CCF Transactions on High Performance Computing (2022) 4:43–52
https://doi.org/10.1007/s42514-022-00093-0

REVIEW PAPER

Survey on chiplets: interface, interconnect and integration methodology

Xiaohan Ma[1,4] · Ying Wang[1,2] · Yujie Wang[1,3] · Xuyi Cai[1,4] · Yinhe Han[1,3]

Received: 1 September 2021 / Accepted: 23 February 2022 / Published online: 31 March 2022

Abstract

With the end of Moore's Law and Dennard scaling, it has become increasingly difficult to implement high-performance computing systems on a monolithic chip. The chiplet technology that integrates multiple small chips into a large-scale computing system through heterogeneous integration is one of the important development directions of high-performance computing. Chiplet-based systems have huge advantages over monolithic chip in terms of design and manufacturing cost and development efficiency. In this survey, we summarized the concept and history of chiplet and introduce the critical technology needed to implement chiplet-based system. Finally, we discuss several future research directions of chiplet-based system.

Keywords Chiplet · Packaging · Interconnection · Die-to-die · Network-on-chip · Deep neural network

1 Introduction

Recently, chiplet-based systems with 2-D, 2.5-D or 3-D integration technology is getting a lot of attention. As shown in Fig. 1, these design methods split the system into smaller chiplets, and then integrate heterogeneous or homogeneous chiplets through advanced packaging technology. A chiplet is a functional integrated circuit block, which is usually independently designed under the most suitable process technology node. Different chiplets are flexibly designed under different environments and integrated according to the application requirements.

Chip design and manufacturing based on advanced process nodes are facing two major problems: cost and performance. In the past 50 years, semiconductor process technology nodes had developed in accordance with the predictions of Moore's Law (Moore 2006). However, in recent years, as Moore's Law and Dennard scaling (Dennard et al. 1974) has slowed down, the time between new process technology nodes has become longer and longer, which skyrockets the advanced design cost. For example, the cost of chip design under the 7 nm process technology node is as high as $298 million, while the cost of the 5 nm process technology node reaches $542 million, causing an increase of 82%. In contrast, the early transition from 65 to 40 nm process technology node only caused an increase of 32% in cost (Lau 2021). On the other hand, in the absence of device scaling, it is becoming more and more difficult to integrate more transistors on a monolithic chip. If the transistor size stays the same, the only way to integrate more functional units is to increase the chip area. However, large chips are more prone to defect in the manufacturing process, which greatly reduces wafer yield and increases chip costs. Therefore, monolithic chips are increasingly unable to deliver sufficient performance for emerging compute-intensive workloads such as machine learning.

✉ Ying Wang
wangying2009@ict.ac.cn

Xiaohan Ma
maxiaohan@ict.ac.cn

Yujie Wang
wangyujie@ict.ac.cn

Xuyi Cai
caixuyi20b@ict.ac.cn

Yinhe Han
yinhes@ict.ac.cn

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[2] State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[3] Research Center for Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[4] University of Chinese Academy of Sciences, Beijing, China

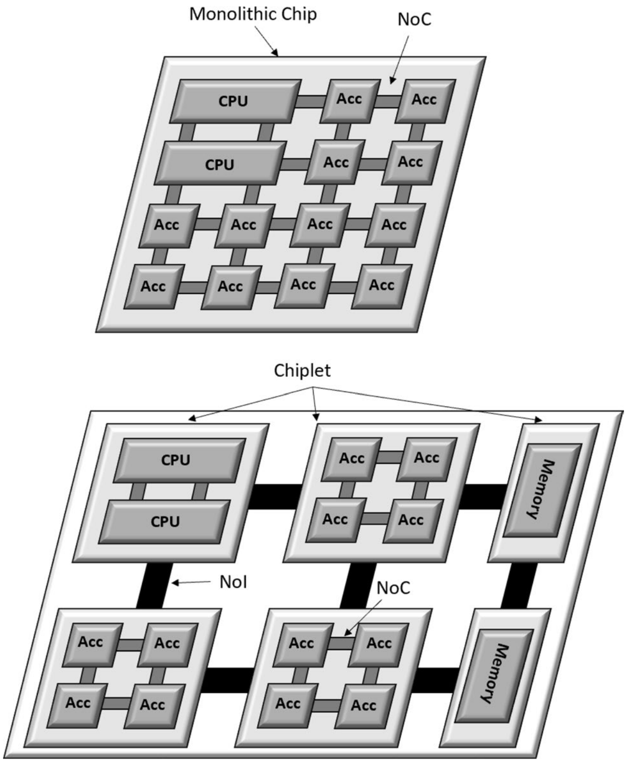


Fig. 1 Monolithic chip and interposer based multi-chiplet system

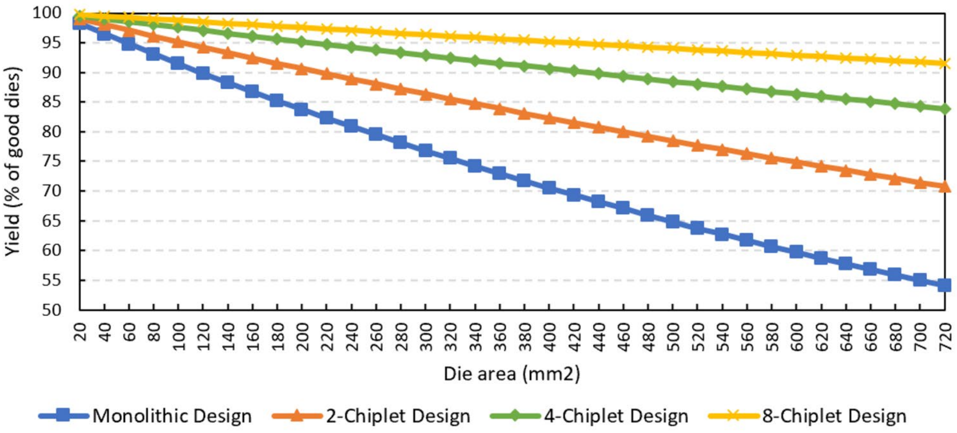
The rise of chiplet-based systems provides a path towards high-performance and cost-effective computing systems. Current chiplet-based systems mainly have the advantages of cost, area and efficiency to solve the above-mentioned problems of monolithic chips.

