Springer Series in Advanced Microelectronics 64

Yan Li
Deepak Goyal *Editors*

3D Microelectronic Packaging

From Architectures to Applications

Second Edition

MOREMEDIA

Springer

Springer Series in Advanced Microelectronics

Volume 64

The *Springer Series in Advanced Microelectronics* provides systematic information on all the topics relevant for the design, processing, and manufacturing of microelectronic devices. The books, each prepared by leading researchers or engineers in their fields, cover the basic and advanced aspects of topics such as wafer processing, materials, device design, device technologies, circuit design, VLSI implementation, and sub-system technology. The series forms a bridge between physics and engineering, therefore the volumes will appeal to practicing engineers as well as research scientists.

More information about this series at http://www.springer.com/series/4076

Yan Li · Deepak Goyal
Editors

3D Microelectronic Packaging

From Architectures to Applications

Second Edition

Springer

Editors
Yan Li
Intel Corporation
Chandler, AZ, USA

Deepak Goyal
Intel Corporation
Chandler, AZ, USA

ISSN 1437-0387 ISSN 2197-6643 (electronic)
Springer Series in Advanced Microelectronics
ISBN 978-981-15-7089-6 ISBN 978-981-15-7090-2 (eBook)
https://doi.org/10.1007/978-981-15-7090-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Contributors

Liangbiao Chen ON Semiconductor, 82 Running Hill Rd, South Portland, ME, USA

Paul Conway The Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, Leicestershire, UK

Yuanyuan Dai Nanyang Technological University, Singapore, Singapore

Indranath Dutta School of Mechanical and Materials Engineering, Washington State University, Pullman, WA, USA

Xuejun Fan Department of Mechanical Engineering, Lamar University, Beaumont, TX, USA

Amlan Ganguly Computer Engineering Department, Rochester Institute of Technology, Rochester, NY, USA

Deepak Goyal Intel Corporation, Chandler, AZ, USA

Ran He The University of Tokyo, Tokyo, Japan

Zhiheng Huang School of Materials Science and Engineering, Sun Yat-sen University, Guangzhou, China

Tengfei Jiang Department of Material Science and Engineering, University of Central Florida, Orlando, FL, USA

Satish G. Kandlikar Mechanical Engineering Department, Rochester Institute of Technology, Rochester, NY, USA

Praveen Kumar Department of Materials Engineering, Indian Institute of Science, Bangalore, India

Antonio La Manna IMEC, Louvain, Belgium

Sangil Lee Invensas Corporation, San Jose, CA, USA

Tae-Kyu Lee Department of Mechanical and Materials Engineering, Portland State University, Portland, OR, USA

Menglu Li Department of Materials Science and Engineering, UCLA, Los Angeles, CA, USA

Yan Li Intel Corporation, Chandler, AZ, USA

Kwang-Lung Lin Department of Materials Science and Engineering, National Cheng Kung University, Tainan, Taiwan, ROC

Jinxin Liu School of Materials Science and Engineering, Sun Yat-sen University, Guangzhou, China

Pilin Liu Intel Corporation, Chandler, AZ, USA

Yang Liu School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China

Yingxia Liu Department of Materials Science and Engineering, UCLA, Los Angeles, CA, USA

Huan Ma Intel Corporation, Hillsboro, OR, USA

Ravi Mahajan Intel Fellow, High Density Interconnect Pathfinding, Assembly Technology, Intel Corporation, Chandler, AZ, USA

Mike Neilsen Sandia National Laboratories, Albuquerque, NM, USA

Erasenthiran Poonjolai Intel Corporation, Chandler, AZ, USA

Bob Sankman Retired Intel Fellow, High Density Interconnect Pathfinding, Assembly Technology, Intel Corporation, Chandler, AZ, USA

Hualiang Shi Intel Corporation, Chandler, AZ, USA

Songhua Shi Medtronic, Tempe, AZ, USA

Sai Vadlamani Intel Corporation, Chandler, AZ, USA

Tadatomo Suga The University of Tokyo, Tokyo, Japan

Chuan Seng Tan Nanyang Technological University, Singapore, Singapore

Peter Tortorici Medtronic, Tempe, AZ, USA

Prithwish Chatterjee Intel Corporation, Chandler, AZ, USA

K. N. Tu Department of Materials Science and Engineering, UCLA, Los Angeles, CA, USA

George Vakanas Intel Corporation, Santa Clara, CA, USA

Paul Vianco Sandia National Laboratories, Albuquerque, NM, USA

Yaodong Wang Department of Materials Science and Engineering, UCLA, Los Angeles, CA, USA

Luhua Xu Intel Corporation, Chandler, AZ, USA

Chapter 1
Introduction to 3D Microelectronic Packaging

Yan Li and Deepak Goyal

1.1 Introduction

Microelectronic packaging is the bridge between the Integrated Circuit (IC) and the electronic system, which incorporates all technologies used between them [1]. Advanced 3D microelectronic packaging technology is the industry trend to meet portable electronics demand of ultra-thin, ultra-light, high performance with low power consumption. It also opens up a new dimension for the semiconductor industry to maintain Moore's law with a much lower cost [1–3].

