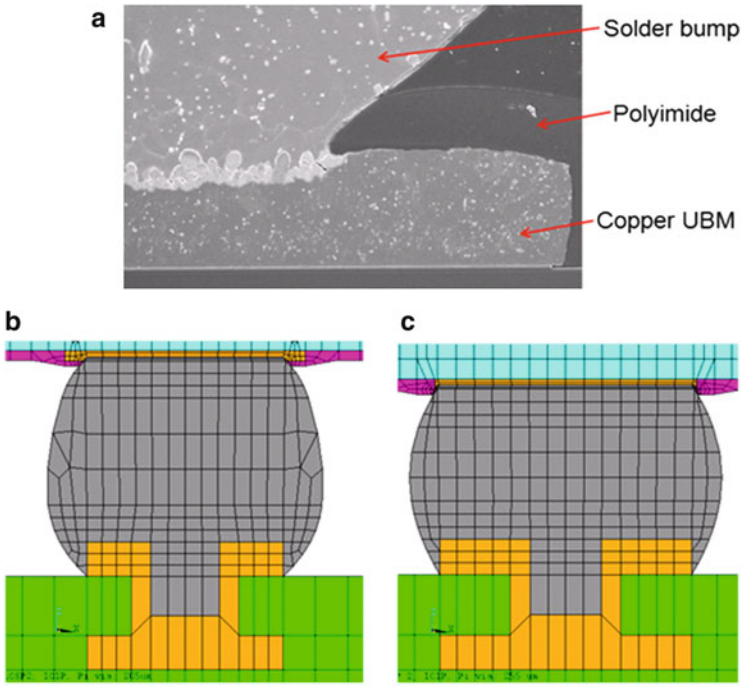


Fig. 10.53 Finite element model of 1Cu1Pi WLCSP with different non-covered UBM diameters. (**a**) New simple polyimide shaper WLCSP, (**b**) Ø uncovered UBM = 205 µm, and (**c**) Ø uncovered UBM = 255 µm

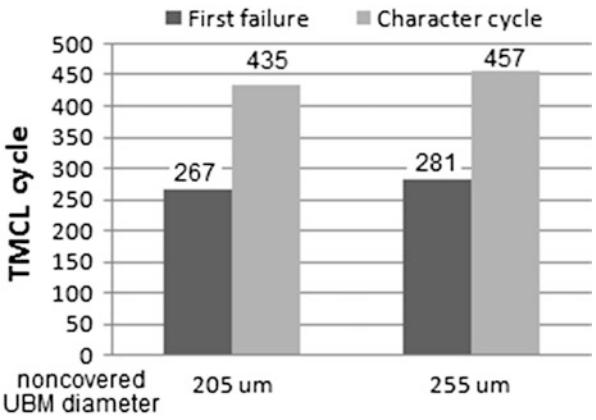


Fig. 10.54 Thermal cycling life cycle comparison of solder at the interface of UBM (different non-covered copper UBM diameters)

copper UBM increases from 205 μm to 255 μm, its solder ball diameter becomes 310 μm; its solder ball height is 150 μm. Both of the two models shown in Fig. 10.53b, c have etched 4 μm deep pocket in copper UBM.

Figure 10.54 shows thermal cycling first failure cycle and character life cycle comparison of solder joint at the interface of UBM. As the non-covered UBM diameter increases from 205 to 255 μm, the first failure cycle increases from 267 cycles to 281 cycles, and the character life cycle increases from 435 cycles to 457 cycles although its solder joint height decreases from 185 to 150 μm.

10.5.3.3 Different UBM Diameter for 2Cu1Pi WLCSP with the Same Solder Ball Volume

Figure 10.55 shows the SEM picture for standard sputtered copper UBM (2Cu1Pi design) WLCSP. A 10 μm thick polyimide coating is above the passivation and the 5.5 μm thick copper RDL. There is a via open in the polyimide layer. The polyimide sidewall angle (between its slope and the bottom surface) is 43°. A 7.5 μm thick copper UBM is sputtered on the non-covered RDL and part of the polyimide coating.

