Low-Profile 3D Silicon-on-Silicon Multi-chip Assembly

P. Andry¹, B. Dang¹ and J. Knickerbocker¹
K. Tamura² and N. Taneichi²

¹IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA
²TOKYO OHKA KOGYO CO., LTD., 1590 Tabata, Samukawa-machi, Koza-gun, Kanagawa 253-0114, Japan

Abstract

The focus of this paper is multi-chip 3D silicon-on-silicon assembly using low-profile lead-free (Sn-Cu) solder interconnects. Thin 3D chips (~70 µm thick) containing tungsten TSVs and Cu wiring links were fabricated, diced and precision bonded to silicon substrates comprising ~20 µm tall lead-free solder “pancake” bumps on Ni UBM’s. Modules with up to 24 thin 3D chips were fabricated and yield tested. The electrical effect of sequential joining was studied by adjusting the batch size of 3D chips joined in a single reflow. Four-point electrical measurements on single bump sites as well as chains having >200 bumps/links show a clear shift in resistance as a function of the number of reflows, and this shift is correlated to the amount of Ni consumption at the solder/UBM interface. Yield chain data show a dramatic difference in the resistance shift depending on whether a chain is statistically “healthy” or “out of spec” after the chip is first joined. The results have significant implications for the cost-effective assembly of 3D silicon-on-silicon MCM and chip-to-wafer 3D chip-stacks.

Introduction

Within the last five years, the development of manufacturable and robust TSV technology has enabled the development of compact and highly integrated 3D Si-on-Si microsystems [1,2]. The drive to increasingly finer pitch interconnection in such 3D microsystems results in reduced solder volume, and thus an increase in the ratio of IMC to solder [3]. Moreover, even in the case of coarser pitch 3D applications e.g. memory stacks, the drive to ever-thinner silicon-on-silicon stacks demands not only the thinning of the die themselves, but also a substantial reduction in the height of the solder interconnects regardless of their pitch. For illustrative purposes, Fig. 1a) shows how a standard “4-on-8” mil solder bump, commonly used for the full thickness chip attach to organic and ceramic packages, is not well-suited. The assembly process employed herein has been called Advanced Chip-to-Wafer (AC2W) bonding [3] as well as collective hybrid bonding [4]. The concept is to separate the C2W process into two distinct stages: 1) a sequential precision alignment stage followed by 2) a single parallel final bonding stage. In the first stage, all die are pre-aligned and tacked or lightly pre-bonded to hold them in place. This stage may make use of any one of a number of methods of fixing the die in place including flux, pre-applied underfill, a pre-bonding adhesive or ultrasonics [3]. Production pick & place tools and/or flip chip bonders are the tools of choice in stage one. In the second stage, a permanent bonding tool is used to apply uniform heat and pressure in a well controlled atmosphere to complete the process. Production wafer bonders are well suited for use in stage two. The net process thus combines the individual strengths of both C2C bonding (exclusive use of known good die, availability of high-volume production equipment) and W2W bonding (a single parallel thermal bonding operation).

As described above, advanced chip to wafer bonding has many attractive attributes. There are, however, a number of situations where one might rightly ask if the technique is well-suited. One such case might be for the production silicon MCMs where dies and chip stacks of varying heights, bump sizes/volumes and metallurgies must be combined on a single silicon carrier. In this situation it might be necessary to permanently bond die of one type, before placing and bonding die or stacks of a second type, etc. If there is no metallurgical hierarchy, how many reflow operations can the first die sustain, and what, if any, effect might this have on yield and reliability? Another case might be where thin die are repeatedly stacked (e.g. memory) on top of another. In this case, can the “tacking operation” be performed...
multiple times before the final stack is permanently bonded? If so, will the final bonding pressure be uniformly applied to all die in the stack and result in even die-to-die gaps? In bonding very low-profile IMC joining as illustrated in Fig. 1c), a large thermocompression force may be applied across multiple thin 3D die without the risk of bridging [5], but in this case, it is also quite unlikely that standard pick & place tools, flux apply or adhesive tacking techniques could be employed in the alignment phase due to the extremely low volume of joining metal. In such cases, 3D chip stacking through the use of a precision chip cavity mold has previously been shown to be a viable solution [6].

