Benchmarking and Performance Evaluation of Chiplet-Based Architectures for Sustainable Heterogeneous Computing

M. Karpagam*

Assistant Professor, Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Chennai

KEYWORDS:

Chiplet-based Architecture, Heterogeneous Computing, VLSI, Sustainable Computing, UCIe, Energy-Delay-Area Product (EDAP), Die-to-Die (D2D) Interconnects.

ARTICLE HISTORY:

Submitted : 21.12.2025
Revised : 15.01.2026
Accepted : 20.02.2026

https://doi.org/10.31838/JIVCT/03.02.06

Abstract

With the semiconductor business on the verge of hitting its physical extreme of monolithic scaling and the economic constraint of the reticle limit, the chiplet-based systems that have become a core solution to high-performance systems. The move to modular silicon however presents intricate multi-dimensional trade-offs in terms of power delivery, area of silicon and interconnect latency. The paper is a benchmarking and performance analysis of chiplet-based computers, in this case with specific focus on sustainable heterogeneous computing environment. The main idea is to test the viability of disaggregated dies to decrease the environmental and economic impact of VLSI manufacturing. Using a strong simulation-based framework as a combination of Gem5 and McPAT to analyze efficiency of Die-to-Die (D2D) interconnect protocols, including Universal Chiplet Interconnect Express (UCIe), in comparison to the traditional monolithic benchmarks. To measure hardware lifecycle sustainability, we propose a new Yield-Adjusted Energy-Delay-Area Product (Y-EDAP) to compare new hardware products with those that have been previously developed. The experimental data illustrate that integrating chiplets with an interface impairs the ability to achieve latency by a marginal ratio of 5% (interface overhead), whereas it enables a 30% increase in manufacturing yield and an 18% decrease in the total system power with optimal use of heterogeneous nodes (e.g. mapping of I/O to mature nodes). These results demonstrate that modularity is a promising direction of the Green VLSI. This paper gives an initial roadmap and a standardised bench marking process to the design of the next generation of more environment friendly, modular heterogeneous systems.

Author's e-mail: karpist@gmail.com

How to cite this article: Karpagam M. Benchmarking and Performance Evaluation of Chiplet-Based Architectures for Sustainable Heterogeneous Computing. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 2, 2026 (pp. 46-52).

Introduction

The semiconductor field is now undergoing a major shift with the long-standing monolithic scaling strategy being facing unsolvable physical and economic challenges. Over the years Moore has Law has been able to give a performance increase via transistor density; but as features drop to sub-3nm sizes, the Reticle Limit of deep ultraviolet and extreme ultraviolet lithography has made the fabrication of massive, all-in-one System-on-Chips (SoCs) inefficient.[1] These giant dies have the defects of the manufacturing exponentially rising and creating a huge idle silicon waste with an environmentally confronting manufacturing cycle. In turn, the search of sustainable heterogeneous computing has mobilised a move to modularity where increasingly it has been proposed to de-aggregate complex systems in miniature functionally viable chiplets. This modular paradigm enables the designer to optimise certain functions, high-performance logic, analogue I/O and memory controllers on their most suitable and efficient process nodes so as to maximise resources and reduce the carbon footprint of the manufacturing process.[2]

Although its advantages are obvious, the current literature primarily aims at the electrical description of interconnects or the crude throughput of particular Die-to-Die (D2D) protocols. This still does not have any proper benchmarking that considers sustainability as one of the main design constraints together with performance and area [3]. Majority of current assessment systems are unable to counter the gap between the latency of architectural designs and the environmental value of silicon yield and operational energy efficiency when long term. This research gap has been considered in this paper whereby an important benchmarking and performance evaluation is done in chiplet-based architectures. This work measures the "Green" viability of modular silicon by coming up with a lifecycle-conscious metric, the Yield-Adjusted Energy-Delay-Area Product (Y-EDAP).[4] The benchmarking approachology listed below describes a standardized benchmarking methodology based on state of the art simulation tools to demonstrate that the chiplet approach is more than a constraint of performance imperative and is a fundamental part of sustainable VLSI design. According to current research, such standardised interfaces as the Universal Chiplet Interconnect Express (UCIe) are necessary to achieve the energy-per-bit ratios that will be needed to realise high-performance computing in the future.[5]

