Chapter 2
3D Packaging Architectures and Assembly Process Design

Ravi Mahajan and Bob Sankman

Acronyms

2D Two dimensional
3D Three dimensional
BEOL Back end of line
BI Burn-In
CMP Chemical mechanical polishing
D2D Die-to-die
D2W Die-to-wafer
ECD Electro-chemical deposition
ECG Deleted in chapter
EMIB Embedded multi-die interconnect bridge
FEOL Front end of line
IP Intellectual property
KGD Known good die
KOZ Keep out zone
MCM Multi chip module
MCP Multi chip package
MEOL Middle end of line
MPM Multi package module
PECVD Plasma enhanced chemical vapor deposition
PVD Plasma vapor deposition
Rx Receiver
SBS Side by side
SIP  System in package
SOC  System on chip
TDP  Thermal design power
TIM  Thermal interface material
TSV  Through silicon via
Tx   Transmitter
W2W  Wafer-to-wafer

2.1 Introduction

Increasing transistor density enabled by Moore’s Law [1, 2] has led to increasingly powerful and pervasive computer systems that enable multiple personal and business applications (see Fig. 2.1 for an estimate of the overall size and growth trends of the semiconductor market). The trend in increased compute capabilities has been fueled by increased wired and wireless connectivity [3] which has led to powerful interconnected computer networks. These computers and associated networks utilize a myriad of computing and communication functions that are implemented within digital circuits (e.g., Microprocessors, Field Programmable Gate Arrays), memory circuits (e.g., SRAM, DRAM), and analog circuits (e.g., power supplies, clocks, RF front end modules, amplifiers, SERDES, USB, PCIe, DDR). Different computing and communication functions can be integrated on a monolithic silicon chip (typically referred to as System On Chip or SOC integration) or on a package (typically referred to as System In Package or SIP integration\(^1\)). Integration on chip has the advantages of improved signal transmission fidelity due to reduced interconnect lengths, lower system power due to efficient on-chip connections between IP blocks, and overall reduced silicon resulting from Moore’s Law scaling. Thus, on-chip integration is usually preferred when:

(a) The integrated functions can easily be implemented on the same silicon fabrication process. For example, digital logic and SRAM can be built using compatible silicon processes. Conversely, high performance digital logic and DRAM are rarely fabricated on similar silicon manufacturing processes, and hence are not commonly included in the same chip.

\(^1\) As a general definition, SIP refers to the on-package integration of multiple heterogeneous and/or homogenous ICs each of which may be in the form of unpackaged die, individually packaged die, or packaged modules. iNEMI (International Electronics Manufacturing Initiative) defines SIP as: System in Package is characterized by any combination of more than one active electronic component of different functionality plus optionally passives and other devices like MEMS or optical components assembled preferred into a single standard package that provides multiple functions associated with a system or subsystem. SIP is considered to be subset of the broader concept of System On Package (SOP) [4] where an entire computer system is built on a package.
(b) The IP\(^2\) blocks desired for the SOC are available in the same silicon fabrication process and the resulting chip meets the cost target required to make the product economically viable.

SIP is preferred for the integration of heterogeneous functions [5] (i.e., functions manufactured using disparate semiconductor technologies) and to help bring products to market quicker when technical and/or business reasons prevent timely SOC integration. As seen in Table 2.1 and Fig. 2.2, there continues to be significant interest in SIP configurations. SIP architectures can be broadly classed in three categories:

(a) Planar configurations where two or more die or packages are placed side by side and connected to each other through lateral interconnects in a multilayer substrate.

(b) Stacked configurations where two or more die, or packages, are stacked one on top of the others and connected through a combination of both lateral and vertical interconnects. The value of implementing stacking in a product

---

\(^2\) An IP (Intellectual Property) block is a reusable circuit block that performs a certain specialized functions and serves as a building block for constructing the SOC.
Table 2.1  Volume forecast of different SIP configurations (Source: Prismark Partners LLC)

