



IMAGE LICENSED BY INGRAM PUBLISHING

Exploring Heterogeneous Integration: Its Essence and Future Path

Christopher D. Nordquist¹ and Stanley S. Chou²,
Sandia National Laboratories

Heterogeneous integration, a process that involves the amalgamation of chipllets, enhances system performance and spurs innovation in various applications, including, but not limited to, artificial intelligence, sensors, and cloud computing.

Heterogeneous integration (HI) is the integration of chipllets using packaging technologies. Whereas a traditional system on chip (SoC) uses semiconductor technologies to integrate functionalities on a single silicon wafer, HI looks

to disaggregate the functionalities of a SoC into smaller chipllets, or use proven intellectual properties (IPs) and older technologies and repackage them into a new product built for specific applications. This approach of producing a system in package (SiP) has proven attractive, especially as the progress of Moore's law has waned. Indeed, for the bulk of our lifetime, Moore's law has allowed transistor densities to double every two years. This fueled the integration of increasingly complex functionalities into a single silicon chip and the aggregation of multiple functionalities, including the

CPU, graphics processing, custom accelerators, memory, and input-output functionalities. However, as the transistor integration approaches billions per chip, the economic and technological drivers are steering manufacturers away from monolithic integration. Here, as monolithic chip sizes swell to the size of a reticle, the yield of a monolithic SoC falls appreciably because of defects in the semiconductor manufacturing process. Consequently,

Digital Object Identifier 10.1109/MC.2023.3339364
Date of current version: 7 February 2024

this has changed the way in which manufacturers approach the design of large and complex systems. Increasingly, the trend has been toward the integration of smaller “chiplets” through packaging technology. In this model, chiplets can be selected based on the desired functionality. As each chiplet can be optimized using different process nodes, this also opens the way for the reuse of older chips and their IPs.

However, while chiplet integration is an elegant solution, it comes with its share of challenges. For example, packaging technologies and their associated problems are now intertwined with the system architecture design. Suddenly,

HI offers the tantalizing possibility of combining chiplets, individually optimized and potentially from different technologies, to construct new amalgamations at the package level.

thermal and latency issues must be managed at the millimeter-length scale, rather than that of the typical microns seen in monolithic chips. Similarly, chiplet interfaces must be standardized across multiple manufacturers, rather than in house. All of these factors complicate the economic and architectural decisions surrounding chip disaggregation and the HI of chiplets at the packaging level.

In anticipation of these integration challenges, the industry has launched several efforts to standardize chip-to-chip communication. Leading the charge are the Universal Chiplet Interconnect Express (UCIe) and Open Compute Project (OCP) standards being developed for interoperability and package-level integration. Both efforts aim to standardize how the industry operates with regard to chip-to-chip interfaces by offering an agreed-upon physical layer and the controller block. Doing so will ensure interoperability and guardrail the expected parameters of the product, including data

bandwidth, power efficiency, and latency, among others. Associated with UCIe and OCP are also the High Bandwidth Memory (HBM) standards, which are evolving to keep pace with the needs of high-performance computing (HPC) and artificial intelligence (AI). Here, the HBM3 standard was unveiled to tightly couple with expanding SiP needs.

Beyond the challenges of die-to-die interface standardization, there are also economic considerations with HI. For example, the time and cost required to architect a SiP are fundamentally different than in the traditional SoC pathway. To understand

these tradeoffs, cost and performance models have been developed. For example, Stow et al. benchmarked the viability of a SoC versus a SiP using 2.5D and 3D integration and found the latter approach to be plausible despite required changes in the design flow and ultimate architecture.¹ Feng and Ma performed a study with added cost for the standardization and optimization of die-to-die and interfaces for interoperability of chiplets from different manufacturers.² And lastly, Ahmad et al. developed an open source cost model to encompass materials, chiplet IP, SoC disaggregation, integration, and development costs versus those of the traditional monolithic chip integration.³ Like Stow and Feng, Ahmad found that SiPs were economically viable and that they allowed users to perform economic tradeoffs based on SoC die-yield tradeoffs. In the end, despite the differences in these cost models, the results consistently indicate that HI produces sufficient gross margins for it to be economically viable.

Of course, SiPs are not an attractive technology unless they offer certain performance advantages over SoCs. In this case, SiPs are distinct in that they can enable HPC by increasing the capacity of chip-to-chip communication.⁴ Here, chiplet integration addresses the die-to-die data interface and bottleneck with higher density integration that allows parallelization and low latency such that intrachip protocols can be adapted for interchip communication, greatly increasing the data throughput. Two key places where this higher throughput provides an advantage are high-speed data converter interfaces and HBM interfaces. High-speed data converter interfaces enable high-speed communication and signal processing capability, while high memory bandwidth and capacity enable HPC capability.

