Electronic packaging for future electronic systems

By Michael Töpper, Tanja Braun, Rolf Aschenbrenner  [Fraunhofer IZM]

Future electronic systems like autonomous systems using high-performance computing (HPC) and edge computing systems, sensor-integrated systems and bio-integrated devices will require more and more functions that cannot be managed by a single chip, even if advanced system on chip (SoC) concepts are used. Heterogeneous integration will be the next step and will pass beyond current system in package (SiP) approaches. This concept of true heterogeneous integration is highly important for next-generation devices based on future CMOS-nodes, SiGe, SiC, III/Vs like GaAs or GaN, and all different kinds of microelectromechanical systems (MEMS).

The digital transformation of society and economy creates an increasing demand to transfer, process and store vast amounts of data generated in the context of technologies such as autonomous driving, artificial intelligence (AI), and the Internet of Things (IoT). For conventional approaches, the amount of data to be stored is too big, the data transfer rates are too low, the available computational power is limiting, and the energy consumption, as well as the heat production of general-purpose computer processing units (CPUs) in a von Neumann architecture, are too high. Therefore, complex calculations, simulations, and decision-making cannot be performed on a practical time scale. A paradigm shift is taking place in many applications, in that data is progressively processed at a more localized level – from the cloud to the edge and down to the sensor node, thereby enabling meaningful information to be extracted, transmitted, stored or acted upon faster. In the era of connected intelligence, fast information and decision-making are important and require effective concepts for low-power, secure, connected, and embedded computing. The new paradigm is systemic efficiency, characterized by multi-parametric optimization: reduction in power consumption, latency, and data transfer by means of preprocessed data and flexible processor architectures. Promising approaches are artificial neural networks and neuromorphic computing.

Status of computing today

Today, a major part of computing is performed in specialized data centers (cloud computing). Computing is executed on highly parallel set-ups with thousands of cores. Similarly, supercomputers are built for large-scale calculations mainly in the scientific and very special industrial domains. These computers consume an enormous amount of electrical energy—in the megawatt region. The main bottleneck of today’s computers is the efficiency of data handling between processing units and memories. Today’s supercomputers run at about 5%, or even less, of their theoretical computing power due to their limited memory bandwidth. New hardware and software architectures are necessary to break down the boundary between pure logic and pure memory domains. This can be achieved by 3D stacking of logic and memory on top of each other and ultra-fast interconnections between the domains. Additionally, the technology of embedding integrated circuits (ICs) (Figure 1) into high-density organic board may be an alternative approach – perhaps in combination with 3D Si-stacking [1,2].

In addition to the aforementioned approaches, optical intra-chip and intra-board communication need to be created in order to dramatically speed up communication and to significantly reduce the power consumption. This can be achieved by the use of artificial neural networks and neuromorphic computing.

Status of packaging today

Today, a major part of packaging is performed in specialized data centers (cloud computing). Packaging is executed on highly parallel set-ups with thousands of cores. Similarly, supercomputers are built for large-scale calculations mainly in the scientific and very special industrial domains. These computers consume an enormous amount of electrical energy—in the megawatt region. The main bottleneck of today’s computers is the efficiency of data handling between processing units and memories. Today’s supercomputers run at about 5%, or even less, of their theoretical computing power due to their limited memory bandwidth. New hardware and software architectures are necessary to break down the boundary between pure logic and pure memory domains. This can be achieved by 3D stacking of logic and memory on top of each other and ultra-fast interconnections between the domains. Additionally, the technology of embedding integrated circuits (ICs) (Figure 1) into high-density organic board may be an alternative approach – perhaps in combination with 3D Si-stacking [1,2].

In addition to the aforementioned approaches, optical intra-chip and intra-board communication need to be created in order to dramatically speed up communication and to significantly reduce the power consumption. This can be achieved by the use of artificial neural networks and neuromorphic computing.

Figure 1: Chip-embedding into organic substrates using 5µm lines and space interconnections.