Cost. Chiplet technology can reduce the cost of chip design and manufacturing from three aspects: wafer yield, mixed process technology nodes and chiplet design reuse. Firstly, The chiplet-based system integrates smaller chiplet dies into a complete chip system through advanced packaging processes. The small area of chiplet causes less manufacturing defect impact and high area utilization on circular wafers. Therefore, the chiplet-based system has a higher wafer yield compared to a monolithic chip. Secondly, monolithic chip design usually requires the entire architecture to be implemented under the same process technology node. But for chiplet-based systems, different chiplet modules can adopt different process technology nodes according to their own requirements, and then these modules are integrated by advanced packaging. This method of mixing process technology nodes reduces the proportion of dies using advanced technology processes nodes in the entire design, thereby reducing costs. Third, different chiplet modules are usually designed separately and independently. The chip designer can select appropriate chiplet modules from the mature designs and integrate them into the target design. This chiplet module reuse method can greatly reduce the cost of design and verification.

Area. High-performance chips need to integrate more transistors, resulting in a larger chip area which is difficult to achieve due to the yield limitations of monolithic design. The mismatch between large-area chips required by high-performance computing systems and low yield of semiconductor manufacturing forms the "area wall" phenomenon. However, the chiplet technology greatly increases the limit of areas of chips by integrating lots of small chiplets into a large chip system. Under Murphy's Model of Die Yield (Murphy 1964), assuming the defect density is 0.09/mm^2, the additional area overhead caused by cross-die communication of chiplets is 10%, Fig. 2 shows a comparison of the yield of monolithic design with 2-chiplet, 3-chiplet, and 4-chiplet designs. From the figure, we can find that when the total area of the chip system reaches 720 mm^2, the monolithic design has a yield of 54%, while the 8-chiplet design increases the yield to 92%. The chiplet design makes super-large-area chip systems possible and alleviates the "area wall" phenomenon.

Efficiency. Through the reuse of mature chiplet modules, the chiplet method can greatly improve the efficiency of chip

Fig. 2 A comparison of yields of monolithic design and chiplet design



development. There are two types of chiplet reuse methods: heterogeneous reuse and homogeneous reuse. For heterogeneous reuse, designers can focus on the implementation of the core chiplet module, and reuse the previous verified chiplet design on other chiplet modules or buy them from other companies. This chiplet reuse method also simplifies the system upgrade by only replacing the chiplet module that has become the bottleneck of the system, while keeping other modules unchanged and avoiding the redesign of the entire system. The representative example of heterogeneous integration is the Zen2 (Naffziger et al. 2020) architecture released by AMD in 2019 which contains two type of heterogeneous chiplets, IO die for interconnection and CCD for computing. The server version and desktop version of the Zen2 architecture share the same CCD design but have different IO die designs. AMD also reuses the IO die as the × 570 chipset. For homogeneous reuse, designers integrates multiple identical chiplet modules in a chip design. This reuse method simplifies the design of scalable systems. To meet the different performance and power requirements from mobile to server, the chiplet method simply increases the number of chiplets integrated into the chip architecture. Many deep learning processors use homogeneous reuse to achieve high scalability, such as Simba (Zimmer et al. 2019) which achieves performance scalability from 0.32 to 128 TOPS by integrating different numbers of homogeneous computing cores.

This paper presents a survey on chiplet-based systems and is structured as follows. Section 2 recalls a summary of the history of chiplet and related technologies. Section 3 reviews the technologies used in chiplets, including interfaces, interconnections, packaging, etc. Section 4 shows the future prospects of chiplet technology. Finally, Sect. 5 concludes the paper.

2 The development of chiplet

In this section, the evolution of the chiplet is summarized starting as 2-D multi-chip modules (MCM) technology on substrates in the 1970s to recent chiplet heterogeneous integration on a silicon interposer.

2-D chiplet integration. Multi-Chip Modules (MCM) technology (Wong et al. 1999) that appeared in the 1970s is one of the sources of chiplet technology. MCM technology integrates multiple dies into a complex chip system on organic or other substrates, as shown in Fig. 3a. In advanced packaging technology taxonomy, this technology is also referred as 2-D chiplet integration. Historically, semiconductor manufacturers such as Intel and IBM have developed a series of high-performance processor products based on MCM technology. Intel's first dual-core × 86 processor, Pentium D (Manusharow et al. 2006), is based on MCM technology. The most famous MCM-based product in recent years is AMD's Ryzen processors (Suggs et al. 2020). By separating the IO logics and the processor core logics and modularizing the processor core logics into CCX modules, the cost of the Ryzen processor has been greatly reduced.

2.5-D chiplet integration. As shown in Fig. 3b, 2.5-D chiplet integration replaces the substrate of MCM technology with a silicon-based interposer which has through-silicon vias (TSVs) (Sunohara et al. 2008) connecting the upper

