HPC Packaging Platform Including Compression Molding

M. Adler[1], K.-F. Becker[1], T. Braun[1], M. Dreissigacker[2], O. Hoelck[1], A. Gaebler[1], U. Maass[1], I. Ndip[1], K. Zoschke[1], M. Schiffer[1], M. Schneider-Ramelow[2]

[1]Fraunhofer Institute for Reliability and Microintegration IZM
Gustav-Meyer Allee 25
13355 Berlin, Germany
Ph: +48-30-46403207
Email: Marius.Adler@izm.fraunhofer.de

[2]Technical University Berlin
Straße des 17. Juni 135
10623 Berlin, Germany
Ph: +49-30-314-23922

Abstract

Efficient high-performance computing (HPC), especially for applications in the field of artificial intelligence, requires the use of specialized computing and memory chips, short transmission paths and high packing densities. A key technology for HPC is therefore the integration of processors and memory stacks in a single System-in-Package (SiP) using silicon-based interposers.

Manufacturing concepts have been developed for such a package, each of which utilizes different process flows to build an HPC module. Following a holistic approach, the process flow is evaluated from the processing of the interposer with high density organic redistribution layers, compression molding and balling. Evaluation methods were defined for the individual process steps and criteria established to enable the selection of suitable process steps and combinations for an optimized package. The process flow options were realized and analyzed for a reference HPC module. An important parameter for all these process steps is warpage, since an exceeding warpage can make processing impossible. For compression molding, a key process step, the influence of different materials and process parameters on the warpage was therefore investigated.

The results of the process development are presented for the backend process flow with an emphasis on compression molding process step, an evaluation of the process variants is carried out according to the previously developed criteria and the most suitable variant of the process step for later production is selected on this basis.

Key words

High Performance Computing, Wafer Level Packaging, Silicon Interposer, Compression Molding, Epoxy Molding Compound, Warpage

I. Introduction

Artificial intelligence, big data and complex modeling require efficient high-performance computing modules. In addition to the development of ICs that can perform individual computing operations with high efficiency and speed, the interconnection of these chips to their environment is also one of the main challenges of high-performance computing (HPC).

For instance, short transmission paths between computing and memory chips and high signal integrity are required to take full advantage of a leading-edge chip architecture. Advanced packaging is therefore a central building block for the realization of HPC modules. One option for the advanced packaging of HPC modules is the integration of computing and memory chips in a single System-in-Package (SiP) [1]-

[3]. In addition to the integration of one computing and memory chip each, such a packaging platform can also enable the integration of a larger number of chips or chiplets and thus also makes a significant contribution to the emerging chiplet infrastructure.
According to this approach of heterogeneous integration in a SiP, one or more ICs that are optimized for specific computing operations such as stencil processing units [4]-[6] or GPUs can be assembled together with High Bandwidth Memories (HBM) on a silicon interposer [7],[8]. Interposers and chips are connected with microbumps with a pitch in the range of 50 µm [9]. In addition to the use of silicon as interposer material, glass and organic interposers, possibly with integrated glass or silicon bridges, can also be considered [10],[11]. Silicon interposers are currently wide-spread due to their ability to support high density wiring and existing process lines.
For the packaging of such HPC modules in a high-volume production, wafer level processes such as redistribution layer (RDL) processing and compression molding are key technologies. While wafer level packaging imposes major advantages such as good scalability, it also includes challenges such as warpage that increase with the size of wafers [12]-[14] which have to be addressed in a proper process evaluation.

II. Process Flows for HPC Modules

In order to investigate the manufacturing of an HPC module with the previously described setup of stencil chip and HBM on silicon interposers and to analyze optimal process control, a simplified demonstrator setup is developed (Figure 1). A stencil processing unit and an HBM are integrated on a silicon interposer. In subsequent development steps, the platform can be extended to integrate a larger number of chips on an interposer using the chiplet approach.

