2024 IEEE 74th Electronic Components and Technology Conference (ECTC) | 979-8-3503-7598-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ECTC51529.2024.00340

Key Technologies and Design Aspects for Wafer Level Packaging of High Performance Computing Modules

Kai Zoschke
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
kai.zoschke@izm.fraunhofer.de

Hermann Oppermann
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
hermann.oppermann@izm.fraunhofer.de

Michael Schiffer
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
michael.schiffer@izm.fraunhofer.de

Ivan Ndip
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
ivan.ndip@izm.fraunhofer.de

Karl-Friedrich Becker
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
karl-friedrich.becker@izm.fraunhofer.de

Marius Adler
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
marius.adler@izm.fraunhofer.de

Alexander Gäbler
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
alexander.gaebler@izm.fraunhofer.de

Uwe Maaß
Fraunhofer Institute for Reliability and Microintegration IZM
Berlin, Germany
uwe.maass@izm.fraunhofer.de

Gianna Paulin
Integrated Systems Laboratory (IIS) - ETH Zurich
Zurich, Switzerland
pauling@iis.ee.ethz.ch

Walter Kocon
GlobalFoundries
Fishkill, NY USA
walter.kocon@globalfoundries.com

***Abstract*— As contribution to projects like European Processor Initiative (EPI) as well as Stencil- and Tensor Accelerator (STX), Fraunhofer IZM has further developed its advanced packaging portfolio with special focus on wafer level packaging of high performance computing (HPC) modules. This includes the further scaling of the well-established multi-layer copper redistribution technology to enable a 4 µm line / space routing (8 µm pitch) over multiple layers with 6 µm thick polymer interlayer dielectric and micro vias of 8 µm diameter. The redistribution layers (RDL) provide the signal routing on top of a TSV interposer device and related RF simulations show the capability for a very high signal integrity and low transmission loss of this routing scheme to be more performant than inorganic routing schemes based on SiO dielectric and copper or aluminum metallization. The RDL technology is based on semi-additive copper structuring and excimer laser ablation for generation of the micro vias. It is further scalable down to 3 µm line / space (6 µm pitch) and beyond.**

Further key elements of the WLP flow for the fabrication of HPC modules are TSV interposer processing including front side RDL and back side pad formation followed by flip chip assembly, underfilling, compression molding and a final back side balling of the TSV interposer device.

As one example of fabricated HPC modules we present Occamy, which is a 2.5D integrated dual-chiplet system designed by ETH Zürich and supported by the Europractice-IC team at Fraunhofer IIS. The system contains 2 compute chiplets fabricated in GlobalFoundries 12 nm FinFet technology as well as 2 high bandwidth memories 2e (HBM2e). Each compute chiplet has a size of 73 mm^2 and includes six groups of four compute clusters, the host CVA6, an HBM2e controller IP from Rambus, as well as a source synchronous serial DDR die-to-die link. All 4 ICs are mounted on a passive 600 mm^2 silicon interposer called Hedwig which is fabricated in GlobalFoundries 65 nm technology. Related packaging work performed at Fraunhofer IZM was interposer TSV back side reveal, front and back side pad formation, flip chip assembly, underfilling, balling and second level assembly to PCB. Further details will be presented in this paper.

Keywords—high performance computing, wafer level packaging, high density redistribution, silicon interposer, through silicon vias (TSVs), chiplet

2377-5726/24/$31.00 ©2024 IEEE
DOI 10.1109/ECTC51529.2024.00340

I. Introduction

Complex simulations for answering difficult questions in science, engineering as well as business and economy increasingly rely on high performance computing (HPC) capabilities. They have in common, that huge amounts of data need to be processed and stored in reasonable time. Examples are climate or weather modelling, as well as modelling in material science and also artificial intelligence-based approaches like processing of deep learning algorithms for image recognition. [1-6]

Core elements of HPC hardware are efficient and fast compute units which are closely connected to related memories units. Depending on the system architecture, one or more compute units are connected to a similar number of memory units to build a so called HPC module. The final performance of an HPC system is then determined by the global system architecture and how many HPC modules are part of it.

Very common for building of HPC modules is the usage of graphics processing units (GPUs) as compute dies in combination with high bandwidth memories (HBMs). For very application specific calculations optimized stencil processing units are used.

Major challenge for implementation of HPC hardware is the efficient packaging of those systems. Imperative are high speed, low loss and high bandwidth interconnections between the particular compute and memory chips and also their connection to the periphery with high signal and power integrity. Based on this, advanced packaging technologies are incorporated for the build-up of HPC systems [7, 8].

A very promising approach for the packaging of HPC modules is utilization of interposers with through vias. A fine pitch multi-layer routing on top side of the interposer device can provide the necessary electrical interconnection between the compute and memory chips which are attached as flip chips here. The input/output signals to/from the module are fed through the interposer to its back side where it is connected to the system board. Depending on system concept and architecture, the approach is scalable with respect to interposer size and complexity and thus the number of mounted ICs [9-11].

Fraunhofer IZM has further developed its silicon interposer technology to provide an integrated process flow for the fabrication of HPC modules using such a chiplet-based approach. The following chapters give an insight into the general process flow concept as well as design considerations and technology adaptions. Finally, an example of a fabricated HPC device is presented.