A wide variety of real products assembled by the advanced 3D packaging technology have been unveiled in the recent years. For example, the Apple A7 inside the iPhone 5S, introduced in September 2013, is a 3D package with Package on Package (POP) configuration [4]. As displayed in Fig. 1.1, the wire bond Elpida (now Micron) memory (low power double data rate type-3 (LPDDR3) mobile random access memory (RAM)) package is stacked on top of the Apple A7 flip chip package to achieve better performance with smaller form factor. In early 2014, SK Hynix announced its high bandwidth memory (HBM) products having higher bandwidth, less power consumption, and substantially small form factor, achieved by stacking up to eight DRAM dice interconnected through Through Silicon Vias (TSV) and micro bumps [5]. In July of 2015, AMD introduced the AMD Radeon™ Fury graphics cards, the first Graphic Processing Unit (GPU) to implement HBM by TSVs and micro bumps [6]. Figure 1.2 shows a top down and schematic cross sectional view of the advanced 3D package. The big GPU die is integrated into the Si interposer

Y. Li (✉) · D. Goyal
Intel Corporation, Chandler, AZ, USA
e-mail: yan.a.li@intel.com

D. Goyal
e-mail: Deepak.goyal@intel.com

Y. Li and D. Goyal (eds.), *3D Microelectronic Packaging*,
Springer Series in Advanced Microelectronics 64,
https://doi.org/10.1007/978-981-15-7090-2_1

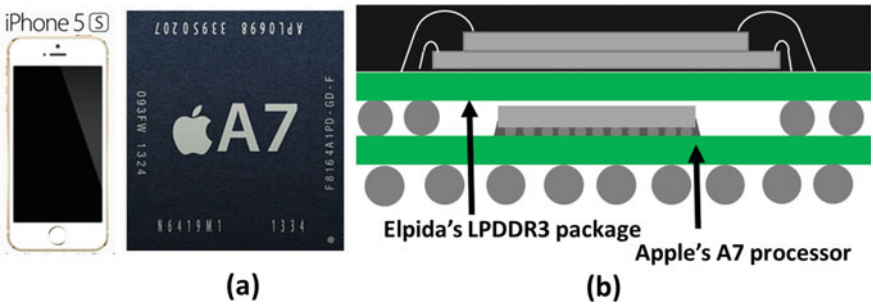

Fig. 1.1 POP inside iPhone 5s. (**a**) Top view of the Apple A7 package. (**b**) Schematic of the cross sectional view (not in scale) (Adapted from Ref. [4])

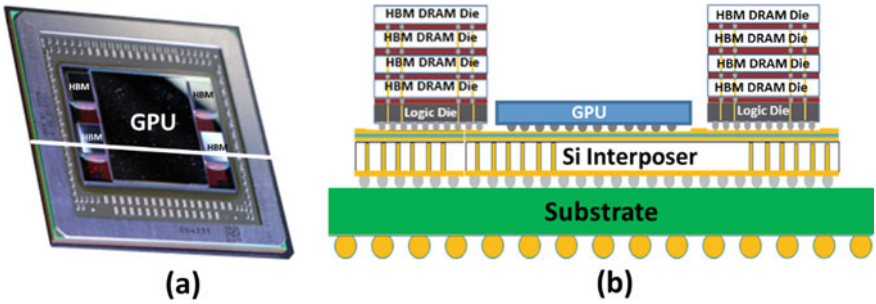

Fig. 1.2 Top view (**a**) and schematic cross sectional view (not in scale) (**b**) of the AMD Radeon™ Fury. The big GPU die is integrated into the Si interposer along with four HBM memory stacks by micro bumps and TSVs (Adapted from Ref. [6])

along with four HBM memory stacks by micro bumps and TSVs to ensure faster and shorter connection between chips [6].

Similar packaging technology is called Chip-On-Wafer-On-Substrate (CoWoS) in TSMC, which achieved volume production in 2016, having multiple, advanced chips integrated on a single Si interposer [7].