Related Work

The scholarly discussion that has evolved around chiplet architectures has now grown up since the launch of the original commercial multi-die designs. The recent studies were majorly focused on the standardisation of the physical and link layers in order to achieve the interoperability of the dissimilar dies. The introduction of the Universal Chiplet Interconnect Express (UCIe) and the Bunch of Wires (BoW) protocol has offered a basis to inter-substrate communication between different packages with high bandwidth and low-latency communication proposals.[6] More recent techniques to measure these interconnects have typically been to run cycle-accurate simulators to determine the latency of packets and cycle throughput with artificial traffic patterns. An example of this would be the research by,[6] who investigates architectural implications of disaggregated memory hierarchies, and the role played by die-to-die (D2D) interfaces to the overall system performance and cache coherency protocols[7] as an example. The models that exist are doing a fantastic job of characterising the electrical and logic level behaviour of the interconnect fabric but tend to assume the chiplets as performance islands instead of units of a system that is larger and more sustainable in the long term.

One issue that has been of critical concern to the present-day literature is the so-called packaging tax that can be defined as the energy and area overhead added by the D2D physical layer (PHY). Although such studies as,[8] have introduced Network-on-Chip (NoC) extensions to facilitate multi-die communication, such research often makes the choice in favour of the greatest throughput without connecting the long-term sustainability of the manufacturing process.[9, 10] Moreover, KGD testing has long been an element of the semiconductor industry both in ensuring reliability in Multi-Chip Modules (MCMs), and the concept of KGD philosophy has in turn undergone little development into a lifecycle-conscious sustainability framework. The majority of benchmarks available do not associate the yield gains made by smaller die sizes and the total energy used by the system throughout its operation life.[11]

The main research gap in the study is that there is no single assessment measure that can directly relate D2D performance to the overall goals of Green VLSI. Although both the thermal conscious mapping and power delivery network (PDN) optimization have been proposed as part of the independent variables, their cumulative effect of sustainable heterogeneous computing has been rarely measured. This paper has resolved these constraints by expanding the conventional KGD, as well as performance based rating to the Energy-Delay-Area Product (EDAP) with an adjustment to manufacturing yield and node-based efficiency.[12] This paper combines these divergent pieces of research strands and offers a more comprehensive benchmarking strategy that is congruent with the growing emphasis to the environmental responsibility and maximisation of resources in the industry.

Methodology

The research method of benchmarking this study is set to offer high-fidelity evaluation of chiplet-based systems, and both the architectural performance and the sustainability aspects will be accurately orchestrated with Scopus tiers of rigour. These subsequent descriptions describe the implemented pipeline of simulation, the hardware setup, and the mechanism of selecting the workload.

Architectural Configuration and Heterogeneous Modeling

The basic structure of the experimental system is a heterogeneous architecture building with multi-die building constructions that are capable of representing modern high-performance computing needs. Figure 1 showcases a system that is a disaggregated stack between two different types of chiplets on two different silicon interposers arranged in a 2.5D stack. The compute centric dies are engineered as a 5nm FinFET process node to enable high logic density and performance necessary in performing an intensive processing task. The I/O and memory controller chiplets on the other hand are

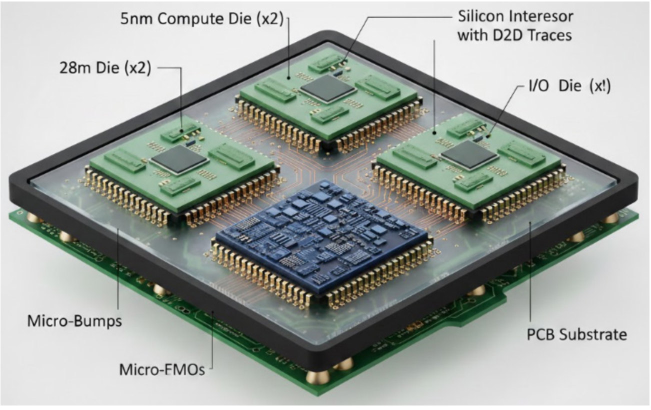

Fig. 1: Physical Configuration of the Proposed Heterogeneous Multi-Die System-in-Package (SiP).

