Integration Through Wafer-Level Packaging Approach

by

Kai Liu, Bernard Adams, and SeungWook Yoon
STATS ChipPAC
Tempe, Arizona, USA

Originally published in the 2015 International Wafer Level Packaging Conference (IWLPC) Proceedings, San Jose, California, Oct 13 – 14, 2015.

INTEGRATION THROUGH WAFER-LEVEL PACKAGING APPROACH

Kai Liu, Bernard Adams, and SeungWook Yoon
STATS ChipPAC
Tempe, Arizona, USA
kai.liu@statschippac.com

ABSTRACT
System-in-Package (SiP) concept has been adopted from a system-on-board practice when board-space is a concern. This is very applicable for mobile and wearable applications. Moving many devices/components into a small area would reduce the interconnection lengths between components/devices, and therefore reduce the overall size. But it does not automatically yield good results, as it may increase EMI (electromagnetic interference) between components/devices. To have short cycle-times for package development, strong electrical skills are required for SiP designs. For wafer-level SiP, as finer design rules and thinner dielectric are available for designs, further size-reduction is greatly expected. But EMI or cross-talk could be much more severe. The performance verification is needed to be more on EMI or cross-talk rather than on interconnection RCL parasitic, after individual components/chips are made as known-good-dies electrically. In an embedded Wafer Level Ball Grid Array (eWLB), as there are no bumps (or Cu-column) at the die side and the dielectric substrate is much thinner than laminate substrate, its thermal performance is typically better than incumbent fc-FBGA, due to short thermal path. Successfully-made eWLB modules can achieve super electrical and thermal performance, where thin profiles and small form factors are required.

In this paper, we describe the advantages of using eWLB for system-level integration through several examples, such as, 2D integration with multiple-dies, and 3D integration for high-speed and high-frequency. Area reductions and performance trade-off, along with other key features, are discussed in these examples.

Key words: WLP, fan-out, integration, SiP.

INTRODUCTION OF FAN-OUT PACKAGING

Wafer-level packaging has been widely used for semiconductor devices. It has some unique features compared to conventional packaging approaches such as lead-frame and substrate based packages. The packaging processes use a lithographical approach in batch-mode, and the tolerance or electrical yield is typically better than what other packaging approaches can provide.

Silicon nodes have advanced consistently as predicted by Moore's law in order to provide more functionality for devices in higher speeds. This requires packaging technologies to match the higher and higher I/O density on chips. Lead-frame and substrate-based packaging technologies have been successfully used in almost every package type. But they also are facing more and more challenges to meet the higher I/O density requirements. Silicon-based packaging solutions, such as TSV, from foundries using advanced process machines can provide seamless and shortest interconnection for electronic devices, however, their high-cost processes are still the main barriers before these technologies become mainstream for packaging applications.

Wafer-level packaging is somewhere between a foundry-based silicon approach and substrate-based approach. It has some similarities with either approach, to some extent. Batch-modes and lithographical steps, which are inherent in foundry processes, are also implemented in wafer-level packaging. A dedicated wafer-level packaging line decoupled from an IC manufacture line would have many advantages in terms of cost and cycle-time.

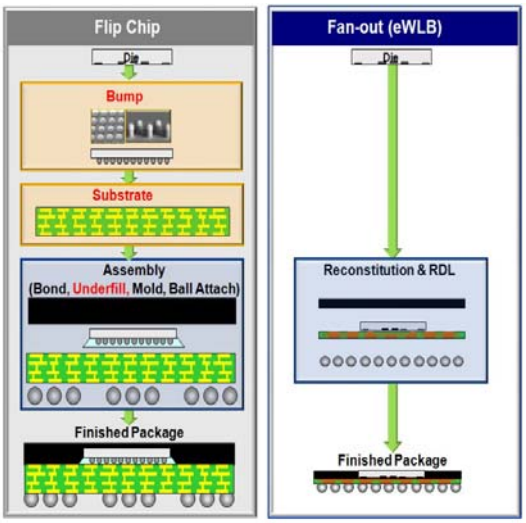

Figure1. Comparison of flip chip packaging and fan-out eWLB packaging.