Since 2018, Intel launched Kaby Lake-G and Stratix 10 series products with the Embedded Multi-Die Interconnect Bridge (EMIB) technology [8]. Instead of using a full silicon interposer, field-programmable gate array (FPGA), transceivers, HBM die stacks, or GPUs in these products connect together through small Si chips embedded in organic substrates and localized high density micro bumps. In 2019, Intel announced Lakefield, the first product using 3D "active interposer" stacking technique called Foveros [9]. The IO, the cores, and the onboard LLC/DRAM can be manufactured as separate dice and get connected together through die stacking 3D packaging technique. The IO die, which is at the bottom of the stack, performs as an "active interposer" that can deal with routing data between the dies on top.

As illustrated in Fig. 1.3, Integrated Fan-Out (InFO) -Package On Package (PoP)

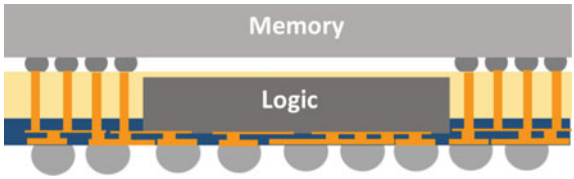

Fig. 1.3 Schematic of Integrated Fan-Out (InFO)—Package on Package (PoP) from TSMC (Adapted from Ref. [10])

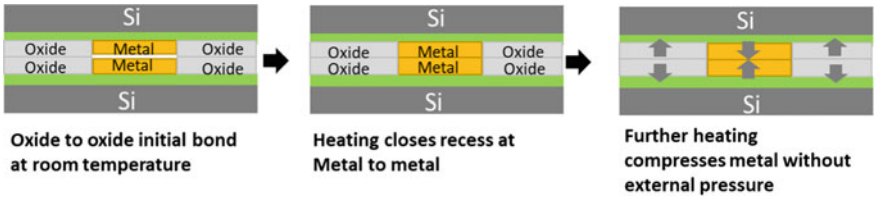

Fig. 1.4 Schematic of Hybrid bonding or Direct Bond Interconnect (DBI) technology from Xperi (Adapted from Ref. [11])

from TSMC, which achieved volume production in 2016, features high density Redistribution Layer (RDL) and Through InFO Via (TIV) to integrate logic die with DRAM package stacking for mobile application [10]. Comparing to Flip Chip PoP, InFO_PoP has a thinner profile and better electrical and thermal performances because of no organic substrate and C4 bump.

Hybrid bonding technology from Xperi, also called Direct Bond Interconnect (DBI), combines a dielectric bond with embedded metal to form interconnect. As demonstrated in Fig. 1.4, the low temperature hybrid bonding solution that allows wafers or die to be bonded with exceptionally fine pitch 3D electrical interconnect enables bump pitch scaling, and has been applied by Sony for 3D stacked back-illuminated image sensors (IMX260) used in Samsung Galaxy S7 Edge in 2016 [11].

These real products bring the 3D packaging techniques from paper to reality, from concept to commercialization, and indicate the extensive applications of 3D packaging technology to microelectronics.

3D packaging technology involves multiple disciplines, for example materials science, mechanical engineering, physics, chemistry, and electrical engineering. A technical book, which could provide a comprehensive scope of 3D microelectronic packaging technology is desirable for graduate students and professionals in both academic and industry area. Current available books on 3D integration typically focus on processing of wafers, especially TSV fabrication, and do not cover other key elements in 3D packaging. This book is proposed to fill in the gap. It presents an thorough extend of 3D packaging, covering the fundamentals of interconnects, bonding process, advanced packaging materials, thermal management, thermal mechanical modeling, architecture design, quality and reliability, and failure analysis of 3D packages, which are critical for the success of advanced 3D packaging.

This chapter provides detailed illustration of motivations as well as various architectures of 3D packaging. Challenges in 3D packaging, including fabrication, assembly, cost, design, modeling, thermal management, material, substrate, quality, reliability, and failure analysis, are reviewed with brief introduction to the chapters addressing these challenges.

1.2 Why 3D Packaging

1.2.1 Moore's Law

Since Intel introduced the world's first single-chip microprocessor, the Intel 4004, in 1971, an exponential growth of ICs has been observed following Moore's law in terms of transistor number per chip [12]. As illustrated in Fig. 1.5, the number of transistors per Si chip doubles approximately every 18 months, resulting in a straight line on a log scale [12, 13]. In 1990, the bipolar transistor technology switched to CMOS in order to reduce thermal power, circuit size, and manufacturing costs, at the same time increase the operating speed and energy efficiency [3]. In the early 2000s, multi-core processors were developed to address the challenging thermal power issue in conventional single-core processors [3]. Since multicore processors require enormous cache capacity and memory bandwidth to achieve the designed performance, 3D packaging becomes one of the viable solutions to provide the required cache and bandwidth with a relatively low cost [3].