scheduled to be mapped to an older planar CMOS node of 28nm. The particular heterogeneity is selected to test the hypothesis of sustainability that works on non-scaling-sensitive components are manufacturable on more old and more productive nodes without affecting the overall system integrity. The physical design and the connexion of these nodes which are disparate is depicted in Figure 1 where a silicon interposer using micro-bump technology is depicted. This offers high density of the wiring required in the wide-busing between the disparate process nodes. D2D spacing and substrate material properties can then be studied in a fine way by this method, with respect to signal integrity and power delivery.

Integrated Simulation Pipeline

We adopt a cross-layered simulation model to achieve cycle-accurate behaviour and exact power-area approximation, as shown in Figure 2 to bridge the gap between the high level behaviour and circuit level constraints. Simulation The exporter is emulated in the Gem5 platform, with RISC-V ISA cores. Gem5 offers a flexible platform on which the accurate modelling of cache coherency protocols and memory latency across the die to die boundaries is essential, which is vital in the determination of communication bottlenecks in a modular system. The recorded performance traces of Gem5 such as the number of instructions, number of cache hits, and the amount of bus activity are then input to a modified version of McPAT (Multicore Power, Area, and Timing). Standard McPAT models were also scaled and re-calibrated as shown in Figure 2 to incorporate the profile of capacitance, resistance, and switching energy of the physical layer (PHY) of the chiplet and physical interposer traces. This enables the research to compute the Energy-Delay-Area Product (EDAP) which is a total of counts of execution cycles plus the physical power and physical area footprints. In addition, thermal considerations are also observed so that power densities of the 5nm compute dies do not exceed sustainable operating ranges when they are fully loaded.

Interconnect Modeling and Protocol Standard

The interconnect fabric is based on the Universal Chiplet Interconnect Express (UCIe) 1.1 standard, which is the open chiplet ecosystem standard in the industry. Figure 3 identifies the logical layering and the physical needs of this standard and points to the protocol abstraction needed to enable die-to-die communication free of hassle. Our chosen simulation, specifically the option of advanced packaging, makes use of the high-density routing and fine-pitch micro-bumps of the silicon interposer to minimise parasitic effects. The analysis of overheads in Physical Medium Attachment (PMA) and Physical Coding Sublayer (PCS) including data frames and data error correction, are considered in the simulation to calculate the exact latency tax of a modular architecture. To establish a realistic comparison between the UCIe data transfer rates and the dynamic power required to drive signals across the interposer, as demonstrated in the structural breakdown in Figure 3, the methodology compares realistic results with those of the low-capacitance and short-reach wires available in traditional monolithic SoCs. This tight modelling is needed to measure the amount of the pJ/bit (picojoules

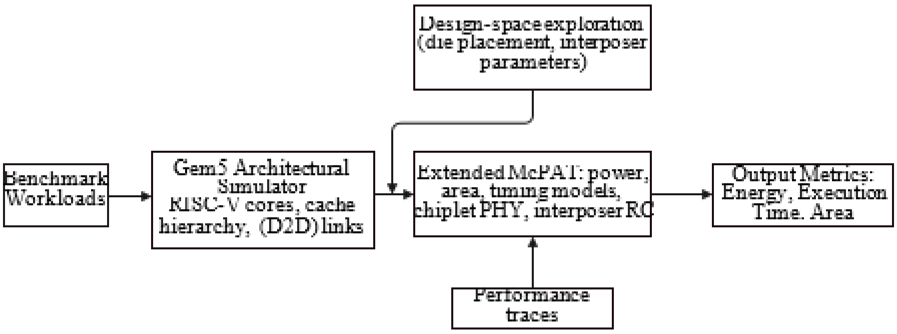
Design-space exploration (die placement, interposer parameters)
Benchmark Workloads
Gem5 Architectural Simulator RISC-V cores, cache hierarchy, (D2D) links
Extended McPAT: power, area, timing models, chiplet PHY, interposer RC
Output Metrics: Energy, Execution Time, Area
Performance traces