II. Technologies and Intended Process Flow

In this chapter, the basic technologies as well as the entire process flow for the fabrication of HPC modules are described and explained. The expectation is, that compute and memory components are available as bare dies or even better as wafers, so that first level advanced packaging technologies can be applied, which is imperative to enable the required interconnection density. The process should be capable for high-volume production and should rely on standard wafer level packaging technologies like 2,5D/3D TSV interposer, redistribution and bumping as well as flip chip assembly, underfilling and molding. The related process flow is shown in Figure 1. In the first step, copper-filled blind holes with SiO_2 sidewall passivation are prepared in blank or even active silicon wafers. The process relies on the well established TSV process technology from Fraunhofer IZM in Berlin. It includes dry etching of vertical holes with high aspect ratio and depth of up to 120 µm, followed by thermal or PECVD oxide deposition as well as copper filling. Detailed descriptions of the TSV process can be found in various publications [12-14].

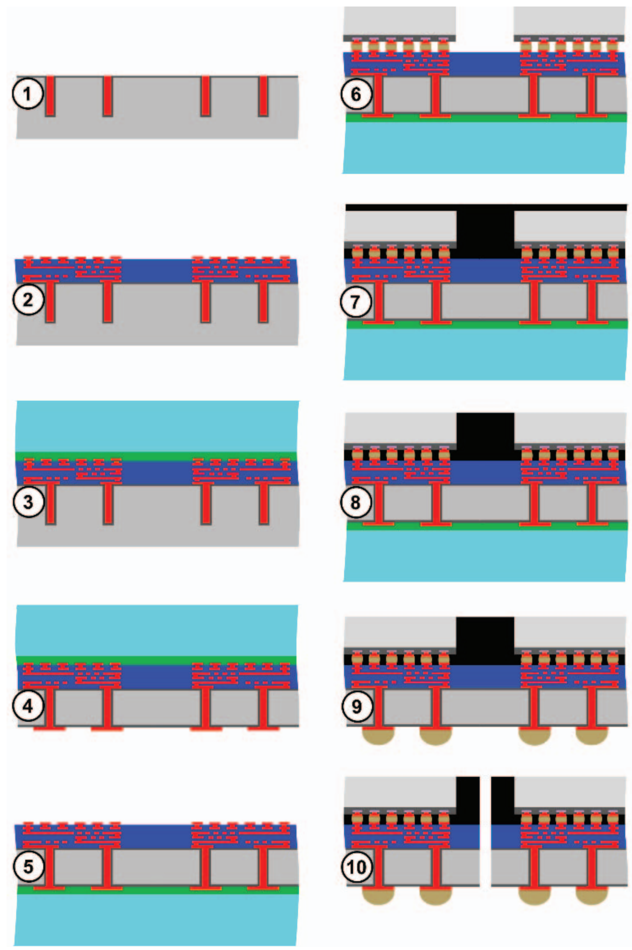

Figure 1: Schematic process flow for wafer level fabrication of HPC modules

In a second process block, multi-layer routing is fabricated, which establishes connections to the TSVs as well as between the IO pads to which the ICs are mounted later on. The multi-layer routing technology is adapted from well-known redistribution technology featuring electroplated copper for the wiring and thin film polymer as intra- and interlayer dielectric. Standard redistribution technology typically has relaxed wiring density with minimum 12 µm lines/space and large vias in the polymer, realized by direct lithographic structuring using photosensitive polymer. An example of an interposer after processing of the TSV plugs as well as the multi-layer routing is shown in Figure 2.

However, for the interconnection of ICs with high IO-density, a higher wiring density is needed. To meet this requirement, the multi-layer routing technology has been adapted during the last five years. The wiring density could be reduced to 7 µm lines/space and the diameter of the polymer vias could be reduced to 10 µm [15]. But even this increased wiring density is not sufficient for the routing requirements of compute and memory components used in HPC, which will be discussed in the next chapter. A further scaling of the multi-layer routing is required and will be presented in the Chapter IV.

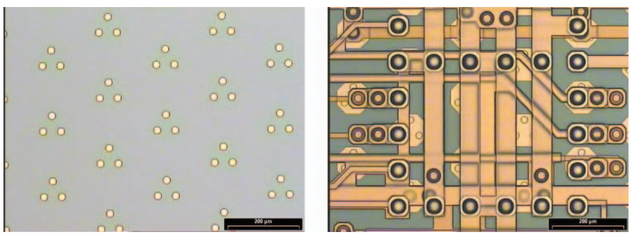

Figure 2: Example of TSV Interposer after processing of TSV plugs (left) and after processing of multi-layer wiring (right)

After processing of the multi-layer wiring and IO pads, a temporary carrier wafer is bonded at the front side to enable the backside processing of the interposer device, as shown in Figure 1, step 3. The backside processing includes grinding/polishing, re-passivation and TSV reveal as well as backside pad formation. For a later balling with AgSn solder, in most cases NiAu is deposited as pad metallization. A backside wiring is also possible, but it is not shown in the schematic image in step 4. After the backside processing is finished, the second temporary carrier is bonded and the first carrier is removed from the front side (step 5). Based on this, the front side of the interposer is accessible again and the components can be flip chip bonded here (step 6). For the assembly different technologies can be applied, but pick and place + reflow soldering utilizing micro bumps at the chips are preferred.

In the step 7 wafer level transfer molding is performed to encapsulate the chips. Ideally, the molding process can also be utilized to fill the gap between ICs and interposer. If not, underfilling has to be done additionally prior to the molding step. In the step 8, the mold material can be removed from the back side of the ICs by special grinding operation to later enable a better thermal connection of a cooling device.