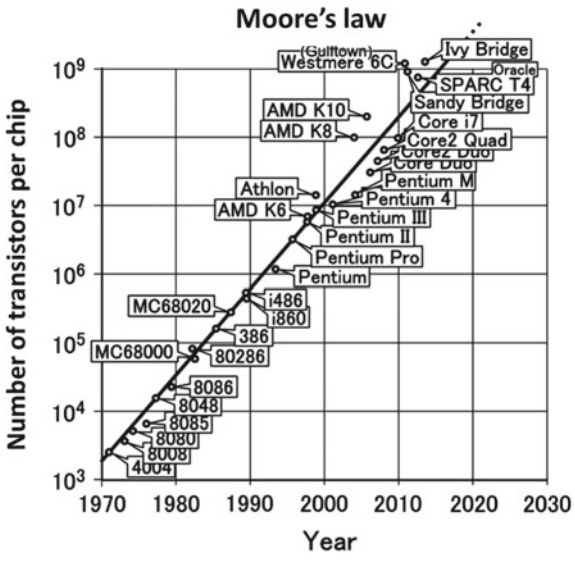

Fig. 1.5 Moore's law predicts the exponential growth of ICs since 1970s (Adapted from Ref. [13])

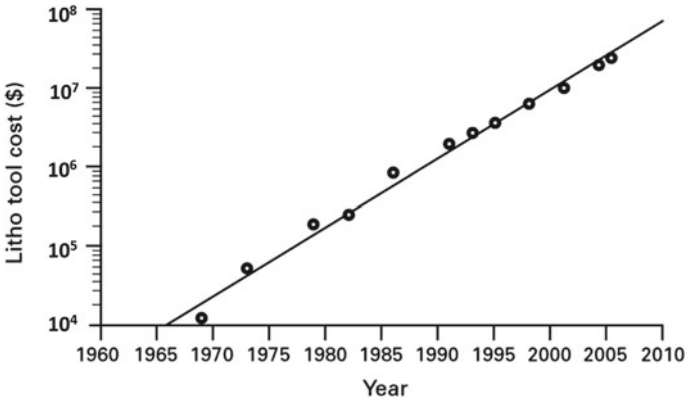

Fig. 1.6 The exponential growth of lithography equipment cost since 1970s (Adapted from Ref. [13])

The conventional method to maintain Moore's law is to decrease the dimensions of components by lithography, which is becoming more and more sophisticated and expensive [13]. Figure 1.6 illustrates the exponential growth of lithography equipment cost since 1970s, which presents an economic challenge as the capital cost rises faster than semiconductor industry revenue [13]. 3D integration technology, which has been recognized as an enabling technology for future low cost ICs, provides the third dimension to extend Moore's law to ever higher density, more functionality, better performance with lower cost [3].

1.2.2 Small Form Factor Requires 3D Packaging

Market demands of small form factor microelectronics head to 3D packages, which are ultra-light, ultra-thin, and with small chip footprint. Si chips in 3D packages are typically 50–100 μm thick, about 90% thinner compared with those in conventional packages. Substrate core thickness of 3D packaging is about 0–100 μm, more than 90% thinner than that of traditional packaging. High density interconnects in 3D packaging are on the order of 5–20 μm in diameter, more than 90% smaller than those in 2D packaging. Thus tremendous reduction in size and weight could be achieved by replacing conventional packaging with 3D technology [2].

Small form factor requires small chip footprint, which is defined as the printed circuit board area occupied by the Si chip, as illustrated in Fig. 1.7 [2]. By stacking multiple dice on top of each other using 3D packaging techniques, the chip footprint could be reduced dramatically. Figure 1.7 schematically demonstrates the difference between conventional 2D packages and 3D packages.

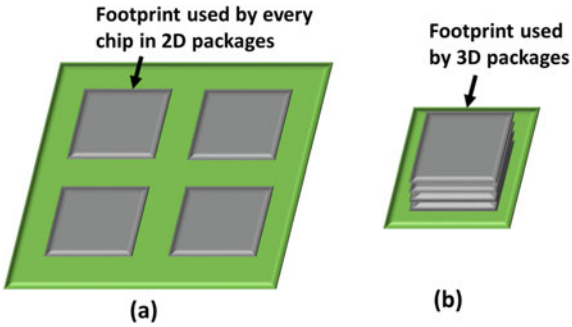

Fig. 1.7 a Schematic illustration of the footprint difference between conventional 2D packages (**a**) and 3D packages (**b**) (Adapted from Ref. [2])