Figure 10.56 shows the finite element models with different UBM diameters of the 2Cu1Pi WLCSP. In all of these three models, the solder volume keeps same. Figure 10.56a shows the finite element model of 2Cu1Pi WLCSP with 230 μm UBM diameter. Its polyimide via diameter is 195 μm. The solder ball height is 156 μm. Figure 10.56b shows the finite element model of 2Cu1Pi WLCSP with 245 μm UBM diameter. Its polyimide via diameter is 205 μm. The solder ball height is 150 μm. Figure 10.56c shows the finite element model of 2Cu1Pi WLCSP with 255 μm UBM diameter. Its polyimide via diameter is 212 μm. The solder ball height is 146 μm.

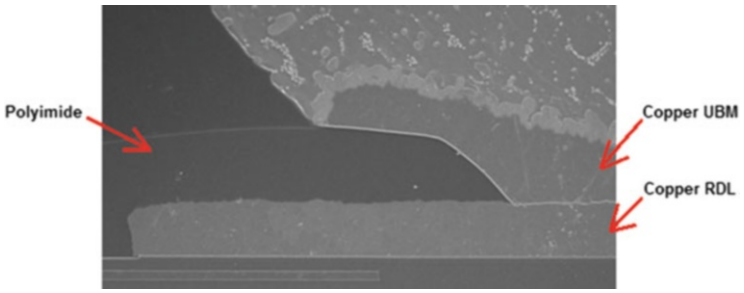


Fig. 10.55 SEM picture of the 2Cu1Pi WLCSP

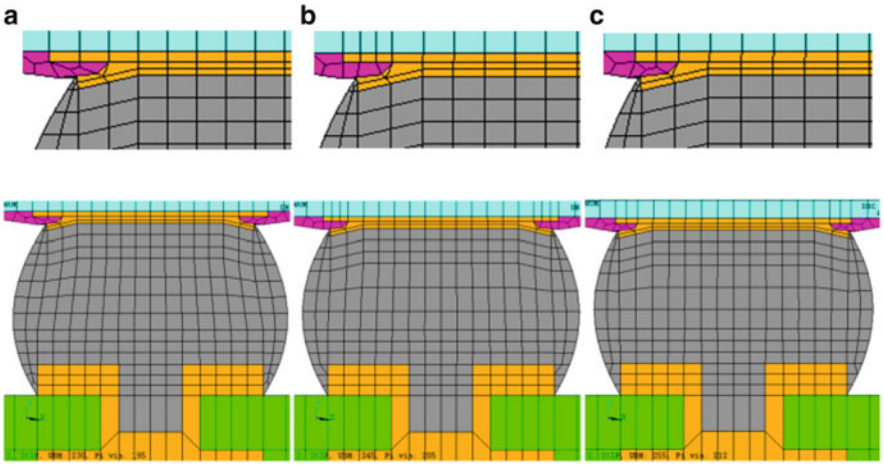


Fig. 10.56 Finite element model 2Cu1Pi WLCSP with different UBM diameters: (**a**) ØUBM = 230 μm, (**b**) ØUBM = 245 μm, and (**c**) ØUBM = 255 μm

Figure 10.57 gives the drop test modeling result. It shows that the first principle stress, von Mises stress, Z component peeling stress, and XZ direction shear stress comparison of solder at interface of UBM decrease as UBM diameter increases.

Figure 10.58 shows thermal cycling first failure cycle and character life cycle comparison of solder joint at the interface to UBM. For the model with 230 μm diameter UBM, the first failure cycle and character cycle of solder are 270 cycles and 439 cycles. As the UBM diameter increases 6.5 % to 245 μm, the solder joint first failure cycle increases 18.5 % to 320 cycles. The character life cycle increases 18.5 % to 520 cycles. As the UBM diameter increases 10.9 % to 255 μm, solder joint first failure cycle increases 31.5 % to 355 cycles. The character life cycle increases 31.7 % to 578 cycles.