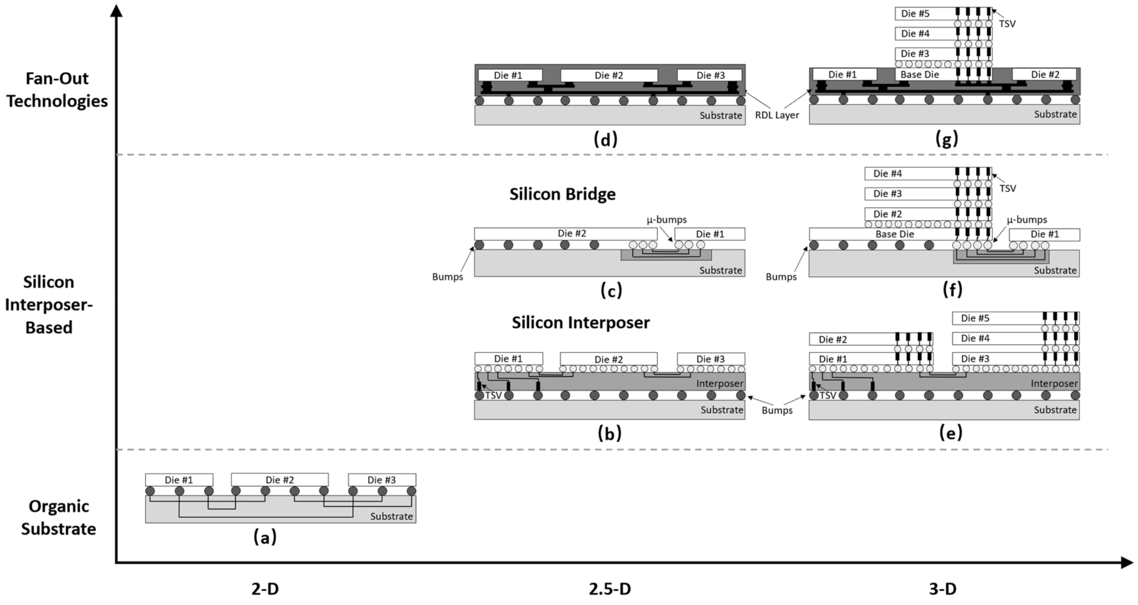


Fig. 3 Classification of chiplet integration technology and chiplet packaging technologies, summarized from (Lau 2021)

chiplets and lower substrate. Compared with the substrate, the interposer can provide high-density die-to-die connections, providing higher bandwidth for data communication between chiplets. Xilinx launched the Virtex-7 2000 T (Lenihan et al. 2013), an FPGA product using 2.5-D chiplet integration in 2011. The Virtex-7 2000 T integrates 4 pieces of FPGA dies implemented at the 28 nm process technology node on a passive silicon interposer implemented at the 65 nm process technology node.

3-D chiplet integration. Unlike 2-D and 2.5-D chiplet integration which only have the die spread on the plane, and the 3-D chiplet integration also stacks dies in the vertical direction, as shown in Fig. 3e. The history of 3-D chiplet integration can be traced back to the "Three Dimensional Circuit Element R&D Project" (Kada et al. 2015) proposed by the Research and Development Association for Future (New) Electron Devices in 1981. Subsequently, NEC, Fujitsu, Intel, TSMC, and other semiconductor manufacturers have launched research on 3-D integration. In 2020, Intel released Lakefield (Gomes et al. 2020), a hybrid × 86 computing architecture based on Foveros 3-D chiplet integration technology. Lakefield integrates a computing die implemented at the 10 nm process technology node and an active interposer base die implemented at the 22 nm process technology node. Different from the passive interposer in 2.5-D chiplet integration, Lakefield's active interposer not only includes TSV connection but also includes CMOS logics for the audio codec, USB, PCI-E, etc.

Chiplet heterogeneous integration. Different from the traditional above-mentioned integration technologies that integrate multiple homogeneous dies, chiplet heterogeneous integration focuses on integrating multiple heterogeneous chiplets into a complete chip system. DARPA (US Department of Defense Advanced Research Projects Agency) has conducted research work on chiplet heterogeneous integration for 15 years in conjunction with industry (Intel, Micron, Cadence, Synopsys, etc.) and academia (Michigan University, and Georgia Institute of Technology). In 2007, DARPA proposed the Compound Semiconductor Materials on Silicon (COSMOS) (Rosker et al. 2008) plan which aims at developing a viable process of fine-scale heterogeneous integration of compound semiconductors within CMOS technology. In 2017, DARPA launched the Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) (DARPA, 2021) project which aims at building an ecosystem of modularity and reusable chiplet and a process to assemble them into a system using advanced integration technologies. In 2021, Intel announced Ponte Vecchio (Intel 2021), XPU for exascale computing and AI-based on Foveros (Ingerly et al. 2019) 3-D chiplet heterogeneous integration. Ponte Vecchio is composed of 47 chiplets which integrate more than 100 billion transistors in total. Foveros technology flexibly supports the integration of chiplets based on different process nodes in a chip system, which greatly reduces the difficulty and cost of chip design.

3 Techniques for chiplet

In order to realize chiplet-based system, academia and industry have proposed and implemented many related technologies. These technologies can be divided into four aspects: interface and protocol, packaging, EDA tool and testing. In this section, we will review these technologies.

3.1 Interfaces and protocols

An important problem in chiplet heterogeneous integration is how to establish efficient communication connections between different dies, which directly determines the performance and functionality of the entire chip system. The technologies of the die-to-die connection include the physical layer interface and the data transmission protocol. The design of interfaces and protocols needs to take into account the process technology, packaging technology, power consumption constraint, and the requirements of upper applications, etc.

3.1.1 Die-to-die physical layer interfaces

Serial interface and parallel interface are two options of die-to-die physical layer interface.

Serial interface. The serial interface only needs a pair of differential connections in the physical layer to realize data transmission. SerDes (Serializer/Deserializer) that serializes and deserializes digital data is the basic build block to realize the serial interface for die-to-die interconnection. The functions of SerDes include parallel-to-serial data conversion, serial-to-parallel data conversion, impedance matching circuitry, and clock data recovery functionality. SerDes contains a series of serial interfaces ranging from long reach to short reach, among which USR (Ultra-Short Reach) (Carusone et al. 2016) focuses on high-speed die-to-die connection in 2.5-D/3-D chiplet heterogeneous integration. USR achieves < 1 pJ/bit power consumption, nanosecond latency, and low error rate through non-return-to-zero (NRZ) signal modulation, multi-lanes data transmission, and other technologies. Rambus (2021), Synopsys (2021), Kandou (2014), and other companies have launched their own IP products of USR serial interface which usually have a signal rate of 112 Gbps. However, USR is designed for ultra-short reach where the distance between the dies is less than 10 mm, limiting its application on larger-scale chips.