1.2.3 Improved System Performance with Reduced Power

Interconnect length in 3D packages can be significantly reduced compared with conventional 2D packaging [2]. Figure 1.8a shows a typical interconnect length of 4 mm in a 2D package. The interconnect length could be reduced to 200 μm in a 3D package as demonstrated in Fig. 1.8b. Additionally 3D packaging can also greatly improve the interconnect usability and accessibility [2]. Figure 1.9 shows a comparison between 2D and 3D packaging in terms of the accessibility and usability of interconnection. In contrast to eight neighbors to the center element in the case of 2D packaging technology, the utilization of a 3D packaging configuration provides access to 116 neighbors within an equal interconnect length to a center element in the stack.

The significant decrease in interconnect length as well as dramatic improvement in interconnect usability and accessibility result in much less delay in 3D devices, which is primarily limited by the time taken for the signal to travel along the interconnects [2]. Low latency and wide buses in 3D microelectronic systems lead to significant improvement to the system bandwidth [2]. Noise in well-designed 3D

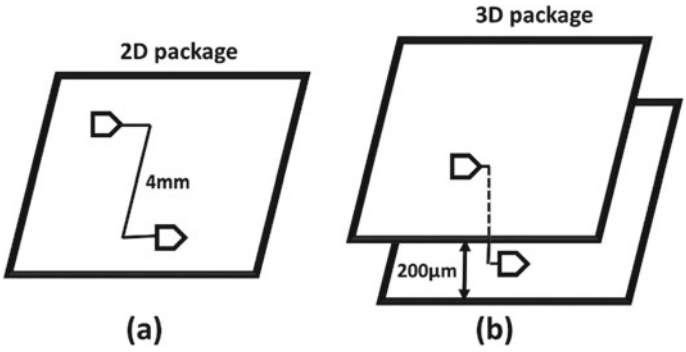

Fig. 1.8 Schematic comparison between the wiring length in 2D packages (**a**) and 3D packages (**b**) (Adapted from Ref. [2])

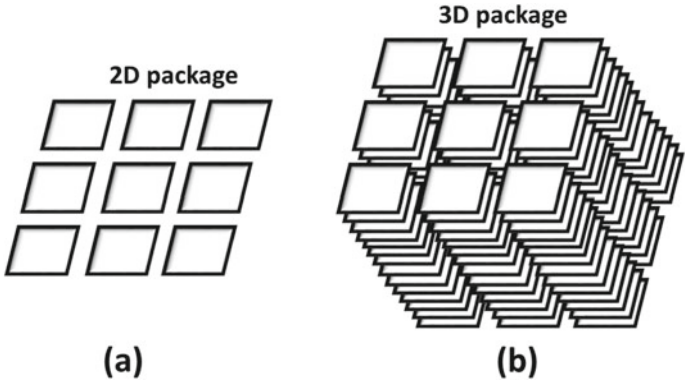

Fig. 1.9 Schematic comparison between 2D packages (**a**) and 3D packages (**b**) in terms of the interconnect accessibility and usability (Adapted from Ref. [2])

microelectronic systems, including reflection noise, crosstalk noise, simultaneous switching noise, and electromagnetic interference, can be reduced as a result of the reduction of interconnection length [2]. Additionally, as the parasitic capacitance in microelectronic packages is proportional to the interconnection length, the total power consumption in 3D packages is also reduced because of the reduced parasitic capacitance [2]. The power saving achieved by 3D technology enables 3D devices to perform at a faster rate or transition per second (frequency) with less power consumption. The overall system performance is greatly improved by applying 3D packaging technology [2].

1.3 3D Microelectronic Packaging Architectures

The various 3D packaging architectures could be divided into the following three categories: die to die 3D integration, package to package 3D integration, and heterogeneous 3D integration combining both package and die stacking [3, 14, 15]. Chapter 2 discusses different 3D packaging architectures in detail along with assembly and test flows.