The focus of this paper is thus to investigate collective hybrid bonding of thin 3D Si die to silicon substrates using low-volume “pancake” bumps, and in particular to physically and electrically evaluate the effects of alignment, tacking, parallel bonding and reflow(s) of simple Pb-free solder/Ni interconnects on 3D Si MCM yield.

**Test Vehicle: Design and Fabrication**

The primary test vehicle used in this investigation was designed specifically for rapid TSV/interconnect yield evaluation of a very large array of interconnects, and has been described in detail previously [2,5]. Illustrated in Fig. 2, a 49 mm x 50 mm silicon substrate was designed to accept up to 24 thin 3D die, where each die comprises a 36 x 60 array of TSVs at 200 μm pitch. Thus, each die contains 2,160 interconnects and the fully populated MCM totals 51,840. The 3D die, which are 7.4 mm x 12.2 mm, include 2 μm Cu wiring links on the topside, and simple 100 μm diameter Cu/Ni/Au joining pads on the bottomside.

![Figure 2. Si MCM test vehicle: up to 24 thin 3D die can be joined for a total test coverage of 51,840 interconnects.](image)

The silicon substrate includes 2 μm thick alternating Cu wiring links and the low-profile Cu/Ni/Pb-free solder pancake bumps to form complete chains at module assembly. With the exception of the corners (denoted as sites 1, 6, 19 and 24 in Fig. 2) each joined die contributes 10 interleaved yield chains, each chain having 204 links. The corner sites contain 8 long chains, two 144-link chains and a number of shorter chains ranging from 4 to 20 links for calibration and debugging. The substrate includes a large collection of 4 pt. probe taps and wiring traces leading to vertically-aligned sets of 2x12 probe-card pads on the left and right edges. Automated 4 pt. step & repeat testing of the 420 probe-card pads along each edge of the substrate allows for a total 232 of long chains (204 links), 96 single TSV/bump sites and multiple substrate wiring calibration structures to be probed in about 30 minutes.

Fig. 3 shows the fabrication sequence used to build the 3D chips. First, fields of annular TSVs (gap = 4 μm, diameter = 50 μm and pitch = 200 μm) are etched into the silicon substrate using Bosch-type deep reactive ion etching (D-RIE). Insulation is formed on the sidewalls by thermal oxidation, after which tungsten is deposited by chemical vapor deposition (CVD). The TSVs are completed by chemical mechanical polishing (CMP). Single damascene Cu pads and wiring links are formed on top of the TSVs, after which a top layer of PECVD silicon nitride is used to passivate the surface. The 3D chip wafers are then bonded to glass handlers for mechanical support during wafer grinding.

![Figure 3. Thin 3D chip fabrication sequence which includes temporary wafer bonding to a perforated glass handler using the TOK Zero Newton system.](image)
In this work, we have used the Zero Newton wafer bonding system from TOK. It has previously been shown that this system, used in combination with TZNR-A2002 adhesive (also from TOK), allows for subsequent thin wafer processing at temperatures up to 280ºC with no decomposition during vacuum steps, while maintaining its ability to dissolve during the solvent debonding step at the very end of the process [7]. Following wafer bonding and grinding to a thickness of ~80 µm, the bottoms of the tungsten-filled vias are exposed by Si RIE. Backside dielectric is deposited using plasma enhanced chemical vapor deposition (PECVD), and a final CMP step exposes the TSV metal cores. Terminal metal pads consisting of 1.0 µm Cu, 1.0 µm Ni and 0.1 µm Au were evaporated through an aligned shadow mask to complete the 3D chip fabrication sequence. The wafers were then attached to standard dicing tape frames, and placed into the complementary TOK Zero Newton debonder/cleaner tools, where the adhesive was dissolved and the perforated glass handlers removed. The system leaves the ~70 µm thin 3D die on tape, clean and ready for dicing, pick and place.

The same standard Cu BEOL processing used to fabricate the 3D chips was used to create the silicon substrates, except that substrate wafers contained no TSVs and remained full thickness. Vias, 85 µm in diameter, were etched through the top PECVD passivation dielectric and landed on the Cu link pads below. In the final stage of substrate fabrication, a thin Ti/Cu seed metal layer was sputtered and patterned with 120 µm diameter openings in 20 µm resist for solder plating. A UBM of 2.0 µm of Cu / 2.0 µm of Ni and ~18 µm of Pb-free SnAg 1.7% solder were sequentially electroplated, after which the resist was stripped and the seed metal etched. Upon reflow, the pancake bumps assumed a partially curved shape and ready for dicing, pick and place.