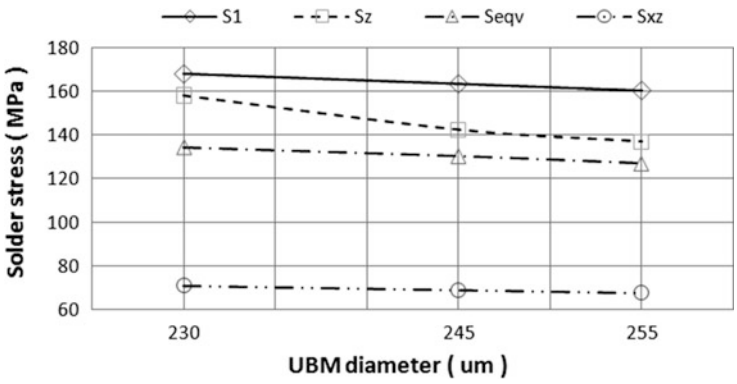


Fig. 10.57 Max drop test-induced stress comparison of solder at the interface of UBM

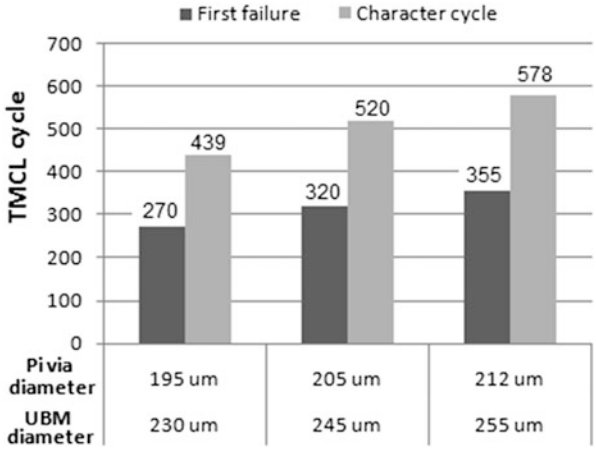


Fig. 10.58 Thermal cycling life cycle comparison of solder at the interface of UBM (different UBM diameters)

10.5.3.4 Different UBM Diameter with Same Solder Ball Height

Figure 10.59 shows different non-covered UBM diameter (Pi via diameter) of 1Cu1Pi WLCSP and different UBM diameter of 2Cu1Pi WLCSP. The solder ball heights of these five models are the same. Solder volume increases as Pi via diameter or UBM diameter increase.

Figure 10.60 gives thermal cycling modeling results. It shows the first failure life and character life of solder near the UBM interface increase as Pi via diameter or UBM diameter increases. As the diameter of UBM increases, the solder joint first failure cycle life and character life increase, and the longest life is the 2Cu1Pi design WLCSP with UBM diameter 255 μm, which has reached the character life over 1,000 cycles.