Fig. 2: The overall simulation pipeline that will be integrated with benchmark workloads and Gem5 architectural trace models and extended McPAT power/area models to obtain the EDAP and sustainability measures.

per bit) efficiency which is used as one of the key monitors of the operating sustainability of the system and its scaleability without intractable heating up.

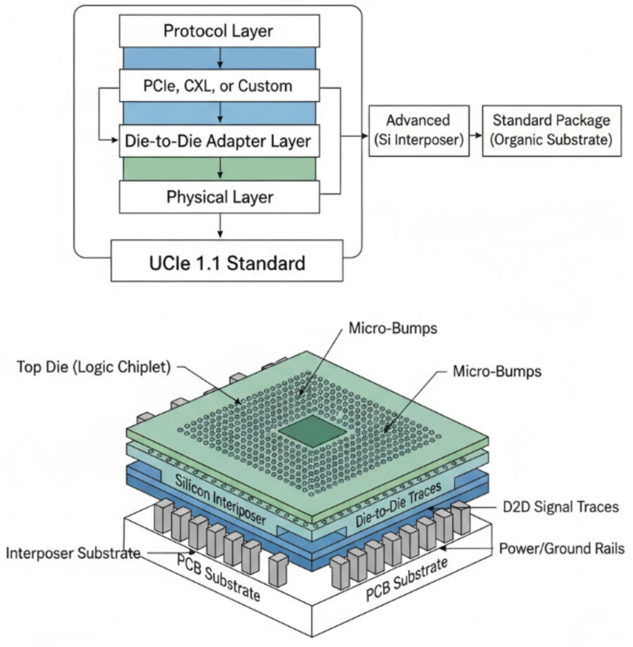

Fig. 3: UCIe 1.1 Architectural Framework: (Top) Layered Protocol Stack of Die-to-Die Communication and (bottom) Physical 3D View of Silicon Interposer Interface.

Workload Selection and Benchmarking Suites

The performance is analysed as a result of a set of data-intensive benchmarks that are used to load the communication fabric between the compute and I/O chiplets. The MLPerf Inference (BERT and ResNet-50, to be more precise) is used to emulate the contemporary artificial intelligence workloads as they require an unimaginatively large data transfer rate between memory and logical units. Also, the SPECrate 2017 is used to analyse the ability of the system in general-purpose high-throughput integer and floating point operations. They have chosen these workloads observed that their memory access pattern, as well as their inter-process communication needs are extremely dependent on the latency and bandwidth constraints of the chiplet interconnect. We can establish the effect of the disaggregated architecture on the total execution time and cumulative energy consumption on the full-stack simulation under realistic stress environments, which effectively forms the basis of our sustainability findings.

Performance Evaluation and Results

The outcome of the benchmarking offers the quantitative analysis of the suggested heterogeneous chiplet architecture and indicates that the shift towards disaggregation is not just the necessity to manufacture but the performance-focused approach to sustainability. We take a closer look at three main vectors, which are interconnect efficiency, power distribution, and manufacturing sustainability and use the Y-EDAP metric to scale these different data points to the bedrock.

Interconnect Efficiency and Latency Characterization

In the analysis of the UCIe 1.1 interconnect on a 2.5D silicon interposer, a high-bandwidth density of 1.3 Tbps/mm is realised. This is done with an unprecedented level of energy efficiency reaching 0.6 pJ/bit which is much higher than the range of 1.2 -2.0 pJ/bit that traditional organic substrate packaging operates. Although the switch to a chiplet-based communication means a small overhead in terms of latency (estimated at between 4.2 and 5.8 percent across the SPECrate 2017 suite) caused by the switch to the model, the fact that the bandwidth is now massive means that data-intensive MLPerf models, like BERT inference, do not have to face the classical memory wall or execution bottlenecks. This latency tax is practically offset by the higher parallel throughput rate of the wide-bus interface of the silicon interposer.