On the other hand, a fan-out process (using eWLB as example) is a simplified fcFBGA package to some extent. As can be seen in Figure 1, the bumping process for fine pitch in fc-process is eliminated. The connection between die-pads and RDL are directly made through via formation in the wafer process. The laminate substrate for fc-process is also removed; as wafer-level RDL processes itself add its own dielectric layers and metals layers. The big difference,

however, is on the molding sequence. In fc-process, the molding is made after chips are assembled (chip-last) on substrate. But in eWLB process, the molding is done when chips are not connected to packaging circuitry (chip-first).

The wafer-level fan-out (eWLB) process flow is depicted in Figure 2. As is evident from Figure 2, the process steps are similar to the more familiar WLCSP (fan-in technology) manufacturing process with the exception of *reconstitution* which is unique to eWLB. To provide a better understanding of reconstitution, the process flow is further broken down into its four sub-steps comprising

(i) application of an adhesive layer to a carrier; (ii) accurate face-down placement of the die on to the carrier; (iii) encapsulation of the die using a proprietary compression molding process; and (iv) removal of the carrier, resulting in a self-standing *reconstituted wafer* with the active surface of the die exposed. Subsequent wafer-level RDL processes are the same as for fan-in WLCSP.

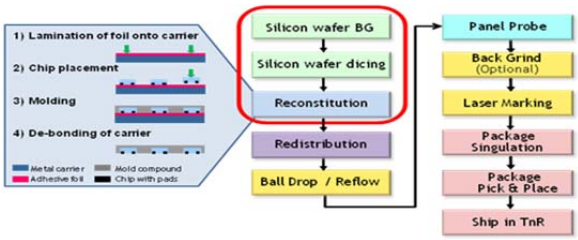

Figure 2. Process flow for fan-in/fan-out wafer level packaging. For fan-in approach, the enclosed portion in red is just replaced by incoming wafers.

For fan-out packaging, back-grinding of the incoming wafers at certain thickness and dicing are required before picking and placing (PnP) the dies on to the carrier.

SYSTEM-IN-PACKAGE

An electronic system is made of individual components either for active or passive function blocks. Historically each of such function blocks was made in one package or one component. Putting these blocks together in a system board was the natural and only way to perform system-level functionalities. This methodology was straightforward and proved to be efficient. As long as individual packages are known good packages, this method could yield fast cycle-time to market.

However, as there is more and more functionality required in modern electronic devices, this approach reaches its limit. More board space for accommodating more individual packages is required, which is against consumers' demands on light, compact, small form-factor requirements, especially for mobile, wearable and IoT applications.

As one of the alternatives, SoC (System on Chip) concept has come into place. SoC contains more system functions in one single-chip. The package for such single chip approaches does reduce the board area significantly. Although SoC has found its applications in certain areas, there are still some insurmountable obstacles for a SoC to contain functions covering the entire RF, analog, digital domains.

First, passive components do not follow Moore's law, which means that they typically cannot be further integrated on a chip through finer and finer silicon nodes, or it is just not a cost-effective approach to do so. Secondly, the best overall performance (both electrical and cost) from different functions needs circuitry of individual functions to be implemented in different silicon (or other types of wafer substrate) nodes. For example, high-speed memory and ASIC are typically made in lower than 40nm node, while logic and RF are mainly made in not less than 65nm node, to be effective. Third, in RF or wireless systems, the portion for RF (e.g., TRX transceiver, LNA) has strong EMI to the analog and digital circuits if these circuits are implemented in a SoC. Shielding in SoC level is not a solution yet. Therefore, these RF functions are still made as standalone packages, on which shielding solutions are available. Fourth, implementing many functions in SoC makes the chip area large, which always reduces the yield proportionally, especially for advanced silicon nodes (<28nm). In order to overcome this, the concept of 'splitting die' or silicon partition has been used for high-speed memory, where several individual dies are integrated through packaging to achieve the same function of a SoC.