Parallel interface. The parallel interface requires multiple connections for data transmission between dies. A parallel interface usually consists of hundreds of connections, hence a bandwidth similar to USR can be achieved at a lower connection rate (up to 16 Gbps). Parallel interface delivers high bandwidth, low latency, and relatively low IO power consumption, which meet the requirements of current high-performance computing applications. At present, the industry has proposed general parallel interfaces, including Intel's AIB (Kehlet 2017), Intel's MDIO (Ramm et al. 2020), TSMC's LIPINCON (Lin et al. 2020), OCP's BoW (Farjadrad et al. 2019), and dedicated interfaces for HBM (High Bandwidth Memory) (Ko et al. 2019) storage. The parallel interfaces of Intel and TSMC are silicon-based and mainly suitable for their own 3-D packaging technology. As a member of the DARPA's CHIPS project, intel is providing a royalty-free license of the AIB interface to CHIPS project participants (Intel 2020). BoW (Bunch-of-Wires) is a parallel interface that focuses on chiplet heterogeneous integration on organic substrates. However, the massive parallel interconnections on interposers make assembly more expensive and more difficult for parallel interfaces.

Serial interface and parallel interface each have their own advantages and disadvantages, as summarized in Table 1. Chiplet designers need to carefully select the optimal interface design based on the overall performance, power requirements, and area constraints of the chiplet-based system.

Table 1 Comparisons between serial interface and parallel interface (Rajendiran 2021, Schor 2020)

PHY type	Serial interface	Parallel interface
Signal Rate	Up to 224 Gbps/line	Up to 16 Gbps/line
Clocking	CDR based	Clock Forwarding
Reach	10–50 mm	Up to 5 mm
Latency	~5–10 ns	<5 ns
Power	>0.7 pJ/bit	<0.4 pJ/bit

3.1.2 Data transmission protocol

Usually, a variety of different data transmission protocols can be adapting to the physical layer interfaces. PCI-E (Mayhew et al. 2003) is a traditional chip-to-chip data transmission protocol. Intel uses PCI-E as the die-to-die data transfer protocol between the CPU and GPU in the i7 8809G MCM processor (GlobeNewswire 2017). However, PCI-E has poor support for cache coherence, which limits its further application in high-performance multi-chiplet heterogeneous integrated systems. In order to solve this problem, SiFive and the University of California, Berkeley proposed TileLink (SiFive 2017), a data transmission protocol designed for cache coherence transactions implementing a particular cache coherence policy. Any cache consistency protocol that obeys the TileLink transaction structure can work with the physical networks and cache controllers provided by TileLink. OpenCAPI (Open Coherent Accelerator Processor Interface) (Stuecheli et al. 2018), CLX (Compute Express Link) (Van Doren 2019) and CCIX (Cache Coherent Interconnect for Accelerators) (CCIX Consortium 2017) are three cache coherent interconnection protocols based on the PCI-E protocol.

3.2 Packaging technology

Building a chiplet-based system requires assembling multiple chiplets that may come from different vendors using different process technology nodes in one package. Different packaging technologies have different support for the interconnection interface, which will affect the performance and power consumption of the system and other design metrics. As shown in Fig. 3, modern 2-D/2.5-D/3-D packaging technologies are mainly divided into three categories: 2-D packaging based on organic substrates, 2-D/2.5-D/3-D packaging based on silicon interposer and 2-D/2.5-D/3-D RDL-based (Redistribution Layer) packaging using fan-out technology.

Substrate-based packaging. Figure 3a shows organic substrate-based 2-D packaging technology which is the most widely used and successfully commercialized. This packaging method is mainly composed of two layers: chiplet and substrate. Chiplets are connected to the substrate with underfill by high destiny solder bumps through SMT (surface mount technology) and flip chip technology. The connection between chiplets is achieved by wiring in the substrate. This method can implement a very large-scale chiplet-based system at a low cost (Suggs et al. 2020).

Silicon interposer-based packaging. Silicon interposer-based packaging technology adds a layer of silicon interposer between chiplet dies and the substrate to accommodate the connection between chiplets, as shown Fig. 3b, e. Interposer uses TSV to connect the upper chiplet dies and the lower substrate. Chiplets are connected to the interposer by high destiny micro bumps through flip chip technology. There are two types of the interposer, passive interposer and active interposer. The difference between them is that there are no CMOS devices in the passive interposer. Because the micro bump pitch can be as small as 0.4 um, silicon interposer-based packaging technology has a much higher IO density than substrate-based packaging technology. CoWoS (Chip-on-Wafer-on-Substrate) (Lin et al. 2013) by TSMC is a famous silicon interposer-based packaging. The larger area of the interposer results in higher manufacturing costs. In order to solve the cost problem of interposer, Intel introduced the silicon bridge technology EMIB (Embedded Multi-die Interconnect bridge) (Mahajan et al. 2016) that reduces the interposer to a tiny piece of silicon at the edge of chiplets, as shown in Fig. 3c, f.

RDL-based packaging. As shown in Fig. 3d, g, the RDL-based package uses a redistribution layer interposer to connect the upper chiplet dies and the lower substrate. The redistribution layer adds metal and dielectric layers onto the surface of the wafer to re-route the I/O layout and carry the die-to-die interconnections. This packaging technology is meant for high-density, high-performance, and large chiplet-based systems. The representative work of RDL-based packaging is TSMC's InFO (Integrated Fan-Out) (Lin et al. 2016) technology.

Table 2 summarizes the characteristics of the three packaging technologies.

3.3 EDA tools

The EDA tools with functions of exploration, design, implementation, validation are essential for the chiplets. Therefore, industry and academia have launched research on EDA tools for chiplet.