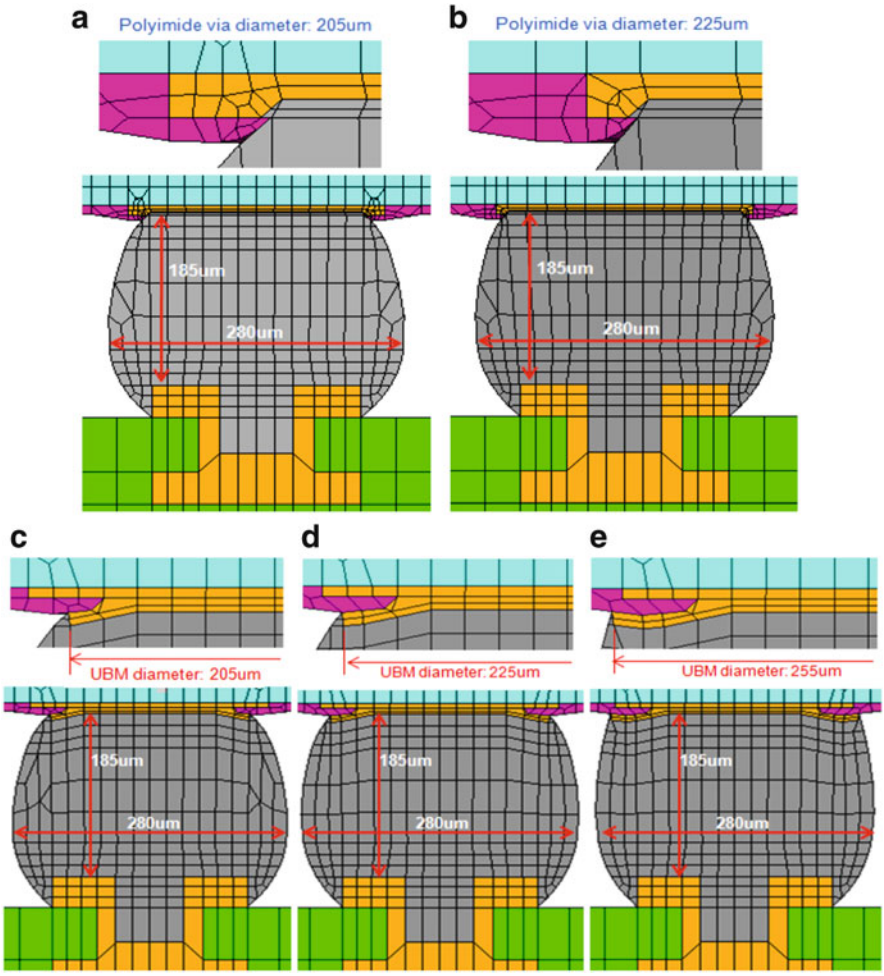


Fig. 10.59 Finite element model of 1Cu1Pi WLCSP with different non-covered UBM diameter (Pi via diameter) and 2Cu1Pi WLCSP with different UBM diameters: (**a**) Ø Pi via = 205 μm, (**b**) Ø Pi via = 225 μm, (**c**) ØUBM = 230 μm, (**d**) ØUBM = 245 μm, and (**e**) ØUBM = 255 μm

10.5.4 Drop Test and Thermal Cycling Test

The drop test was done based on JEDEC standard JESD22-B111. The test condition is 1,500 g with half sine wave in 0.5 ms. Figure 10.61 shows the drop test SEM pictures with the solder crack and modeling stress cross-section view of the 1Cu1Pi design of WLCSP. The stress cross section shows that the max solder stress locates at the junction of solder, polyimide, and UBM. Drop test results show that solder crack originates at the same position as modeling. Figure 10.62 shows the drop test FA picture and the modeling stress cross-section view of the 2Cu1Pi design of