Bond and Assembly

A SET FC150 precision flip chip bonder was used in both stages of the collective hybrid bonding work. The FC150 bonding head accepts a variety of different sized "tools" designed for very small chip areas of only a few square millimeters, up to a flat plate large enough to simultaneously cover all 24 thin die. One of the first tasks was to determine the applied bonding force per 3D die required to achieve the desired interconnect height target of ~20 µm. Table 1 shows the bump height results as a function of bonding pressure.

Table 1. Final “pancake” solder bump heights as a function of bonding pressure.

<table>
<thead>
<tr>
<th>3D chip sample</th>
<th>Bonding force (grams)</th>
<th>Final bump height (microns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>2000</td>
<td>14</td>
</tr>
<tr>
<td>#2</td>
<td>100</td>
<td>28</td>
</tr>
<tr>
<td>#3</td>
<td>200</td>
<td>21</td>
</tr>
<tr>
<td>#4</td>
<td>200</td>
<td>20</td>
</tr>
</tbody>
</table>

With a baseline process established to join 3D die to Si MCM substrates with the desired low-profile bump height, the next tasks were to establish the electrical yield of the process, and to investigate the effect of batching in the AC2W process i.e. to investigate how the bonding of additional die to a partially populated MCM effects the yield and resistance of the previously bonded die. An initial batch size of 4 die was chosen. The first 4 die were aligned and tacked with flux (in chip sites 1, 2, 7 and 8) and then bonded in parallel according to the optimal conditions established previously. The module was electrically tested, and then another 4 die were aligned and tacked (in sites 13, 14, 19 and 20) after which all 8 die underwent the bonding operation. Figure 5 shows three different optical micrograph views of the left side of the partially populated module. Figure 5a) shows a top-down view of the corners of two die and the 200 µm spacing between them. The chips are delineated by deep-etched crackstops which are responsible for the precisely cut edges and 45 degree bevel at the corners. The chip topside wiring links are visible, as well as the substrate wiring which runs along the horizontal gap between the chips connecting the yield chains to probe pads which are just visible along the
left side of the image. Fig. 5b) shows a similar view but at a slightly tilted angle, making it easier to see the corners of the two top die in relief, and the dense substrate wiring off to the left. Fig. 5c) shows a highly tilted view at still higher magnification. In this image two low-profile bumps along the edges of the two die can now be resolved. The highly reflective chip edges and the bumps themselves are mirrored in the silicon substrate, but it is clear that the bump height is considerably less than the ~70 μm die thickness.

Figure 5. Optical micrographs showing the corners of two thin 3D die bonded to a Si MCM substrate. Images were taken a) top-down, b) at a slight tilt angle, and c) at a steep tilt angle and high magnification to show the joins.

The electrical chain yield results for the first partially populated module are summarized in Table 2. Since it takes only one open connection (bump, TSV or wiring strap) to knock out an entire 204-link chain, assuming open defects are randomly distributed, the interconnect continuity yield for both batches of joined die is actually close to 99.9%. Note that the continuous chains for both batches have been classified as “healthy” and “out of spec”. Figure 6 best illustrates the difference between these two chain classifications. Here the yield chain resistance distribution for the first batch of four die is shown a) after they were first joined, and b) again after the next batch of four die were joined in adjacent positions. During the joining of the second batch, the first batch Pb-free pancake bumps were reflowed in place. Fig. 6 shows that healthy chains are clustered fairly tightly around the peak of the distribution, in most cases within +/- 2%, and they essentially shift resistance in lockstep during the second reflow operation. Note that the distribution shifts fairly uniformly to the right, with a peak at 46.2 Ω before and a peak at 46.8 Ω afterwards. Essentially, any chain originally in the range from 45 - 48 Ω increased in resistance by 0.6 Ω after the joining of the second batch.

Table 2. Summary of chain yield for the partially populated 3D Si MCM. A first batch of 4 thin die were bonded, followed by a second batch of 4 die. All chains contain 204 interconnect links.

