

Original Article

Design and Integration of HBM/LPDDR Subsystems in Chiplet SoCs

VenkataKrishna Brahmam Pogadadanda

Senior RTL Design Engineer at AMD-Xilinx, USA

Received: 08-06-2024

Revised: 08-20-2024

Accepted: 09-07-2024

Published: 09-12-2024

ABSTRACT

Chiplet-based SoC architectures are basically game changers to deliver the increasing needs of performance, scalability, and energy-efficiency of modern-day computing devices. Breaking down the highly complex SoCs into small functional chiplets not only helps in increasing yield and reducing the overall manufacturing cost but also allows the integration of different types of processors, accelerators, and memory subsystems in the form of chiplets. This way, High-Bandwidth Memory (HBM) and Low-Power Double Data Rate (LPDDR) memories continue to be very instrumental in delivering the bandwidth, latency, and power efficiency that are essential for AI, HPC, mobile platforms, and edge device applications. Nevertheless, bringing HBM and LPDDR subsystems in chiplet architectures not only increases interconnect and thermal management complexities but also poses challenges such as signal integrity, power delivery, memory coherence, and packaging constraints. The proposed methodology in this paper for HBM/LPDDR subsystem integration in chiplet SoCs is based on the use of highly optimized die-to-die interfaces, advanced packaging technologies, hierarchical memory management, and adaptive power-control strategies. A case study is carried out that shows a better bandwidth utilization, smaller access latencies, and higher energy-efficiency than the traditional monolithic designs. These results emphasize scalable memory-centric chiplet architectures that can serve as the base for future heterogeneous.

KEYWORDS

Chiplet SoC, HBM, LPDDR, Memory Subsystem, Advanced Packaging, Interposer, Die-to-Die Communication, Power Optimization, AI Accelerators, Heterogeneous Integration.

1. INTRODUCTION

1.1. Background and Evolution of Chiplet SoCs

Demand for better computational performance, energy efficiency and functional scalability has been a key driver behind the evolution of system-on-chip (SoC) architectures. For a long time, the focus was on producing monolithic SoCs with all processing, memory, and communication modules consolidated on a single silicon die. While this offered very high integration density, transistor scaling became increasingly difficult and expensive as Moore's Law started to slow down. In addition, the complexity of manufacturing advanced technology nodes made things worse. As a result, the chip industry has shifted its focus towards heterogeneous computing and modular chiplet-based architectures, where a number of smaller dies are combined in a single package performing specialized functions. Chiplet ecosystems allow integration of different modules such as CPUs, GPUs, AI accelerators, I/O controllers, and memory by employing advanced packaging techniques like interposers and 3D stacking. At the same time, the intensity of data movement in AI, ML, cloud computing, and mobile systems have led to a serious memory bottleneck. HBM and LPDDR have been created as main memory solutions to address these problems. HBM is characterized by high bandwidth and parallelism which is highly appreciated in AI accelerators and HPC systems, whereas LPDDR is a low-power technology that is ideal for mobile and edge devices. In fact, these technologies are becoming the building blocks of chiplet SoCs of the future.

1.2. Challenges in HBM and LPDDR Integration

Although HBM and LPDDR offer many benefits, combining the memory subsystems with chiplet-based SoCs still has several technological issues to be resolved. Bandwidth and latency management is the main concern since high-speed communication between compute chiplets and memory chiplets demands efficient die-to-die interconnects with little transmission delay. Besides that, the data transfer at multi-gigahertz operating frequencies is made reliable with great difficulty due to signal integrity problems like crosstalk, electromagnetic interference, timing skew, and high-frequency losses. On the other hand, power consumption is the third important aspect to be considered since memory interfaces account not only for dynamic but also leakage power. At the same time, temperature control is also a major problem in compactly integrated packages and in multiple-stacked HBM architectures, especially where thermal coupling and heat concentration between chiplets lead to reliability and performance degradation. Besides, manufacturing complexity through-silicon vias (TSVs), interposers, and assembly reliability constraints is increased when advanced packaging methods like 2.5D and 3D integration are introduced. Besides, another important aspect that needs to be addressed is the scalability of interconnects as the future chiplet ecosystems will neither run without communication frameworks standardization (UCIe) nor without efficient Network-on-Chip (NoC) architectures to support the seamless integration of heterogeneous components.

1.3. Problem Statement and Research Motivation

Existing monolithic memory architectures are failing to meet the expanding performance and efficiency requirements of AI, HPC, and data-heavy applications. Chiplet-based designs, while offering scalability and flexibility, still face significant challenges when it comes to integrating HBM and LPDDR memory subsystems without compromising on bandwidth, latency, power consumption,

thermal efficiency, silicon area, and manufacturing cost. More importantly, there is no current comprehensive method that can align system integration of memory across different chiplet setups and application areas in an optimized manner. These constraints lead to a necessity for research on effective memory-centric chiplet structures that can deliver a good performance/energy efficiency trade-off. The increasing number of companies adopting chiplet technology such as AMD, Intel, NVIDIA, Apple, and TSMC, emphasizes the need for scalable memory integration methods in practice. Fresh applications like AI accelerators, autonomous systems, cloud data centers, and edge computing platforms all call for memory architectures offering high bandwidth, low power, and sustainable thermal levels. The major reasons for doing this research are the scalable cost-effectiveness, enhancing manufacturing yield by chiplet partitioning, heterogeneous integration flexibility, and next generation computing systems' performance per watt optimization.