SiP approach is something between board-level solution and SoC solution. It can achieve much smaller system size than board-level solution. As it contains chips which are optimized with respective silicon-node technologies, its performance and cost can be optimized as well.

Using laminate substrate for SiP is a natural transition from board-level implementation to package-level integration. Most of PCB board materials are the same or similar to the ones used as substrate materials in SiP. Therefore, transferring from a board-level design to a substrate design usually does not encounter big challenges, and it would not alter the system's performance very much if same dies are presented in both options. However, as in SiP, the interconnections are much shorter with less parasitic, electrical performance can be better through optimal design.

Wafer-level integration has come into place in the last few years for some specific applications. Since multiple components/dies are to be implemented in one package, it requires fan-out technology to do this. Although it has not been used for mainstream electronic devices, it shows some unique features that other packaging technologies cannot easily come up with. In the following subsections, we will introduce some eWLB integration examples for mobile and RF applications.

(1) eWLB with Side-by-Side Dies

The first example is a RF power amplifier with integrated passive device. A RF power amplifier typically requires matching, transforming, harmonic filtering function blocks,

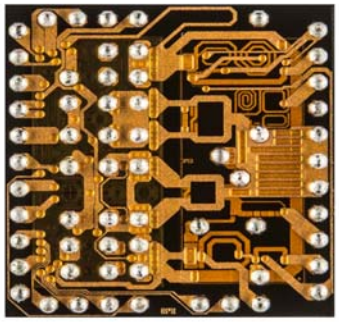

Figure 3. eWLB with two dies. One is power amplifier die and another is IPD die.

which are mainly passive devices. These devices can be formed or made by discrete SMD components in ceramic forms. In a traditional SiP approach, amplifier die is either in wire-bondable format or fc-format. Inductors and capacitors are made in SMD format. Assembly process includes pick-and-place (PnP) of these SMD components on a laminate substrate along with die-attach of the amplifier die. The clearance between these discrete components in SiP design often ends up with large substrate area. PnP of multiple SMD components also reduces throughput of a SiP assembly.

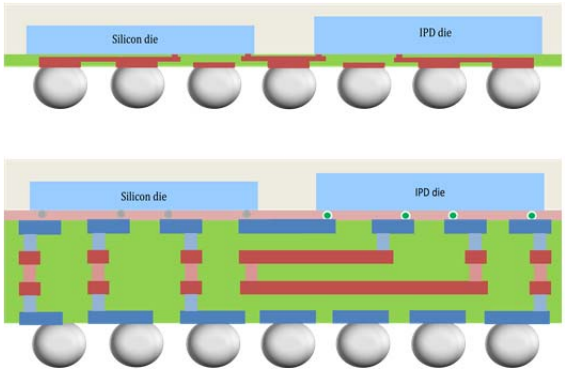

Figure 4. Illustration of eWLB approach (top) and fcBGA approach (bottom) for SiP.

One way to improve the throughput and reduce the substrate area is to use integrated passive devices (IPD). Most of RLC components for RF matching, transforming, and filtering can be made in IPD effectively. The clearance distance between RCL components is reduced about 10 times. As much finer design rules are available for Silicon-based IPD than for discrete components made of ceramic substrate, the intrinsic RCL components in IPD are typically smaller. As a result, to achieve same function, an IPD size can be 2-3 times smaller than the area used to implement multiple SMD parts on substrate.

After most of SMD functions have been realized by an IPD, both amplifier die and the IPD die can be designed into laminate substrate to form a SiP as illustrated in Figure 4. Sometimes large value capacitors, usually for DC decoupling purposes, are needed in a SiP module. Making these capacitors into IPD is not cost or area effective. But, they can be simply treated as a standalone SMD in the SiP.