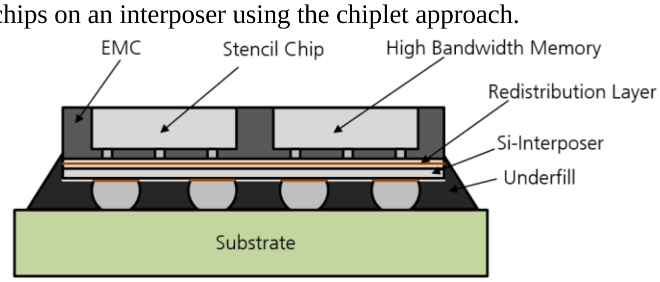


Figure 1: Schematic structure of the demonstrator

To set up the demonstrator module, the HBM and the stencil chip, each with a thickness of 725 µm, are mounted on an interposer. The gap defined by the bump height with a distance of 35 µm between interposer and chips. It can either be filled with an underfill for thermo-mechanical stabilization or the underfilling is integrated into the encapsulation process by compression molding in the so-called molded underfill. The interposer wafer initially has a thickness of 725 µm. After thinning the wafer and exposing the through silicon vias (TSV), the final thickness is 100 µm. The thinning and complete processing of the interposer can take place either before or after compression molding. These two possible process variants are shown in Figure 2 and Figure 3.

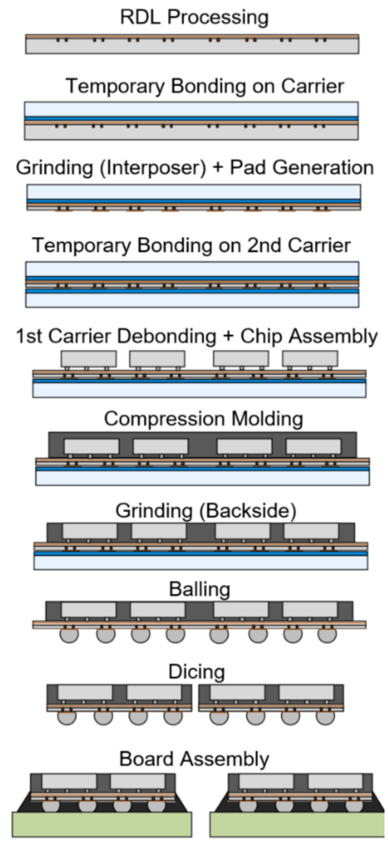


Figure 2: Process flow with interposer wafers thinned at the beginning

The finished module has a size of 25x15 mm² and is assembled to an organic substrate after solder balls have been applied. The ball grid array layout is based on current standards; typical ball sizes are around 300 µm and typical pitches are 500 µm [15]. Here, too, underfilling and the mounting of a heat sink or lid with suitable thermal interface material on the back of the module for optimized heat dissipation follow.
The limit for further processing of the encapsulated wafers in the balling process at IZM is typically 2 mm total warpage for a mold wafer with d = 200 mm; the aim of process development is to remain below this threshold value. A holistic assessment of the whole process chain is carried out to secure a high performance and reliability of the HPC packaging platform.