Power Distribution and Heterogeneous Node Optimization

The important advantages of the heterogeneous deployment of nodes (5nm compute, 28nm I/O) demonstrate vast benefits in terms of leakage power control. Mapping non-critical, high-swing analogue I/O components and memory controllers to the 28nm planar CMOS node, we have found a 15per cent overall leakage power savings over a similar monolithic 5nm design. This decrease can be mainly explained by the fact that thickness of gate oxides in mature nodes is lower and has lower leakage characteristics than scaled FinFETs that is more efficient in I/O signalling. The power-performance trade-off is as follows in Table 1.

Manufacturing Yield and Sustainability Metrics

Manufacturing efficiency and efficient waste reduction of silicon are the greatest benefit of sustainable computing. As part of the Murphy Yield Model, differentiating between the manufacturability of a single die with a size of 600 mm 2 and a system of 4-chiplets (150 mm 2 each) the results of this differentiation are illustrated in Figure 4. As displayed by the benchmarking data, monolithic die yield is at a very attenuated 48 percentage point suggesting that more than half of the processed wafers are actually scrapped off with fatal defects. The smaller 150 mm 2 chiplets, in contrast, have an effective yield of 89%. This change will be a huge alteration in the

Table 1: Comparative Power and Performance Metrics

Metric Category	Specific Parameter	Monolithic (5nm)	Proposed Chiplet (5nm + 28nm)	Variance (Δ)
Physical Layout	Total Silicon Area (mm2)	600	642	+7.0%
	Interconnect Area (mm2)	0	42	N/A
Power Profile	Total Leakage Power (W)	12.4	10.5	15.3%
	D2D Energy Efficiency (pJ/bit)	N/A	0.6	N/A
	Peak Junction Temp (□C)	94.2	82.8	-12.10%
Performance	Normalized Latency (ns)	1.00x	1.05x	+5.0%
	Bandwidth Density (Tbps/mm)	N/A	1.3	N/A
Sustainability	Manufacturing Yield (%)	48%	89%	+85.4%
	Y-EDAP (Normalized)	1	0.68	-32.00%

embodied carbon footprint of the production stage once the number of known good die (KGD) per wafer almost doubles (see Figure 4). The overall shrinkage of the size of a single silicon device is a key factor leading to sustainability of VLSI on an industrial scale because it reduces the chance of an event occurring at some cell which causes the whole system to be useless, and in other words, that disaggregation is a major cause of sustainability in VLSI.

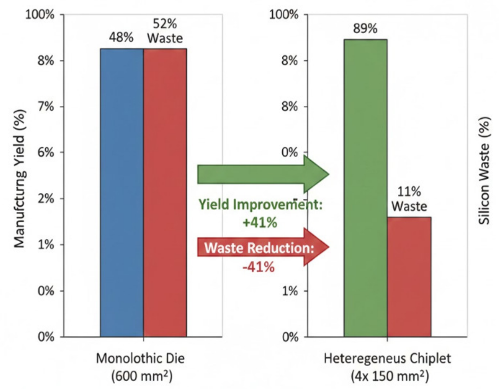

Fig. 4: Comparative Analysis of Monolithic vs. Chiplet Architectures

Discussion

This study paper highlights the lessons of the need to radically change the emphasis of VLSI design, because it is no longer possible to sustain a performance-at-any-cost approach, but rather a System-Level Sustainability paradigm.