<table>
<thead>
<tr>
<th>Batch</th>
<th>Total continuous chains</th>
<th>Interconnect yield (%)</th>
<th>“Healthy” chains</th>
<th>“Out of spec” chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>First 4 die</td>
<td>30 / 38</td>
<td>99.88</td>
<td>21 / 38</td>
<td>9 / 38</td>
</tr>
<tr>
<td>First 4 die (after next batch joined)</td>
<td>30 / 38</td>
<td>99.88</td>
<td>21 / 38</td>
<td>9 / 38</td>
</tr>
<tr>
<td>Last 4 die</td>
<td>30 / 40</td>
<td>99.86</td>
<td>15 / 40</td>
<td>15 / 40</td>
</tr>
</tbody>
</table>

Figure 6. Resistance distribution of 204-link yield chains for the first batch of 4 thin 3D die a) after they were initially joined, and b) after another batch of 4 die were aligned and bonded adjacent to the first batch.
Figure 6 shows the 21 (out of 38) chains from the first batch considered to be healthy, and it also shows 1 of the 9 continuous but “out-of-spec” chains (circled in red). This chain shifted a total of 1.4 Ω, a value which is more than double the shift for the healthy chains. In fact, the average difference between the healthy and out-of-spec chains is much more dramatic as can be seen in Table 3. Here one can see that the average shift of the 9 out-of-spec chains was 4.9 Ω compared to the uniform 0.6 Ω of the healthy distribution. This suggests that the unhealthy chains are dominated by a multiplicity of resistive defects, whereas the healthy population appears to be controlled by some small, but uniformly evolving component of resistance.

Table 3. Summary of average chain resistance and single bump values for the partially populated 3D Si MCM. All chains contain 204 interconnect links. Each batch of 4 die also contains 6 individual single bump sites.

<table>
<thead>
<tr>
<th>Batch</th>
<th>Mean “healthy” chain resistance (Ω)</th>
<th>Mean “out of spec” chain resistance (Ω)</th>
<th>Average of single bump sites (mΩ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First 4 die</td>
<td>46.5</td>
<td>150.4</td>
<td>16.3</td>
</tr>
<tr>
<td>First 4 die (after next batch joined)</td>
<td>47.1</td>
<td>155.3</td>
<td>17.7</td>
</tr>
<tr>
<td>Last 4 die</td>
<td>46.4</td>
<td>121.8</td>
<td>16.2</td>
</tr>
</tbody>
</table>

Table 3 gives an indication as to what the shifting component might be. Dispersed among each batch of 4 die, are six “single bump” measurement sites. In fact, each of these sites accurately measures the 4 pt. resistance of one single TSV plus one pancake bump including the UBM. Previous investigations with these same structures yielded values of ~15 mΩ for the tungsten TSV and ~5 mΩ for a full-sized 100 μm diameter eutectic C4 solder ball (including the UBM) [2]. The pancake joins of this study are a factor of four shorter so one would expect a corresponding reduction in the solder component of the resistance. However, as noted above, the reduction in solder volume enhances the net relative contribution of any IMC formed during reflow, since the ratio of IMC to solder has increased by the same factor of four. Table 3 clearly shows that the extra reflow experienced by the first batch of 4 die during the joining of the second batch causes a shift in both long chain and individual bump sites. Since both the Cu wiring and the tungsten TSVs are demonstrably robust against such resistive shifts at Pb-free reflow temperatures, IMC growth would appear to be the likely cause.

In order to verify the hypothesis, the partial module was taken through an additional 2 solder reflow cycles, and the yield chains and single bump sites were once again electrically measured. In parallel, a partial module containing 4 die joined sequentially (in four separate bonding operations) was analyzed by SEM cross-section in order to estimate the amount of IMC grown on the substrate UBM and the 3D chip Ni/Au pad as a function of the number of reflows. Since the substrates are reflowed once before the first chip is joined, IMC growth begins at the solder/UBM even before the first die is joined. The results are shown in Table 4. The healthy chain population remained constant throughout the multiple reflow experiment, as did the trend of slowly increasing chain resistance and single bump resistance. The IMC thicknesses measured on both sides of the bump are averaged estimates; the IMC/solder interface is a collection of phases made up of Sn, Ni and Cu, and not of uniform thickness (as can be seen in Fig. 4) but rather highly scalloped in its formation. An average thickness was calculated by making measurements at multiple locations across the width of the pad of each SEM image. Each reflow increases the total proportion of IMC in the interconnect, from less than 20% of the join height after the first reflow, to about 35% after four reflows. The fact that most of the common IMCs have resistivities which are substantially higher [8] than Pb-free solder (e.g. Ni₃Sn₄, Cu₆Sn₅) explains the slow increase in net resistance.