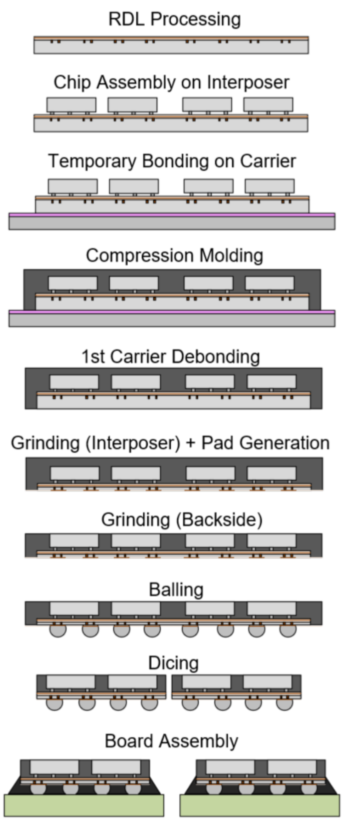


Figure 3: Process flow with initially unthinned interposer wafers

III. Interposer Layout and Processing

To provide sufficient high wiring density for the interconnection of computing and memory devices on interposer, Fraunhofer IZM has further developed its multi-layer routing technologies. The multi-layer routing technology is originally adapted from well-known redistribution technology for fan in wafer level packaging featuring electroplated copper for the wiring and thin film polymer as intra- and interlayer dielectric. Standard redistribution technology typically has relaxed wiring density with 5 µm thick copper lines having minimum 12 µm lines/space and large vias in 10 µm thick polymer, realized by direct lithographic structuring using a photosensitive polymer.
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 scaled towards higher density during the last couple of years. Routing on an interposer device for HPC applications has to reach densities of 4 µm lines and space and beyond to meet the interconnection requirements of typical compute and memory chiplets. The required upscaling of routing density was 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 4 were defined. Based on that, a line pitch of 8 µm with symmetric width and spacing of 4 µm is possible over multiple layers. The polymer vias, which are generated by laser ablation using a pre-defined hard mask, were defined with a diameter of 8 µm. Secondary layout rules such as overlaps and via spaces can be also found in Figure 4. 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 the next technology release which can support a routing scenario of 6 µm pitch (3 µm lines and space) to support a 15 µm TSV pitch in the interposer device.

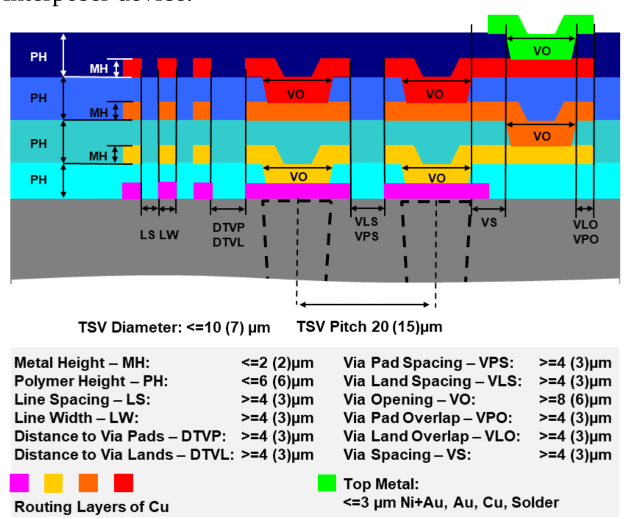


Figure 4: Design rules for fine pitch multi-layer wiring with 4 µm lines and space and 4 routing layers on TSV interposer

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 [16]. Details of the process, especially related to the direct laser structuring approach of the polymer vias are described in [17].

IV. Encapsulation by Compression Molding

A. Compression Molding and Warpage

Compression molding is a well-established process for large-area encapsulation in the area of wafer level and panel level packaging in volume production [18]. A variety of mold system suppliers are offering equipment for these processes, for this study a Towa compression molding system in cavity down configuration has been used. For encapsulation at wafer level, an epoxy molding compound (EMC), an epoxy-based polymer filled with SiO_2 particles, is dosed into a cavity in wafer form (lower half of the tool). The components to be encapsulated are then placed on a carrier in the open mold (upper mold half), the mold is closed, the cavity is evacuated and the EMC is pressed into the cavity by moving the mold onto the component carrier (see Figure 5). The EMC hardens under pressure and temperature during the In-Mold Cure (IMC) and achieves sufficient mechanical strength to be removed from the cavity. The final cross-linking takes place in a subsequent post-mold cure (PMC) in an oven outside the cavity.