1.4. Objectives and Organization of the Study

This paper aims to explore the architectural aspects of HBM and LPDDR memory subsystems, as well as their integration in chiplet-based SoCs, by analyzing the problems related to bandwidth, latency, signal integrity, thermal behavior, power management, and advanced packaging technologies. Subsequently, a framework for the integration of HBM and LPDDR subsystems into heterogeneous chiplet environments is provided. The proposed approach is firstly tested by simulation and then corroborated by case studies of bandwidth utilization, latency reduction, power efficiency, and thermal performance. Meanwhile, the paper provides a detailed review of interconnect technologies and packaging solutions contributing towards simultaneous scalability and reliability of chiplet systems. The remaining part of this paper is sectioned into a number of chapters that expound upon the memory subsystem architectures, the integration challenges, the proposed methodologies, the experimental analysis, case studies, results, and future avenues of research in AI accelerators, HPC platforms, mobile processors, and edge computing devices.

2. LITERATURE REVIEW

2.1. Evolution of Memory Architectures

The progress in memory designs is a key factor in enabling computing at a massive scale and with excellent response time, eventually supporting advanced domains of computer science such as artificial intelligence (AI), high-performance computing (HPC), cloud systems, and mobile devices. Refined Double Data Rate (DDR) memory technologies have led the way from DDR to DDR5, offering at every turn increased bandwidth, faster operation frequencies, and more reliability. DDR5 supplements channel efficiency and significantly escalates data transfer rates which is a great match for server and enterprise workloads. On the downside, DDR memories consume quite some power and are subjected to scalability problems for data-intensive applications.

Low-Power Double Data Rate (LPDDR) memory technologies have developed from LPDDR3 to LPDDR5X in order to meet the requirements of energy efficiency in mobile and edge devices. LPDDR memory layouts rely on low-power operating modes, highly selective refresh, and Dynamic Voltage and Frequency Scaling (DVFS) for energy reduction without compromising performance. LPDDR5X

also raises transfer speed and efficiency substantially making it very suitable for smartphones, edge AI platforms, and automotive systems.

High-Bandwidth Memory (HBM) came into existence to break the bandwidth barriers in AI and HPC environments. The move from HBM to HBM2E and HBM3 has led to the development of stacked memory architectures featuring Through-Silicon Vias (TSVs) and wide I/O interfaces. HBM3 offers a very high bandwidth at the same time lowering power per bit transferred which is why it is very efficient for AI accelerators and large-scale computing platforms. While HBM3's bandwidth may be much better than DDR5 and LPDDR5, at the same time, it needs advanced packaging technologies along with higher manufacturing costs.

Table 1: Comparison of Memory Technologies

Memory Technology	Key Feature	Advantages	Limitations	Main Applications
DDR5	High-speed conventional DRAM	Improved bandwidth and reliability	Higher power consumption	Servers, Enterprise Systems
LPDDR5/LPDDR5X	Low-power memory	Energy efficient, low standby power	Lower bandwidth than HBM	Smartphones, Edge AI, Automotive
HBM3	Stacked high-bandwidth memory	Very high throughput, low power per bit	High packaging cost and thermal complexity	AI Accelerators, HPC

2.2. Chiplet-Based SoC Architectures

Due to the slowing down of Moore's Law and the increased complexity of building chips, chiplet-based System-on-Chip (SoC) designs have become an alternative to the usual monolithic designs. In a chiplet architecture, several smaller chips are combined within one package to carry out different functions, e.g., a chip for processor, chip for memory management, chip for graphics acceleration, and chip for input/output communication. This design with smaller parts not only helps in producing good chips more frequently but also reduces the cost of development and allows for different components to be built using different technology nodes to be integrated seamlessly.

Scalability is one of the major advantages of chiplet architectures as designers can not only reuse but also combine chiplets to make custom systems tailored for specific uses. A number of top-notch industry technologies are a clear testimony to this. AMD's Infinity Fabric is a medium which facilitates fast communication between the computing and memory chiplets. Intel's Foveros is a technology that allows 3D stacking of logic and memory dies for compact integration. TSMC's CoWoS technology employs silicon interposers for high-bandwidth communication in AI accelerators, NVIDIA's Grace Hopper architecture is a combination of CPUs, GPUs, and HBM for very large-scale AI processing. Such innovations emphasize the role of chiplet ecosystems in coming heterogeneous computing platforms.