1.3.1 Die to Die 3D Integration

Die to die 3D integration is enabled by through silicon via (TSV) interconnections and thinned die to die bonding [3]. As illustrated in Fig. 1.10, two memory dice are stacked on top of a logic die with TSVs and micro-bumps. Fist Level Interconnect (FLI) solder joints connects the logic die with the substrate, while Second Level

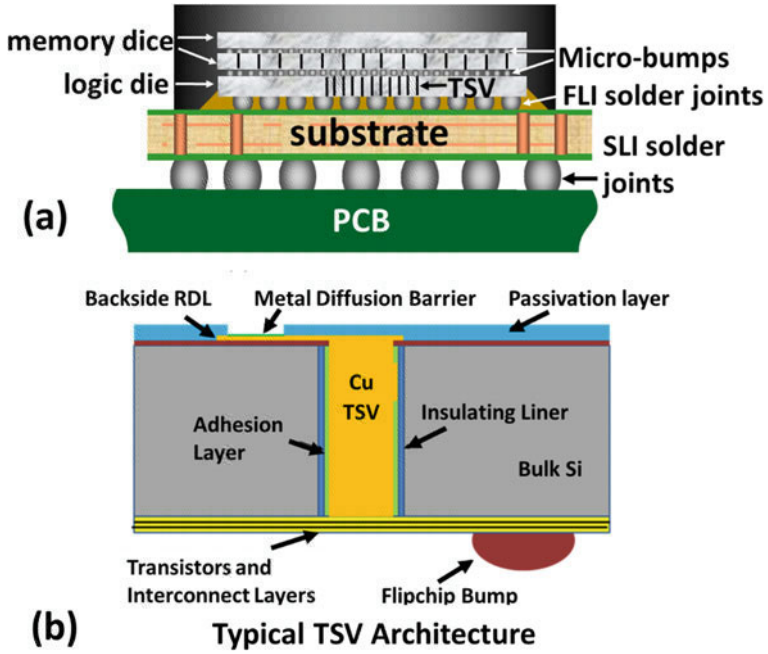

Fig. 1.10 Schematic illustration of die to die 3D integration enabled by TSV and thinned die to die bonding (Adapted from Ref. [15])

Interconnect (SLI) solder joints provides the connection between this 3D package to the Printed Circuit Board (PCB) [15]. The TSVs are formed by laser drilling or deep reactive ion etching, followed by liner deposition and copper fill. There are three typical manufacturing process for TSVs, Via First, Via Middle, and Via Last [16–18]. The detailed process flow of each process as well as the Pros and Cons of each process are discussed in Chaps. 2 and 3.

Die to die bonding is implemented by either Thermal Compression Bonding (TCB) process for solder micro bumps or other alternative bonding process, for example Cu-Cu bonding. Conventional solder mass reflow process in 2D packages, which includes flux dispensing, die attaching, and solder reflow in ovens, is not able to assemble advanced 3D packages having much thinner dice and organic packages, along with much smaller and denser interconnects. As the extend of warpage from both dice and substrates at the reflow temperature overcomes solder surface tension, leading to die misalignment, and results in die tilting, solder joint non-contact opens, and solder bump bridging [19]. TCB bonding process is developed to replace the conventional solder mass reflow process for solder based micro bump assembly in 3D packages. As illustrated schematically in Fig. 1.11, the substrate with pre-applied flux is kept flat on hot pedestal under vacuum to eliminate the substrate warpage. The die is picked up by the bond head, secured and kept flat with vacuum to remove any incoming die warpage. After the die is precisely aligned to the substrate, the

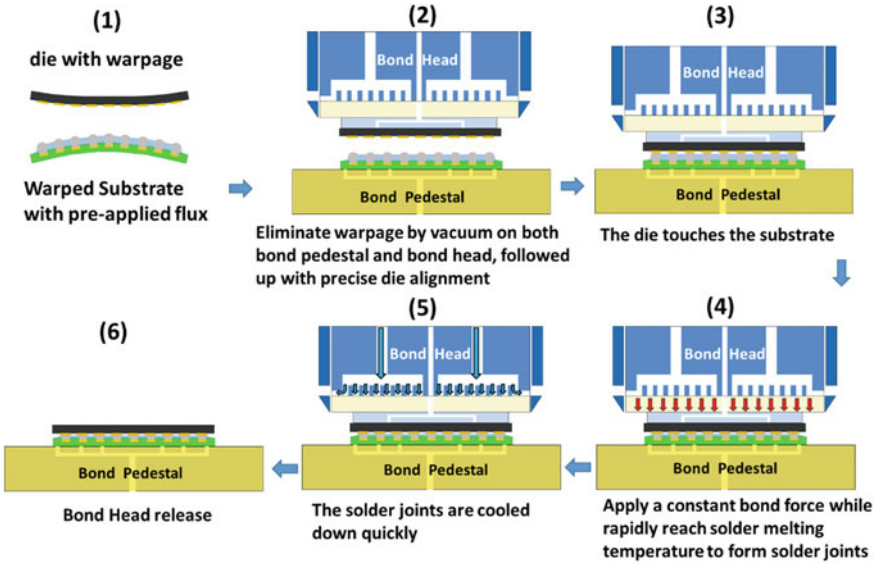

Fig. 1.11 Schematic illustration of a typical Thermal Compression Bonding process (Adapted from Ref. [19])