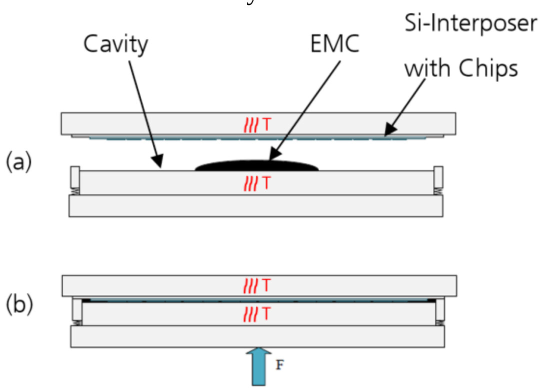
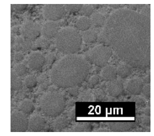
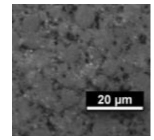
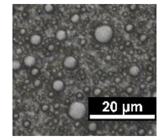


Figure 5: Principle of compression molding: (a) placing interposer and EMC in cavity; (b) closing cavity for In-Mold Cure (IMC)

The epoxy molding compound must have a sufficiently high flowability for processing in compression molding in order to fill the entire cavity and all gaps between the components without voiding. In particular, in order to achieve the narrow gaps between chips and interposer as a molded underfill, small filler sizes and low viscosity are required at the same time. The materials shown in Table 1 were selected for the process discussed here. Material C in particular is characterized by a small filler size, but at significantly higher costs than materials A and B.

In addition to the electrical insulation of the individual contacts (e.g. whisker formation, short circuits due to conductive contamination), the EMC also serves to mechanically reinforce the assembled components under operating conditions, both during further processing and in the subsequent application. Therefore, sufficient material stiffness is required for stabilization and good thermo-mechanical adaptation of the EMC to the other materials present in the package (CTE, T_g). However, excessive stiffness can be critical, as warped wafers can no longer be processed if the stiffness is too high, for example if vacuum chucking and therefore processing of the warped wafer is not possible.

Table 1: Material properties of the selected EMCs

Parameters	**EMC A**	**EMC B**	**EMC C**
Micrograph	 20 µm	 20 µm	 20 µm
EMC type	Granular	Liquid	Liquid
Filler content [wt%]	82	82	83,5
Maximum filler size [µm]	20	10	5
CTE1 [ppm/K]	9	15	13
T_g [°C]	190	138	150
G' [GPa]	19	16	21
In-Mold Cure IMC	10 min @ 130 °C	10 min @ 110 °C	10 min @ 120 °C
Post-Mold Cure PMC	60 min @ 150 °C	60 min @ 150 °C	60 min @ 150 °C

In order to keep warpage generally low, the CTE of the EMC should deviate as little as possible from the CTE of the silicon interposer (ca. 2.6 ppm/K). A difference in the CTE creates internal stresses when cooling from the processing temperature, resulting in a warping of the wafer at handling temperature. Since polymers generally have a significantly higher CTE than semiconductors, a low CTE is an important target value when selecting the EMC. Since the exact warpage and its limits depend on a variety of parameters such as stiffness, CTE and component structure, the warpage after compression molding and the processability of the warped wafers are analyzed in the experiment.

B. *Experimental Design of Molding Study*

To compare the process variants presented above, the process is run through as a demonstrator production. In order to also consider warpage as an important quality criterion, a test plan is developed in which the molding compound and the overmold thickness are examined as parameters in addition to the two variants of the process flow. The choice of material and the overmold thickness can potentially have a decisive influence on warpage, as they determine the degree of contraction of the upper layer of the structure and thus the internal stresses introduced.

The parameters of the full factorial test plan are shown in Table 2. The more cost-effective material variants EMC A and B are examined first.