In the final step 9, the carrier wafer is removed from the back side of the molded interposer wafer so that the backside pads are accessible. A wafer level balling process is now executed to apply solder to the pads. For the balling, different technologies like ball drop of preforms or printing can be used. Finally, a singulation by mechanical dicing is done to create the single HPC modules (step 10).

Whereas the first part of the process up to wafer level flip chip assembly and underfilling was executed for several projects in the past, the introduction of wafer level compression molding [16, 17], mold grinding as well as final balling represents a consequent process finalization at wafer scale and requires some setups and feasibility tests for seamless integration to the overall process flow. Of special interest are handling of the wafers after molding with respect to mold grinding, carrier release and solder balling. To answer these questions, a simple dedicated test vehicle was processed.

The test vehicle for this feasibility study is shown in Figure 3. This vehicle is based on a 200 mm silicon monitor wafer with a PI layer, representing the interposer RDL. This interposer wafer is already thinned down to 100 µm and attached to a glass carrier, thus resembling the process step 5 of Figure 1. The wafer is then populated with an array of Si dies, of which two neighbouring ones represent an HPC module with one processing unit (8 x 11 mm²) and one HBM (10 x 11 mm²). As the target is to assess the mechanical behaviour, mechanical dummy dies are used and bonded to the interposer wafer by adhesive die attach and no electrical functionality is attained.

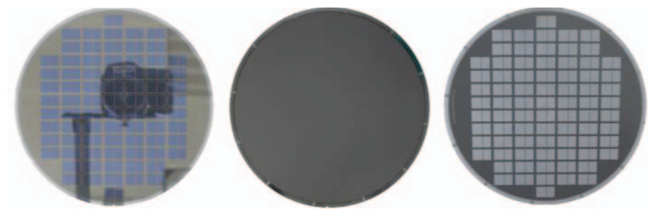

Figure 3: Test vehicle processing from left to right – mounting of dies, molding, mold grinding to revel chip back sides

As general process feasibility has been proven before, an evaluation of the warpage in the course of the process is aim of the study. Warpage in combination with the stiffness of the wafer is a critical parameter for the processability in wafer level packaging, as well as an indicator of inner tension and thus possibly the reliability of final packages.

Compression molding for these test vehicles is carried out on a TOWA molding system with a 190 mm cavity to ensure sufficient sealing on the 200 mm wafers. After molding and each subsequent process step, the warpage of the wafer is monitored. Two different Epoxy Molding Compounds (EMC) and overmold thicknesses, respectively, are investigated with regard to their effect on warpage. Both compounds are highly filled with SiO_2 spheres (82 wt%), while differing in Young's modulus: 19 vs. 16 GPa (A/B), CTE: 9 vs. 15 ppm/K (A/B), Tg: 190 vs. 138 °C (A/B).

The results of the warpage measurements show warpage in the range of 1.5 mm after molding and curing (Figure 4). This warpage is caused by a combination of cure shrinkage and a greater contraction of the upper layers (EMC, EMC+Si) of the structure in comparison to the lower layers (Si, glass) during cooling from cure temperature. Additionally, the non-populated rim of the molded wafer with a higher content of EMC in comparison to the central part of the wafer with Si dies causes inner tension that also contributes to warpage. As expected, the warpage decreases with removing the upper layer of EMC by grinding and increases again after debonding of the glass carrier due to a resulting lower stiffness of the stack. However, for all tested combinations of parameters, the warpage has proven to be below 1,5 mm peak warpage. This is sufficiently low for the described process steps and especially for final balling on wafer level (step 9 in Figure 1) as implemented at IZM using a Wagenbrett WB300 system for wafer balling. Balling process comprises printing of tacky flux, solder ball drop, reflow and cleaning, followed by a singulation by mechanical dicing. Solder balling is typically conducted by using SAC preforms, but also LTS preforms, e.g. SnBi are possible, reducing thermal mismatch during board assembly.

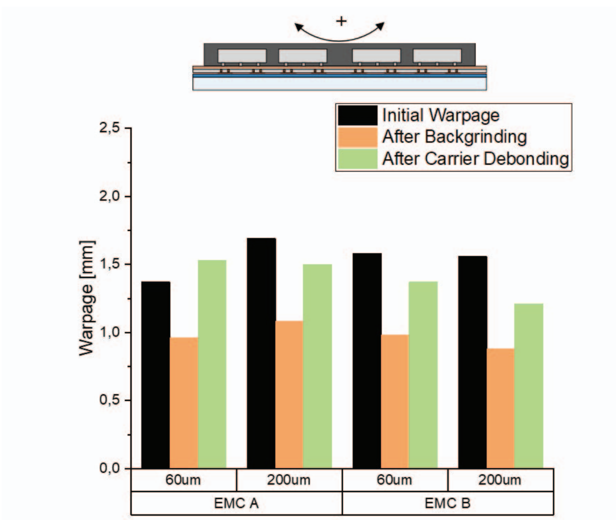

Figure 4: Warpage after each process step, depending on EMC and overmold thickness