2.3. HBM and LPDDR Subsystem Design

HBM subsystem design focuses on TSV-based stacked memory and wide I/O communication channels. Multiple DRAM dies are vertically stacked to achieve massive memory bandwidth with improved energy efficiency. Silicon interposers are commonly used to connect HBM stacks with compute chiplets, enabling high-speed communication. The primary advantages of HBM include very high throughput, reduced power per transferred bit, and efficient support for AI and HPC workloads. However, HBM integration also creates thermal hotspots, increases packaging complexity, and raises manufacturing cost due to advanced TSV and interposer technologies.

LPDDR subsystem design emphasizes low-power operation and energy-efficient data transfer. LPDDR memories use DVFS, sleep modes, and optimized signaling techniques to reduce power consumption in battery-powered devices. These features make LPDDR suitable for smartphones, edge computing systems, and automotive electronics. Although LPDDR provides lower bandwidth than HBM, it offers excellent energy efficiency for mobile-oriented applications.

2.4. Interconnect Technologies and Research Gaps

Chiplet and memory subsystem communications can work seamlessly if resilient efficient interconnect technologies are deployed. SoC Network Architectures (NoC) can be thought of as scalable data routing techniques that can be used to route data traffic within the chiplet ecosystem. Universal Chiplet Interconnect Express (UCIe) is a totally new communication standard for die-to-die communication which is designed to enhance the interoperability of chiplets. In addition to that, to increase the bandwidth of advanced packaging systems and levels of integration, other emerging technologies like Embedding Multi-Die interconnect Bridge (EMIB), silicon interposers, and organic substrates are being widely used.

In spite of a few recent advances, many research challenges still lie ahead. The current HBM and LPDDR studies are largely limited to each of the memory technologies separately. Limited studies exist that propose unified integration frameworks with chiplet SoCs combining both memory technologies. Thermal-aware optimization techniques are far from perfect for extremely integrated systems. Therefore scalable orchestration methods for multi-chiplet memory communication are also lacking. Overall, these issues must be resolved to invent the future energy-efficient yet high-performance chiplet architectures.

3. PROPOSED METHODOLOGY

3.1. Overview of the Proposed Architecture

Basically, the proposed approach involves using a hybrid High-Bandwidth Memory (HBM) and Low-Power DDR (LPDDR) subsystems integrated into a chiplet-based System-on-Chip (SoC) architecture. The main aim of the design is to offer very high computational performance, the scalability to expand capacity over time, and low energy consumption for a variety of applications including artificial intelligence (AI), high-performance computing (HPC), cloud systems, edge computing as well as mobile platforms. The architectural idea is to combine the very high bandwidth

feature of HBM with the extremely low power consumption feature of LPDDR to create a well-balanced and versatile memory hierarchy.

The endeavor is trying to build a complex combination of processing chiplets, memory controllers, HBM stacks, LPDDR modules, and a very fast connecting fabric. Computing chiplets comprise CPUs, GPUs, and AI accelerators that are the main units capable of carrying out the execution of the workloads. HBM stacks are located extremely close to the computing chiplets using highly efficient packaging methods such as the silicon interposer to provide communication that is not only very fast but also very wide in terms of bandwidth. LPDDR modules represent the low-power memory resources that are both physically and logically optimized for mobile and very energy-efficient operations. A main memory controller oversees and manages data exchange and the allocation of the workloads between the processing units and memory subsystems.

Besides that, the architecture is capable of workload-aware memory management too. Specifically, data-heavy AI and HPC workload products will be forwarded to HBM, whereas low-power, background tasks will be performed using LPDDR memory. Such a hybrid format leads to significantly improved performance-per-watt and considerably lower power consumption while still being very scalable and flexible.

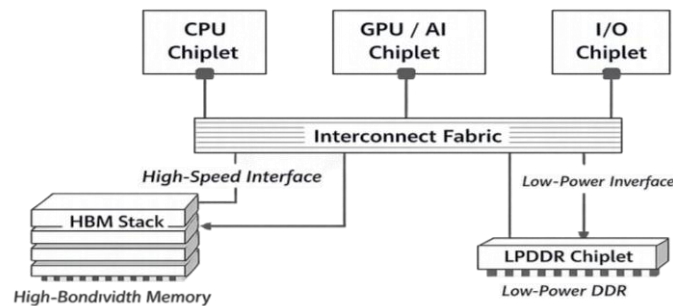


Fig 1 - Hybrid HBM and LPDDR Integration in Chiplet-Based SoC Architecture

3.2. System-Level Design Flow and HBM Integration

The system-level design flow starts with performing a requirement analysis to identify the application needs such as bandwidth, latency, power efficiency, and thermal limits. Then, based on these, a SoC could be divided into multiple chiplets to improve modularity, reduce manufacturing complexity, and increase yield. Subsequently, different memory allocation methods are designed to distribute HBM and LPDDR resources according to the prioritization of the workloads.