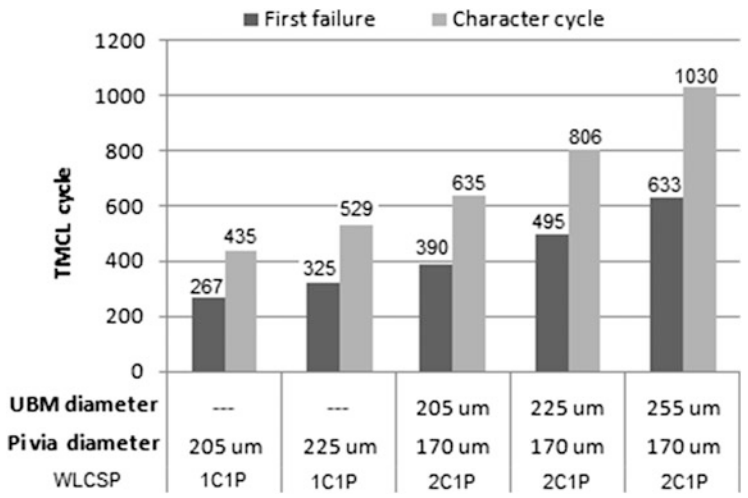


Fig. 10.60 Thermal cycling life comparison of solder near the UBM interface

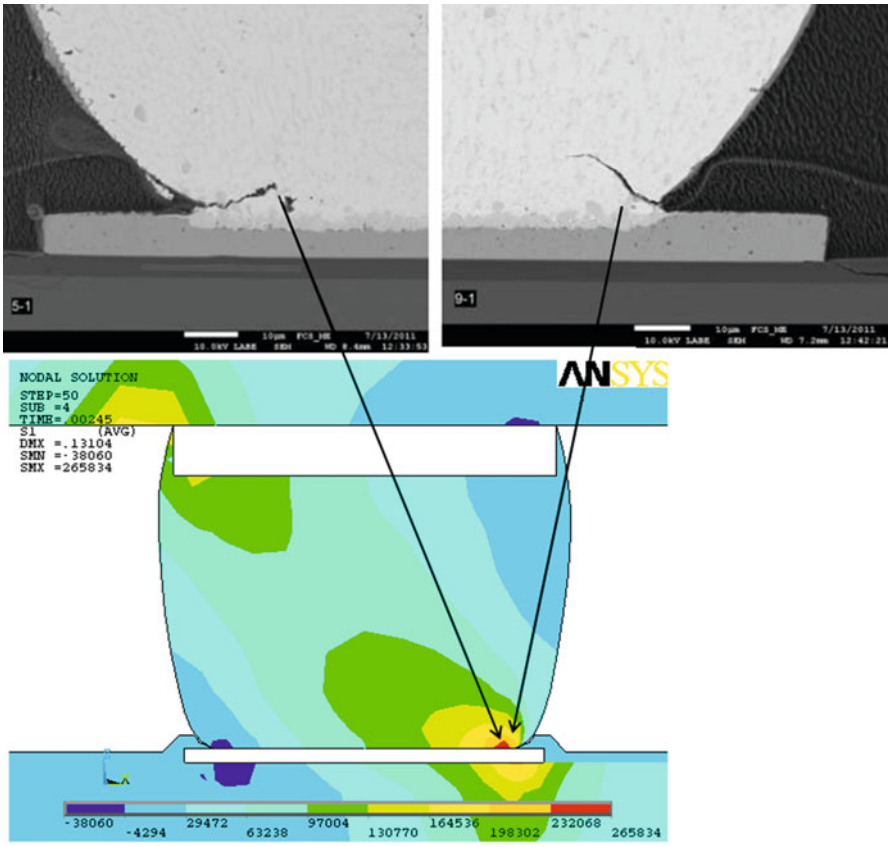


Fig. 10.61 Drop test failure mode of 1Cu1Pi design

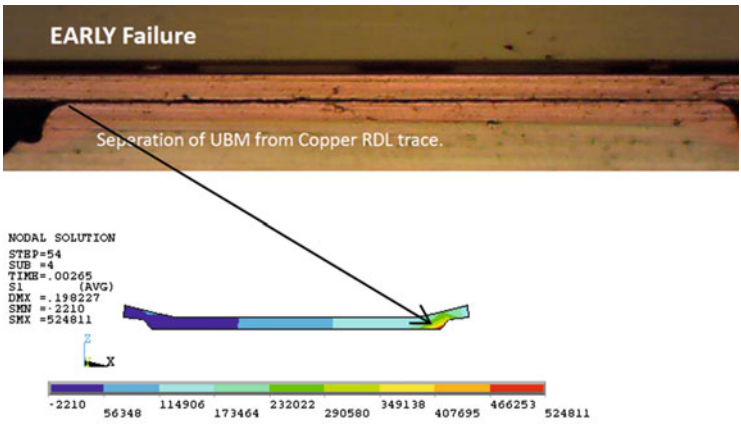


Fig. 10.62 Drop test failure mode of 2Cu1Pi design

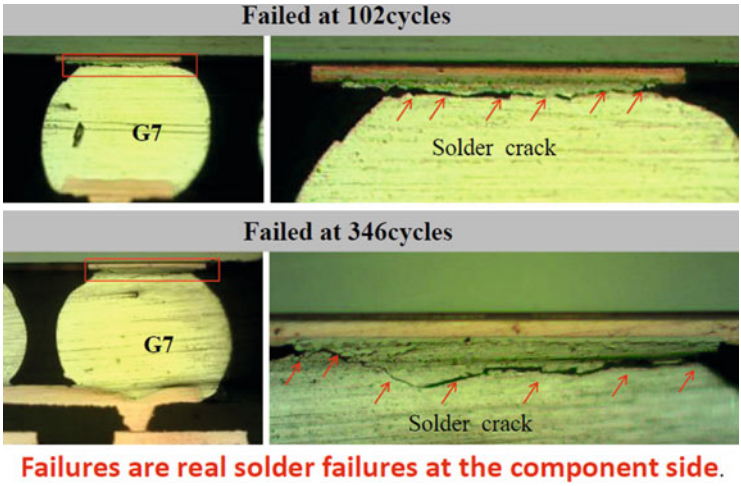


Fig. 10.63 Solder's thermal cycling failure of 1Cu1Pi design

WLCSP. Modeling result shows that the max UBM stress locates at the edge of UBM which is junction to copper RDL. Drop test result shows that the failure happens at the interface of UBM and copper RDL pad, which is consistent with the simulation result.

Figs. 10.62 and 10.63 show the thermal cycling test results. The WLCSP packages were subjected to thermal cycle testing condition according to JEDEC specifications. The temperature range was −40°C to 125°C. Figure 10.63 shows the corner solder failure for 1Cu1Pi WLCSP with UBM non-covered diameter 205 μm and polyimide layout without contact solder (PI layout 3), in which the first failure appears at the 102 cycles and 346 cycles in two group of samples. It shows that

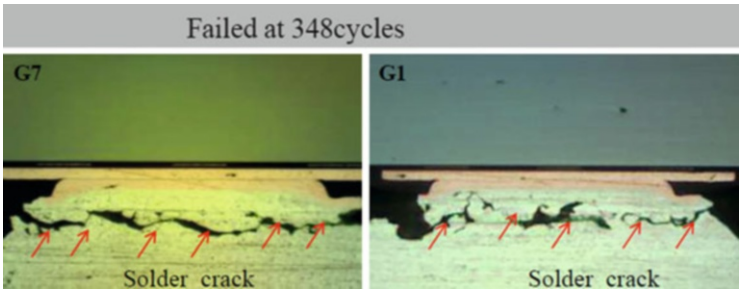


Fig. 10.64 Solder joint thermal cycling failure of 2Cu1Pi design

solder crack locates at the solder area near the UBM interface. The modeling result gives the first failure 265 cycles for the same 1Cu1Pi construction of WLCSP. Figure 10.64 shows the corner solder joint failure for 2Cu1Pi WLCSP with UBM diameter 245 μm and PI open diameter 205 μm. The solder cracks originate at the solder joint near the UBM interface; the first failure appears at the 348 cycles. The modeling result predicts the first failure life 320 cycles for the same design of 2Cu1Pi WLCSP.