<table>
<thead>
<tr>
<th>Product/package type volume (Bn units)</th>
<th>2014</th>
<th>2019 F</th>
<th>Leading suppliers/players</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stacked die in package and memory card</td>
<td>8.3</td>
<td>10.5</td>
<td>Samsung, Micron, SKHynix, Toshiba, SanDisk, PTI, ASE, SPIL, Amkor, STATS ChipPAC</td>
</tr>
<tr>
<td>Stacked package on package: bottom package only</td>
<td>0.95</td>
<td>1.2</td>
<td>Samsung, Apple, Qualcomm, MediaTek, Amkor, STATS ChipPAC, ASE, SPIL</td>
</tr>
<tr>
<td>PA centric RF module</td>
<td>4.5</td>
<td>5.9</td>
<td>Qorvo, Skyworks, Anadigics, Avago, Amkor, ASE, Inari, HEG, JCET, Unisem, ShunSin</td>
</tr>
<tr>
<td>Connectivity module (bluetooth/WLAN)</td>
<td>0.6</td>
<td>0.9</td>
<td>Murata, Taiyo Yuden, Samsung, ACSIP, ALPS, USI</td>
</tr>
<tr>
<td>Graphics/CPU or ASIC MCP</td>
<td>0.2</td>
<td>0.2</td>
<td>Intel, AMD, Nvidia, Xilinx, Altera</td>
</tr>
<tr>
<td>Leadframe module (power SiP)</td>
<td>3.2</td>
<td>4.7</td>
<td>NXP, STMicro, TI, Freescale, Toshiba, Infineon/IR, Renesas, ON Semi</td>
</tr>
<tr>
<td>MEMS and controller</td>
<td>5.4</td>
<td>8.2</td>
<td>ST, Analog, Bosch, Freescale, Knowles, InvenSense, Denso</td>
</tr>
<tr>
<td>Camera module</td>
<td>3.7</td>
<td>5.3</td>
<td>LG Innotek, SEMCO, Hon Hai, Lite-on, Toshiba, Sunny Optical, Sharp, Cowell</td>
</tr>
<tr>
<td>Fingerprint sensor</td>
<td>0.35</td>
<td>1.5</td>
<td>Apple, Synaptics, Fingerprint Cards, Cypress/IDEX, Silead, Goodix, NEXT Biometrics, Qualcomm</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>27.2</td>
<td>38.4</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2.2  2015 SIP Market by device type (Source: TechSearch International)
design, and thus leveraging the vertical dimension in a package, has been discussed in detail in [6].

(c) Hybrid configurations that combine both the planar and stacked configurations.

There are a number of innovative ways to construct SIP architectures using planar and stacked structures. Some of the more well-known SIP architectures are shown in Fig. 2.3. Figure 2.3 illustrates how different architectures have evolved to meet specific needs for different markets. Figure 2.3 is not intended to be a complete list of all the possible architectures. It is clear from the configurations shown in Fig. 2.3a–f that a number of SIP combinations are possible, offering feature, size, configuration, and cost flexibility.

One class of SIP configurations that has driven the most significant technology changes is through silicon vias (TSV)-based SIP. This chapter focuses on 3D stacks that are enabled by TSV technology. TSV-based SIPs have been the subject of considerable research for more than two decades and numerous papers have explored applications, architecture and design opportunities, as well as process, material, and equipment complexities [11–14]. Since it is difficult to comprehensively discuss all these aspects in a general overview, this chapter will attempt to provide a broad perspective of the architectural and process opportunities and complexities. The process of TSV formation has been previously discussed in depth in Chap. 3 and will not be repeated here, except for a brief reference in Sect. 2.3.

The most commonly used interconnect between stacked die for currently available products with TSVs is solder based (Fig. 2.4a, b) with interconnect pitches as low as 40 μm. Solder-based interconnects have an advantage of being compliant and hence they are more tolerant to misalignment and lack of coplanarity between bonded surfaces during assembly3. However as the joints become increasingly small, with decreasing interconnect pitches projected below 40 μm for future 3D stacks, the available solder volume will be reduced and a greater proportion of the solder joint will become intermetallic compounds [15], thus decreasing its compliance. Additionally, with shrinking interconnect pitch, there is an increasing risk of solder bridging during the assembly process since the joints are closer to each other. Various research groups have suggested the need for alternate interconnects; the most common among these are Cu-Cu bonding [17–21], a subject covered in detail in Chap. 6. In 2016, there are two types of widely available products with TSVs. These are DRAM memory stacks [7, 22, 23] and image sensors [24–26].

---

3 In most applications, Thermo-Compression Bonding (TCB) is used to create the fine pitch interconnect typically needed between two stacked die because of its superior alignment capability over reflow based flip-chip bonding [16].
Fig. 2.3  (a) Multi Chip Module (MCP)—Two or more die are attached in a planar configuration to a package substrate (red line in schematic drawing indicates that the two or more die are electrically connected using lateral on-package interconnects). Image shows the Intel Iris Pro processor with a CPU (larger of the two die) and a DRAM chip. (b) Multi Package Module (MPM)—One or more die are packaged before being attached to the final package. Image above shows the Intel’s Knights Landing Processor which uses the MC-DRAM Memory Module. The MC-DRAM memory module (a memory stack that conforms to the Hybrid Memory Cube specification [6]) has four stacked memory die connected using TSVs, placed on top of a logic chip. (c) Embedded SIP modules—Embedded SIP modules have solder-less die-to-package interconnects and are usually built using reconstituted wafer level or panel level processes [7, 8]. (d) Package-on-Package (POP) SIP module—Two packages are stacked on top of each other and connected using peripheral interconnects. The peripheral interconnects can be made using solder balls, Cu posts, or solder-coated Cu balls. Example of an MCeP® (Molded Core embedded Package) is shown above. MCeP® is the registered trademark of SHINKO Electric Industries CO., Ltd. (e) EMIB (Embedded Multi-Die Interconnect Bridge) [9]—EMIB-based SIPS use embedded silicon for localized high-density interconnects. Figure above shows a Stratix-10 part from Altera [5]. (f) Silicon interposer-based SIP Module—A silicon interposer is used to for fine feature interconnects between the different chips or chip stacks. The interposer has TSVs to connect the chips to the package substrate. See [10] for a product implementation. (g) 3D TSV-Stacked module—Multiple die are stacked using TSV-based interconnects. Image above shows a HMC memory stack from Micron. (h) SIP Variants that include combinations of wirebond and flip-chip interconnects. The Possum die is from AMKOR.
Fig. 2.3 (continued)
2.2 3D TSV-Based Architectures: Advantages and Limitations