The selection of an interconnect topology is done to guarantee the efficient communication between the compute chiplets and the memory modules. Along with that, advanced packaging technologies like 2.5D integration and using a silicon interposer help to reduce the communication delay and at the same time, to increase the bandwidth density. Moreover, thermal-aware floorplanning techniques are also utilized in this stage to spread the heat over the package and thus reduce the heat concentration. Also, power optimization techniques are planned and included at this point to help in

reducing the power consumption while at the same time ensuring performance is not negatively impacted.

The aim of the HBM integration is to provide high memory throughput and at the same time, to keep the communication overhead at a minimum level. As the HBM uses memory dies stacked one on top of another and connected via Through-Silicon Vias (TSVs), special attention should be paid to the TSV placement in order to avoid routing congestion and at the same time, to raise the electrical performance. In addition, efficient interposer routing is a must to meet the requirements of high-speed communication between the compute chiplets and the HBM stacks.

Wide I/O communication channels are required for parallel data transfer in AI accelerators and HPC systems. Therefore, dynamic bandwidth allocation schemes are introduced to efficiently allocate memory resources according to workload demand. This way, memory utilization is improved and performance deterioration due to bottlenecks is prevented even during the highest operating conditions.

3.3. LPDDR Integration and Interconnect Optimization

The core of LPDDR integration approach centers on energy-efficient and low power operation. LPDDR memories serve as secondary memory resources which are combined with the high performance HBM subsystem. Through Dynamic Voltage and Frequency Scaling (DVFS) methods, changes in voltage and operating frequency based on workload are achieved, and this leads to cutting down on power consumption when it is not required.

The method involves adaptive refresh mechanisms and low-power standby modes to continue the trend of energy efficiency improvements. These characteristics make LPDDR the right choice for the smartphone, automotive electronics, edge AI and other battery-powered systems markets. While LPDDR offers lower bandwidth than HBM, it is very power-efficient which makes it very suitable for mobile-oriented applications.

A strong set of interconnects is very essential in bringing together various chiplets and memory modules in general communication. The line of action mentioned here takes advantage of chiplets interconnecting through UCIe together with the network-on-chip (NoC) architecture for very fast data transfer. The use of packet scheduling algorithms is a way to prioritize the most latency-sensitive traffic as well as to keep the network load balanced dynamically. Optimized routing methods result in reduced congestion and increased bandwidth availability for the entire chiplet environment.

3.4. Thermal-Aware Design and Reliability Verification

Thermal management remains a critical challenge in the development of chiplet SoCs, most notably in stacked HBM architectures. This work presents various thermal-aware design strategies that, combined, have the potential to alleviate heat accumulation and enhance the overall system reliability over the lifetime. Temperature measuring units are dispersed without a particular pattern on both the compute chiplets and the memory stacks to record the temperature changes during

operation. Similarly, the concept of using heat spreading materials and cooling methods are also a part of the overall design to improve the overall heat removal performance.

Methods of dynamic thermal management adjust the operating frequency along with spreading the workload based on the changes in temperature that have been detected instantaneously. On the other hand, thermal-aware floorplanning is an effective way of decreasing the thermal coupling between neighboring chiplets thereby enhancing the stability of the entire system.

Signal integrity checks, error correction, fault tolerance are examples of techniques to confirm reliability of the system. Evaluating crosstalk, electromagnetic interference, timing skew, and high-frequency losses allows a crosschecking of the signal path for reliable high-speed communications. Performance and timing closure are validated by simulating different workload scenarios and testing them for performance, thermal behavior, and power efficiency. This kind of verification gives confidence that the proposed hybrid HBM and LPDDR memory subsystem can deliver scalable, reliable, and energy-efficient performance for future computing systems.

4. CASE STUDY

4.1. Case Study Overview

This study shares the story of the new System-on-Chip (SoC) design using chiplets for artificial intelligence (AI) accelerator application and assesses its performance. The proposed system comprises four computing chiplets, two High-Bandwidth Memory version 3 (HBM3) modules, and one Low-Power DDR5 (LPDDR5) unit, all combined in a heterogeneous chiplet environment. This case study aims to assess the efficacy of the hybrid memory fusion in aspects like throughput enhancement, latency minimization, power expenditure improvement, and thermal condition preservation in AI-related work.

Compute chiplets contain AI processing cores, vector processors, and also dedicated accelerators for deep learning and matrix-intensive operations. Two HBM3 stacks are combined to provide the very high memory bandwidth that AI training and inference require. On the other hand, the LPDDR5 subsystem serves as the energy-efficient memory for background tasks, control operations, and less important things. The chiplets are connected with a UCIe-based die-to-die interconnect network which uses 2.5D packaging technology.

