2.5D FPGA-HBM Integration Challenges

Jaspreet Gandhi, Boon Ang, Tom Lee, Henley Liu, Myongseob Kim, Ho Hyung Lee, Gamal Refai-Ahmed, Hong Shi, Suresh Ramalingam
Xilinx Inc.
2100 Logic Dr., San Jose, CA 95124 USA
Ph: 408-879-3645; Fax: 408-371-0541
Email: jgandhi@xilinx.com

Abstract
FPGA partitioning and high density integration using interposer based 2.5D stacked silicon interconnect technology (SSIT) has been the pioneering work at Xilinx for several years enabling advanced applications in high performance computing, networking, hyper scale data center and cloud services etc. With the insatiable demand for acceleration workloads, FPGAs need to be coupled with in package memory to enable higher bandwidth, lower power and smaller form factor architecture. 3D stacking based high bandwidth memory (HBM) has paved the way to realize such applications providing 10X higher bandwidth, 4X lower power vs DDR4. This paper provides an overview of unique challenges involved with 2.5D FPGA-HBM SSIT integration from design, process/package development, test and reliability point of view

Key words
2.5D, FPGA, HBM, machine learning, packaging, SSIT, TSV

I. Introduction
There is little doubt that this is a new era for FPGAs (Field Programmable Gate Array). While it is not news that FPGAs have been deployed in many different environments, particularly on the storage and networking side, there are fresh use cases emerging in part due to much larger datacenter trends. Energy efficiency, scalability, and the ability to handle vast volumes of streaming data are more important now than ever before. At a time when traditional CPUs are facing a future where Moore's Law is less certain and other accelerators and custom ASICs are potential solutions with their own sets of expenses and hurdles, FPGAs are being deployed as CPU accelerators for an ever-growing range of workloads in machine learning, search engine indexing, encryption, and data compression providing high performance/watt [1]. These parallel data processing and compute-intensive applications are relying on high density FPGA architecture integrated with in package memory to provide faster speed, higher bandwidth at lower power.

Xilinx already has a leap forward in new and high growth markets and applications such as datacenter, storage and networking creating high-capacity FPGA devices with second generation 3D SSIT Fig.1. This technology enables multiple super-logic regions (SLRs) to be combined on a passive interposer layer, using proven manufacturing and assembly techniques from industry leaders, to create a single device with several thousand low-power inter-SLR connections. Dedicated interface tiles within the SLRs provide ultra-high bandwidth, low latency connectivity to other SLRs.

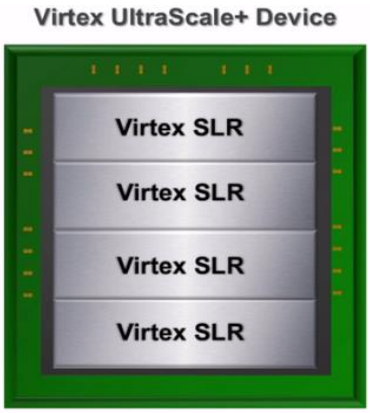

Figure 1: Virtex Ultra Scale+ Device using 3D SSIT

On the DRAM side, thanks to the advent of High Bandwidth Memory (HBM), the memory wall has been broken to achieve multi-Tbps memory bandwidth at significantly lower power. HBM is a new type of memory integration technology that vertically stacks

memory chips via TSVs (thru silicon vias) providing low power consumption, ultra wide communication lanes, faster speed and smaller form factor. Xilinx recently announced HBM-enabled 16nm UltraScale+ FPGAs Fig. 2 built using 3rd generation CoWoS technology co-developed by TSMC and Xilinx and now the industry standard assembly for HBM integration. These heterogeneously integrated packages can deliver 10X bandwidth per HBM stack, 4X lower power vs. DDR4.