In the past few decades, there has been considerable interest in 3D TSV-based SIP architectures [27–30] because of their compact construction and electrical advantages. The advantages include reduced signal latency, and lower signaling power dissipation due to reduced interconnect lengths. Figure 2.5 shows the energy consumption and dissipation during a 0 → 1 and a 1 → 0 bit transition based on a simple capacitance model. Assuming the same Tx and Rx capacitance, energy draw, and dissipation is directly proportional to the capacitance of the interconnect, and this is where TSVs have an advantage over traditional planar wire connections. Since TSVs are short length interconnects, they can exhibit a significantly lower (typically 4–20 times lower [31]) interconnect capacitance compared to a side-by-side or planar interconnect. In Fig. 2.6a, a memory is connected to the processor
Fig. 2.4 (a) Solder joints @ 45 μm pitch at the GPU to silicon interposer (AMD Fury [10]). (b) HBM Stacks. Leftmost image shows solder bumps on the HBM stack, center image shows the top view of a HBM package with side mold. Image on the right shows the Memory–Memory Interconnect (MMI) and the TSVs in the memory die. The MMI pitch is 55 μm

![Diagram](image)

**Fig. 2.5** Simple capacitance model describing the energy consumption in a Tx (Transmitter)—Rx (Receiver) interconnect link

- **0 → 1**
  - Tx draws power (~CV^2) to charge all the capacitors
  - Energy stored in capacitors is ~\(\frac{1}{2}CV^2\). Energy dissipated is also ~\(\frac{1}{2}CV^2\).

- **1 → 0**
  - Capacitors discharge; No power is drawn
  - Energy dissipated is ~\(\frac{1}{2}CV^2\).
through a planar on-package interconnect and in Fig. 2.6b the connection uses TSVs. A comprehensive calculation [32] shows that the interconnect power efficiency (mW/Gbps) dropped from 15.65 mW/Gbps (for a bounding case where a DDR memory in a DIMM slot was accessed by the CPU) to 0.55 mW/Gbps when the CPU accessed a WIO4 memory through TSVs). This is more than a $28 \times$ improvement5. This power efficiency, multiplied by the number of wires in the interface, combined with reduced data latency due to reduced interconnect lengths is what has motivated designers to consider TSV-based architectures [33]. The power efficiency advantage is appealing across the application spectrum from handheld devices where longer usage times between battery recharging can be enabled to servers where lower energy costs make 3D TSV stacks attractive.

While energy efficiency and latency are the significant advantages for 3D TSV-based architectures, a key limitation is that the maximum Thermal Design Power (TDP) can be significantly lower than with comparable 2D configurations. This is due to the thermal resistances in a 3D stack which are in series and hence additive (Fig. 2.7). Consider the case where a DRAM memory is stacked on top of a logic processor6. The Processor–DRAM interface, the bulk DRAM silicon, and in the case of multiple DRAM die, the DRAM-DRAM interfaces all add thermal resistances to the heat flow. This compares to a planar case where heat from the processor has to only flow through the silicon of a single chip and a Thermal Interface Material (TIM) before reaching a Heat Spreader or the base of a heatsink. Thus, the thermal impact of stacking is that it reduces the total Thermal Design Power (TDP) available for the system designer to utilize for the processor. This is

---

5 Power efficiency quoted for the ESD case. See [Ref Hazkazemi paper] for a detailed review of power and performance differences between LPDDR and Wide-IO.

6 This is architecturally a more likely scenario for CPU-DRAM 3D stacks, since the CPU typically needs more bump interconnects than a DRAM and power delivery to a DRAM would require fewer TSVs compared to the converse case, where power is delivered to the CPU through the DRAM.
quantitatively shown in Fig. 2.8 using a simple case study comparing TDP between the 3D case (where 1 and 4 DRAMs are stacked on top of a processor) and the 2D case where 1 DRAM or a 4 DRAM stack is placed side-by-side to the processor. In this case study, the following assumptions are made:

1. The processor and DRAM silicon are both 100 $\mu$m thick.
2. Processor power is uniform across the processor die and is time independent.
3. Junction temperature limits for both the processor and DRAM are 105 °C.
4. Three different system cooling solutions are considered, i.e.,

   (a) High thermal resistance cooling solutions which depend mainly on conduction (Typically encountered in laptop environments where active airflow is not possible)

   (b) Medium thermal resistance cooling solutions which use both conductive and convective heat transfer (Typically encountered in desktop environments that use cost-performance optimized cooling solutions)

---

7 System cooling refers to the cooling solution attached to the SIP.
Low thermal resistance cooling solutions which use both conductive and convective heat transfer (Typically encountered in server environments that use performance optimized solutions to manage high TDP envelopes).