Comparing the cross-section illustration for eWLB and fcFBGA approaches for a SiP, one can see there are a few major differences. First, in eWLB approach, there are no bumps needed at the die side, and no-underfill process is required either. Secondly, eWLB is made in a wafer lithographic process which can provide finer feature dimensions. Typically, eWLB design rules are 2-3 times finer, both for line width/line spacing and for via sizes. As a result, the same I/O routing can end up with a 2-3 time area reduction in eWLB. Translating this into layer count, routing for a 6-layer laminate design may be converted to 2-layer eWLB design. Through an actual project as shown in Figure 3, a 1-RDL eWLB SiP was realized in a 4-layer laminate SiP. Thirdly, the metal thickness and dielectric thickness in eWLB are smaller. For 1-RDL stack, the total thickness is around 20um (8um RDL and 12um PSV), while in laminate it is more than 60um (20um and 40um

Figure 5. Exposed-die eWLB with two dies.

dielectric). The smaller metal cross-section does not necessarily mean its metal0-loss will be higher. Since the RDL length can be routed much shorter in a smaller X-Y size and smaller Z height, its DC resistance can still be equivalent or smaller.

As shorter interconnections are presented in eWLB, its electrical performance should be better. For applications that need lower profiles, eWLB approach appears to be superior. To further reduce the package height, an exposed-die version for eWLB is available which can be as thin as 0.31mm including ball height.

For the same package, warpage (measured in um) is inversely proportionate to the package height. To investigate warpage behavior of thin eWLB, an about 6.0 x 7.0 mm eWLB was backgrinded to around 200um including ball height. The eWLB body thickness is around 120um. For such a very thin body size, the maximum warpage is 90um (Figure 6).

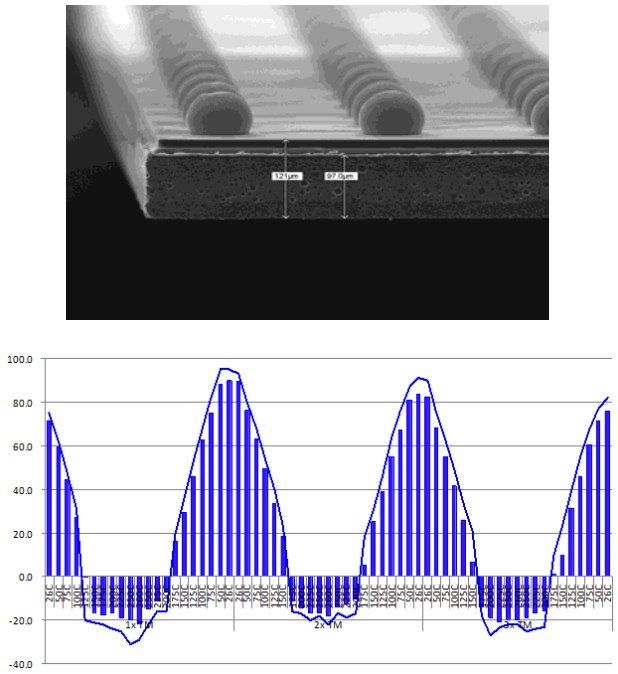

Figure 6. Warpage performance of a very thin eWLB (about 6.0 x 7.0 x 0.2 mm including ball height of 80um).

The exposed-die version is better for thermal dissipation. If mechanical requirements (e.g., warpage) or reliability requirements (e.g., drop test) are more important, a sheet of organic material can be backside-coated in a thickness of around 25um.

(2) eWLB Used in 3D Configuration for High-Speed Package

To reduce footprint of a SiP package, Z-direction asset can be used. An easy partition is to have all ICs in one package and all SMD components in another package. Figure 7 shows an example of this. The top package is a laminate substrate interposer, which contains several discrete components. The vertical interconnections are made available in the periphery of the bottom eWLB package through PCB bars. The bottom package can be a high-speed ASIC. The PCB bars are treated as IC during PnP process for reconstitution wafer, and can be as thin as 200um.

This partition arrangement helps to increase the throughput from assembly point of view. However, for electrical performance reasons, it may not be optimal. For example, decoupling capacitors are typically required to be implemented near the power nets, which the decoupling capacitors are assigned to in circuit schematic. Long distance of routing to the power-nets may not achieve desired decoupling performance. Since the vertical interconnects are made in the periphery of the bottom eWLB package, the best placement for a decoupling capacitor is around the periphery of the top interposer.