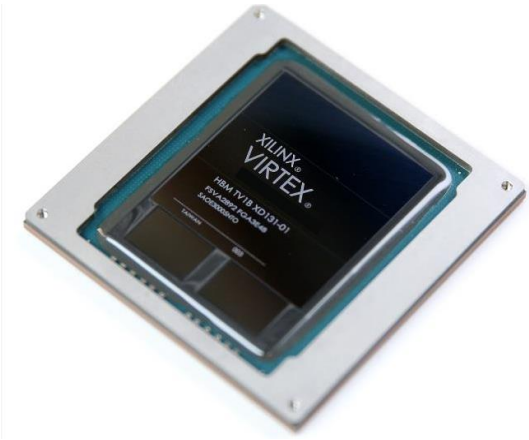

Figure 2: Xilinx Ultrascale+ FPGA-HBM Integration

Several publications recently have highlighted heterogeneous integration approaches for HBM in SiP applications. AMD's 2.5D Si interposer approach talks about the importance of stiffener ring materials, adhesive, thickness and substrate core type to improve package co-planarity [2]. Cisco/e-Silicon [3] used organic interposer for ASIC and memory integration. The low loss dielectric material used for the organic interposer allows ultra-fine line spacing (Line/Spacing = 6µm/6µm), low transmission loss and high insulation reliability. Intel's Multi-Die Interconnect Bridge (EMIB) [4] is an interposer less approach with claims to have no practical limit to die size and can leverage existing organic substrate manufacturing. Regardless of the pros and cons of each approach, heterogeneous integration of HBM with ASIC, GPU, CPU, FPGA is real and progressing very fast. To author's knowledge Xilinx is the first company to attempt HBM integration with partitioned FPGAs in 2.5D format. This paper provides an overview of unique challenges involved with 2.5D FPGA-HBM SSIT integration from design, process/package development, test and reliability point of view.

II. Interposer Design

The integration of HBM cubes and FPGA dice into a SSIT package begins with the design of the silicon interposer. Physical layout considerations and electrical requirements are of paramount importance. The first physical consideration is the placement of the FPGA dice and HBM cubes on the interposer. This is determined by several factors such as underfill dispensing, thermo-mechanical stress on interposer, interposer size and inter-die signal integrity. There is an optimal die-to-die distance to achieve a uniform and void-free underfill during package assembly which further means the gap has to be small (for single underfill dispense) or large (for multi-pass dispense). As for the stress on interposer, it can be influenced by the amount of open space left on the interposer after the dice are placed. The underfill or mold compound occupying this open space will impact interposer warpage. If the HBM cube dimensions differ from one memory supplier to another, a different interposer may be necessary to satisfy the FPGA to HBM die gap requirements

Another physical consideration is the ubump pitch on FPGA PHY and HBM PHY. While there is an industry specification (JEDEC) for HBM ubump pitch, the ubump pitch on FPGA is vendor specific. For ease of interposer routing, the ubump pitches across these 2 dice should match in such a way that an integer number of inter-die signal lines can be routed in a uniform fashion between a pair of ubumps. Such uniform routing is also desired from signal timing point of view. Since the signals between FPGA PHY and HBM PHY are operating at Gbps level, the layout of these interconnect routings in the interposer has to be done carefully so as not to compromise the signal integrity. To optimize these routings, their length has to be kept as short as possible and with low resistance. This usually means placing FPGA PHY and HBM PHY side-by-side facing each other and with a small die gap as discussed earlier. Besides this, allocating sufficient metal routing layers in the interposer will help to achieve optimal routing. In addition to minimizing routing length and resistance, careful shielding of the high speed signal lines is required to minimize electrical cross-talk. Meticulous post-layout simulations of these inter-die signals must be carried out to confirm the electrical specifications. HBM cube comes with a set of direct access (DA) ports which have to be routed to BGA balls. As the name imply, they offer a direct electrical access to the HBM memory and are used by memory vendors to debug any HBM issues (RMA). The design of these ports are vendor-specific and some may operate at high frequencies. Therefore, attention should be given when routing these DA signals through the interposer to BGA terminals. A straight forward routing will be to use a set of stacked vias in the interposer to connect the DA ubumps to the BGAs.