The architecture employs a hybrid memory hierarchy where high-bandwidth transactions are routed to HBM3 while low-power transactions are served by LPDDR5. Such a strategy gives the ability to distribute workload effectively over memory resources and results in a significant leap in the overall system performance. Case study shows that advanced packaging, interconnect optimization, and thermal-aware design techniques reveal how these aspects dictate the performance and reliability of the chiplet SoC.

4.2. System Specifications and Architecture

The chiplet SoC, which was implemented, utilizes 5 nm process technology to achieve a high transistor density while at the same time an improvement in energy efficiency. The HBM3 subsystem delivers a memory bandwidth of around 1.2 TB/s, which makes it possible to have very fast data transfers for AI and high-performance computing workloads. The LPDDR5 subsystem operates at a data rate of 8533 MT/s and is intended for low-power communication with very low standby energy consumption.

The link between compute chiplets and memory subsystems is based on the Universal Chiplet Interconnect Express (UCIe) protocol that allows for standardized high-speed die-to-die connectivity. From the architectural point of view, it is manifested in a 2.5D Chip-on-Wafer-on-Substrate (CoWoS) type of packaging where silicon interposers are used to combine compute and memory chiplets into one single package. This packaging method improves signal integrity, reduces communication latency, and supports high-density interconnect routing.

Architecturally, there are dedicated memory controllers that, along with the scheduling of workloads, also handle memory allocation and bandwidth management between the HBM3 and LPDDR5 resources. The HBM3 stacks are located close to the compute chiplets to minimize communication overhead and maximize data throughput, on the other hand, LPDDR5 modules are connected as secondary low-power memory resources. Besides that, the system is equipped with thermal sensors and power management modules that keep track of operating conditions and adjust energy efficiency on the fly.

Table 2: System Specifications of the Proposed Chiplet SoC

Component	Specification	Purpose/Application
Process Technology	5 nm	High transistor density and energy efficiency
Compute Chiplets	4 AI Compute Chiplets	AI processing and acceleration
Memory Type	2 × HBM3 + 1 × LPDDR5	Hybrid high-performance and low-power memory
HBM3 Bandwidth	~1.2 TB/s	High-speed AI and HPC workloads
LPDDR5 Data Rate	8533 MT/s	Low-power background operations
Interconnect Protocol	UCIe	High-speed die-to-die communication
Packaging Technology	2.5D CoWoS with Silicon Interposer	Improved signal integrity and integration
Target Applications	AI Accelerators, HPC, Edge Computing	High-performance heterogeneous computing

4.3. Implementation Flow and Experimental Setup

The proposed chiplet SoC method first partitions the chiplets identifying the processing units, memory controllers, and memory subsystems as separate, independent functional chips. This partitioning method not only enhances scalability and manufacturing yield but also provides the capability to integrate heterogeneous components in a flexible way. After partitioning, memory

mapping is used to decide which part of the workload is to be handled by HBM3 or LPDDR5, considering their bandwidth and power characteristics.

Then, routing optimization is carried out to minimize inter-chiplet communication delay and enhance bandwidth utilization in the interposer network. The routing plan is designed to reduce congestion and enable efficient packet transmission through UCIe communication links. Also, thermal simulation run at floorplanning stage assists in finding possible heat concentration areas which can be mitigated by chiplet placement optimizing heat dissipation.

Various commercially available tools are chosen for implementation and study purposes. Cadence and Synopsys are used for chiplet design, floorplanning, timing analysis, and interconnect verification. ANSYS RedHawk comes into play when thermal and power integrity analysis is to be performed and MATLAB and SystemC are chosen for performance simulation and workload modeling. The setup of the experiments measures a variety of system parameters including throughput, communication latency, power consumption, and thermal profile during AI inference and training workloads.

The system behavior is analyzed under different workload conditions simulating peak processing activity and low-power modes. The work experiments the hybrid memory subsystem of HBM3 and LPDDR5 against the traditional monolithic memory architectures in terms of performance. This is one of the ways to assess the role of hybrid memory integration in enhancing system-level efficiency and scalability.

4.4. Observations and Performance Analysis

The experimental analysis presents evidence of HBM3 integration resulting in a drastic enhancement of the overall system throughput performance on the basis of its ultra-high bandwidth and parallel communication features. AI tasks like matrix manipulation, DNN training, and extensive data movements run faster on HBM3-based systems than on regular DDR systems. Additionally, the communication delay of memory accesses is as well minimized with the HBM3 stacks residing closely with the compute chiplets.

The LPDDR5 part of the memory system is a major factor in reducing the standby and background power consumption. Background and standby tasks that are given to the LPDDR5 are not only less energy-consuming but also quite capable of delivering a decent level of performance. Besides, a dynamic division of work between HBM3 and LPDDR5 is done towards overall power efficiency improvement, and at the same time, it makes balanced utilization of memory resources possible.

The findings of the research are that the communication latency and bandwidth utilization are significantly affected by interposer routing strategies. Great routing leads to a reduction of congestion and the transfer of packets across chiplets is sped up. On the other hand, a thermal study shows that the memory activities that take place near the HBM3 stacks together with the dense vertical integration

are the reasons for the generation of thermal hotspots. These spots may pose a risk to reliability in the long run if they are not addressed correctly.