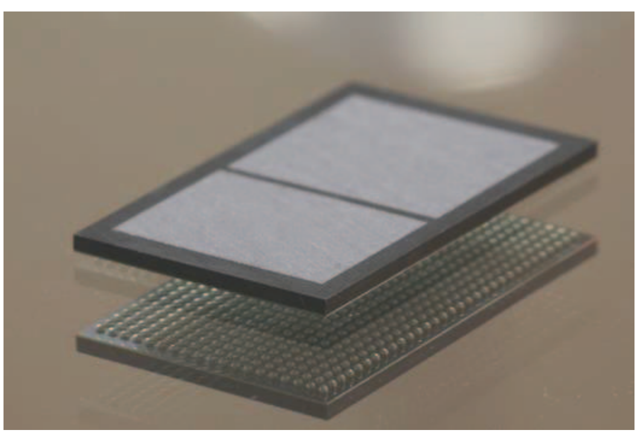

Figure 5: Balled mechanical sample based on 100 µm-thick silicon with polymer layer, assembled silicon dies with back ground mold encapsulation

The encapsulation on wafer level by compression molding hereby shows to be integrable in the overall process flow. Compression molding thus gives the advantage of improving the handling of thin interposer wafers during the process and enabling balling without an additional carrier by mechanically supporting the delicate interposer, besides its general function to improve the reliability of the embedded dies and the interconnections between dies and interposer.

III. RF Simulation of Fine Pitch Multi Layer Wiring and Extracted Requirements

To realize very high levels of data line parallelization within an interposer for HPC applications, beside the realizable technology driven line / space reduction to enable the number of required IO lines, also the electrical requirements regarding power and signal integrity needs to be met. Regarding the latter, major concerns are signal loss and cross talk between neighboring lines. Using a given routing area and minimal number of redistribution layers, this consequently implies a certain required line density. Hence, reducing signal loss (e.g. by using wider IO traces) and mitigating cross talk by placing more ground or shielding traces between the signal lines is eventually conflictive. Therefore, signal integrity aware design aims for a proper balance between those goals which depends on the technology specific design rules and material related constrains.

To obtain initial line pitch requirements to meet the electrical specifications, different signal / ground line arrangements and line dimensions of polymer-based RDLs have been investigated (two examples are shown in Figure 6). These configurations have been analysed by 2D quasi-static and 3D full wave EM field simulations. An appropriate starting point to fulfil the electrical specifications of HBM2e turned out to be an interposer with four redistribution layers and a line / space density of 4 µm.

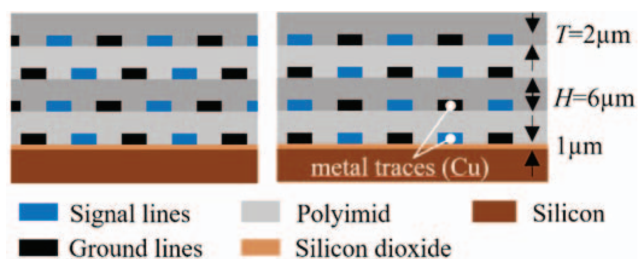

Figure 6: Cross-section of the four-layer interposer stack with different signal and ground line arrangements. The left side shows a staggered configuration and the right a chess field like signal and ground arrangement. The lines are pointing into this section.

No significant differences in terms of the cross-coupling has been observed for similar signal / ground arrangements between this approach and state-of-the-art BEOL technology with inorganic redistribution layers based on silicon dioxide or silicon nitride. However, due to the possibility of finer line pitches in inorganic-RDLs, the higher realizable trace density enables more comprehensive shielding concepts. But as the required line dimensions for those concepts shrink, the signal loss becomes more and more dominant, overcompensating this advantage. Generally expressed in simplified terms: a non-perfect but acceptable cross talk while focusing on loss mitigating line dimensioning is a beneficial solution for those applications. In that regard, the here considered technology using 2 µm-thick copper and a polymer dielectric with a layer thickness of 6 µm gives a significant advantage compared with inorganic RDLs which are limited in metal thickness (around 1 µm) and dielectric layer height (typically around 2 µm). Simulation results of signal transmission lines for different line dimensions as shown in Figure 7 confirm the above-mentioned behaviour. Here, a configuration with 4 µm as well as 3 µm line width and spacing (L/S) using 2 µm-thick metal and 6 µm-thick polymer dielectric results in comparable moderate losses around 0.2 dB per millimetre up to signal frequencies of 5 GHz (black and blue curve). Corresponding configuration for inorganic RDLs with minimal L/S and metal thickness of 1 µm and a dielectric layer height of 2 µm (red curve in Figure 7) increases the loss considerably. Due to their small dimensions, all line losses are dominated by the metallic losses which increase over frequency with a square root behaviour according to the skin effect.

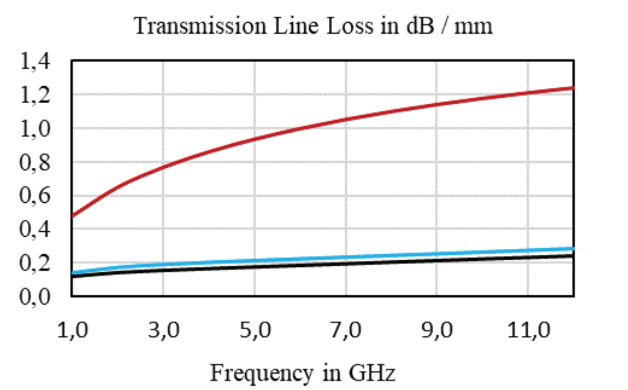

Figure 7: Signal loss simulation results for different conductor dimensions according to the right arrangement in Figure 6. Black curve: 4 µm L/S, 2 µm metal thickness and 6 µm polymer dielectric layer height; Blue curve: 3 µm L/S, same metal and polymer thickness; Red curve: 1 µm L/S, 1 µm metal height and 2 µm thick SiO_2 dielectric according to typical BEOL technologies