In industry, Synopsys launched 3DIC Compiler (Synopsys 2020) which is a unified platform for end-to-end multi-die integration in a package. It can efficiently handle thousands of inter-die interconnections of chiplet-based systems that traditional EDA tools cannot handle. Cadence proposed 3D-IC Design Solutions (Cadence 2021) for the planning, implementation, and signoff of 3D-IC designs. In addition to traditional integrated circuit EDA tool manufacturers, start-ups such as MZ Technologies (2014) for 3-D chiplet package co-design have also appeared.

In academia, Kim et al. (2020) from the Georgia Institute of Technology proposed an EDA flow for interposer-based 2.5-D chiplet integration. The EDA flow which encompasses architecture, circuits, and package is built on a series of commercial tools. The usability of the process has been demonstrated through the complete implementation of a 64-core RISC-V chiplet system. Kabir et al. (2020) from the University of Arkansas proposed a Chip-Package Co-Design flow for implementing 2.5-D systems using existing commercial chip design tools. Lan et al. (2020) from the Agency for Science, Technology and Research proposed an automated chiplet-based codesign flow for 2.5-D system-in-package. Park et al. (2020) from the Georgia Institute of Technology proposed a complete EDA flow and design strategies targeting active interposer-based 2.5-D ICs. The EDA processes proposed by these research institutes are basically based on commercial tools.

Table 2 Comparisons of chiplet packaging technologies, summarized from (Lau 2021)

Type	Substrate-based	Silicon interposer-based	RDL-based
IO density	Low	High	High
Latency	High	Low	Low
3-D Extensibility	Low	High	High
Power	High	Low	Low
Yield	High	Low	Low
Cost	Low	High	High
Routing layers	High	Medium	Low

3.4 Testing

A key challenge in the chip design process is to find defects early in the production phase. For the traditional monolithic chip, the semiconductor industry has developed a complete product testing process. However, the discrete design of chiplet-based system brings new challenges to the test process. A set of standardized testing infrastructures and methodologies is the key to efficient testing of chiplet dies from different vendors.

The IEEE Std 1838 (IEEE 2020) published in 2020 defines a set of standards for 3-D IC testing. IEEE Std 1838 defines two test interfaces: serial test interface and parallel test interface. The parallel test interface is optional, which provides a high-bandwidth data access mechanism. IEEE Std 1838 supports the test of (1) intra-die circuitry and (2) inter-die interconnects in both (a) pre-stacking and (b) post-stacking situations in a hierarchy of testing. For the connection between dies, IEEE Std 1838 mainly focuses on TSV.

4 Future of chiplet

Although there have been many commercial products, commercial tools, and academic research on chiplets, the design and manufacture of chiplets still face many challenges. This section will look forward to the future and show some unresolved problems and potential opportunities.

4.1 Standards

As mentioned earlier, the current chiplet-based system has many different technical roadmaps in terms of interconnection interfaces, data transmission protocols, and packaging technologies. These technologies often lack interoperability and standardization, which hinders the further development of chiplet technology. Fortunately, many companies and organizations have seen this and started to work on the standardization of chiplet technology.

The Open Compute Project (OCP) community (OCP 2011) promoted by Facebook and other companies proposed the Open Domain-Specific Architecture (ODSA) (Vinnakota et al. 2020) project that aims to establish open physical and logical die-to-die interfaces for chiplets. The ultimate goal of ODSA is to establish a set of open interface standards for interoperable chiplet so that dies from different vendors can be freely integrated. Optical Internetworking Forum

(OIF 1998) focuses on optical networking and has proposed a variety of serial interfaces for chiplet systems in recent years. Joint Electron Device Engineering Council (JEDEC) specified standards for HBM (JEDEC, 2021) storage and proposed OpenHBI (ODSA Wiki 2021) for chiplet die-to-die communication in conjunction with ODSA.

These standards are still under development, and the standardization of chiplet has a long way to go.

4.2 NoC and NoI of chiplet

In the chiplet-based system, each chiplet has network-on-chip (NoC) to connect the resources in the chip, and there is also a network-on-interposer (NoI) in the interposer to connect the chiplets. The two-level interconnection of NoC and NoI increases the complexity of system interconnection design. Therefore, designing an efficient network on a chip becomes one of the challenges of chiplet-based system design. Jerger et al. (2014) from the University of Toronto proposed the NoC architectures for active silicon interposers. Kannan et al. (2015) from the University of Toronto proposed NoC architectures for passive silicon interposers. Yin et al. (2018) from AMD proposed an NoC design method to avoid deadlock routing in multi-chiplet systems. Pano et al. (2019) from the Drexel University proposed an NoC architecture for 3-D chiplet system with an active interposer. Kite (Bharadwaj et al. 2020) introduced a family of chiplet topologies which decouple NoI from NoC and higher throughputs than previous works. Zheng et al. (2020) from George Washington University designed a versatile and flexible chiplet NoC architecture that can be dynamically reconfigured into disjoint sub-NoCs. Wang et al. (2021) proposed an NoI architecture that can adapt to the communication mode of the target neural network workload.

4.3 DNN on chiplet

Deep neural networks are one of the most important compute-intensive workloads in the past few years. As a chip design method that can effectively improve performance, chiplet technology is quickly applied to the design of neural network accelerators. Simba (Shao et al. 2019, Zimmer et al. 2019, 2020) is a scalable deep learning processor with MCM-based architecture proposed by NVIDIA. By integrating different numbers of chiplets on the substrate, Simba can achieve high scalability from 0.32 to 128 TOPS. Manticore (Zaruba et al. 2020) is a chiplet architecture with 4096 risc-v cores for ultraefficient floating-point computing. Manticore contains four computing chiplets and four 8 GiB HBM storage chiplets. Centaur (Hwang et al. 2020) is a chiplet-based accelerator for sparse recommendation systems.

Performing efficient DNN calculations on chiplet-based systems requires reasonable allocation of computation tasks and data among different chiplets. This requires the co-design of the tiling and scheduling scheme for computing tasks and the chiplet hardware architecture. NN-Baton (Tan et al. 2021) is an automatic tool that aims to guide the chiplet granularity exploration and workload orchestration on different computation levels.