5. The Processor–DRAM and DRAM–DRAM interfaces, comprising metallic interconnects and protective underfill, are key interfaces that limit heat transport since the effective thermal conductivity (K) across that interface tends to be in the approximate range (0.3–3.0 W/m °C), which is significantly lower than the thermal conductivity of silicon (110 W/m °C) or Cu (390 W/m °C). Two different interface conditions are considered for the analysis including a low interface with \( K = 0.3 \) W/m °C and high interface with \( K = 3.0 \) W/m °C.

In cases where the system thermal resistance is high relative to the package resistances, the impact of changes in package resistance has a lesser influence on the overall TDP. In these cases, a system designer can increase memory capacity and still take advantage of improved memory capacity and power savings. As the performance of the system thermal cooling solution improves, increasing package resistance (with increasing number of stacks) will have a greater influence on the product’s TDP capability. In this scenario, the TDP degradation and increasing the number of stacks is more significant. It can be seen from Fig. 2.8 that since the bulk of the heat transfer in the package is conduction based. For this case, improvements in the effective thermal conductivity of the interfaces between the stacked silicon chips are critical to improving the overall TDP capability. The effective thermal conductivity of the interface can also be improved by increasing the number of micro-bumps between the die, and by increasing the effective thermal conductivity of the underfill, or polymeric encapsulant, used to increase the reliability of the micro-bumps. Additionally, there is a need to improve the quality of system thermal solutions. Chapter 10 summarizes the various thermal solution strategies developed to address the thermal management problem in 3D stacking.

It should also be noted that the assumption of uniform power distribution in the processor is simplistic. In most cases, the processor power distribution is nonuniform. The presence of the additional thermal interfaces from the DRAM stack will exacerbate the hot spots compared to a side-by-side configuration. Design approaches that minimize the thermal burden on the package have been shown to be capable of achieving thermal equivalence with 2D configurations [34].

Another key consideration is the impact of TSVs on the stress state in silicon. Cu and W are typical materials used for TSVs and polysilicon TSVs have also been mentioned in literature [35]. TSVs are typically filled using Electro-Chemical Deposition (ECD) at temperatures greater than 200 °C. The ECD temperatures are significantly higher than the typical operation temperature of a processor or a memory device which tend to be in the (90 °C–110 °C) range. At the deposition temperature, the TSV is in an equilibrium (i.e., stress-free) state with the

\[^8\] 200 °C quoted as a typical lower temperature bound. A number of FEOL and MEOL processes have deposition temperatures significantly higher than 200 °C.
surrounding silicon, however at operating temperatures, the TSV induces radially tensile and tangentially compressive stresses in the surrounding silicon (see Fig. 2.9 for a schematic illustration) resulting from the CTE differential between the Cu/W fill and the surrounding silicon. These stresses have an impact on the electron and hole mobility in the transistors and hence an impact on transistor performance and reliability [36–38]. Keep out zones (KOZ), i.e., regions in the silicon surrounding the TSVs where transistors cannot be placed, are specified for silicon designers so that performance and reliability impact is reduced to an acceptable level. The consequence is that there is a silicon area increase due to both the TSV and its associated KOZ.

In summary, the power efficiency advantage due to short TSV interconnects must be balanced against the disadvantages in TDP reduction and increased chip area due to the TSV integration. A designer must ensure that the disadvantages don’t adversely affect the overall performance or value of the product being designed.

### 2.3 Methods of Fabrication and Other TSV Attributes

The structure and location of a TSV with reference to the transistors and back-end interconnect stack in a typical wafer is shown schematically in Fig. 2.10. TSVs are typically created using a process with the following steps:

- A photoresist is coated on the silicon wafer and lithographic exposure is used to define the TSV locations.
- The silicon is etched, typically using the Bosch process [39], which uses multiple sequences of *etch and coat* to create the via holes.
- The photoresist is then stripped and the wafer surface is cleaned.
The inner walls of the TSV are coated with a dielectric liner (typically SiO₂ though Si₃N₄ has also been reported [40] using thermal oxidation or PECVD (Plasma Enhanced Chemical Vapor Deposition). The liner electrically isolates the TSV from the bulk silicon.

Next an adhesion layer, typically Ti or Ta, is added on the inner lining of the TSV using PVD (Physical Vapor Deposition).

A seed layer is deposited in the via hole next and the vias are ready for deposition of the TSV metal.

Electro-Chemical Deposition (ECD) is used to fill the via holes with Cu or W.

Vias are annealed to stabilize the TSV microstructure and to relieve stresses within the TSV.