Table 4. Summary of average healthy chain and single bump resistance as well as IMC growth for a batch of 4 joined die as a function of the number reflow cycles.

<table>
<thead>
<tr>
<th>Solder reflows (die / substrate)</th>
<th>Mean “healthy” chain resistance (Ω)</th>
<th>Average of single bump sites (mΩ)</th>
<th>IMC thickness (microns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 / 2</td>
<td>46.5</td>
<td>16.3</td>
<td>1.2 / 2.5</td>
</tr>
<tr>
<td>2 / 3</td>
<td>47.1</td>
<td>17.7</td>
<td>1.5 / 3.5</td>
</tr>
<tr>
<td>4 / 5</td>
<td>47.7</td>
<td>19.0</td>
<td>2.0 / 5.0</td>
</tr>
</tbody>
</table>

The overall goal of this work was not to determine the limiting number of chip reflows, the maximum number of die to be placed, nor the maximum acceptable chain yield or resistance shift; all of these depend strongly on the UBM materials chosen, the solder composition and volume, the AC2W tooling and process parameters, and ultimately on the 3D application itself. The goal was to define a reasonable starting process space in which to test the applicability of collective hybrid bonding in the fabrication of useful low-profile 3D multichip Si-on-Si assemblies. Towards that end, a collection of 24 thin 3D die, randomly selected from across the source wafer, were joined in two batches of 12, to create a fully populated 24-die Si-on-Si MCM. An electrical output map showing the yield and resistance of each 204-link chain is shown in Figure 7. Out of a total of 232 chains, 34 were open and 198 were continuous for an interconnect continuity yield of 99.92%. The continuous chains were composed of 177 healthy chains and 21 out-of-spec chains. In figure 7, the healthy chains are shown in two shades of green, indicating values within +2% and -2% of the mean value of 46.2 Ω, while the out-of-spec chains are shaded yellow, and full opens are shaded red. The four extreme corners of the MCM, shaded in grey, each contain two 144-link chains, which are all continuous and whose values are also shown, but which are not included in the long chain interconnect statistics.

It is worth noting that each individual die is represented by a 2x5 block of cells in Figure 7, so there are two die (both...
adjacent to the upper right corner die) which are electrically dead. Rather than indicating a problem with the bonding, this highly correlated pattern suggests a more pervasive defect in the chip itself, most likely an open connection between the TSV tungsten core and the backside pad metal. Such a defect may occur near the very edges of the 3D die wafer due to non-uniformities during RIE and CMP steps. The 3D die used in this study were not tested prior to bonding due to the time limitations involved in probing such a large number of linked TSV pairs.

Potential Future Work

Future work will include investigations on the rework of individual die within an MCM, as well as standard reliability assessments of such assemblies. Further, the applicability of the AC2W method for producing multi-layered 3D chip stacks will be examined. Research may include thermal measurements which would provide quantitative data on the enhanced thermal conductivity benefits of this approach due to the large area low-resistance thermal path provided by these metallic connections. Moreover, while this study utilized a single diameter “pancake” interconnection throughout, future work may target mixed diameter I/O interconnection at each layer in the die stack. In this scenario, larger diameter joins could support power/ground interconnections with excellent vertical thermal conductivity, and small diameter connections could lower the electrical parasitics which are known to limit vertical signal transmission at very high frequencies.

Acknowledgments

The authors would like to acknowledge support from the Microelectronic Research Laboratory at IBM T.J. Watson Research Center; D. Dimilis and R. Trzeciński of Central Scientific Services; J. Rosa, F. Dicesare, M. Jones and J. Advocate of the IBM S&TG analytical group in East Fishkill, NY; Steve Wright and Bob Polastre of the 3D Test Group at T.J. Watson Research Center; K. Nakada of TOK Japan; and Chris Rosenthal of TOK America. We would also like to acknowledge management support and encouragement from R. Wisnieff, G. Shahidi and T.C. Chen.

References