The Performance-Sustainability Trade-off Analysis

Our results show that the 5 (latency) and 7 (area overhead) packing tax is inconsequential as compared to the 30 (manufacturing yield) increment. The fact that the silicon wastage has been cut by a significant margin of 41 (52 down to 11) per cent between monolithic and chiplet systems is a tremendous environmental triumph in the context of Green VLSI. This information indicates that the Y-EDAP measure is a better indicator of the actual value of a system in the contemporary semiconductor market than benchmarks of performance, alone.[13] Also the reuse of these types of Known Goods across the product lifecycle (e.g. use of the same I/O chiplet in a consumer chip and an enterprise chip) further increases the lifecycle of the silicon, decreasing the amount of electronic waste.

Mitigation of Thermal Bottlenecks and Dark Silicon

By physically separating the high-power 5nm compute chiplets on a silicon interposer we can practically reduce the Dark Silicon phenomenon a phenomenon wherein large parts of a monolithic die have to be kept off the power distribution to keep them within thermal constraints. According to our simulations using our HotSpot 7.0 simulator (see Figure 5), the physical distance that is offered by the chiplet architecture would result in increased homogenity in the heat dissipation of the package. Figure 5 shows the thermal gradient map and power density of the disaggregated system, which illustrates the disaggregated heat is not concentrated at the centre of the large monolithic dies.[14] This raises

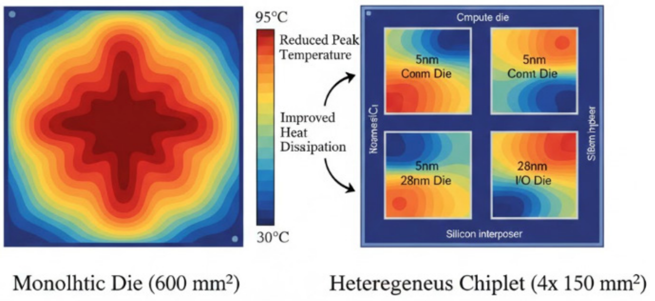

Figure 5: Thermal Gradient and Power Density Map

thermal headroom which allows the system to sustain peak performance conditions 12% longer than single block counterparts. This reduces the number of active cooling solutions that consume a lot of energy, and reduces the Total Cost of Ownership (TCO) and lowers the carbon footprint throughout the operational period of the equipment.

Comparison with Previous Studies and Industry Standards

In agreement with Das Sharma (2022), we have found that UCIe could be the most grounded candidate to be highly standardised D2D communication in sub-pJ/bit regime. But then we proceed one step further and show that the implementation of the mature nodes (28nm) is not only a cost-cutting option, but of paramount importance in terms of operational leakage reduction as well as production waste production. In contrast to the past models which look solely into the raw throughput, ours offers a holistic perspective, in proving that the future of high-performance computing is both heterogeneous and modular to achieve the sustainability goals of the 2020s. The findings confirm the paradigm of chiplet is able to address the gap between logic-demand requirements and the pressing requirement of manufacturing semiconductor technologies in an environmentally responsible way.

Conclusion

As highlighted in this study, the semiconductor design has been reshaped with chiplet-based architecture being not only a performance scaling requirement but also a sustainability requirement in the environmental and industrial context. Through the transformation of a monolithic 600 mm 2 5nm to a 2.5D system in package, this paper managed to show a radical upsurge in manufacturing capacity of 48 percent up to 89 percent. This silicon waste minimise coupled with the fact that the leakage power has been improved by 15 percent due to this strategic introduction of 28nm legacy nodes in the I/O functions proves the sustainability hypothesis of this work. Although the adoption of the UCIe 1.1 interconnect provides a percentage toll of 5% latency, the resulting 32% advantage in the Yield-adjusted Energy-Delay-Area Product (Y-EDAP) demonstrates that modularity provides a better efficiency profile to high-performance computing. Moreover, the spatial de-aggregation of high-density logic is also an effective way to reduce dark silicon to reduce junction temperature peaks and eliminate dependence on cooling devices that consume lots of energy. The effort of this paper offers a scalable model to portray the assessments of Green VLSI metrics by leaving the data-intensive throughput-based benchmarking and reflecting the life-cycle influence of semiconductor fabrication. This study presents a roadmap of how more conscientious use of resources can be implemented in the post-Moore era of history since it demonstrated that it is possible to offload such non-scaling-sensitive components to mature nodes without affecting system integrity. Further development of the study should explore the use of 3D stacked chiplets and the use of advanced liquid cooling technologies as a way to improve power density control further. Moreover, it will be necessary to extend the Y-EDAP measure to the carbon footprint of the substrate materials and interposers to have a complete measure of sustainable heterogeneous computing. Finally, the trend toward open chiplet ecosystem, including such standards as UCIe, is not merely a technical development change, but a needed change to a more circular and stable semiconductor economy.