10.5.5 Discuss

A comprehensive modeling is carried out to investigate the drop test performance and thermal cycling reliability for the critical design of 1Cu1Pi WLCSP and 2Cu1Pi WLCSP:

1. Modeling results have shown the failure mode of the solder crack at the interface between solder and UBM for 1Cu1Pi design in both drop test and thermal cycling test. While for 2Cu1Pi, the drop test failure happens at the interface of copper RDL and UBM, the thermal cycling failure appears at the solder joint under the UBM.
2. For the 1Cu1Pi design of WLCSP, thicker UBM and larger air gap between polyimide and solder can improve its thermal cycling reliability performance. Deeper etched UBM seems to have better thermal cycling performance. But its effect is not significant.
3. Increasing the non-covered UBM diameter of 1Cu1Pi design of WLCSP can improve solder joint's thermal cycling reliability performance. As the non-covered UBM diameter increases from 205 to 255 μm, the first failure cycle increases from 267 cycles to 281 cycles, and the character life cycle increases from 435 cycles to 457 cycles.
4. Increasing UBM diameter of 2Cu1Pi WLCSP can increase solder's thermal cycling life cycle significantly. As the UBM diameter increases 6.5 %, solder ball's first failure cycle and character cycle increase 18.5 %. As the UBM

diameter increases 10.9 % to be 245 μm, solder ball's first failure cycle and character cycle increase 31.5 %.

5. Increasing UBM diameter of 2Cu1Pi WLCSP can decrease the stress of solder joint at the interface of solder and UBM. As UBM diameter increases from 230 to 255 μm, the max first principle stress, the von Mises stress, Z component peeling stress, and max shear stress decrease.