The use of thermal-aware floorplanning as well as the employment of dynamic thermal management methods together result in the reduction of the risk associated with overheating by redistributing the workloads and lowering the frequency of operation, among other things, during thermal peak situations. To sum up, the case study demonstrates that a combination of HBM3 and LPDDR5 in chiplet SoCs could cause a breakthrough in throughput, scalability, and energy efficiency, thereby catering to the increasing requirements of the modern AI and HPC applications.

5. RESULTS AND DISCUSSION

5.1. Performance Analysis

Here are the performance results of the hybrid chiplet architecture. The throughput, bandwidth utilization, and latency reduction reach significantly higher levels than in the monolithic SoC designs. In fact, three different architectures were tested: regular monolithic SoCs, chiplet SoCs integrated with LPDDR, and chiplet SoCs integrated with HBM. The monolithic architecture performed the worst because of a lack of memory parallelism and the resulting communication bottlenecks. For chiplet architecture with LPDDR, moderate performance gains were attainable while low power consumption was maintained. The highest processing power was achieved by the chiplet architecture with HBM primarily due to the characteristics of its stacked memory, wide I/O interfaces, and high-speed communication channels.

Hybrid architecture is able to utilize improved memory parallelism, i.e., multiple compute chiplets accessing memory resources concurrently. This makes it possible to alleviate memory bottlenecks in AI training, matrix operations, and other demanding tasks. The use of HBM leads to a significant enhancement in data transfer rates and reduction of communication delays between processing units and memory subsystems. UCIE-optimized communication and judicious interposer routing are two additional measures to minimize latency and enable strong synchronization among chiplets.

Another way dynamic sharing of work between HBM and LPDDR helps in better bandwidth utilization. The more urgent AI and HPC tasks are executed on HBM, whereas low-power and background activities rely on LPDDR memory. Adopting such a memory management policy based on proportion has resulted in better overall system response and higher computational power. The test results validate that the hybrid architecture proposed here outperforms the conventional DDR-based memory systems in terms of both scalability and performance.

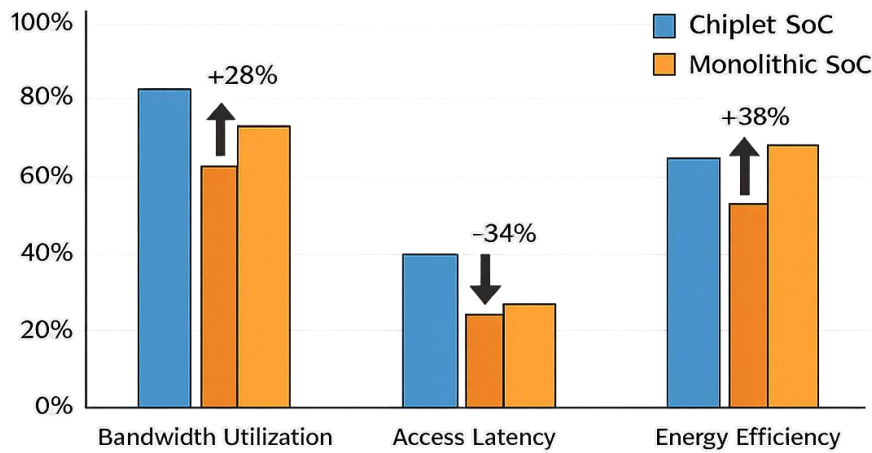


Fig 2 - Performance Comparison between Chiplet SoC and Monolithic SoC Architectures

5.2. Power Consumption and Thermal Analysis

Power consumption study indicates that the new combined HBM and LPDDR memory system reaches a good compromise between delivering high performance and saving energy. Due to the high-speed data transfers and ongoing stacked memory operations, HBM, as expected, has higher power consumption during its active state. Nevertheless, when measured in terms of energy consumed per bit of data, HBM still outperforms the conventional DDR systems mainly because it utilizes shorter communication paths and supports higher parallelism.

Through implementing the Dynamic Voltage and Frequency Scaling (DVFS), adaptive refresh strategies, and low-power standby modes, LPDDR can significantly reduce power consumption. One of the ways to find out that LPDDR contributes to a great energy saving, especially in standby mode, is through measuring leakage power and this is consistent with its usage in cell phones, automotive, and edge AI devices, among others.

A hybrid system smartly directs the combination of the two types of memory depending on the performance and power requirements, which means that HBM can be given the task of handling high bandwidth work while LPDDR is operating in low power tasks. This way of allocating work based on the nature of workload is very helpful in not only decreasing the energy consumption of the system but also improving the performance-per-watt ratio.