III. Package Design & Process

Packaging solutions for 2.5D FPGA SSIT devices have been developed and demonstrated in Xilinx's last two generations product (28nm and 20nm). Thermal-mechanical simulation & test vehicle evaluation have provided greater insight into package design with optimized thermal interface material, lid structure, lid adhesive, substrate structure and core/build-up material for robust reliability performance in both component and board levels [6]

With 2.5D FPGA-HBM integration in the current 16nm product, additional design and process requirements have presented new packaging challenges. The major changes are (i) C4 bump structure moving from eutectic solder bump to Cu pillar bump (CPB) with Pb-free solder (ii) lid type moving from regular copper forged lid to stainless steel stiffener ring. CPB has advantages for fine pitch interconnect, bump reliability, and pkg. thermal performance, however, high Tg underfill is required which often increases package stress and reliability concern due to increased die-to-package interaction (DPI). Typical failure modes include C4 underfill delamination, FPGA die gap delamination, FPGA ubump voiding, etc. Continued process optimization such as proper underfill material selection (for both ubump and C4 bump), C4 underfill curing, interposer dicing, etc. can help to improve DPI performance. Stainless-steel based stiffener ring is required for FPGA-HBM SSIT device due to more stringent thermal budget. The implementation of stiffener ring not only has cost advantages, it also eliminates the thermal interface resistance between lid and die and provides a direct thermal path to the heat sink or other effective cooling solutions.

The combination of CPB and stiffener ring has an adverse impact on package coplanarity also. Efforts are made to not only reduce package coplanarity at room temperature, also reduce package warpage at high temperature (240-260C) in order for system house to properly attach the SiP component on the board. Using a thicker and lower CTE core material for the substrate results in major improvement for package coplanarity, however may have adverse effect on BGA ball board level reliability due to increased CTE mismatch between package substrate and system board material. More studies are currently underway to evaluate this concern. A proper stiffener ring design along with an adequate adhesive material is required to improve overall package coplanarity. In general, thicker ring with wider foot will help to increase the rigidity of the package and therefore reduce its warpage, however, the thickness of the ring may impact heat sink assembly, while the ring foot width will affect the keep out area between ring and chip capacitors. All these factors needs to be incorporated into overall package design

Precise control of bare die parallelism and flatness in the FPGA-HBM package is also required to enable adequate heat sink (or other cooling solutions) attachment and minimize second-level thermal interface resistance. In the current industrial practice, assembly houses often only measure and control lid tilt and package coplanarity as part of the assembly specifications. Parallelism and flatness (either on lid for lidded package or on bare die for stiffener ring package) are mostly being ignored. Efforts are needed with assembly houses to prioritize this measurement and incorporate in their key performance indicators (KPI) to meet product specifications

Challenges for Package Coplanarity

The stiffener ring scheme can reduce package coplanarity (COP), but it is not enough for FPGA-HBM integrated package. The following approaches are used to reduce coplanarity more; lower CTE substrate, stiffener ring type and size etc. Estimated coplanarity results are presented with various parameters. The coplanarity value is predicted by the simulated package warpage result. Figure 3 is the lidless package image used for the simulation. There are 3 FPGA dies and 2 HBM dies on the interposer. The simulated package coplanarity results are presented in Table1. Wider ring width or thicker ring is better to reduce the coplanarity. Partitioned FPGA die type (Figure 4) is better for lower coplanarity than a big single FPGA die as shown in Table 2.

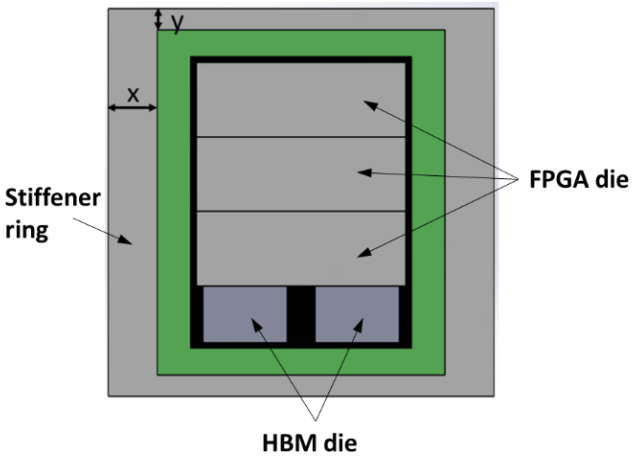

Fig 3: Sample image of FPGA-HBM integrated package. X dimension is the wider side and Y the narrow side of stiffener ring

Table 1: Simulated coplanarity results by stiffener ring width and thickness

Ring thickness (Z, mm)	Ring thickness A- 0.2mm	Ring thickness A	Ring thickness A+ 0.2mm
COP (mil)	12.4	11.5	11.1
Ring width (X, mm)	Ring width A- 1mm	Ring width A	Ring width A+ 1mm
COP (mil)	12.5	12.1	11.5