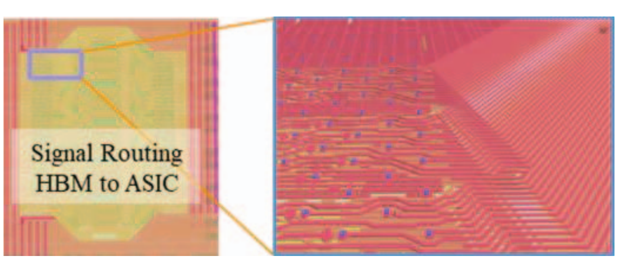

Figure 8 Layout view of TSV interposer signal routing for ~1700 signal line HBM2e data bus to ASIC (left), detail view of multi-layer layout (right)

Based on the investigations above, a polymer/Cu RDL with 4 µm L/S and proper signal / ground configuration enables the required line density between the HBM and ASIC. Though, some compromise has to be made in the region where the chip IOs and power connections are located because the larger via pads and lands reduce the available routing space (Figure 8).

Due to the ongoing demand for higher computing power for AI, simulations, big data, and other services, further increases of data rates e.g. between memory and CPUs are in progression. To fulfill the electrical requirements of increasing signal data rate and bandwidth for the upcoming next generation chips and modules as HBM3, a further increase of the polymer/copper RDL density to 3 µm L/S is envisioned as the next technology scaling step. This enables a better behavior in the start and end locations by reducing unwanted cross talk. Further performance improvement can be achieved by a lower and more homogeneous signal latency due to a simplified and straight signal routing. The signal loss can still be kept at a similar level due to shorter routing paths, allowing the realization of interconnections with very high signal quality and eventual fulfilment of the upcoming requirements.

IV. Scaling of Multil Layer Wiring

As discussed in the previous chapter, the multi-layer routing on an interposer device has to reach densities of 4 µm lines and space and beyond to meet the interconnection requirements of typical compute and memory chiplets. The current routing capabilities rely on up to 5 µm high semi-additive structured copper wires with 7 µm lines and space and laser structuring of vias with 10 µm diameter in 10 µm-thick polymer interlayer dielectric. Details of the process, especially related to the direct laser structuring approach of the polymer vias are described in the reference [18].

All pattern definition steps are done by mask aligner lithography and thin photo resists. Lithography based on mask aligner is a feature of this technology, because it is especially useful for realization of large devices which exceed typical stepper reticle sizes, so that no reticle stitching is required. In extreme cases full wafers can be re-routed and finally used as one device [19].

The required upscaling of routing density can be enabled by reduction of layer thickness for copper and polymer. Thinner layers allow use of thinner photo resists and reduce final warpage, so that finer structures are possible for both wiring and vias. With a restriction of the maximum copper height to 2 µm per layer and maximum polymer thickness to 6 µm per layer, the layout rules shown in Figure 9 were defined. A line pitch of 8 µm with symmetric width and spacing of 4 µm should be possible over multiple layers. The polymer vias, which are generated by laser ablation using a pre-defined hard mask, were defined with the diameter of 8 µm. Secondary layout rules such as overlaps and via spaces can be also found in Figure 9. As can be seen in the picture, the design rules support a possible TSV pitch of 20 µm for the interposer device. All numbers in brackets refer to e next technology release which may support a routing scenario of 6 µm pitch (3 µm lines and space) to support a 15 µm TSV pitch in the interposer device.

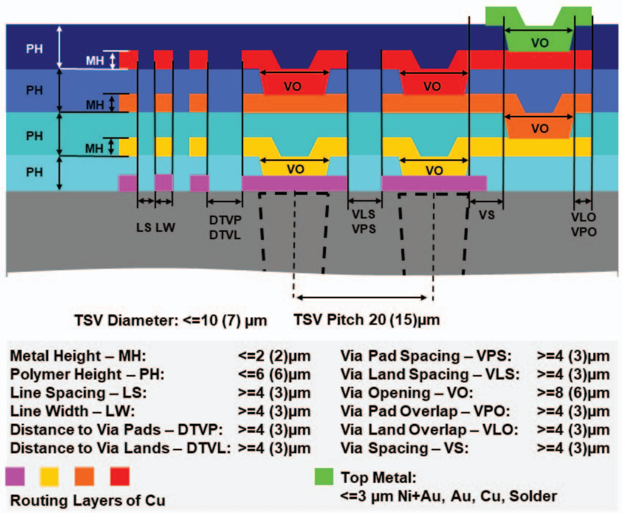

Figure 9: Design rules for fine pitch RDL with 4 µm lines and space and 4 routing layers on TSV interposer

Based on the fixed design rules shown in Figure 9, a technology test layout for 3 routing layers was prepared. Some details of the test layout are given in Figure 10. A reticle containing 3x24 straight lines of 2/4/6 mm length each as well as with 24 lines in zigzag configuration was placed 2460 times on a 200 mm wafer layout (wafer layout not shown). The purpose of these test structures is to enable full wafer optical inspection and electrical test later on to check the capability of the technology. Electrical test will give indication if connectivity of the wiring lines and vias as well as isolation between adjacent lines can reach sufficient yield or if position-based effects or repeatability issues occur.