5 Conclusions

This paper introduces the overview and development history of chiplet. As an important design method for high-performance computing in the post-Moore law era, chiplet technology has received more and more attention in recent years. Both academia and industry have launched researches on chiplet technology. Although these researches have achieved certain results and supported the development of a large number of commercial products, chiplet technology still faces many challenges. This paper summarizes the interconnection interfaces, data transmission protocols, packaging, simulation, testing and EDA tools that aim at chiplet-based systems. Many companies, organizations and academic institutions are working on the standards of these technologies to build an ecosystem of interoperable chiplets. This article also surveys the important research directions of chiplet in the future.

References

Bharadwaj, S., Yin, J., Beckmann, B., Krishna, T.: Kite: a family of heterogeneous interposer topologies enabled via accurate interconnect modeling. In: Proceedings of the 57th ACM/IEEE Design Automation Conference, pp. 1–6. (2020)

Cadence: 3D-IC Design Solutions (2021). https://www.cadence.com/en_US/home/solutions/3dic-design-solutions.html

Carusone, A.C., Dehlaghi, B., Beerkens, R., Tonietto, D.: Ultra-short-reach interconnects for package-level integration. In: Proceedings of the IEEE Optical Interconnects Conference, pp. 10–11. (2016)

CCIX Consortium: Cache Coherent Interconnect for Accelerators (2017). http://www.ccixconsortium.com

DARPA: Common heterogeneous integration and ip reuse strategies (chips) (2021) https://www.darpa.mil/program/commonheterogeneous-integration-and-ip-reuse-strategies.

Dennard, R.H., Gaensslen, F.H., Yu, H., Rideout, V.L., Bassous, E., LeBlanc, A.R.: Design of ion-implanted MOSFET's with very small physical dimensions. IEEE J. Solid State Circ. **9**(5), 256–268 (1974). https://doi.org/10.1109/JSSC.1974.1050511

Farjadrad, R., Kuemerle, M., Vinnakota, B.: A bunch-of-wires (BoW) interface for interchiplet communication. IEEE Micro **40**(1), 15–24 (2019)

GlobeNewswire: AMD delivers semi-custom graphics chip for new intel processor (2017). http://www.nasdaq.com/press-release/amd-delivers-semicustomgraphics-chip-for-new-intel-processor-20171106-00859

Gomes, W., Khushu, S., Ingerly, D.B., Stover, P.N., Chowdhury, N.I., O'Mahony, F., Balankutty, A., Dolev, N., Dixon, M.G., Jiang, L., Prekke, S.: 8.1 Lakefield and Mobility Compute: A 3D Stacked

10nm and 22FFL Hybrid Processor System in 12× 12mm 2, 1mm Package-on-Package. In: Proceedings of the IEEE International Solid-State Circuits Conference, pp. 144–146 (2020)

Hwang, R., Kim, T., Kwon, Y., Rhu, M.: Centaur: A chiplet-based, hybrid sparse-dense accelerator for personalized recommendations. In: Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture, pp. 968–981. (2020)

IEEE: IEEE standard for test access architecture for three-dimensional stacked integrated circuits. (2020). https://doi.org/10.1109/IEEESTD.2020.9036129.

Ingerly, D.B., Amin, S., Aryasomayajula, L., Balankutty, A., Borst, D., Chandra, A., Cheemalapati, K., Cook, C.S., Criss, R., Enamul, K., Gomes, W.: Foveros: 3D integration and the use of face-to-face chip stacking for logic devices. In: Proceedings of the IEEE International Electron Devices Meeting, pp. 19–6 (2019)

Intel: intel/aib-phy-hardware (2020). https://github.com/intel/aib-phy-hardware

Intel: New Intel XPU Innovations Target HPC and AI (2021). https://www.intel.com/content/www/us/en/newsroom/news/new-intel-xpu-innovations-target-hpc-ai.html

JEDEC: High Bandwidth Memory (HBM) DRAM|JEDEC (2021). https://www.jedec.org/standards-documents/docs/jesd235a

Jerger, N.E., Kannan, A., Li, Z., Loh, G.H.: Noc architectures for silicon interposer systems: why pay for more wires when you can get them (from your interposer) for free?. In: Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture, pp. 458–470. (2014)

Kabir, M.A., Peng, Y.: Chiplet-package co-design for 2.5 D systems using standard ASIC CAD tools. In: Proceedings of the 25th Asia and South Pacific Design Automation Conference, pp. 351–356. (2020)

Kada, M.: Research and development history of three-dimensional integration technology. In: Three-Dimensional Integration of Semiconductors, pp. 1–23 (2015)

Kandou: XSR/USR interface analysis including chord signaling options (2014). https://kandou.com/assets/downloads/presentation-XSR-USR-Interface-Analysis-Including-Chord-Signaling-Options.pdf

Kannan, A., Jerger, N.E., Loh, G.H.: Enabling interposer-based disintegration of multi-core processors. In: Proceedings of the 48th Annual IEEE/ACM International Symposium on Microarchitectur, pp. 546–558 (2015)

Kehlet, D.: Accelerating innovation through a standard chiplet interface: the advanced interface bus (AIB). Intel White Paper (2017).