Table 2: Full-factorial experimental design

Factor	-	+
EMC	EMC A	EMC B
Overmold Thickness	60 µm	200 µm
Process Flow	Unthinned Interposer	Thinned Interposer

For this demonstrator run to investigate the mechanical behavior, dummy chips made of silicon with a die attach material are mounted on the silicon wafers, which serve as interposers (Figure 6).

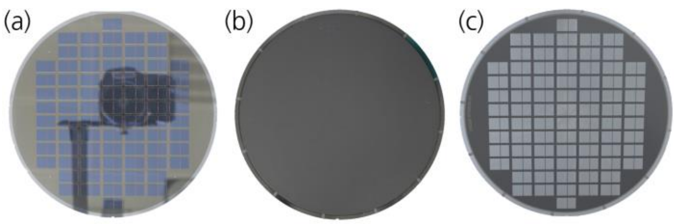


Figure 6: Demonstrator setup: (a) mounting the chips on interposer; (b) molding; (c) backgrinding to reveal chip backsides

Compression molding is carried out using a Y-120 wafer molding machine (TOWA, Kyoto, Japan). The material manufacturer's recommendations according to Table 1 are used as the processing temperature and time. Directly after molding and after each process step, the warpage is measured using a precision 3-coordinate measuring system (OMS 600, Mahr GmbH, Göttingen, Germany).

C. *Evaluation*

After molding, a warpage of the wafers in the range of 1.1 mm to 1.7 mm results for all process variants. The most significant effect on warpage can be observed for the factor process flow (Figure 7). A significant difference is understandable due to the different mechanical structure. However, since additional steps for processing the interposer follow for the unthinned interposers, even a lower warpage can also be critical here.

The other parameters examined, EMC and overmold thickness, have a less significant influence on warping. The observed slightly higher warpage for greater overmold thickness is generally plausible, as more EMC also means more stress due to the greater contraction of the upper layer of the build-up when cooling from the processing temperature.

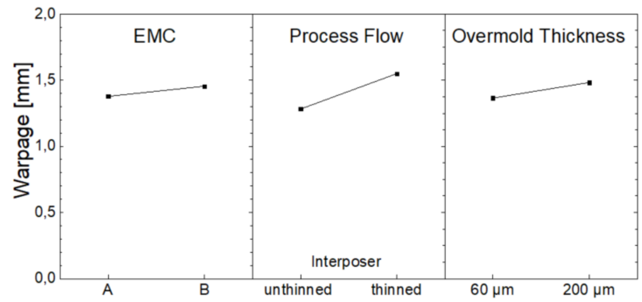


Figure 7: Main effect plot showing the largest effect for process flow

In the further investigation of the first process flow, in which not yet fully processed interposers are overmolded, an increase in warping can be seen for almost all of the wafers examined after the interposer has been thinned and the backside subsequently ground.

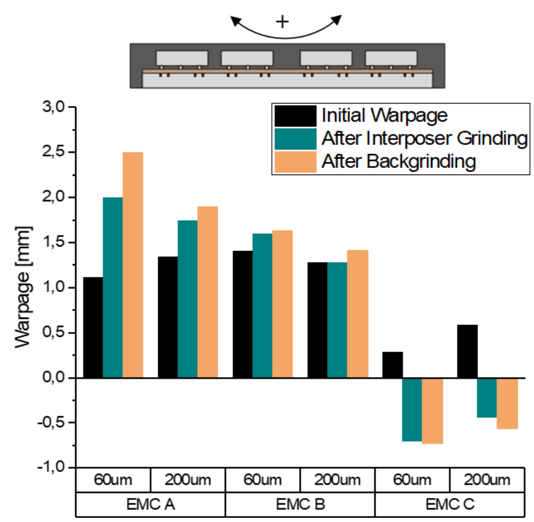


Figure 8: Warpage with initially unthinned interposer wafers after each process step

This can be explained by the decrease in the stiffness of the overall structure. The fact that the grinding of the rear side causes an increase in warpage can be explained by internal stresses generated by the surrounding ring of EMC around the interposer. Further processability proves to be problematic with this process variant; for example, the suction of the wafer to oxidize the underside of the interposer is not possible with the system used.