Finally, Chemical Mechanical Polish (CMP) is used to planarize the wafer.

TSVs are typically manufactured using three different processes (Fig. 2.11)

1. **Via First Process (Also referred to as the Front End Of Line (FEOL) Via Process):** In this process, the TSVs are created before the transistors [40, 41]. Key steps in this process are

   (a) The wafer surface is patterned and TSVs are etched and filled on the transistor side (front or active side) of the wafer.

   (b) After the TSV formation, the wafer is planarized and readied for transistor creation.
Transistor creation is followed by the Back End of Line (BEOL) process which creates the multiple metal/insulator layers on the silicon wafer. Finally, the wafer backside is thinned to reveal the vias and creates a Redistribution Layer (RDL) (Fig. 2.12).

An important advantage of the Via First approach is that since the vias are created before transistor creation, only known good wafers with defect-free TSVs can be used for subsequent steps. Additionally, some authors [40] have claimed that the Via First process allows for a fewer design constraints and a higher density of vias compared to the Via Middle and Via Last processes. A key limitation of the Via First approach is that all the TSV materials and processes need to be compatible with CMOS processes and temperatures (~1200 °C). Polysilicon vias are mostly used in the Via First process [41]. Bauer et al. [40] discuss a process where high density vias are created in a Via First process using a silicon as a sacrificial material during the initial via formation step. The silicon is replaced with W after the transistor formation processes are completed.

2. Via Middle (Also referred to as Mid End of Line (MEOL) Via Process: In this process, TSV formation is after front-end device creation and before the creation of all the back-end metal layers [41–47]. Key steps in this process are as follows:

(a) Front-end devices are fabricated including transistors and several lower metal layers.
(b) Next, TSVs are created from the active device side typically at temperatures in the 400 °C–450 °C.
(c) The TSV is then plated with Cu or W.

While this statement seems intuitively feasible, the authors are not aware of an authoritative study that establishes the design advantages of the Via First process over the Via Middle or Via Last options.
Fig. 2.12 Bumping process flow post TSV formation, shown for the MEOL/FEOL TSV process
(d) The wafer is then annealed at \(\sim 400 \, ^\circ\text{C}\) to relieve stresses in the TSV and to stabilize the TSV metal structure.

(e) Next, the wafer is planarized with a typical damascene process and readied for the remaining BEOL metallization.

(f) After BEOL metallization is complete, the wafer is mounted on a carrier and thinned to reveal the TSVs on the inactive (backside) of the wafer. Via reveal is usually (optionally) followed by creation of a Re-Distribution Layer (RDL) and backside bump formation processes that is a critical structure for die stacking (See Fig. 2.12).

An advantage of the Via Middle process is that subsequent chip fabrication processing temperatures are lower than for the Via First process, and hence the thermal stresses induced in silicon due to the CTE mismatch between the silicon and TSV metal are lower.

3. **Via Last (Also referred to as Back End Of Line (BEOL) Via Process):** In this process, TSVs are created after the transistors and interconnects are created \([41,48]\). Key steps in this process are as follows:

   (a) The transistors and the entire BEOL process is first completed.

   (b) The wafer is then temporarily attached to a carrier using an adhesive (usually referred to as a wafer bond/debond adhesive) on the active (front side).

   (c) The wafer is thinned and the TSVs are formed using a similar process to the Via Mid described earlier. This is followed by manufacturing of the RDL and backside bumping. The backside RDL and bumping process are essentially the same as described in Fig. 2.12.

In the Via Last process, TSV processing temperatures are typically lower than \(200 \, ^\circ\text{C}\). This is important since the fully formed wafer is mounted on a carrier using a low temperature adhesive. An advantage of this approach is that the thermal stress issues are lower than with the Via First and Via Middle process options. However, there are two significant concerns with the Via Last process. First since TSVs are created after the BEOL, the TSVs need to land at the right metal layer in the chip metal wiring. Special landing surfaces, also called catch cups, are included in silicon metal layers to precisely locate the TSVs. Special care needs to be taken that the TSVs land on catch cups and don’t accidentally punch through the catch cups. Integration of catch cups requires careful design and restricts the placement of the TSVs. Secondly since the TSVs are formed after complete wafer processing, yield loss due to TSV formation could result in loss of valuable silicon.

Some of the key design attributes of TSVs are their diameter, pitch, aspect ratio (i.e., the ratio of TSV diameter to its depth), electrical characteristics (including resistivity, inductance, and frequency dependent capacitance), and stress-related KOZs. TSV aspect ratios are influenced by a few key parameters such as the ability to get good insulation coverage as a function of via depth and the ability
Table 2.2 Some recent TSV dimensions reported in the literature