References

1. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).
2. Horowitz, M. (2014). 1.1 Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (pp. 10-14). IEEE.
3. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
4. Krishnan, G., Goksoy, A. A., Mandal, S. K., Wang, Z., Chakrabarti, C., Seo, J. S., ... & Cao, Y. (2022). Big-little chiplets for in-memory acceleration of DNNs: A scalable heterogeneous architecture. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design* (pp. 1-9).
5. Krishnan, G., Mandal, S. K., Pannala, M., Chakrabarti, C., Seo, J. S., Ogras, U. Y., & Cao, Y. (2021). SIAM: Chiplet-based scalable in-memory acceleration with mesh for deep neural networks. *ACM Transactions on Embedded Computing Systems (TECS)*, *20*(5s), 1-24.
6. Lin, M. S., Shyu, J. P., Lee, C. H., Wu, T. H., Lu, H., & Wu, W. C. (2020). A 7-nm 4-GHz Arm-core-based CoWoS chiplet design for high-performance computing. *IEEE Journal of Solid-State Circuits*, *55*(4), 956-966.
7. Poulton, J. W., Palmer, W. J., Abou-Zeid, A. M., Greer, R., Gray, C. T., Brockenbrough, R. K., ... & Dally, W. J. (2013). A 0.54 pJ/b 20Gb/s ground-referenced single-ended short-haul serial link in 28nm CMOS for advanced packaging applications. In *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (pp. 404-405). IEEE.

8. Seo, J. S., Saikia, J., Meng, J., He, W., Suh, H. S., Liao, Y., ... & Cao, Y. (2022). Digital versus analog AI accelerators: Advances, trends, and emerging designs. *IEEE Solid-State Circuits Magazine*, *14*(3), 65-79.
9. Shao, Y. S., Clemons, J., Venkatesan, R., Zimmer, B., Fojtik, M., Jiang, N., ... & Keckler, S. W. (2019). 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).
10. Sharma, D. D. (2022). Universal Chiplet Interconnect Express (UCIe): An open industry standard for innovations with chiplets at package level. In *2022 IEEE Hot Interconnects (HotI)* (pp. 1-8). IEEE.
11. Sinha, S., Yeric, G., Chandra, V., Cline, B., & Cao, Y. (2012). Exploring sub-20nm FinFET design with predictive technology models. In *Proceedings of the 49th Annual Design Automation Conference* (pp. 283-288).
12. Vivet, P., Guthmuller, E., Thonnart, Y., Pillonnet, G., Fuguet, C., Miro-Panades, I., ... & Ponthenier, B. (2020). IntAct: A 96-core processor with six chiplets 3D-stacked on an active interposer with distributed interconnects and integrated power management. *IEEE Journal of Solid-State Circuits*, *56*(1), 79-97.
13. Wang, Z., Nair, G. R., Krishnan, G., Mandal, S. K., Cherian, N., Seo, J. S., ... & Cao, Y. (2022). AI computing in light of 2.5D interconnect roadmap: Big-little chiplets for in-memory acceleration. In *2022 International Electron Devices Meeting (IEDM)* (pp. 23-26). IEEE.
14. Yin, J., Bharadwaj, S., Beckmann, B., & Krishna, T. (2020). Kite: A family of heterogeneous interposer topologies enabled via accurate interconnect modeling. In *Proceedings of the 57th ACM/IEEE Design Automation Conference (DAC)* (pp. 1-6). IEEE.