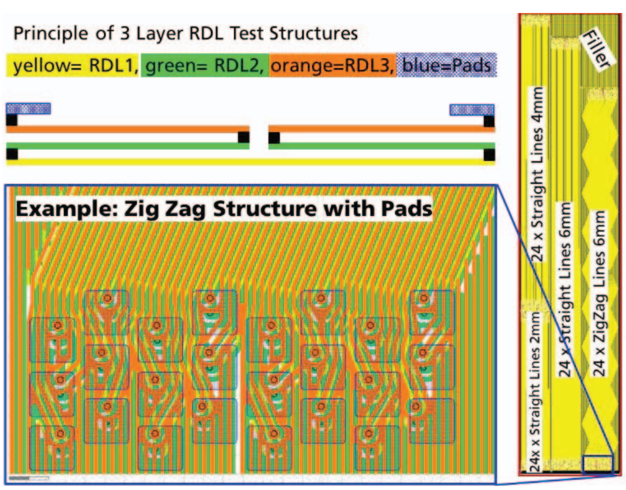

Figure 10: Overview of reticle with RDL test structures having different line lengths (right) and example of zigzag structure with pads (left)

Pictures in Figure 11 show exemplary microscopic images with different magnification and focus of the multi-layer test structures during the final fabrication steps. The upper three images show the vias with 8 µm diameter in the last polymer layer after the laser structuring was done. The thin metal hard mask used for the ablation is still present on top of the polymer. The three images in the middle show the same locations after the hard mask was removed. Now also the underlaying fine pitch wiring with 4 µm lines and space is visible through the optical transparent polymer. The three images at the bottom show the same location after the fabrication of the NiAu test pads was finished. The small 8 µm vias connecting the pads with the 3rd metal routing layer appear as black dot in each test pad.

Figure 12 shows a cross-sectional view of the RDL layer stack with the 8 µm laser structured polymer vias. The total RDL stack has a height of 26-28 µm.

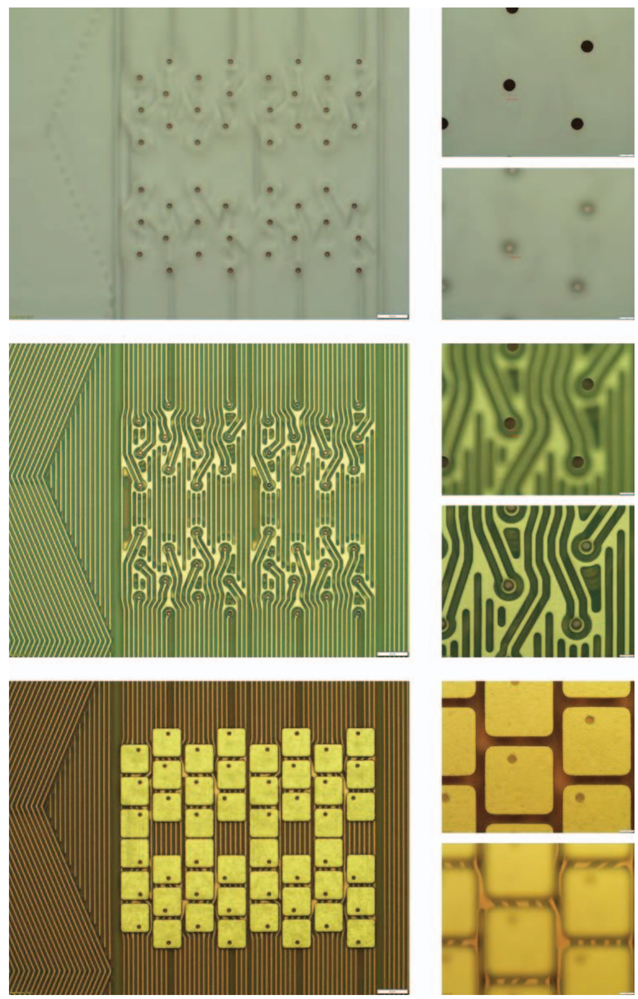

Figure 11:Images of fabricated test layout after via ablation of last polymer layer with hard mask still on (top), with hard mask removed (middle) and after pad formation (bottom)

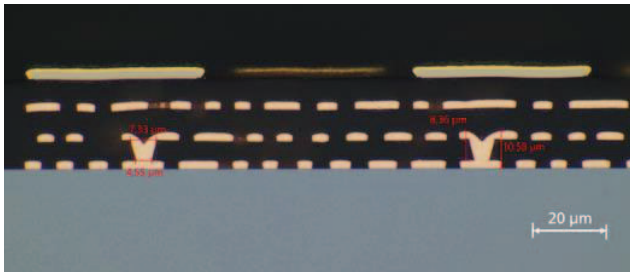

Figure 12:Cross sectional view of fabricated test layout with 3 routing layers and NiAu pads

For electrical testing, we used a Cascade ELITE300 automated probing station with a Keithley SMU4200-SCS. In order to reduce the efforts for electrical testing, 4 out of all 24 available 2/4/6 mm and zigzag lines were measured per reticle, which resulted in 9840 DC tests per structure type and a runtime of >30 h per wafer. As can be seen from the box plots in Figure 13, quite stable and consistent results could be obtained. The diagram shows exemplary the test results from one wafer. With a very few exceptions all test structures show resistance values in the expected range with quite low scattering. Further testing for shorts between adjacent lines were done using 40 V maximum voltage. Again 4 test per structure type and reticle were done. Based on these 9840 tests per structure type and wafer a 100% isolation yield was determined. This indicates capability and stability of both the fine pitch semi-additive copper technology as well as the polymer via formation process.