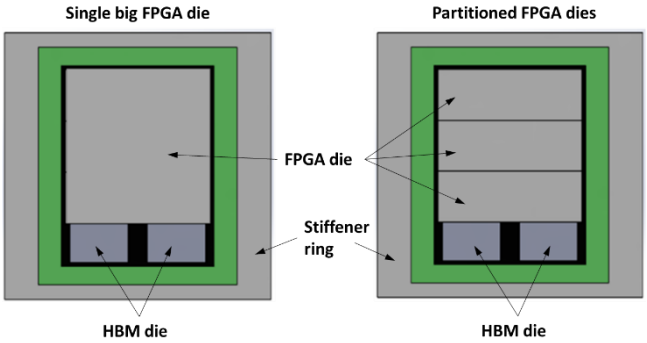

Fig 4. Simple image of FPGA-HBM integrated package by FPGA die types

Table 2. Simulated coplanarity results by FPGA die type

FPGA die type	Single die	Partitioned dies
COP (mil)	11.9	11.5

Coplanarity can be also changed by substrate core CTE. Table 3 is the comparison table with substrate core CTE. Substrate with lower CTE core shows lower coplanarity. This result is aligned with measured data (Figure 5) where lower CTE substrate core B shows 1.1mil ~ 1.3mil lower coplanarity than the higher CTE core A.

Table 3: Simulated coplanarity result by substrate core type

SBT core	Higher CTE	Lower CTE
COP (mil)	12.1	11.5

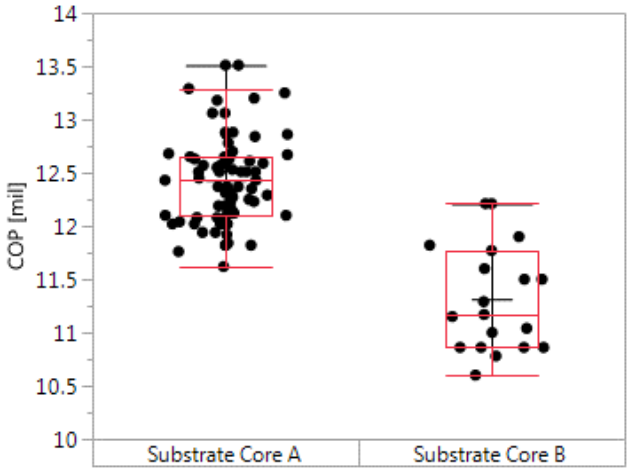

Fig.5 Coplanarity comparison by substrate core

The coplanarity can also be reduced by ring type. Figure 6 shows the measured coplanarity data with different ring patterns. Pattern A is a general flat pattern while pattern B is specially designed ring pattern to enhance adhesion. Package with pattern B reduces coplanarity by 0.5mil ~ 0.7mil. Maximum coplanarity can be reduced to 10.8mil