bond head with die touches the substrate. A constant bond force is then applied on the die through the bond head, while the die is heated up rapidly beyond the solder melting temperature, with a ramping rate higher than 100 °C/s. As soon as the solder joints melt, the die is moved further down to ensure all the solder joints are at the same height. The die is held in this position long enough so that the solder joint forms between the die and the substrate. While the solder is still in the molten state, the bond head with die could retract upwards to control the solder joint height. Subsequently, the solder joints are cooled abruptly below the solidus temperature, with a cooling rate of more than 50 °C/s, followed up with die release from the bond head [19]. Unlike the traditional solder mass reflow process, with up to 10 min of process time for units in batches, the TCB bonding process assemble units one by one with about a couple of seconds per unit [19]. Additionally, thermal ramping rates during both the heating and cooling cycles are much higher than the conventional method. These higher rates result in solder grain size and orientation differences that can affect mechanical properties as detailed in Chap. 10.

Solder based micro bumps are more compliant, thus could compensate bump height variations, lack of co-planarity, and misalignment issues during high volume manufacture. However the TCB process peak temperature needs to be higher than the melting point of solder material, typically in the range of 250–300 °C, which brings more assembly and reliability challenges. Additionally, solder bridging risk gets much higher as bump pitch shrinks from more than 100 μm to less than 40 μm. Alternative bonding process, like hybrid bonding, which could address interconnects with less than 5 μm bump pitch, and assemble at relatively lower temperature is very

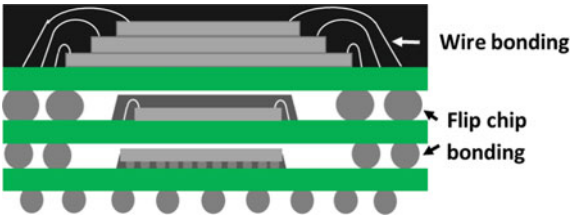

Fig. 1.12 Schematic illustration of package to package 3D integration (Adapted from Ref. [3])

promising. Various types of alternative bonding process is reviewed in Chap. 8 as well as Pros and Cons comparing with solder based TCB process.

1.3.2 Package to Package 3D Integration

System in Package (SIP) and Package on Package (POP) are typical configurations of package to package 3D integration, which is enabled by stacking packages through wire bonding or flip-chip bonding [3]. Comparing to die to die stacking, package to package stacking technique has a shorter development cycle, thus help bring products to market faster with a low price. As displayed in Fig. 1.12, a wire bonding package is stacked on top of the other wire bonding package by flip chip bonding. The two packages are then stacked on a flip chip package to form a POP. The conventional solder mass reflow process could still be used in package to package stacking if the package warpages are within control, and the interconnect size and density is comparable with traditional 2D packages. However market demands require packages to be ultra-thin, which limits the number of packages that could be stacked together. Additionally, solder joints and package materials in SIP and POP need to go through multiple cycles of reflow, bringing process and reliability challenges, like solder joint open, delamination between multiple layers in packages, and moisture control between each reflow process. Chapter 17 discusses in detail the processing and reliability of stacked packaging technique, as well as the Pros and Cons comparing with die stacking.

1.3.3 Heterogeneous 3D Integration

Heterogeneous 3D integration assembles separately manufactured components into the same compacted package, thus provides enhanced functionality, lower power consumption, higher system performance, smaller size, and lower costs [9, 20]. Comparing with the monolithic die technology, the heterogeneous 3D integration allows the use of the best available technology node for each chiplet with different functions to maintain maximum performance. It also greatly improves the Si yield due to much smaller die size for each chiplet. The product development cycle can

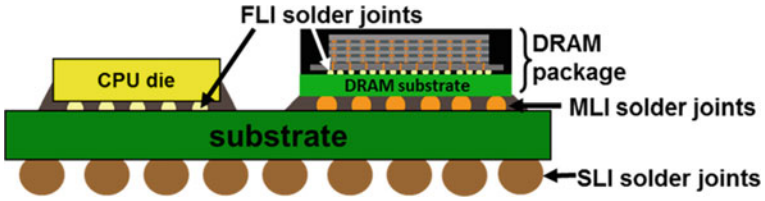

Fig. 1.13 Schematic illustration of 3D packaging architectures having the combination of both die stacking and package stacking (Adapted from Ref. [15])

also be dramatically reduced by integrating commercially available components into the new packages [9, 20].