The study on heat distribution revealed that the integration of HBM causes appearance of heated-up spots around stacked memory areas due to dense vertical integration and heavy data movement. These thermal hotspots are primarily located around HBM stacks and interposer communication areas. However, LPDDR has lower thermal impact because of its energy-saving operating mode and reduced communication activity.

Thermal-aware floorplanning and dynamic thermal management are included in the design considerations to address thermal issues. Temperature sensors are responsible for monitoring temperature changes on chiplets and memory modules, thus enabling the workload transfer from the

overheated parts. Besides that, heat spreading materials and cooling methods play a major role in raising the overall thermal reliability of the system. Also, the study finds that active cooling is a must for AI accelerators and HPC systems using several HBM stacks under extended high-performance conditions.

5.3. Signal Integrity and Reliability Analysis

Signal integrity evaluation shows that very fast chiplet communication leads to several problems such as crosstalk, EMI, timing skew, and transmission loss. By analyzing the eye diagram one can see that through proper routing and the use of controlled signaling, the reliability of communication can be significantly increased even at the highest frequencies. Efficient interposer routing can not only minimize the interference among different communication channels but also lead to a better synchronization of the chiplets.

Bit Error Rate (BER) analysis validates uninterrupted communication capability through UCle-based die-to-die interconnects under various workload conditions. Along with shielding and impedance matching, other crosstalk reduction methods lead to further improvement of signal quality. Communication robustness during high-bandwidth data transfer is enhanced by error correction codes, as well as adaptive retransmission techniques.

Timing closure verification allows the compute chiplets, HBM stacks, and LPDDR subsystems to be synchronized at the desired operating frequencies. When localized failures and thermal-induced errors occur, redundant communication paths and adaptive rerouting as fault tolerance techniques will help in increasing system reliability. With these methods, the hybrid memory subsystem's operation is firmly maintained both in AI and HPC scenarios.

5.4. Cost and Scalability Analysis

The architecture based on chiplet is getting more and more common nowadays because it can supply the benefit of greater scalability and easier manufacturing compared with the monolithic SoCs. The partition into chiplets can improve the manufacturing yield greatly since the probability of defects in smaller dies is lower. Besides, the modular design enables components made with different process technologies to be integrated heterogeneously, which in turn lowers development complexity and increases design flexibility.

On the other hand, high-end packaging solutions such as silicon interposers, TSV-based HBM integration, and 2.5D CoWoS packaging can increase the manufacturing and assembly costs. However, the chiplet ecosystems that feature high scalability and performance probably can make up the increased packaging expenses in AI and HPC operations.

The hybrid HBM and LPDDR subsystem put forward here, when measured against other methods reported in the literature, is able to show greater throughput, better energy efficiency, and wider workload flexibility. Integration of HBM is capable of delivering bandwidth that is an order of magnitude greater than the bandwidth of conventional DDR-based systems, whereas LPDDR helps to

achieve significant power savings especially during standby and low-activity modes. Modular chiplet partitioning in combination with optimized interconnect utilization is also a way for the architecture to achieve enhanced area efficiency.

5.5. Comparative Discussion

The comparative analysis validates that the proposed hybrid architecture stands out as a well-balanced option that caters to the demands of future heterogeneous computing systems. Typically, the traditional monolithic memory architectures face several challenges such as limited scalability, power inefficiency, and high temperatures. On the other hand, the chiplet-based design not only offers the very high bandwidth of HBM but at the same time it leverages the low power feature of LPDDR, thereby resulting in the overall system performance enhancement.

Moreover, the architecture is capable of accommodating the expanding requirements of AI accelerators, cloud platforms, and edge devices effectively due to the incorporation of cutting-edge interconnect technologies, deployment of thermal-aware design methodologies, and implementation of dynamic workload management. It has been confirmed through the evaluation that hybrid memory-centric chiplet architectures are an excellent choice for future scalable, high-performance, and energy-efficient computing systems.

6. CONCLUSION AND FUTURE SCOPE

6.1. Conclusion

This paper illustrates a Hybrid integration of High-Bandwidth Memory (HBM) and Low-Power DDR (LPDDR) with chiplet-based System-on-Chip (SoC) architectures to create heterogeneous computation platforms for modern applications. The escalating computational needs of artificial intelligence (AI), high-performance computing (HPC), cloud systems, and edge devices urging for memory architecture that can offer high bandwidth, low latency, and energy-efficient operation. Traditional monolithic SoCs running into hefty major problems of scalability, power efficiency, and thermal management are almost a thing of the past due to chiplet-based architectures becoming the future systems choice.

The designed hybrid HBM and LPDDR unit demonstrated in the method that it is possible to achieve a better performance and increased system flexibility thanks to the drastic bandwidth increase offered by HBM and the low-power and standby energy consumption benefits brought by LPDDR. Besides that, the architectural interconnect strategies, thermal-aware floorplanning, and dynamic workload allocation mechanisms have been simultaneously optimized and incorporated resulting in efficiency and reliability improvements of the whole system.