<table>
<thead>
<tr>
<th>Reference</th>
<th>TSV diameter (μm)</th>
<th>Silicon thickness (μm)</th>
<th>Aspect ratio (diameter:depth)</th>
<th>TSV material</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>[40]</td>
<td>2</td>
<td>45</td>
<td>1:22.5</td>
<td>W</td>
<td>FEOL</td>
</tr>
<tr>
<td>[41]</td>
<td>5</td>
<td>150</td>
<td>1:30</td>
<td>Poly-silicon</td>
<td>FEOL</td>
</tr>
<tr>
<td>[42]</td>
<td>3</td>
<td>100</td>
<td>1:33.3</td>
<td>W</td>
<td>MEOL</td>
</tr>
<tr>
<td>[48]</td>
<td>10</td>
<td>50</td>
<td>1:5</td>
<td>Cu</td>
<td>BEOL</td>
</tr>
<tr>
<td>[43]</td>
<td>6</td>
<td>55</td>
<td>1:9</td>
<td>Cu</td>
<td>MEOL</td>
</tr>
<tr>
<td>[44]</td>
<td>10</td>
<td>50</td>
<td>1:5</td>
<td>Cu</td>
<td>MEOL</td>
</tr>
<tr>
<td>[49]</td>
<td>5</td>
<td>50</td>
<td>1:10</td>
<td>Cu</td>
<td>MEOL/BEOL</td>
</tr>
<tr>
<td>[50]</td>
<td>20</td>
<td>50</td>
<td>1:2.5</td>
<td>Cu</td>
<td>BEOL</td>
</tr>
<tr>
<td>[51]</td>
<td>2</td>
<td>30</td>
<td>1:15</td>
<td>Cu</td>
<td>Not specified</td>
</tr>
<tr>
<td>[46]</td>
<td>3</td>
<td>50</td>
<td>1:17</td>
<td>Cu</td>
<td>MEOL</td>
</tr>
</tbody>
</table>

to get void-free via filling. Table 2.2 shows some of the published data on TSV diameters and depth. Electrical and stress characteristics are determined by the TSV materials choices (including via metal, liner, and adhesion materials) and process choices (deposition and anneal temperatures).

2.4 Assembly Process Flows

3D stacks can be assembled using three different approaches. They are commonly referred to as Wafer-to-Wafer (W2W) attach [52, 53], Die-to-Wafer (D2W) attach [53–56], or Die-to-Die (D2D) attach.

(a) Wafer-to-Wafer Attach (W2W)

In this process, entire wafers are aligned and then bonded, followed by the singulation of individual die stacks. Two W2W flows are schematically illustrated in Fig. 2.13.

- Back-to-Face Flow: where the inactive side (back) of the top wafer is bonded to the active (front) side of the bottom wafer. This approach can enable multiple wafers with TSVs to be bonded one after the other, while retaining the back-to-face connectivity between the die.
- Face-to-Face Flow: where the active side of the top wafer is attached to the active side of the bottom wafer. This approach is less valued for stacking more than 2 wafers.

A significant advantage of a W2W process is that both bonding surfaces are extremely flat, and thus there can be excellent wafer-to-wafer alignment (<3 μm across the wafer) and hence very fine interconnect pitches can be achieved. In the case of solder-based interconnects, pitch scaling in W2W attach is limited by solder...
bridging between interconnects and the implications of intermetallic compound formation. Another advantage of the W2W process is that it is a batch process and with high manufacturing throughput. Key limitations of the W2W process include: (1) all the die have to be of the same size; (2) the process precludes the ability to bond known good die together; (3) yield loss due to misalignment during this process can be significantly expensive, especially in the case where multiple wafers are stacked.

(b) Die-to-Wafer Attach (D2W)

In this process, individual die (with or without TSVs) are bonded on a base wafer incorporating TSVs mounted on a carrier (Fig. 2.14). This process has similar advantages of alignment as the W2W process, and it is not limited by the requirement for the die to be the same size. The top stacked die used in successive stacking steps can be the same size or smaller than the corresponding die below it. If only known good die are stacked, better yields than the W2W process can be expected. Additionally, top die with different functionality than the bottom die can be stacked, allowing for increased heterogeneous integration. Unlike the W2W flow, D2W is a sequential process which will have considerably slower process throughput times.

(c) Die-to-Die Attach (D2D)

In this process flow, bottom die is first assembled to a package substrate and subsequently other die or die stacks are then stacked on the die connected to an assembled package (Fig. 2.15). This process resolves the die size limitations of
the W2W and D2W flows through careful alignment methods. In this way, die larger than the bottom die can be stacked on top. Since the bottom die can be fully tested prior to committing the top die, this process has best chance among the three process of creating known good stacks. However, a key disadvantage of this process is that since the bottom die is fully assembled to a package substrate
(typically to an organic package), it can become quite warped because of the CTE mismatch with the package. Hence during the attach of the top die to the bottom package, the die to package yield can be compromised due to the alignment challenges that are created. In general, the 3D stacking process is an exercise in precise tolerance and process control and requires very good characterization of the surfaces being assembled as a function of temperature. Some of this is schematically illustrated in Fig. 2.16.