Figure 2: Vision of the Mass Manufacturing of Transceivers for Terabit/s Era (MASTART) project.
lower energy consumption. To those ends, we are working in the EU-founded L3MATRIX and Mass Manufacturing of Transceivers for Terabit/s Era (MASSTART) projects offering technological innovations in the fields of silicon photonics and the integration of 3D devices [3-5]. The aim of the L3MATRIX project partners is to develop a novel matrix with a larger scale than any other device, and to integrate more than 100 modulators on a single chip, and laser sources with a logic chip (Figure 2).

So today, the main bottleneck to the realization of next-generation computing systems for all large, secure data applications/industries, including SiP and SoC, is the lack of off-chip/off-core interconnects with low latency, low power, high bandwidth, and high density. One of the most promising approaches to overcome these challenges is the use of photonics. This will be key to the introduction of disruptive computing technologies and photonics-enabled architectures, leading to faster, cheaper, power-efficient, secure, denser solutions for industrial applications. Furthermore, generic co-integration with all building blocks of computing technology will be possible, as photonic-based standard interfaces between building blocks are introduced and implemented.

**Next-generation computing solutions**

A solution for next-generation computing might be that part of the computing will migrate down to the edge and the node level with substantial advantages such as less required overall bandwidth, less latency, and therefore faster decisions enabling new services, and the possibility of sensor fusion. But next-generation computing can also be based on substantially more efficient architectures that consume fractions of the energy that is needed today. The radical improvement of energy consumption of high-performance computers (edge and cloud) is, therefore, one of the major challenges of next-generation computing.

Creating very fast computing systems, overcoming the current memory bottleneck, and significantly improving energy efficiency will require 3D stacking of memories and logic ICs for high-bandwidth memories (HBMs) [6]. With a dual-sided cooling approach, we have successfully demonstrated the cooling of a large 20x20mm² heater die with targeted 672W heat dissipation by newly developed fluidic interposers with up to a 27x27mm² footprint (Figure 3).

**Chiplets gaining traction**

The increasingly higher costs for further node miniaturization in the IC manufacturing process will also promote the interconnection of chiplets. This means that intellectual property (IP) blocks made in different technology nodes will be combined on an active interposer to reduce cost by increasing the production yield (smaller chips) and reuse across applications. High-density organic substrates with or without embedding technologies may also be a solution. To that end, Fraunhofer IZM started the PLC-consortium working on different R&D topics related to high-density embedding technologies in organic substrates [7,8]. A close coordination between the different R&D thrusts was established to produce assembled panels in order to ensure a continuous process flow. A major work package in the beginning of the project was to collect data about the available materials provided by the partners. Several test layouts were designed for process development. The implementation and improvement of the die shift compensation, as well as the warpage and thickness measurement and evaluation, was a joint effort necessary for the high-density redistribution layer (RDL). For this metal routing process, a fusion of wafer-level (thin-film) processing and printed circuit board (PCB)-like substrate processing was developed. Different material setups and process combinations were assessed and improved and finally, were successfully applied to demonstrator packages. The work focus of the first year of the PLC was to establish and to optimize the entire process chain from assembly via molding and RDL formation to device singulation on a half format 457mmx229mm² panel. This also included the design of test vehicles and demonstrators, as well as first reliability tests. The focus in the second year was on design modification based on first year results, technology scaling from half format to full format 610mmx305mm², as well as the integration of vertical interconnect elements (VIE) and passive components.

The optimization of single process steps, but also the alignment and optimization of the interfaces along the manufacturing flows was an ongoing activity throughout the entire project. Reference process flows based on photosensitive dry-film dielectric material and non-photosensitive materials were developed including a semi-additive plating (SAP) process. Die shift and warpage control to handle and process large molded fan-out panels were of key importance, including a smart alignment strategy development from assembly to RDL processing. This enabled successful evaluations of newly developed materials for the PLC partners. In addition to the technical progress, a complex cost modeling has been successfully implemented, which allowed the cost calculation and analysis to be done with fine granularity of reference applications taking into account process, material and design options, panel utilization, and scalability. The development of the data collection tables has been the result of a collaborative work with the consortium. Different application scenarios have been analyzed, and a comparison between a mold-first and a RDL-first process flow was calculated.