EMC C is tested as a possible alternative material. Figure 8 shows that a significant reduction in warpage can be achieved. After thinning the interposer, the measured warpage is negative, i.e. the wafer is warped in the direction of the interposer. This can possibly be explained by the expansion caused by the absorption of moisture. A higher maximum moisture absorption is supported by material analysis, that shows a significantly higher saturation concentration of water in EMC C compared to the other materials, eventhough the diffusion coefficient, indicating the absorption kinetics, is lower. Figure 9 displays these values, gained from dynamic vapor sorption analysis with a Discovery SA (TA Instruments, New Castle, DE), see [19] for meterology.

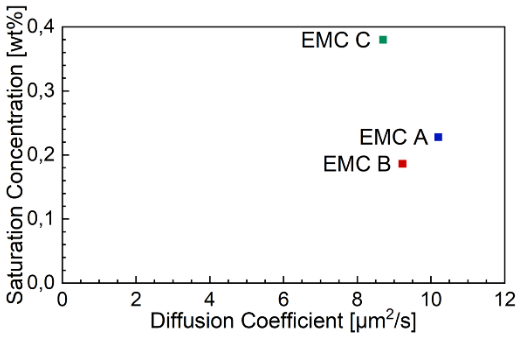


Figure 9: Saturation concentration and diffusion coefficient as parameters of moisture absorption of EMC A-C

For the process variant in which already processed interposer wafers are overmolded, there is a clear reduction in warpage after the chip backsides have been exposed by grinding (Figure 10). This decrease is to be expected, as the tension from the "pulling" EMC layer is reduced. After debonding the encapsulated wafer from the glass substrate, on which the thinned interposer is initially applied for processing, there is again an expected increase in warpage due to the decrease in stiffness of the entire structure. There are no problems with the suction of the wafers in any of the process steps.

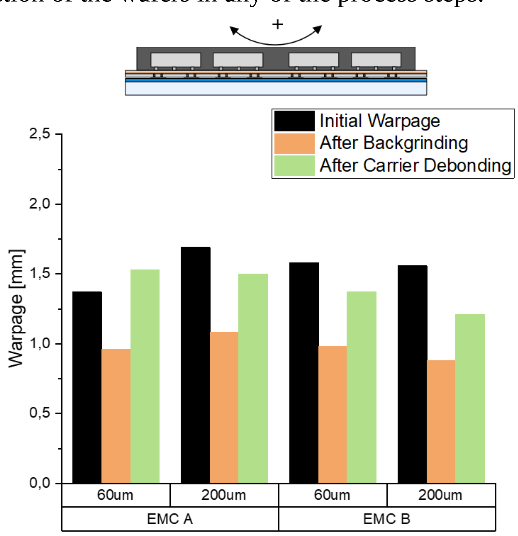


Figure 10: Warpage with initially thinned interposer wafers after each process step

The process variant with first thinned and finished interposer fulfills the warpage criterion < 2 mm for balling for both molding compounds used and is also a lower-risk variant. The process with first thinned interposer was successfully demonstrated in the tests with material A and B as well as with both investigated overmold thicknesses (60 µm and 200 µm). To further reduce the warping without considering the material costs, the use of EMC C is also possible here, but not necessary.

Figure 11 shows an example of a module after molding, backgrinding, application of solder balls and dicing.