Depending on product needs, complex heterogeneous 3D packages can have the combination of both die stacking and package stacking [15]. As shown in Fig. 1.13, a 3D Dynamic Random Access Memory (DRAM) package formed by stacking four memory dice on top of the logic die through TSVs and micro bumps is integrated along with a flip CPU chip to form a 3D package by package stacking. FLIs between the CPU chip and the 3D package, MLIs between the DRAM package and the 3D package, as well as interconnects in the substrate provide connection between the CPU chip and the DRAM package [15]. Smaller and denser interconnects between chips and packages are highly desired for better performance, higher bandwidth, and lower power consumption. Figure 1.2 demonstrates the Silicon interposer technology, which could provide better connection between the GPU chip and the HBM die stacks through the Si interposer with dense Cu interconnects and TSVs [15, 21]. Embedded Multi-Die Interconnect Bridge (EMIB) technology is an alternative approach to provide localized high density interconnects between chiplets without Si interposer having TSVs [8]. As illustrated in Fig. 1.14, the link between dice is provided by fine Cu interconnects in Si bridges embedded in the organic substrate and confined denser FLIs between Si bridges and chips. Comparing with Si interposer technology, EMIB technology is able to provide similar performance with

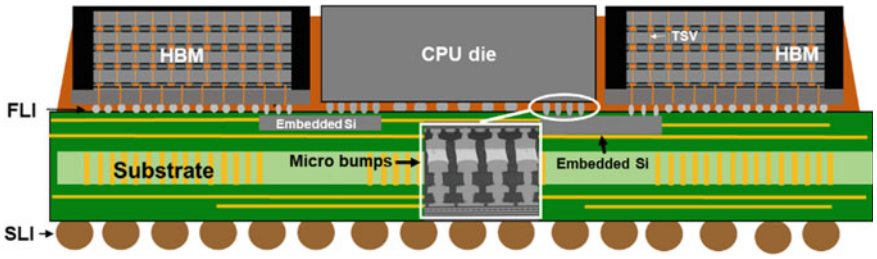

Fig. 1.14 Schematic illustration of Embedded Multi-Die Interconnect Bridge (EMIB) technology providing localized high density interconnects between CPU die and HBM die stacks (Adapted from Ref. [8])

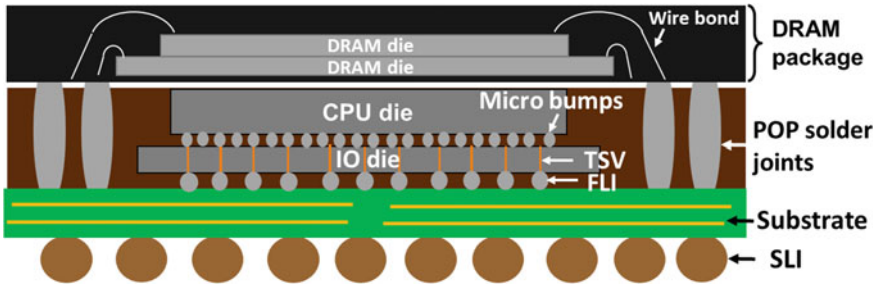

Fig. 1.15 Schematic illustration of a heterogeneous 3D package with Foveros technology providing Face to Face chip stacking between the CPU chip and the IO chip through TSVs and micro bumps. A POP configuration is used to stack a DRAM package to top of the CPU and IO die stack (Adapted from Ref. [9])

a much lower cost, thus opens up new opportunities for heterogeneous 3D packaging [8]. Figure 1.15 illustrates a heterogeneous 3D package for small form factor products, where the package size is constrained. The CPU chip and the IO chip are stacked with Foveros technology, which provides Face to Face connection between chiplets through fine Cu interconnects in chips and micro bumps between chips. TSVs inside the bottom chip are used to connect the organic substrate through FLIs. A POP configuration is used to stack a DRAM package on top of the CPU and IO die stack [9]. The heterogeneous 3D integration approach provides an unprecedented flexibility to chip architects in new product design with various form factor requirements, very aggressive development timeline, better system performance, minimum power consumption, and lowest possible cost.

1.4 3D Microelectronic Packaging Challenges

1.4.1 Assembly Process, Yield, Test, and Cost Challenges

3D packaging involves more challenging assembly steps than conventional packaging, such as TSV wafer fabrication and die singulation process (reviewed in Chap. 7), TCB of micro-bumps (discussed in Chap. 15), multiple solder reflow process for POP (refer to Chap. 17). The complicated process results in yield, test, and cost challenges [2], which could be addressed by redundancy or fault tolerant designs, through-put time (TPT) improvement of assembly process, and minimize process steps based on product quality and reliability requirement [2, 3].