An important consideration in all three attach processes is the ability to successfully underfill the die–die interconnects which is especially challenging at very fine
pitches. Underfill is needed to improve the reliability of the die-to-die and die-to-package interconnects. Since the die–die interconnect pitches are significantly lower (typically ≤ 55 μm) compared to typical die–package interconnect pitches (typically ≥ 100 μm) the gaps or empty space between bumps become smaller for die–die interconnects. This creates another disadvantage of the D2W process. Since it is a sequential stacking process, a dispense underfill process is possible, but underfill bleed-out and unintentional cure in adjacent unattached chip area is a concern. Dispense of underfill after all the die are attached is a possibility that needs to be carefully planned ahead to ensure there is a sufficient gap between the die for adequate underfill flow between the solder-based bump interconnects. Pre-applied film underfills are a better choice for D2W stacking, however if filled underfills are used to deliver the improved thermal performance (Sect. 2.2) filler entrapment in the solder joints during chip attach becomes an important concern.

2.5 Manufacturing Yields and the Role of Test

One of the key goals in manufacturing multi-die modules is to maximize manufacturing yields. Minimizing the number of good die scrapped in modules that fail when tested, has a large impact on average product cost. This proposition can be quantified using a simple example. Consider a module that has $n$ number of die, each of which costs $a$, and is attached to a single package substrate that costs $b$. For simplicity, assume that each die is attached to the package in a single process step (i.e., the entire assembly process is a single integrated step), hence the overall SIP assembly process has $n$ steps—one step for each die. If the yield in each
assembly step is $z$, then the yield after $n$ steps is $z^n$. The yield loss is plotted as a function of number of steps and individual process step yield in Fig. 2.16. The cost due to scrapped modules that fail due to a flaw in the assembly process is $(na + b) \times (1 - z^n)$. There are two primary ways of reducing the cost of scrapped units.

(a) **Increasing individual step yield**
Increasing yield of each individual step in the assembly process is the main focus of packaging technology development. Process, materials, and design parameters for each assembly step, and the impact to upstream and downstream steps of the process are carefully studied and optimized for maximal yield.

(b) **Ensuring each key component “Known Good”**
If only known good components are assembled, the chances that the overall module will perform as intended is significantly increased. Components and subassemblies need to be fully tested prior to assembly to ensure that they are known good and cause no yield loss when modules are tested. Designing an efficient test flow that maximizes quantity of known good components while minimizing any added test cost is a key focus area.

Before discussing the challenges involved in testing SIP modules with 3D TSV stacks, it is necessary to understand the different steps involved. A high-level assembly process flow for a single component is illustrated in Fig. 2.17 to show the typical points in the flow where the wafer, die, and assembled package are tested to check for manufacturing quality or performance. Key steps involved in testing products include:

(a) **E-Test (or Electrical-Test)**—This test step is a process characterization step used in the wafer fab to check the manufacturing quality and device parametric values. E-Test can be used after an intermediate process step to check the quality of that step and/or at the end of wafer processing to assess overall quality. Note that this test step does not influence the SiP Module yield, but can affect Module costs indirectly as lower wafer yields will translate into higher die cost.

---

10 It should be pointed out that this model while illustrative is simplistic in a number of ways. In real life situations, the number of process steps is higher than n; individual process step yields vary and are not always independent of each other.

11 In industry parlance, the acronym KGD (Known Good Die) is used to describe the need for working pretested die.
(b) **Wafer Level Sort**—This is a wafer level test where each individual die is probed and functionally tested to see if it is good enough for assembly. Contact is accomplished using a probe card that has individual probes to make electrical contact with the interconnect pads on the die. Typically, a probe card can contact several die simultaneously with each one individually tested. The test results are stored in a database that tracks each wafer, and die indexed by its position on the wafer. For SiP Modules, Sort is expected to deliver Known Good Die to the assembly process.

(c) **Burn-In**—The function of Burn-In is to accelerate the failure of latent defects (such as fab and assembly process-induced defects or design marginalities) in an assembled SiP unit to become detectable at a Test step. The Burn-In step helps remove Modules that would have contributed to early customer failures typically called “Infant Mortality.” By removing these units, product failures observed by customers are limited to “wearout” mechanisms which occur far in the future. Acceleration of latent failures is accomplished by subjecting the units to higher voltages and temperatures then they would experience in the field, and at long stress times, usually in the range of many minutes to many hours. The goal of Burn-In is to aid in identifying process and design factors responsible for latent failures. As manufacturing processes mature, Burn-In times typically are reduced.

(d) **Package Level Functional Test**—This step is the most comprehensive of all test steps and is used to ensure that the assembled units are defect free, meet performance requirements, and to grade the performance level of the assembled units. Since this is the final and most comprehensive of the Test steps, the level of Test coverage (i.e., how much of the product functionality is evaluated) and the precise control of test parameters are critically monitored.

