1. Introduction

Mobile computing has evolved beyond PC computing capabilities and can carry out tasks ranging from office productivity and communication to HD media and gaming. There are three broad trends in computing over the past few years that have a significant impact on the processor-memory architecture and its implementation. They are: 1) The transition from single core to multi-core CPUs, which dramatically increased the need for multiple memory channels, the channel bandwidth and the amount of memory that is accessible to each core; 2) Low power computing that emphasizes physically short interconnects along with a wide bus at lower speeds for high bandwidth, and 3) Cloud computing, which benefits from having both hardware and software optimized and centrally managed for power and usage.

The most critical feature to keep increasing the performance is the processor-memory interconnect. Figure 1 shows that the CPU and memory cycle time gap is increasing, which means that it takes far longer to get data to the processor than the time taken to use it.[1] This problem is typically addressed by optimizing across various memory hierarchies. As shown in Fig. 2, knowing that storage is not suitable due to its very high latency, processor-DRAM subsystem receives the most attention for improvement.

Another factor is the importance of low power computing, which has led to an explosion in mobile platforms such as phones and tablets. This has direct implication to processor-memory subsystem since approximately 50% of the memory power is used to drive the IO between the

---

**Abstract**

3D processor-memory packages potentially offer very high performance due to short interconnects between the two chips. Current Package-on-Package (PoP) technology offers less than 300 interconnects between the processor and memory. To meet future bandwidth requirements of greater than 25.8 GB/s bandwidth at low power, wide IO memory in x512 configuration is expected. This memory requires more than 1,000 interconnects and current PoP technologies do not scale to meet these requirements. To address this problem, a new PoP technology called Bond Via Array (BVA) PoP is presented that offers very fine pitch (0.24 mm and lower) and high height/diameter aspect ratio (8:1 and higher).

This is achieved by forming free-standing wire-bonds along the periphery of the processor chip and encapsulating the package leaving miniature posts projecting from the top of the package to be connected to the memory package. More than 1,000 interconnects can be formed within the same footprint as current packages. The BVA PoP process development, assembly and reliability test results are presented. The assembly and all reliability tests including Moisture Sensitivity Level (MSL) testing, on-board temperature cycling, high temperature storage, and drop tests were successfully completed. These results demonstrate that the BVA PoP is ready for high volume manufacturing.

**Keywords:** Bond Via Array (BVA), Fine-pitch Package-on-Package (PoP), Free Standing Wire-bond Interconnects, Wide IO Memory, High Bandwidth 3D Package, Film Assist Molding, Palladium Coated Copper Wires

---

**Fig. 1** Processor and memory cycle trends.
two. Figure 3 shows typical power efficiency values for different types of memory IO.[2]

The physical layout of process-memory subsystem has evolved over the past few decades as shown in Fig. 4. Currently the memory is in the form of Dual Inline Memory Modules (DIMMs) for desktops and servers, and many tablets have multiple memory packages placed next to the processor. The phones have PoP where the memory is on top of the processor. The current PoP modules have limited IO (32-64) and hence a TSV solution is proposed to meet very high IO (128-512) requirements. Since the TSV technology is not mature yet, a PoP based on conventional processes would be very appealing.

An order of magnitude increase in the processor-memory IO in a PoP form would make it a universal candidate, from mobile to high computing. For example, a 32-bit wide memory of current PoP found in today’s mobile devices offers 6.4 GB/s bandwidth at 1,600 MegaTranfers/second (800 MHz DDR). Operating memory at high frequency uses lot of power, which is not desirable for mobile devices. If 512-bit wide memory is used, even at 800 MegaTransfers/second (400 MHz DDR), a 51.2 GB/s bandwidth can be achieved. Hence it is seen that by using a wide data path while using slow low power memory, high bandwidth can be achieved. Bond Via Array™ (BVA™) Package-on-Package (PoP) offers ultra-high bandwidth between multi-core CPU-GPU SoC processors and wide IO low power memory chips utilizing conventional wire-bond technology and existing materials and infrastructure.

2. Bond Via Array (BVA) Package-on-Package (PoP)

Figures 5–6 illustrate the Bond Via Array (BVA) wire-bond array interconnect concept. The main feature is that the BVA interconnects (free-standing wire-bonds) extend from the bottom substrate to the top surface of the bottom package to be connected to a package mounted on top.

The mature wire-bonding technology offers very fine pitch, and free-standing wires are formed using proprietary processes utilizing conventional wire-bond equipment. As the wire-bonds can be done at a pitch as small as 50 μm, and can extend in length to any desired value, high aspect ratio (height to diameter ratio greater than 10) interconnects can be achieved. This interconnect technology lends itself to a wide variety of 3D packaging, including PoP, wafer level, embedded, etc. Preliminary and exploratory work done for BVA technology has already been reported.[3, 4]

The interconnect scaling capabilities are shown in Fig. 7. For a given 14 mm × 14 mm package and assuming a 1 mm peripheral width for IO, up to 1,440 interconnects can be formed at 0.2 mm pitch. These numbers of IO are enough to meet future wide IO memory requirements. Here, a 14 mm × 14 mm package size is chosen because it is the most common size for PoP. An IO area width of 1 mm is assumed to match the IO width of current solder ball stack PoP, which has only 2 rows in this width.
Figure 8 shows that BVA offers the smallest pitch and highest IO compared to BGA (Ball Grid Array) PoP and Through Mold Via (TMV) PoP. TSV also offers fine pitch and high IO, but it is not a mature technology yet.