Electrical characterization was completed on a 14x14 mm test vehicle. The PCB bar height was 200um. The maximum RDL routing length is less than 5mm in the eWLB package. From measurement, the largest parasitic capacitance for RDL traces was less than 1pF, and the largest parasitic inductance for RDL traces was less than 2.5nH.

The top package may not be only limited to laminate substrate interposer. A pre-made package, like a memory package, can be the top package and be integrated with a bottom package (such as a high-speed processor), as long as the top package ball-arrangement is made along the periphery.

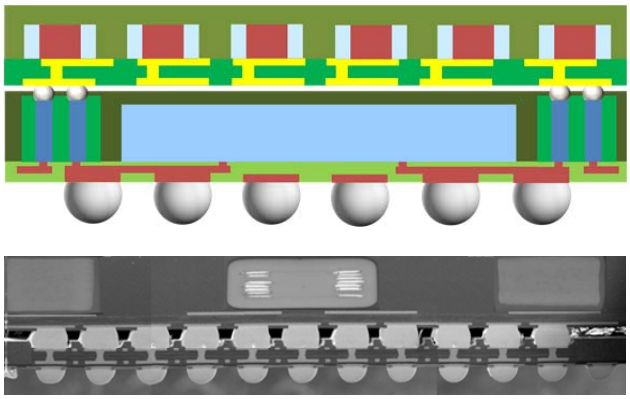

Figure 7. Top figure: illustration of PoP package with interposer as top package and eWLB as bottom package. Bottom figure: SEM picture showing discrete SMD components in PoPt package which is assembled onto an eWLB with PCB bars for vertical interconnections.

If the top package has restriction for balls arranged in periphery, and has to make balls in anywhere, a double-side RDL process for the bottom eWLB can be used to meet this PoP interconnection requirement. In next example, we will describe the double-side RDL application in more details.

(3) Antenna Module Using eWLB for Very-high Frequency Applications

Antenna for a wireless communication is typically in the size of ½ wavelength at its operation frequency. The higher the operation frequency is, the smaller the antenna is. An antenna used for cellular application is bulky, large in size, and it is mainly implemented on phone board. For certain higher frequency applications, such as 60 GHz WLAN, and 77 GHz automotive radar, the antenna's feature-size can be less than 10mm, which may therefore be realized in a package.

Making such antenna in a package, as a standalone component, does not fundamentally change its effectiveness, as most of antennas in applications are already in component (standalone) forms. One benefit from a packaged antenna is that now it can be implemented with

other devices (packages) for further integration and performance improvement.

At such a high frequency (>60GHz), substrate-loss and metal-loss for interconnection (usually designed as transmission lines) is much higher than in cellular bands. Reducing interconnection length is a key to retain good performance.

Wire-bonding packages have been widely and cost-effectively used for cellular and tablet applications, and there may be still some room before this technology reaches its limits. But for 60 GHz and above applications, wire-bonding is nearly prohibited. There are a few reasons for this. First, at such high frequency, the wires may show large reactance or may be in resonating stage, which is not desired for antenna application. Secondly, the wires (seen as inductors) would talk to other ICs or components in its proximity, which creates a lot of EMI noise. Thirdly, the tolerance of wire-bonding (e.g., less than 50um in wire-length), would be magnified to some point (due to very high frequency), where severe deterioration of electrical yield could occur. In other words, a tighter tolerance control in assembly and manufacture is needed for such high-frequency packages, where wafer-level packaging is a better candidate.