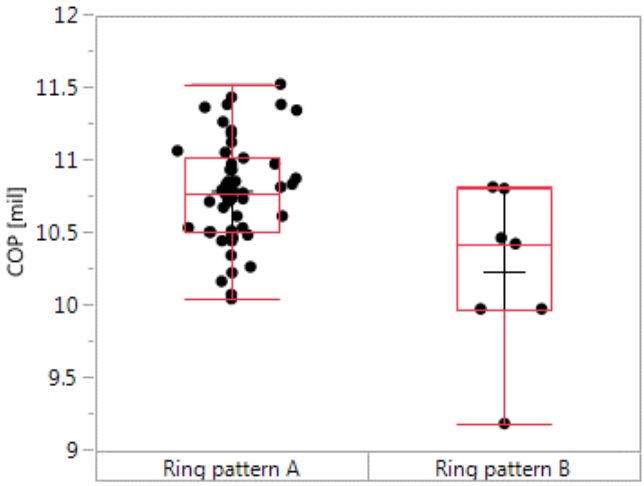

Fig. 6 Coplanarity comparison by stiffener ring pattern

Challenges for C4 joining

One of the major process challenges for 2.5D HBM is C4 joining. With HBM and the empty area around HBM on the interposer, the CoW or CoC die with HBM and FPGA combined has higher warpage than the die with multiple FPGA dies without HBM. This higher die warpage in turn results in either C4 cold joint or C4 bridging. Compared to the other 2.5D HBM-SoC integrations reported so far, Xilinx 2.5D HBM-FPGA integration has the unique challenge that FPGA slice may cover 2 corners of a super-large interposer (Figure 1) with tighter C4 pitch than HBM. Limited number of C4 are required for HBM DA (direct access) ports and power/ground pins connected directly to substrate BGA pins. Thus, the C4 pitch under HBM can be more relaxed. On the other hand, FPGA slice may have tighter C4 pitch at corners. In addition to the C4 pitch challenge, another different factor is the die warpage behavior. FPGA-2 HBM CoW or CoC die has different warpage curvature than a SoC-4 HBM die. A typical 2.5D HBM-SoC integration either has 4 HBM at 4 corners in a super large interposer, or has 2 HBM in a relatively smaller interposer since there is no partition in SoC die.

To address the C4 joint challenge within the confines of CoW or CoC die warpage, the C4 bump and substrate pre-solder size need to be optimized first. A

smaller C4 bump size reduces the risk of C4 bridging, but increases the risk of cold joint, so certainly there is a trade-off here. Reducing CoW or CoC die warpage is another way to prevent C4 joining issues. Table 4 shows the difference in die warpage due to different ubump underfill material. Figure 7 shows typical CoW warpage profile at room and reflow temp.

Table 4: Actual CoW die warpage results from two different ubump underfill materials

uBump underfill	UF#1	UF#2
Die warpage (um) at 250C	70	50

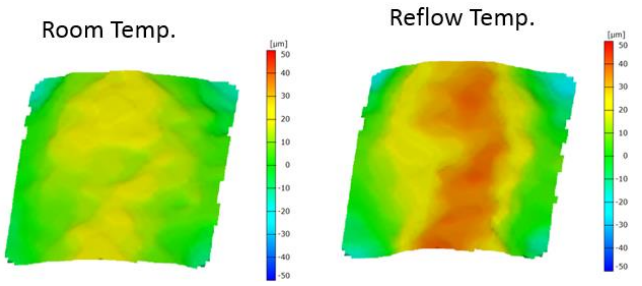

Fig. 7: Typical FPGA-HBM CoW die warpage profile at room and reflow temp.

IV. Test Hardware & Mechanical KPIs

FPGA-HBM package under development even though has similar size as earlier generations, it requires several modifications for the test hardware. Optimum stiffener ring footing is required for the test handler to impart sufficient contact force to the package pins during testing. The tradeoff could be reduced space for decoupling capacitors and higher package stress. Another important consideration as touched upon in the interposer design section above is BGA pin mapping. HBM DA port functionality and assignment vary from vendor to vendor so even though HBM is a JEDEC standard, a different substrate design may be required while switching vendors to ensure DA ports are routed and assigned correctly through interposer connecting to substrate. These considerations must be given a serious thought before releasing a test board design

With a stiffener ring package, die flatness and parallelism are new mechanical KPIs that need to be considered to ensure thermal enhancement. The thickness of TIM (thermal interface material) between component and thermal management solution (heat sink) is one of the critical factors for thermal extraction performance which further depends on package flatness and parallelism. These new KPIs are measured with components before BGA attachment process. Sample is placed on the flat surface and using either optical sensor or mechanical probe, 9 data points of the top die surface profile are obtained. Flatness is defined by subtracting minimum from maximum value among 9 data points and parallelism is defined same way among 5 data points lying on the vertical and horizontal centerlines (data points 2,4,5,6 and 8 in Figure 8 below). These new KPIs also help external customers understand tolerances required for heat sink design.