Kim, J., Murali, G., Park, H., Qin, E., Kwon, H., Chekuri, V.C.K., Rahman, N.M., Dasari, N., Singh, A., Lee, M., Torun, H.M.: Architecture, chip, and package codesign flow for interposer-based 2.5-d chiplet integration enabling heterogeneous ip reuse. IEEE Trans Very Large Scale Integr (VLSI) Syst 28(11): 2424–2437 (2020)

Ko, H.G., Shin, S., Kye, C.H., Lee, S.Y., Yun, J., Jung, H.K., Lee, D., Kim, S., Jeong, D.K.: A 370-fJ/b, 0.0056 mm 2/DQ, 4.8-Gb/s DQ receiver for HBM3 with a baud-rate self-tracking loop. In: Proceedings of the Symposium on VLSI Circuits, pp. C94–C94. (2019)

Lan, J., Nambiar, V.P., Sabapathy, R., Dutta, R., Chong, C.T., Rotaru, M.D., Lin, K.K., Bhattacharya, S., Chai, K.T.C., Do, A.T.: An automatic chip-package co-design flow for multi-core neuromorphic computing SiPs. In: Proceedings of the IEEE 22nd Electronics Packaging Technology Conference, pp. 77–80. (2020)

Lau, J. H.: Semiconductor advanced packaging. Springer (2021)

Lenihan, T.G., Matthew, L., Vardaman, E.J.: Developments in 2.5 D: The role of silicon interposers. In: Proceedings of the IEEE 15th Electronics Packaging Technology Conference, pp. 53–55. (2013)

Lin, M.S., Huang, T.C., Tsai, C.C., Tam, K.H., Hsieh, K.C.H., Chen, C.F., Huang, W.H., Hu, C.W., Chen, Y.C., Goel, S.K., Fu, C.M.: A 7-nm 4-GHz Arm[1]-core-based CoWoS[1] chiplet design for high-performance computing. IEEE J. Solid-State Circ. **55**(4), 956–966 (2020)

Lin, M.S., Tsai, C.C., Chang, C.H., Huang, W.H., Hsu, Y.Y., Yang, S.C., Fu, C.M., Chou, M.H., Huang, T.C., Chen, C.F., Huang, T.C.: An extra low-power 1Tbit/s bandwidth PLL/DLL-less eDRAM PHY using 0.3 V low-swing IO for 2.5 D CoWoS application. In: Proceedings of the Symposium on VLSI Technology, pp. C16–C17 (2013)

Lin, M.S., Tsai, C.C., Hsieh, C.H., Huang, W.H., Chen, Y.C., Yang, S.C., Fu, C.M., Zhan, H.J., Chien, J.Y., Li, S.Y., Chen, Y.H.: A 16nm 256-bit wide 89.6 GByte/s total bandwidth in-package interconnect with 0.3 V swing and 0.062 pJ/bit power in InFO package. In: Proceedings of the IEEE Hot Chips 28 Symposium, pp. 1–32. (2016)

Mahajan, R., Sankman, R., Patel, N., Kim, D.W., Aygun, K., Qian, Z., Mekonnen, Y., Salama, I., Sharan, S., Iyengar, D., Mallik, D.: Embedded multi-die interconnect bridge (EMIB)—a high density, high bandwidth packaging interconnect. In: Proceedings of the IEEE 66th Electronic Components and Technology Conference, pp. 557–565. (2016)

Manusharow, M., Hasan, A., Chao, T.W. Guzy, M.: Dual die Pentium D package technology development. In: Proceedings of the 56th Electronic Components and Technology Conference, pp. 7 (2006)

Mayhew, D., Krishnan, V.: PCI Express and advanced switching: evolutionary path to building next generation interconnects. In: Proceedings of the 11th Symposium on High Performance Interconnects, pp. 21–29. (2003).

Moore, G.E.: Cramming more components onto integrated circuits, reprinted from electronics. IEEE Solid State Circ. Soc. Newslett. **11**(3), 33–35 (2006). https://doi.org/10.1109/N-SSC.2006.4785860

Murphy, B.T.: Cost-size optima of monolithic integrated circuits. Proc. IEEE **52**(12), 1537–1545 (1964)

MZ Technologies: Monozukuri-MZ Technologies Genio (2014). https://www.monozukuri.eu/

Naffziger, S., Lepak K., Paraschou M., Subramony M.: AMD chiplet architecture for high-performance server and desktop products. In Proceedings of the IEEE International Solid- State Circuits Conference, pp. 44–45. (2020).

OCP: Home » Open Compute Project (2011) https://www.opencompute.org/

ODSA Wiki: Server/ODSA–OpenCompute (2021). https://www.opencompute.org/wiki/Server/ODSA

OIF: OIF (1998). https://www.oiforum.com/

Pano, V., Kuttappa, R., Taskin, B.: 3D NoCs with active interposer for multi-die systems. In: Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip, pp. 1–8. (2019)

Park, H., Kim, J., Chekuri, V.C.K., Dolatsara, M.A., Nabeel, M., Bojesomo, A., Patnaik, S., Sinanoglu, O., Swaminathan, M., Mukhopadhyay, S., Knechtel, J.: Design 25-D ICs and study of RISC-V architecture with secure NoC. IEEE Trans Comp Pack Manuf Technol **10**(12), 2047–2060 (2020)

Rajendiran K.: Die-to-die interface PHY and controller subsystem for next generation chiplets (2021). https://semiwiki.com/semiconductor-services/openfive/298127-die-to-die-interface-phy-and-controller-subsystem-for-next-generation-chiplets/

Rambus: 40G USR and C2C SerDes PHYs - Interface IP | Rambus (2021). https://www.rambus.com/interface-ip/serdes/40g-usr-and-c2c-serdes-phys/

Ramm, P., Franzon, P., Garrou, P., Swaminathan, R., Vivet, P., Badaroglu, M.: Heterogeneous integration and chiplet assembly–all between 2D and 3D. (2020)

Rosker, M.J., Greanya, V., Chang, T.H.: The DARPA compound semiconductor materials on silicon (COSMOS) program. In: Proceedings of the IEEE Compound Semiconductor Integrated Circuits Symposium, pp. 1–4. (2008)

Schor D.: OCP bunch of wires (2020). A new open chiplets interface for organic substrates. https://fuse.wikichip.org/news/3199/ocp-bunch-of-wires-a-new-open-chiplets-interface-for-organic-substrates/

Shao, Y.S., Clemons, J., Venkatesan, R., Zimmer, B., Fojtik, M., Jiang, N., Keller, B., Klinefelter, A., Pinckney, N., Raina, P., Tell, S.G.: Simba: Scaling deep-learning inference with multi-chip-module-based architecture. In: Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, pp. 14–27. (2019)