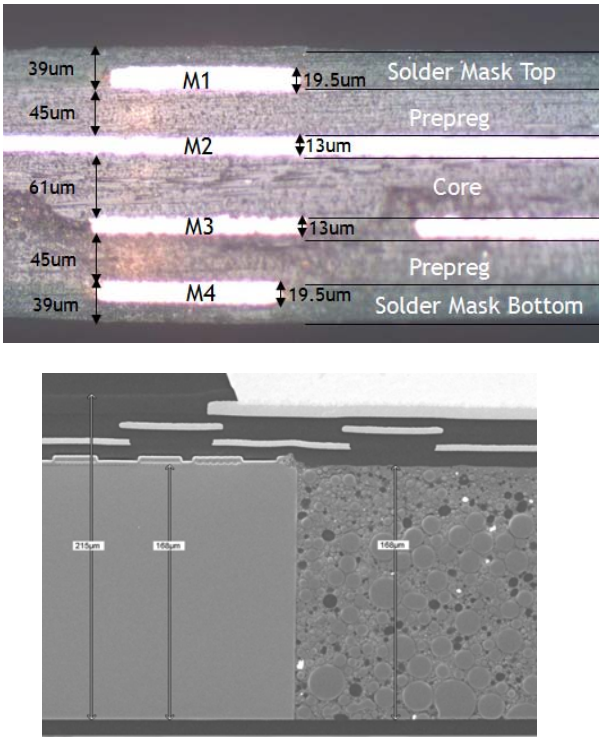

Figure 8. Surface roughness comparison between eWLB (bottom) and laminate package (top), where the surface from wafer-level processing is much smoother.

Surface-roughness effect plays a significant role in metal loss at very high frequency (e.g., >30 GHz). Compared to smooth metal surface, a rough surface can result in double insertion-loss at >30 GHz range. A typical metal surface in laminate and in eWLB is depicted in Figure 8. In laminate substrate technology, the surface is on purpose made in rough condition for better adhesion between BT material and metal traces, but it will yield much higher insertion-loss at larger than 30 GHz high-frequency applications. For typical laminate substrates, Cu surface roughness, in terms of RMS (root-mean-square), can be up to 3um, while Cu metal surface roughness from a wafer-process is typically less than 0.3um. At 60 GHz, typically-made RDL thickness from a wafer-level process (>4um) is much larger than the skin-depth, therefore, having thicker metal (like 20um in laminate) does not gain much. But smoother metal surface from wafer-level processing can reduce transmission loss from metal-trace by about a half.

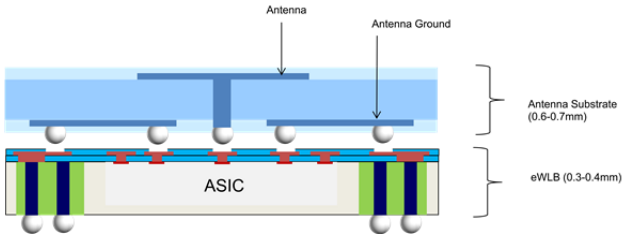

Figure 9. Antenna package stacked on to an eWLB transceiver package for larger than 60 GHz application.

Figure 9 shows PoP stack for an antenna-in-package approach, where the top package is a pre-made antenna structure working at larger than 60 GHz, and the bottom package is an eWLB for transceiver. The top antenna-package can be made in other substrates (like ceramic, glass, or low-loss laminate). Sold-ball finish or OSP finish on the top-package can be used for subsequent stacking on the bottom eWLB package.

In this approach, the interconnection between the transceiver and the antenna is the shortest, only through direct vertical connection. The antenna ground plane in the top package also serves a shielding structure between the antenna and the transceiver.

A double-side RDL process has been developed which can eliminate the PoP stacking process. As can be seen from Figure 10, RDL layers are made in each side of the eWLB. The vertical interconnections are made available through the PCB bars. Depending on low-current or high-current application, the PCB-bars can be provided in either shallow or solid via format.

The two RDL layers on one side facilitates routing for complex design and can provide some passive functions (like eWLB inductors). As the mold material used in eWLB has very low loss-tangent (~0.004), two Cu-layers process

can make inductors having Q-peak from 45-70 depending on inductance value from 2 nH to 10 nH.