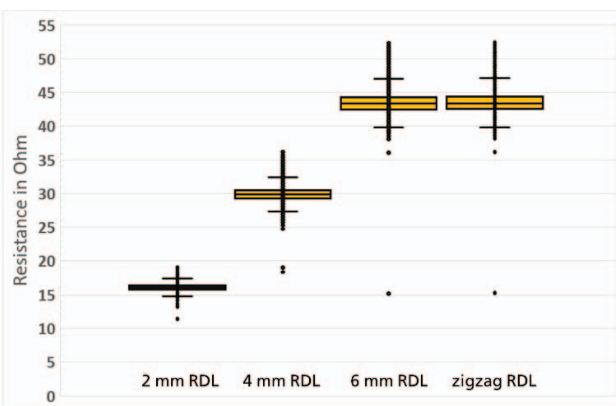

Figure 13: Boxplots of resistance measurements from RDL test structures based on 9840 measurements per wafer

V. Example: Occamy Integrated Dual Chiplet System

One example for realization of HPC-modules is the Occamy integrated dual chiplet system. The system contains two compute chiplets called Occamy as well as two 16 GB high bandwidth memories HBM2e, which are assembled on a silicon interposer with through silicon vias (TSVs). The Occamy dies have a size of 73 mm^2 and are fabricated in GlobalFoundries 12 nm FinFet technology. Each die includes six groups of four compute clusters, the host CVA6, an HBM2e controller IP from Rambus, as well as a source synchronous serial DDR die-to-die link. The system was designed by ETH Zürich and can reach a maximum performance of 6.144 TFlop/s in 8-bit floating-point formats [20].

For the assembly and packaging of the modules, the Occamy ICs were delivered on multi-project wafers and post-processed at Fraunhofer IZM, Berlin to apply copper pillar bumps onto their IOs with minimum pitch of 55 µm. The HBM2e were available as bare dies with similar copper pillar bumps. The silicon interposer called Hedwig, was fabricated in GlobalFoundries 65 nm technology including copper filled TSV blind plugs, which were prepared to reach 100 µm deep into the bulk silicon wafer for a later backside access. Those interposer wafers were postprocessed at Fraunhofer IZM, Berlin including wafer thinning and TSV reveal, as well as front- and back-side pad formation to enable the later assembly processes. Copper pads and nickel/gold pads were electroplated on the front and back side of the interposer, respectively. In the following steps, the interposer wafers were diced to obtain single Hedwig's. Now, the four ICs were assembled by pick & place and reflow soldering as well as underfilling. For the final assembly of the interposers to the system boards, the related PCBs had to be prepared with passives first, using standard surface mount technology of SMD components. After attach of 400 µm large solder balls to the array of 1399 pads on the back side of the interposers, their assembly to the PCBs could be done by another reflow soldering step.

Figure 14 shows a cross sectional image through an assembled Hedwig interposer. Here, the intention was to show the TSVs and their front and back side connection. The TSVs have a diameter of 10 µm and a depth of 100 µm. In this design multiple TSVs are grouped for one signal, which can be also recognized in the picture by the common backside pad. The cross section was taken before the solder balls were attached to the back side pads.

Figure 15 shows images of a Hedwig interposer with the 2x2 assembled Occamy and HBM2e ICs before and after final assembly to the system board. The system board hosts also 453 SMD capacitors of different sizes down to 0201.

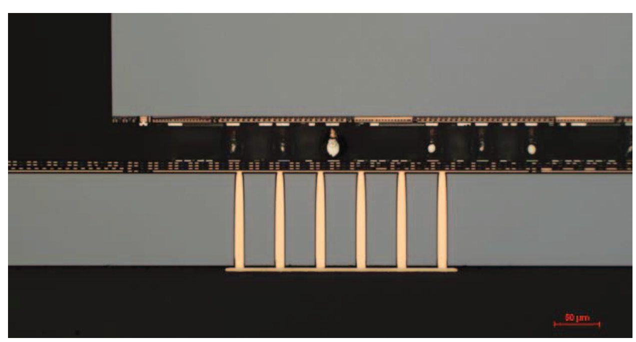

Figure 14: Cross section of Occamy chiplet system with view on the TSVs in the interposer device, TSVs are grouped to one back side contact here, back side solder ball not present yet

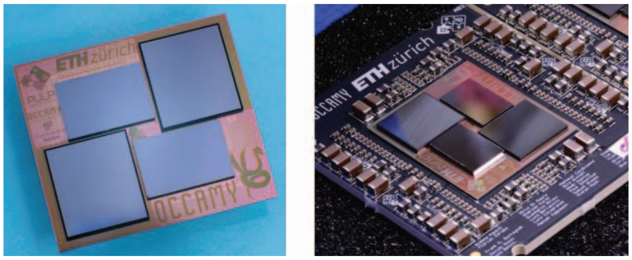

Figure 15: Occamy chiplet system with 2 Occamy and 2 HBM dies on TSV Interposer (left) and after assembly to system board with SMD capacitors (right)

VI. Conclusion

An overall process flow for wafer level packaging of high-performance computing modules was presented. It relies on TSV silicon interposer technology with high density routing layers for chiplet interconnection. The chiplets are assembled by reflow soldering and encapsulated by compression molding on the thin interposer wafers, which are still mounted on temporary carrier wafers. A moderate wafer warpage allows a final solder balling at the back side of the interposer wafer after the carrier wafer is removed. Thus, a wafer level processing is possible through the entire packaging sequence.