Now that the key test steps have been described, specific challenges related to the testing of 3D TSV stacks can be discussed with reference to Fig. 2.18 which shows a two chip stack although the conclusions can be extended to higher number stacks. Chips 1 and 2 will be tested at Sort before the wafer thinning and TSV reveal process steps for both the FEOL and MEOL flows, and before TSV formation in the BEOL flow. Only the active side of the die are probed at Sort, and the die backside is ignored. In the W2W and D2W flows, each chip is similarly sorted before TSV and RDL formation. The chips go through the assembly process and package stacks are created. After the entire die stack is mounted on a package, it is functionally tested.

---

12 Determining viability for assembly requires a careful optimization of cost (i.e., cost of probing die on a wafer needs to be balanced against the cost of package waste and need for additional test steps later in the flow) and test coverage (while checking a greater degree of die functionality before packaging is financially viable, it can also require more sophisticated Sort technology).

13 One of the key considerations in manufacturing costs is the time it takes to test units. The greater the amount of time, the lower the throughput and hence higher the costs. The goal in E-Test, Sort, and Package Level Test steps is to focus on test time minimization without impacting quality. Burn-In processes, on the other hand, are designed to run longer so that latent defects are screened out.
The advantage of the D2D flow is that Chip 1 can be tested for full functionality before Chip 2 is attached. Thus, only known good packaged Chip 1 die will be used for assembly. Full functionality of Chip 2 can only be verified through connectivity through Chip 1. This constraint would not exist if the two die were assembled side-by-side in a package. Additionally, since the heat path for Chip 1 is impeded by Chip 2, precise temperature control of Chip 1 during full functional test is a more challenging than in the corresponding side-by-side case. Both of these constraints imply that there are more challenges with stacked die testing vs. side-by-side, and the potential for higher compound yield loss is greater too.

2.6 Challenges with 3D TSV Architectures

It should be clear from the discussion so far that TSV-based architectures, while attractive from a performance and power efficiency perspective, can be thermally limited. Increasing the thermal envelope of 3D stacks is an important challenge. Considerable research in enhancing conduction and convection modes of heat transfer are described in Chap. 10. 3D TSV stacks break existing paradigms and require new design tools that accommodate the area and stress impacts on transistor performance while taking advantage of the newly available vertical TSV interconnects.

From a manufacturing perspective, TSV manufacturing processes create greater overlap between back-end silicon fabrication and assembly technologies. Additionally, it requires considerable investment in new equipment such as deep via etchers, high aspect ratio Cu plating, ultra-thin wafer handling, and stacked die assembly. Stacked die assembly drives the need for new materials such as fluxes that work in narrow chip gaps with minimal residue, and high thermal conductivity underfills that are compatible with different assembly flows. 3D stacks drive new Test paradigms including fine pitch test and a very detailed understanding of KGD and KGSD.
2.7 Summary

A broad overview of the value and importance of SIP packages has been provided along with motivation for TSV-based 3D architectures. TSV-based 3D stacking has generated considerable interest in the past two decades and the considerable research and development effort in architecture and manufacturing has resulted in a detailed understanding of multiple aspects of the technology.

Acknowledgments

The authors would like to acknowledge Prismark Partners LLC, TechSearch International Inc., Shinko Electric Industries Co., Ltd, Amkor Technology®, ASE Group, SK Hynix, for their generous permission to use their pictures. Thanks are also due to Dr. Zhiguo Qian (Intel Corporation) for his help on the section on IO power dissipation, Upendra Sheth (Intel Corporation) for his help in compiling information on different MCP technologies, Dr. Arnab Choudhury (Intel Corporation) for help with thermal analysis and Prasad Ramanathan (Intel Corporation) for help with images. Guidance from Chris Nelson (Intel Corporation) on test processes is also gratefully acknowledged. Finally, thanks are due to Dheeraj Reddy (Intel Corporation) for a thorough review of this chapter.

References

1. G. Moore, Cramming more components onto integrated circuits. Electronics 38(8), 114 (1965)
15. G.G. Vakanas, O. Minho, B. Dimcic, K. Vanstreels, B. Vandecasteele, I. De Preter, J. Derakhshandeh, K. Rebibis, M. Kajihara, I. De Wolf, Formation, processing and


32. M.A. Karim, P.D. Franzon, A. Kumar, Power comparison of 2D, 3D and 2.5D interconnect solutions and power optimization of interposer interconnects, in *ECTC 2013*, pp. 860–866

33. M.H. Hajkazemi, M.K. Tavana, H. Homayoun, Wide I/O or LPDDR? exploration and analysis of performance, power and temperature trade-offs of emerging DRAM technologies in


die-level 3D integration and backside TSV technologies. in *Proceedings of the 64th Electronic Components and Technology Conference (ECTC)*, pp. 304–308, May 2014


53. T. Ohba, Wafer level three-dimensional integration (#DI) using bumpless TSV interconnects for tera-scale generation, in *2013 IEEE*, 2013, pp. 1–4


3D Microelectronic Packaging
From Fundamentals to Applications
Li, Y.; Goyal, D. (Eds.)
2017, IX, 463 p. 331 illus., 253 illus. in color., Hardcover
ISBN: 978-3-319-44584-7