3. Manufacturing Process

A 432 IO BVA PoP daisy-chain test vehicle was designed and fabricated that measured 14 mm × 14 mm with two perimeter rows of Palladium coated Copper wires at 0.24 mm pitch with a wire diameter of 50 μm and a height of 0.4 mm, as shown in Fig. 9. This test vehicle has an interconnect aspect ratio (height/diameter) of 8 and pitch ratio (height/pitch) of 1.7, which is better than any existing PoP technology.

The process flow for the BVA PoP is shown in Fig. 10. The top (memory) package is similar to current memory packages, including high IO BGA. For example, four memory chips with each being x32 can be packaged to form a x128 BGA package. With higher IO, wide IO memory can
also be used. The bottom package has the logic device in conventional flip-chip format, with the BVA wires around the periphery. The molding with wires is also different. Finally, stacking is done using conventional Surface Mount Technology (SMT) with the condition that the top memory package has fine pitch BGA. The four unique process steps are explained below.

4. Forming the BVA Copper Interconnect

The free-standing wire-bonds are the most important feature of BVA PoP. Forming the wire-bonds with the tips having good positional accuracy (x and y) and uniform height (z) are important in enabling very fine pitch and high yield package assembly. Figure 11 shows bottom package substrate with the flip-chip attached logic chip and BVA around the periphery. These free standing wire-bonds were done by first bonding to the copper pad, which is a very conventional process, and then by a proprietary cutting mechanism. This cutting mechanism makes a wedge-like wire tip and it may be slightly bent depending on the cutting direction. The wedge shape and its bending angle are artifacts of cutting process and may vary for different wires for the same wire bonding process.

The x, y and z positional accuracy data is shown in Fig. 12. The results given in these graphs depict data from 43 packages with each data point on the graph representing all the wire-bonds in one package. The positional accuracy for all directions was found to be within ± 15 μm (3σ). The wires were also bonded well, and an average of greater than 0.15 N was obtained from bond pull test. Figure 13 shows the wires as bonded. The bonding speed was about 10 wires per second using industry standard wire-bonders with similar tools and tool wear as for conventional wire-
bonding. The only change is wire-bonding software that was developed based on the wire-bonding process development.

5. Molding and Exposing the Interconnect
The next step is to mold the logic package while exposing the BVA tips with a consistent desired height. Film assisted mold technique as shown in Fig. 14 was used to expose the tips. It is a mature technology commonly found in many packaging assembly operations and stable supplier support. The process uses a mold chase design with mold cavities only slightly deeper than the formed Cu wires. When the mold is clamped to the substrate, the Cu wires are pushed into the mold film. The mold cavity is filled with the molding compound, the molding compound is cured, the mold is opened and the mold film is pulled away from the package exposing the wire tips. The mold film thickness determines the wire tip exposure. No special molding parameters were needed to provide repeatable wire tip exposure results. Figure 15 shows the exposed wire tips. The target value was 0.12 mm and this height was obtained within a tolerance of ± 10 μm.

6. Cleaning Wire Tips with Robust Surface Finish
Wet etch processes to clean the tips of the wires were presented published earlier,[5] Since the wet etch equipment is not very common in a traditional packaging assembly line, plasma etch was developed since plasma etch equipment is commonly found. The recipes shown in Table 1 were used to develop the plasma etch process.

The process recipes P1 and P2 did not yield clean wire tips as the mold residue was still visible. But two process recipes, CF₄ + O₂ with enough pressure (659 mT) or CF₄ + O₂ at low pressure (215 mT) with 15 minutes of Ar were successful in cleaning the wire tips. Figure 16 shows the initial wire tips with residue and the cleaned ones after the plasma process. To ensure that the Palladium coating on the wires is still intact after plasma process, EDX analysis was carried out and the results are shown in Fig. 17. It was found that the Palladium coating is largely unharmed and

![Fig. 14 Film assist mold to expose the wire-tips.](image)

![Fig. 15 The BVA interconnects exposed on the top surface of the bottom package.](image)

![Fig. 16 (a) Initial wire tips with mold residue and (b) clean wire tips after processes P3 or P4.](image)

<table>
<thead>
<tr>
<th>Process</th>
<th>Description</th>
<th>Rinse</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO</td>
<td>No plasma</td>
<td>2 min water blast + 2 min air blast</td>
<td>Residue present</td>
</tr>
<tr>
<td>P1</td>
<td>15 min Ar</td>
<td></td>
<td>Residue present</td>
</tr>
<tr>
<td>P2</td>
<td>30 min CF₄+O₂ @ 28.7 Pa</td>
<td>Residue removed</td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>30 min CF₄+O₂ @ 87.8 Pa</td>
<td>Residue removed</td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>30 min CF₄+O₂ @ 28.7 Pa + 15 min Ar</td>
<td>Residue removed</td>
<td></td>
</tr>
</tbody>
</table>
is present even after the plasma process. Note that the presence of Si is due to the silica particles in the mold material that surrounds the wire bonds.