Key enabler for the presented approach is a fine pitch multi-layer routing technology, which enables low loss / high performance signal feeding at the top side of the interposer, where the chiplets are mounted. Simulations have shown that a routing density of at least 4 µm lines and space is necessary to fulfill the requirements of HBM2e interconnection with related compute chiplets. Based on that, the existing copper / polymer multi-layer RDL technology at Fraunhofer IZM was scaled to meet these requirements by using a nominal metal track height of 2 µm and laser via ablation into polymer dielectric thickness of 6 µm. A related test layoutt for technology development has revealed a very high electrical yield for both connectivity and isolation. The technology was approved with four routing layers and is further scalable to 3 µm lines and space and beyond to meet further increasing requirements in routing density.

As an example for fabricated high performance computing modules the Occamy chiplet system was presented. It is a four die chiplet system designed by ETH Zürich and assembled on a TSV silicon interposer fabricated by GlobalFoundries and post processed with thinning, TSV reveal and pad formation at Fraunhofer IZM, Berlin.

Acknowledgment

The authors thank all employees from Fraunhofer IZM, Berlin and Technical University of Berlin who contributed to this work. Special thanks to Manuel Bäuscher for electrical testing.

References

[1] T. Sterling et al., “High Performance Computing”, ISBN: 9780124201583

[2] J. Krueger et al., "Hardware/software co-design for energy-efficient seismic modeling," Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, Seattle, WA, USA, 2011, pp. 1-12

[3] D. Donofrio et al., “Energy-Efficient Computing for Extreme-Scale Science”, IEEE Computer, vol. 42, 2009, S. 62-71

[4] X. Jia et al., “GPU-based high-performance computing for radiation therapy”, Phys Med Biol. 2014 Feb 21;59(4):R151-82. doi: 10.1088/0031-9155/59/4/R151. Epub 2014 Feb 3. PMID: 24486639; PMCID: PMC4003902

[5] J.-P. Correa-Baena et al., “Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing”, Joule, Volume 2, Issue 8, 2018, Pages 1410-1420, ISSN 2542-4351

[6] Yole Intelligence, “Artificial Intelligence Computing for Automotive”, Market and Technology Report 2020

[7] Yole Intelligence, “High-End Performance Packaging 2023”, Market and Technology Report 2023

[8] J. Shi et al., “Interconnect Technologies and Advanced Packaging for High Performance Computing”, 2010 ECS Trans. 27 819

[9] A. Usman et al., “Interposer Technologies for High-Performance Applications”, IEEE Transactions on Components, Packaging and Manufacturing Tech-nology, vol7, 2017, S. 819-828

[10] L. Li et al., “3D SiP with Organic Interposer for ASIC and Memory Integration”, ECTC 2016

[11] S. Bhattacharya and V. S. Rao, "Multi-Chiplet Heterogeneous Integration Packaging for Semiconductor System Scaling," 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 2023, pp. 1-4

[12] K. Zoschke et al., “TSV based Silicon Interposer Technology for Wafer Level Fabrication of 3D SiP Modules”, 61st Electronic Components and Technology Conference, May 31 – June 3, 2011, Orlando, Florida, USA, pp. 836-843

[13] K. Zoschke et.al., „Wafer Level 3D System Integration based on Silicon Interposers with Through Silicon Vias”, 2012 IEEE 14th Electronics Packaging Technology Conference, 5 – 7 December 2012, Singapore, pp.11-16

[14] K. Zoschke et al., “Fabrication of 3D Hybrid Pixel Detector Modules based on TSV Processing and Advanced Flip Chip Assembly of Thin Read Out Chips”, 67th Electronic Components & Technology Conference, May 30th – June 2nd, 2017, Orlando USA, pp. 917-924

[15] K. Zoschke et al., “High Density Flex and Thin Chip Embedding Technology for Polymeric Interposer and Sensor Packaging Applications”, IWLPC2019, Oct. 22nd-24th 2019, San Jose, USA

[16] T. Braun et al., “Trends on Large Area Advanced Packaging”, 40th International Conference of Photopolymer Science and Technology; Chiba, Japan

[17] M. Adler et al., “Robuste High Performance Computing Module durch Compression Molding”, Proc. EBL Fachtagung 2024; Fellbach, D

[18] K. Zoschke et al.,High-Density Flexible Substrate Technology with Thin Chip Embedding and Partial Carrier Release Option for IoT and Sensor Applications”, 69th Electronic Components & Technology Conference, May 28th – May 31st, 2019, Las Vegas, USA, pp. 1475-1484

[19] K. Zoschke et al., “Full Wafer Redistribution and Wafer Embedding as Key Technologies for a Multi-Scale Neuromorphic Hardware Cluster”, 19th Electronics Packaging Technology Conference, 6th Dec – 9th Dec 2017, Singapore, pp. 1-8

[20] https://pulp-platform.org/occamy/