This ability for high-capacity chip-to-chip data transmission enables the disaggregation of integrated circuit functions, allowing for the use of microelectronics technology that is application specific. With chiplets, a processor can be realized using the latest CMOS technology, while surrounding functions, such as memory and peripherals, may be realized using lower cost and higher yielding technologies. As alluded to previously, this allows the reuse of existing chips and IPs while also allowing chips to be sourced from different foundries and suppliers, providing the opportunity for cost and schedule savings.

Chiplet approaches have been used in modern computing. One example is the Intel Ponte Vecchio GPU, which incorporates 100 billion transistors across 47 separate chiplets into a single module to serve as a building block for future HPC architectures.⁴ Each module contains microelectronics technologies realized in nodes as small as 5 nm and is capable of 45-teraflop 32-bit operation. The same manufacturing capabilities used to produce this module are also being used to produce microprocessors for consumer applications, bringing the


capability of chiplet integration to mass production.

With the capability for high-throughput chip-to-chip data transmission, chiplet technology provides the potential for incorporating hardware accelerators for machine learning, AI, or neuromorphic computing.^{5,6} Chiplets enable the hardware accelerators to be designed and manufactured using ideal and custom technologies while still maintaining low-latency communication with the central processor. Research in this area is active with universities and industry groups, who report chiplet-based accelerators for machine learning and signal processing, taking advantage of interchip data rates exceeding 1 Tb/s.^{7,8,9}

The enhanced data rates derived by improved die-to-die interfaces also offer advantages for communication between data converters and processors. In the radio-frequency (RF) domain, analog-to-digital converters (ADCs) now exceed 10 Gb/s and can produce more than 100 Gb/s of data for downstream signal processing, providing opportunities for direct sampling of RF signals with broader bandwidths. Furthermore, an ADC demonstrated operation at 12 Gb/s with 12 bits of data, for a total data rate of 144 Gb/s.^{10,11} Using the parallelization of a chiplet HI interface, these data are offloaded to the signal processing chip in 32 12-bit channels operating at 375 Mb/s. These types of data rates will be required for future radar and 6G communication.

In the even faster case of optical interconnects, data rates >8 Tb/s have been achieved by combining multiple optical interface chips with a field-programmable gate array.^{12,13,14,15} Beyond the obvious telecommunication applications, this type of optical interface can be used to enable chip-to-chip, module-to-module, and rack-to-rack networking for massively parallel computing architectures. In all, these exemplars serve to demonstrate the architectural advantages of disaggregation and individual optimization over traditional SoC designs.

Chiplet technologies present several key challenges that must be addressed. First, chiplet assemblies require simulation and co-design across multiple technologies, domains, and vendors. Second, standardization and compatibility of protocols, interfaces, and design kits across the industry are required for maximum flexibility and accessibility. Third, methods for identifying a known good die prior to assembly and the capability to replace or rework the individual die within chiplet assemblies will improve the yield and affordability of chiplet approaches. The chiplet community is working to address these and other challenges to realize the promise of chiplet technologies.

To conclude, HI offers the tantalizing possibility of combining chiplets, individually optimized and potentially from different technologies, to construct new amalgamations at the package level. These SiPs, if properly architected, can provide individually customized, application-specific products that outperform traditional SoCs. Given the slowdown of Moore's law, HI is likely the most viable path for continued progress in microelectronics. As HI technology and its surrounding technological ecosystem mature, there will likely be new technologies and applications that we do not foresee. We envision that these technologies will improve communication and interactivity among the populace, and that these advancements will improve the well-being of our society. 

REFERENCES

1. D. Stow, I. Akgun, and Y. Xie, "Investigation of cost-optimal network-on-chip for passive and active interposer systems," in *Proc. ACM/IEEE Int. Workshop Syst. Level Interconnect Prediction (SLIP)*, Jun. 1-2, 2019, 2019, pp. 1-8, doi: 10.1109/SLIP.2019.8771333.
2. Y. Feng and K. Ma, "Chiplet actuary: A quantitative cost model and