7. Connecting Top Package to Exposed Wire Tips

The last step is to stack the memory package on top of the logic package, as shown in Fig. 18. This process is very similar to conventional PoP assembly where solder paste is printed on the main board, the logic package is placed on the board, the memory package is dipped in solder flux and placed on top of logic package, and the stack is refloved along with other components on the board. The solder was lead free with 96.5% Sn, 0.5% Cu and 3% Cu composition. An underfill was dispensed between the packages of the stack for high reliability under dynamic loading. Good joints were obtained across all interconnects over the whole package, as shown in Fig. 19. The quality of the joints was confirmed through electrical continuity test for the whole package. The tip of the wires has a wedge-like appearance and its orientation depends on the direction of the wire cutting mechanism. These shapes are artifacts of wire cutting process and have no impact on performance or reliability.

8. Reliability

A full suite of reliability tests were conducted and the results are summarized in Table 2. In situ monitoring of the daisy-chain resistance was done for all tests and no failures were detected. The drop test was continued to 128 drops and no failures were observed. This is to be expected as the underfill was present in the PoP.

Extensive failure analysis was carried out for the samples under test and no failures were detected. The 3D X-Ray images are shown in Fig. 20.

To study the effect of copper-tin diffusion between the
copper wire and the solder ball, accelerated testing was done by evaluating the amount of intermetallic formation under high temperature conditions. The accelerated testing was done through by subjecting the samples to 150°C for a period of 1,000 hours. To inhibit diffusion and intermetallic formation, the copper wires were coated with Palladium. Figure 21 shows that BVA interconnect after assembly and before the high temperature storage test. Figure 22 shows the BVA interconnect after the high temperature test. The Copper-Tin intermetallic compounds have grown from approximately 3 μm in thickness before the test to approximately 7.5 μm in thickness after the test.

This growth is within the range observed in typical Nickel barrier coated copper-tin interfaces such as solder balls on Nickel coated copper pads. Hence, the Palladium coating has served its role as barrier between Copper and Tin and has largely inhibited the growth of intermetallic compounds. If Palladium is not present, extensive intermetallic

### Table 2 Reliability test results.

<table>
<thead>
<tr>
<th>Test</th>
<th>Standard</th>
<th>Test condition</th>
<th>Number of samples tested</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moisture sensitivity Level 3</td>
<td>IPC/JEDEC-J-STD-020C</td>
<td>125°C for 24 hours; 30°C /60%RH for 192 hrs, 3X Pb-free reflow</td>
<td>22 logic and 22 memory packages</td>
<td>Pass</td>
</tr>
<tr>
<td>High temperature storage</td>
<td>JESD22-A103D-condition B</td>
<td>150°C, 1,000 hours</td>
<td>22 PoP off-board</td>
<td>Pass</td>
</tr>
<tr>
<td>Unbiased autoclave</td>
<td>JESD22-A102D-condition D</td>
<td>121°C/100%RH/2atm for 168 hours</td>
<td>22 PoP off-board</td>
<td>Pass</td>
</tr>
<tr>
<td>Drop test</td>
<td>JESD22-B111</td>
<td>&gt;30 drops, 1,500 G, 0.5 msec of half sine pulse</td>
<td>20 PoP on board with underfill</td>
<td>Pass (no failures till 128 drops)</td>
</tr>
<tr>
<td>Temperature cycling (board level)</td>
<td>JESD22-A104D Condition G</td>
<td>–40°C to 125°C, 1,000 cycles</td>
<td>45 PoP on board with underfill</td>
<td>Pass</td>
</tr>
</tbody>
</table>
growth is seen with the copper almost completely disappearing thereby impacting the reliability of the interconnect.[5]

9. Conclusions
Mobile computing requirements driven by multi-core, low power and cloud computing trends have placed a high premium for high processor-memory bandwidth through a very large number of interconnects with short physical length. To meet this challenge, a novel interconnect is presented that is based on wire-bonds and can be utilized in different applications such as 3D packaging, embedded packaging, wafer level packaging, etc. This technology offers >1,000 logic to memory interconnects at 0.24 mm or lower pitch in a PoP form to enable high bandwidth and reliability at low cost using conventional PoP processes. By enabling 1,000 interconnects, up to 512 bit wide memory can be used to obtain very high bandwidths even while operating memory at slow speeds for low power.

The assembly results show that free-standing wire-bonds can be formed at fine pitch with very good consistency having 3-directional positional accuracy within ±15 μm. Plasma cleaning process was developed to obtain clean wire-tips after molding. The stacking process was done with very high yields at 0.24 mm pitch. Full suite of reliability tests was done and there were no failures, including minimal intermetallic growth under high temperature storage test. These results demonstrate that the BVA PoP is ready for implementation at a high volume manufacturing facility.

References

Ilyas Mohammed
Hiroaki Sato
Yukio Hashimoto