However, this layer-stack is not enough for antenna application mentioned earlier. Antenna design has some specific requirement on its antenna metal-pattern to its ground-plane. The separation between two RDL layers in eWLB is about 8um-12um. For antenna application over 60 GHz, the separation between its antenna metal-pattern and its ground-plane is in sub-millimeter range. Obviously, eWLB cannot provide such antenna structure in one side of the package.

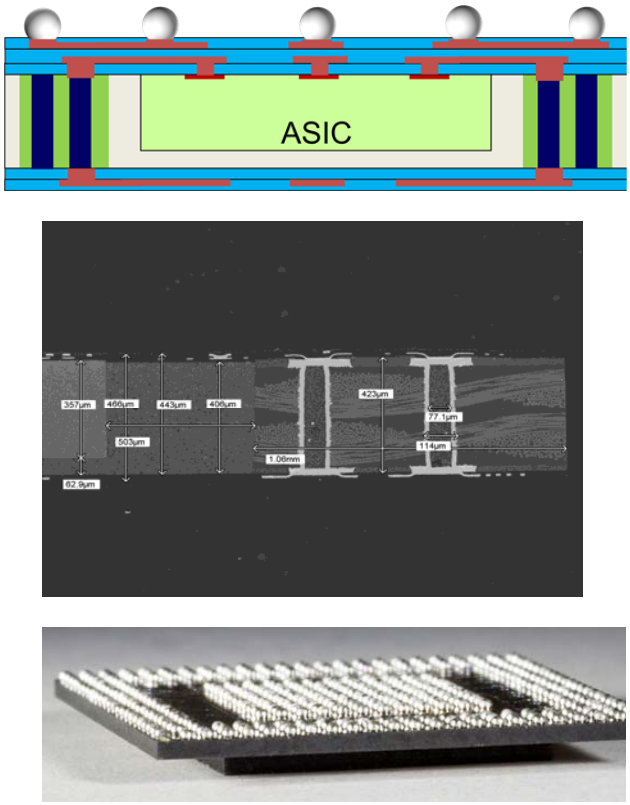

Figure 10. Top: concept illustration of double-side eWLB. Middle: SEM picture of a double-side eWLB with PCB-bars. Bottom: a double-side eWLB with a smaller package stacked onto it (upside down).

The needed separation between antenna metal pattern and its ground plane can be provided by the PCB. As with the incumbent eWLB technology, a total package height less than 1.0 mm can be conventionally manufactured, antenna-in-package integrating a transceiver chip can be realized for larger than 30 GHz application, using this approach.

Figure 11 shows a concept of such high-integration antenna package for larger than 60 GHz applications. In this approach, the antenna ground-plane is made in the top-side RDL. All the connections to the transceiver are made through the top-side RDL as well, with solder-balls for the second-level interconnection. Antenna metal-pattern itself is made in the bottom RDL. The connection to the bottom RDL is through the PCB bars in the periphery.

The PCB-bars serve two major purposes. First, the PCB-bars provide specific height (antenna to ground-plane distance, in sub-millimeter range) required for proper operation of such antenna. Secondly, they provide shielding between antenna and transceiver (ASIC). This is a critical requirement when a radiation element (antenna) and RF transceiver are placed in close proximity.

A similar implementation is to have the antenna made in the top RDL, and to make its ground plane on the bottom RDL. Solder-balls for second-level interconnections are also made on the bottom RDL to facilitate the radiation upward (not shown in Figure 11).

Figure 11. Concept illustration of double-side eWLB with one RDL in each side. Antenna pattern is made in bottom RDL and antenna ground is made in top RDL.

CONCLUSIONS

Compared to other substrate-based packaging solutions, eWLB technology enables smaller form-factors and better integrations, through finer routing capabilities achieved from lithographic steps. The much tighter process-tolerance and the smoother metal-surface from wafer-level packaging are especially suitable for very-high frequency and very-high speed applications in terms of electrical yield. Standard eWLB process and material set can routinely make double-side RDL with total height in sub-millimeter range, which paves the way for eWLB used for antenna-in-package for larger than 60 GHz applications.