10.6 Summary

This chapter discusses the typical WLCSP reliability and tests. The basic reliability tests are listed first to show the different reliability requirement and test standards in Sect. 10.1. Then the WLCSP solder ball shear performance and failure mode are studied with both experiment and simulations in Sect. 10.2. The results showed that more brittle fracture takes place in IMC layer when the impact speed increases. The simulated and experimental load–displacement response curves reveal that the required fracture energy in high speed is less than that in low speed; however, the peak traction is evidently improved. In Sect. 10.3, the reliability of WLCSP assembly reflow process and PCB design is investigated. The simulation results clearly indicated high stresses can be introduced by the placement of the copper plated through vias. In the case where through vias are only under some of the solder joints, excessive stress are expected on solder joints with copper plated through vias under. To avoid the early failures seen in 25 ball WLCSP qualification, design changes in test PCB have to be made. If routing is not an issue, PCB design without through vias is recommended. If more than one layer is needed for signal/power/ground connection, blind vias should be in favor of through vias. If through vias are designed due to other considerations, it is recommended to place through vias under every solder balls. The board level drop test of WLCSP is presented in Sect. 10.4, in which different WLCSP design parameters, geometry, and materials versus the dynamic response in both simulation and the drop test experiment are studied. More comprehensive designs for reliability studies with next-generation WLCSP design in both board level temperature cycling and drop test are presented in Sect. 10.5; these studies give the detailed design layout for metal stacking and polyimide for 1 Cu1 Pi and 2Cu 1Pi design layouts.

References

1. Liu, Y.: Power electronic packaging: Design, assembly process, reliability and modeling. Springer, Heidelberg (2012)
2. Chai, T. C., Yu, D. Q., Lau, J., et al.: Angled high strain rate shear testing for SnAgCu solder balls. In: Proceedings of 58th Electronic Components and Technology Conference, pp. 623–628. (2008)
3. Zhang,Y., Xu, Y., Liu,Y., Schoenberg, A.: The experimental and numerical investigation on shear behavior of solder ball in a Wafer Level Chip Scale Package, ECTC 60, (2010)

4. Dugdale, D.S.: Yielding of steel sheets containing slits. J. Mech. Phys. Solids **8**, 100–108 (1960)
5. Barenblatt, G.I.: The mathematical theory of equilibrium of crack in brittle fracture. Adv. Appl. Mech. **7**, 55–129 (1962)
6. Needleman, A.: A continuum model for void nucleation by inclusion debonding. ASME J. Appl. Mech. **54**, 525–531 (1987)
7. Tvergaard, V., Hutchinson, J.W.: The influence of plasticity on mixed mode interface toughness. J. Mech. Phys. Solids **41**, 1119–1135 (1993)
8. Tvergaard, V., Hutchinson, J.W.: On the toughness of ductile adhesive joints. J. Mech. Phys. Solids **44**, 789–800 (1996)
9. Liu, Y., Qian, Q., Qu, S., Martin, S., Jeon, O.: Investigation of the assembly reflow process and PCB design on the reliability of WLCSP. ECTC62. (2012).
10. Liu, Y., Qian, Q., Kim, J., Martin, S.: Board level drop impact simulation and test for development of wafer level chip scale package. ECTC 60. (2010)
11. Dhiman, H.S., Fan, X.J., Zhou, T.: JEDEC board drop test simulation for wafer level packages (WLPs). ECTC59. (2009)
12. Liu, Y., Qian, Q., Ring, M., et al.: Modeling for critical design and performance of wafer level chip scale package. ECTC62. (2012)
13. Liu, Y.M., Liu, Y.: Prediction of board level performance of WLCSP. ECTC63. (2013)
14. Liu, Y.M., Liu, Y., Qu, S.: Bump geometric deviation on the reliability of BOR WLCSP. ECTC 64. (2014)

Index

S. Qu, Y. Liu, *Wafer-Level Chip-Scale Packaging*,
DOI 10.1007/978-1-4939-1556-9