- multi-chiplet architecture exploration," in *Proc. 59th ACM/IEEE Des. Autom. Conf.*, San Francisco, CA, USA, 2022, pp. 121-126, doi: 10.1145/3489517.3530428.
3. M. Ahmad, J. DeLaCruz, and A. Ramamurthy, "Heterogeneous integration of chiplets: Cost and yield tradeoff analysis," in *Proc. 23rd Int. Conf. Thermal, Mech. Multi-Phys. Simul. Experiments Microelectron. Microsyst. (EuroSimE)*, Apr. 25-27, 2022, pp. 1-9, doi: 10.1109/EuroSimE54907.2022.9758914.
4. W. Gomes et al., "Ponte Vecchio: A multi-tile 3D stacked processor for exascale computing," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 20-26, 2022, vol. 65, pp. 42-44, doi: 10.1109/ISSCC42614.2022.9731673.
5. S. S. Iyer, S. Jangam, and B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for AI systems," *IBM J. Res. Develop.*, vol. 63, no. 6, pp. 5:1-5:16, 2019, doi: 10.1147/JRD.2019.2940427.
6. S. Mukhopadhyay et al., "Heterogeneous integration for artificial intelligence: Challenges and opportunities," *IBM J. Res. Develop.*, vol. 63, no. 6, pp. 4:1-4:1, 2019, doi: 10.1147/JRD.2019.2947373.
7. M. D. Rotaru, W. Tang, D. Rahul, and Z. Zhang, "Design and development of high density fan-out wafer level package (HD-FOWLP) for deep neural network (DNN) chiplet accelerators using advanced interface bus (AIB)," in *Proc. IEEE 71st Electron. Compon. Technol. Conf. (ECTC)*, 1 Jun./4 Jul. 2021, pp. 1258-1263, doi: 10.1109/ECTC32696.2021.00204.
8. J. A. Stevens, T. H. Pan, P. P. Ravichandiran, and P. D. Franzon, "Chiplet set for artificial intelligence," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, May 10-12, 2023, pp. 1-5, doi: 10.1109/3DIC5717.2023.10154953.
9. Z. Tan, Y. Wu, Y. Zhang, H. Shi, W. Zhang, and K. Ma, "A scalable multi-chiplet deep learning accelerator with hub-side 2.5D heterogeneous integration," in *Proc. IEEE*

- Hot Chips 35 Symp. (HCS), Aug. 27–29, 2023, pp. 1–17, doi: 10.1109/HCS59251.2023.10254703.
10. C. Hornbuckle, E. Mrozek, T. Krawczyk, and M. Lugthart, “Low-power K/Q-band digital phased array chiplet,” in *Proc. IEEE Int. Symp. Phased Array Syst. Technol. (PAST)*, Oct. 11–14, 2022, pp. 1–7, doi: 10.1109/PAST49659.2022.9975074.
 11. S. Shumarayev, A. Chan, T. Hoang, and R. Keller, “Heterogenous integration enables FPGA based hardware acceleration for RF applications,” in *Proc. IEEE Hot Chips 34 Symp. (HCS)*, Aug. 21–23, 2022, pp. 1–20, doi: 10.1109/HCS55958.2022.9895615.
 12. K. Hosseini et al., “8 Tbps co-packaged FPGA and silicon photonics optical IO,” in *Proc. Opt. Fiber Commun. Conf. Exhib. (OFC)*, Jun. 6–10, 2021, pp. 1–3.
 13. K. Hosseini et al., “5.12 Tbps co-packaged FPGA and silicon photonics interconnect I/O,” in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technology and Circuits)*, Jun. 12–17, 2022, pp. 260–261, doi: 10.1109/VLSITechnologyand-Cir46769.2022.9830221.
 14. R. Mahajan et al., “Co-packaged photonics for high performance computing: Status, challenges and opportunities,” *J. Lightw. Technol.*, vol. 40, no. 2, pp. 379–392, 2022, doi: 10.1109/JLT.2021.3104725.
 15. M. Wade et al., “TeraPHY: A chiplet technology for low-power, high-bandwidth in-package optical I/O,” *IEEE Micro*, vol. 40, no. 2, pp. 63–71, Mar./Apr. 2020, doi: 10.1109/MM.2020.2976067.

CHRISTOPHER D. NORDQUIST

is with Microsystems Engineering Science and Applications, Sandia National Laboratories, Albuquerque, NM 87123 USA. Contact him at cdnordq@sandia.gov.

STANLEY S. CHOU is with

Microsystems Engineering Science and Applications, Sandia National Laboratories, Albuquerque, NM 87123 USA. Contact him at schou@sandia.gov.

IEEE Annals of the History of Computing

From the analytical engine to the supercomputer, from Pascal to von Neumann, from punched cards to CD-ROMs—*IEEE Annals of the History of Computing* covers the breadth of computer history. The quarterly publication is an active center for the collection and dissemination of information on historical projects and organizations, oral history activities, and international conferences.

www.computer.org/annals