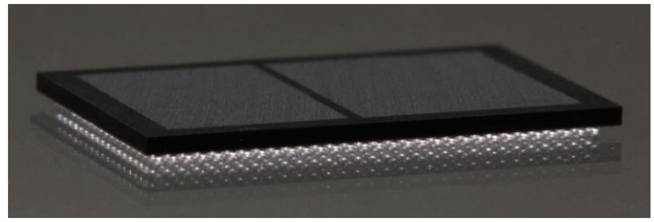


Figure 11: Demonstrator module after balling and separation ready for PCB assembly

V. Conclusion and Outlook

HPC modules require a suitable packaging platform to enable communication between multiple chips and to protect the components. Such a platform ranges from interposers with high-density electrical redistribution layers and solutions for encapsulation to balling and mounting on a substrate. The work presented here deals with suitable processes for the production of HPC modules with a focus on encapsulation. Encapsulation by compression molding is suitable for increasing the robustness of the interposer wafers during the process and for meeting the reliability requirements for assembly and operation in the application. However, the encapsulation process also introduces a warping that influences handling in the subsequent process steps or makes it impossible if the warpage is too large. With sufficient knowledge about the factors influencing the warpage, about the change in warpage during and after the respective process steps and about the limits of tolerable warping, optimized parameters for the encapsulation process can be found. In the investigations carried out for this purpose with non-functional components, the process variant in which compression molding takes place on a thinned interposer wafer was successfully demonstrated up to balling and separation, and a suitable encapsulation material was qualified.

The next step is to investigate the molded underfill of the gap between the chip and interposer. In parallel, work on the simulation, design and construction of functional interposers with high-density redistribution layers is ongoing.

References

[1] Mitsuo, M. et al: Silicon interposer technology for high-density package. ECTC 2000

[2] Lau, J.: Semiconductor Advanced Packaging. Springer Nature, Singapore 2021

[3] Bhattacharya, S. and Rao, V.S.: Multi-Chiplet Heterogeneous Integration Packaging for Semiconductor System Scaling. IEEE VLSI Technology and Circuits 2023

[4] Datta, K. et al: Stencil computation optimization and auto-tuning on state-of -the-art multicore architectures. SC 2008
[5] Krüger, J. et al.: Hardware/software co-design for energy-efficient seismic modeling. SC 2011
[6] Donofrio, D. et al: Energy-Efficient Computing for Extreme-Scale Science. IEEE Computer, vol. 42, 2009, pp. 62-71
[7] Chidambaram, V.: Active Silicon Interposer for Heterogeneous Integration: System Scaling and Cost Effectiveness. IEEE Electron Devices Technology and Manufacturing Conference 2019
[8] Wang, H. et al: A Review of System-in-Package Technologies: Application and Reliability of Advanced Packaging. Micromachines 2023, 14(6), 1149
[9] Lee, J. et al: Micro Bump System for 2nd Generation Silicon Interposer with GPU and High Bandwidth Memory (HBM) Concurrent Integration. ECTC 2018
[10] Li, L. et al: 3D SiP with Organic Interposer for ASIC and Memory Integration. ECTC 2016
[11] Usman, A. et al: Interposer Technologies for High-Performance Applications. IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 7, 2017, pp. 819-828
[12] T. Braun et al.: How to Manipulate Warpage in Fan-out Wafer and Panel Level Packaging. ECTC 2024
[13] T. Braun et al.: A Closer Look to Fan-out Panel Level Packaging. EDTM 2023
[14] T. Braun, O. Hoelck: *Manufacturing for Reliability of Panel-Level Fan-out Packages* in: W. D. van Driel, M. Yazdan Mehr (eds.), Reliability of Organic Compounds in Microelectronics and Optoelectronics, Springer, Zurich, 2022, pp. 517-531
[15] IPC 7095 - Design and Assembly Process Implementation for BGAs, www.ipc.org
[16] 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
[17] 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, 2019, Las Vegas, USA, pp. 1475-1484
[18] T. Braun et al.; Trends on Large Area Advanced Packaging; 40th International Conference of Photopolymer Science and Technology; Chiba, Japan, 2023
[19] T. Braun et al.: Influence of Humidity on Reliability of Plastic Packages, European Microelectronics Packaging Conference (EMPC), Friedrichshafen, Germany, 2015