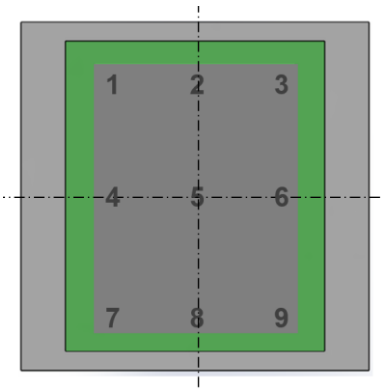

Fig. 8: Schematic for die parallelism and flatness measurement

V. Reliability

FPGA-HBM package was subjected to JEDEC MSL4 pre-conditioning where it passed 1000 hrs. HTS (High Temperature Storage) at 150C and 1000 cycles TCB (Temperature Cycle condition-B) at -55 to125C. ubump and C4 bump shape and profile looked normal after these tests (Figure 9, 10 & 11) with no physical abnormalities in the package. More tests are currently in progress

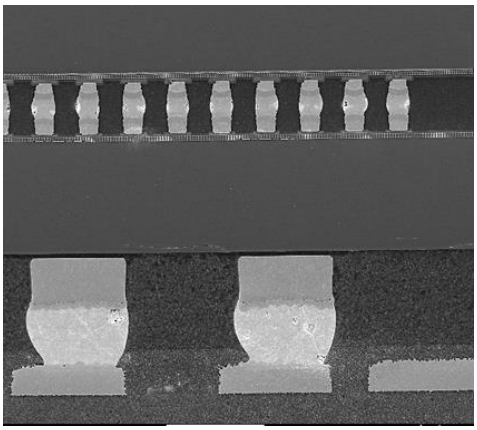

Fig. 9: ubump and C4 bump after HTS 1000 hours

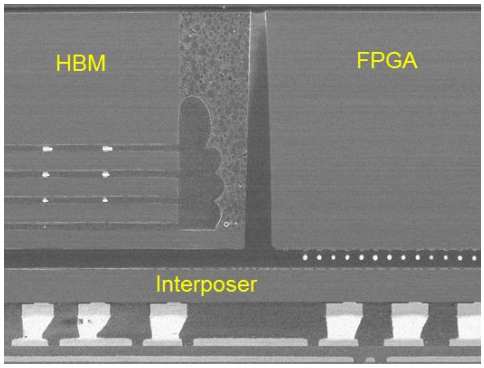

Fig. 10: FPGA-HBM cross-sectional view after HTS 1000 hours.

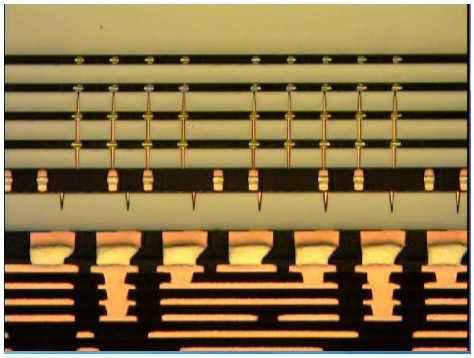

Fig. 11: Cross-sectional view after TCB 1000 cycles.

VI. Conclusion

2.5D FPGA-HBM integration involves unique challenges which demand close knit collaboration between memory vendor, design/process/test and external customers. Several important factors were highlighted to achieve high yield, low coplanarity and robust reliability for 55X55mm2 FPGA-HBM heterogeneous package. Importance of design considerations for HBM swap & new KPIs such as package parallelism and flatness measurement were also underscored for a large size stiffener ring only package

Acknowledgment

The authors would like to acknowledge Xilinx's Global Operations and Quality team for their timely Test, reliability and failure analysis support. Fruitful discussions with die designers and process engineers is also deeply appreciated.

References

[1] Nicole Hemsoth, Timothy Prickett Morgan, "FPGA Frontiers: New Applications in Reconfigurable Computing" 2017. Available: https://www.nextplatform.com/2017/01/16/fpga-frontiers-new-applications-reconfigurable-computing/

[2] Chang-Chi Lee etal, "An Overview of the Development of a GPU with integrated HBM on Silicon Interposer" in *Proc. 66th Electronic Components and Technology Conference*, Las Vegas, 2016, pp. 1439-1444

[3] Li Li etal, "3D SiP with organic interposer for ASIC and Memory Integration" in *Proc. 66th Electronic Components and Technology Conference*, Las Vegas, 2016, pp. 1445-1450

[4] Ravi Mahajan, "Embedded Multi-Die Interconnect Bridge (EMIB)-A High Density High Bandwidth Packaging Interconnect" in *Proc. 66th Electronic Components and Technology Conference*, Las Vegas, 2016, pp. 557-565