SiFive: SiFive TileLink specification (2017). https://sifive.cdn.prismic.io/sifive%2Fcab05224-2df1-4af8-adee-8d9cba3378cd_tilelink-spec-1.8.0.pdf

Stuecheli, J., Starke, W.J., Irish, J.D., Arimilli, L.B., Dreps, D., Blaner, B., Wollbrink, C., Allison, B.: IBM POWER9 opens up a new era of acceleration enablement: OpenCAPI. IBM J. Res. Dev. **62**(4/5), 8–1 (2018)

Suggs, D., Subramony, M., Bouvier, D.: The AMD "Zen 2" processor. IEEE Micro **40**(2), 45–52 (2020)

Sunohara, M., Tokunaga, T., Kurihara, T., Higashi, M.: Silicon interposer with TSVs (through silicon vias) and fine multilayer wiring. In: Proceedings of the 58th Electronic Components and Technology Conference, pp. 847–852. (2008)

Synopsys: 3DIC Compiler (2020). https://www.synopsys.com/implementation-and-signoff/3dic-design.html

Synopsys: DesignWare Die-to-die PHY IP solutions|Synopsys (2021). https://www.synopsys.com/designware-ip/interface-ip/die-to-die.html

Tan, Z., Cai, H., Dong, R., Ma, K.: NN-Baton: DNN workload orchestration and chiplet granularity exploration for multichip accelerators. In Proceedings of the ACM/IEEE 48th Annual International Symposium on Computer Architecture, pp. 1013–1026 (2021)

Van Doren, S.: HOTI 2019: compute express link. In: Proceedings of the IEEE Symposium on High-Performance Interconnects, pp. 18–18. (2019)

Vinnakota, B., Agarwal, I., Drucker, K., Jani, D., Miller, G.L., Mittal, M., Wang, R.: The open domain-specific architecture. IEEE Micro. (2020)

Wong, C.P., Michelle, M.: Wong: recent advances in plastic packaging of flip-chip and multichip modules (MCM) of microelectronics. IEEE Trans. Compon. Packag. Technol. **22**(1), 21–25 (1999)

Wang, M., Wang, Y., Liu, C., Zhang, L.: Network-on-interposer design for agile neural-network processor chip customization. In: Proceedings of 58th ACM/IEEE Design Automation Conference (2021)

Yin, J., Lin, Z., Kayiran, O., Poremba, M., Altaf, M.S.B., Jerger, N.E., Loh, G.H.: Modular routing design for chiplet-based systems. In: Proceedings of the ACM/IEEE 45th Annual International Symposium on Computer Architecture, pp. 726–738 (2018)

Zaruba, F., Schuiki, F., Benini, L.: Manticore: A 4096-Core RISC-V chiplet architecture for ultraefficient floating-point computing. IEEE Micr. **41**(2), 36–42 (2020)

Zheng, H., Wang, K., Louri, A.: A versatile and flexible chiplet-based system design for heterogeneous manycore architectures. In: Proceedings of the 57th ACM/IEEE Design Automation Conference, pp. 1–6 (2020)

Zimmer, B., Venkatesan, R., Shao, Y.S., Clemons, J., Fojtik, M., Jiang, N., Keller, B., Klinefelter, A., Pinckney, N., Raina, P., Tell, S.G.: A 0.11 pj/op, 0.32–128 tops, scalable multi-chip-module-based deep neural network accelerator with ground-reference signaling in 16nm. In: Proceedings of the Symposium on VLSI Circuits, pp. C300–C301 (2019)

Zimmer, B., Venkatesan, R., Shao, Y.S., Clemons, J., Fojtik, M., Jiang, N., Keller, B., Klinefelter, A., Pinckney, N., Raina, P., Tell, S.G.: A 0.32–128 TOPS, scalable multi-chip-module-based deep neural network inference accelerator with ground-referenced signaling in 16 nm. IEEE J Sol State Circ **55**(4), 920–932 (2020)

Xiaohan Ma received the B.S. degrees from the Huazhong University of Science and Technology, Wuhan, Hubei, China, in 2016. Currently, he is working toward a Ph.D. degree in the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. His current research interests include domain-specific accelerators, deep learning acceleration, and edge computing.

Ying Wang (M'14) received the B.S. and M.S. degrees in Electrical Engineering from Harbin Institute of Technology, in 2007 and 2009 respectively, and the Ph.D. degree of computer science from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2014. He is currently an associate professor at ICT, CAS. His research interests include computer architecture and VLSI design, specifically memory system, energy-efficient architecture, and machine learning systems. He has authored and coauthored more than 100 papers in refereed journals and conferences. Dr. Wang has received Best Paper Award in ICCD 2019 and ITC-ASIA 2019, and Best Paper Nominations in ASPDAC15. He currently serves as the Associate Editor for the ACM SIGDA ENewsletter.

Yujie Wang (Member, IEEE) received B.S. and Ph.D. degrees from Nankai University in 2007 and 2017, respectively. He is currently an Associate Professor with the Center of Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences. He visited Texas A&M University as a visiting Ph.D. Student from 2014 to 2017. He has published papers in the scope of VLSI design. His research focuses on low power accelerator design, EDA tools for physical design, and hardware security.

Xuyi Cai received B.S. degrees from Wuhan University, Wuhan, China, in 2018. She is currently pursuing a Ph.D. degree with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. Her current research interests include architecture design and compilation framework of deep learning systems.

Yinhe Han (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), in 2003 and 2006, respectively. He is currently a Professor at ICT, CAS. His main research interests are microprocessor design, integrated circuit design, and computer architecture. In these research fields, he has published 100 papers. He is a Senior Member of the Chinese Computer Federation (CCF). He serves on the technical program committees of several conferences, such as DAC and HPCA. He is an Associate Editor of IEEE TRANSACTIONS ON COMPUTERS.