Standardization of panel sizes has been discussed at various public events and a standardization group has been installed within the framework of SEMI with guidance and involvement of Panel Level Consortium members. Last, but not least, a couple of high-level events at major international conferences have been organized by the PLC 1.0. The Panel Level Consortium has been widely visible during the last two years and is considered as the best consortium on this topic. PLC 2.0 will continue the R&D work of PLC 1.0, but will also focus on some specific targets. Additionally, Fraunhofer IZM will
for high-performance computing. Future electronic systems will be based on increased functionality requiring not only the integration of different technologies, but also optimization with respect to reduced carbon footprint and low energy consumption. The highest reliability and long lifetime for autonomous systems will be key for economic success — keeping an optimized balance between cost and performance.

Looking ahead to neuromorphic computing and photonics

Neuromorphic computing may replace the von Neumann architecture with the so-called neuromorphic architecture for some applications. Inspired by the neural network of the human brain, a neuromorphic computer processes tasks in a flexible, artificial neural network, minimizing computer processes tasks in a flexible, artificial neural network, minimizing process and transfer times, as well as the amount of data that needs to be transferred. To build a neuromorphic computer requires new hardware concepts. Computer performance will be increased by the use of specialized and reprogrammable processors, non-volatile memory circuits, high-speed optoelectronic or optic communication, and flexible self-learning algorithms. Together with the Kirchhoff-Institute for Physics, Fraunhofer IZM has developed a full-wafer redistribution and embedding technology as the basis for a large-scale neuromorphic hardware system. In the first phase of the project, standard redistribution technologies from wafer-level packaging were adapted to enable a high-density reticle-to-reticle routing on 200mm CMOS wafers [11]. Neighboring reticles of the CMOS chips on the wafer were interconnected using a polymer/Cu-RDL across the scribe lines with an 8µm pitch routing with final Ni/Ai I/O-pads (Figure 5).

For final electrical connection to the board, the wafers were placed into mechanical fixtures and the I/Os of all reticles were touched by elastomeric connectors. With that concept, neuromorphic systems based on full wafers could be assembled and tested. In order to

Figure 4: Panel-level packaging line (457mm x 305mm, some machines are 600mm x 600mm) at Fraunhofer IZM in Berlin.

Figure 5: Fabrication of neuronal networks based on wafer-scale technology: wafer with 159744 chip-to-chip connections (Fraunhofer IZM).
allow an upscaling of the system size to a large number of wafers with reasonable effort, a full-wafer embedding concept for printed circuit boards was developed. The wafers were thinned to 250µm and laminated with additional prepreg layers and copper foils into a core material (Figure 6). After lamination of the PCB panel, the reticle I/Os of the embedded wafer were accessed by microvia drilling, copper electroplating, lithography and subtractive etching of the PCB wiring structure. Hardware/software co-designs are a prerequisite for generating an assembly design kit (ADK) for new complex packages like wafer- and panel-level embedding for further miniaturization, enhanced functionality, and increased energy efficiency (Figure 7).

Photonics is expected to play an important role in the near future through optical interconnect between chips, boards, and subsystems. Such optical interconnects are a prerequisite for extreme high-speed, low-latency, high-bandwidth and highly energy efficient data transfer between chips. Some examples include high-performance computing (HPC) and edge computing systems (e.g., autonomous driving or digital industries), and for data transfer within the backbone of 5G infrastructures. This development is also an essential prerequisite for the realization of an all-optical computer. Such a computer will require the integration of photonics components like solid-state III-V lasers, photonics integrated circuits (PIC), and the development of high-performance optical input/output (I/O) technologies. The technological basis for this is – although partially already in existence – still far from industrial maturity.

References


Figure 6: Bridge from the wafer-to-the-board technology: 200mm wafers with a thickness of 250µm have been laminated with the printed circuit board process in a stack of FR4 frame, two prepreg layers and two copper foils.

Figure 7: Assembly design kit (ADK) for new complex packages like wafer- and panel-level embedding.