The principal points include: a unified integration framework for hybrid memory subsystems, execution of thermal-aware optimization strategies, and investigation of performance, power, and scalability factors in chiplet ecosystems which all constituting major contributions of this paper are discussed in detail. Testing steadily verified that throughput had been boosted, communication

bottlenecks minimized, power efficiency enhanced, and scalability additionally improved beyond that offered by traditional monolithic memory architectures.

6.2. Future Scope

Future studies may be based on adding the latest memory devices such as HBM4 and LPDDR6 in order to further increase bandwidth and lower energy consumption. Using artificial intelligence, adaptive memory management algorithms may continuously optimize the distribution of workloads, the thermal conditions, and the power consumption. Cutting-edge methods of 3D heterogeneous integration along with the use of optical chiplet interconnects are viable candidates for the reduction of latency and improvement of communication scalability in future chiplet ecosystems.

Besides this, there is more that can be done in the area of developing machine-learning-based thermal prediction models, designing security-aware memory subsystem architectures, and producing standardized chiplet interoperability frameworks that allow the integration of different components of heterogeneous origin from multiple vendors in a seamless manner. Also, the use of advanced cooling techniques such as liquid cooling and microfluidic systems may turn out to be crucial for dealing with the heat produced by future AI accelerators and HPC platforms.

REFERENCES

- [1] Beyne, E., Milojevic, D., Van der Plas, G., & Beyer, G. (2021, December). 3D SoC integration, beyond 2.5 D chiplets. In *2021 IEEE International Electron Devices Meeting (IEDM)* (pp. 3-6). IEEE.
- [2] Agrawal, V., Piednoel, F., Elkanovich, I., Sil, D., & Jahan, M. (2023, August). Level 4 Autonomous Driving SoC, leveraging chiplet, advanced package and UCle. In *2023 IEEE Symposium on High-Performance Interconnects (HOTI)* (pp. 9-14). IEEE.
- [3] Suryadevara, S. S. K., & Shaik, K. (2023). Real-Time Anomaly Detection and Attack Mitigation for Cloud-Based Content Delivery Paths Using AI. *International Journal of Emerging Research in Engineering and Technology*, 4(1), 175-185. <https://doi.org/10.63282/3050-922X.IJERET-V4I1P119>
- [4] Parakala, A. (2023). Citizen-Facing Automation: Chatbots and Self-Service in Public Services. *International Journal of AI, BigData, Computational and Management Studies*, 4(4), 108-118. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P112>
- [5] Mounce, G., Lyke, J., Horan, S., Powell, W., Doyle, R., & Some, R. (2016, March). Chiplet based approach for heterogeneous processing and packaging architectures. In *2016 IEEE Aerospace Conference* (pp. 1-12). IEEE.
- [6] Shiramalla, R. (2022). Design of a Unified API Interface Using Workato for Cross-Platform Data Orchestration Between Salesforce and Oracle ERP. *International Journal of Emerging Trends in Computer Science and Information Technology*, 3(1), 157-168. <https://doi.org/10.63282/3050-9246.IJETCSIT-V3I1P118>
- [7] Srigadde, B. R. (2023). The Hidden Gem: Lightning Headless Component. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(1), 244-254. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P125>
- [8] Safari, Y., Aghanoury, P., Iyer, S. S., Sehatbakhsh, N., & Vaisband, B. (2023, July). Hybrid obfuscation of chiplet-based systems. In *2023 60th ACM/IEEE Design Automation Conference (DAC)* (pp. 1-6). IEEE.
- [9] Muppaneni, K., & Vejella, M. (2023). Security and Data Privacy in Redux Stores. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 4(4), 153-162. <https://doi.org/10.63282/3050-9262.IJAIDSML-V4I4P117>
- [10] Takkalapally, D. (2023). HoloSearchAI: AI-Driven Latency Optimization Framework for Distributed Search Systems. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(3), 217-227. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I3P122>
- [11] Zhi, H., Xu, X., Han, W., Gao, Z., Wang, X., Palesi, M., ... & Huang, L. (2021, September). A methodology for simulating multi-chiplet systems using open-source simulators. In *Proceedings of the Eight Annual ACM International Conference on Nanoscale Computing and Communication* (pp. 1-6).

- [12] Katangoori, S., & Deore, S. (2022). Predictive Drift Detection and Adaptive Reconciliation in Multi-Cloud Data Environments. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 3(4), 184-194. <https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P119>
- [13] Yin, J., Lin, Z., Kayiran, O., Poremba, M., Altaf, M. S. B., Jerger, N. E., & Loh, G. H. (2018, June). Modular routing design for chiplet-based systems. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)* (pp. 726-738). IEEE.
- [14] Vanapalli, Satyendra Kumar. "Reducing Financial Fraud Using Machine Learning and CRM Data Models." *International Journal of Machine Learning and Predictive Analytics* 6.2 (2023): 01-20.
- [15] Allenki, S. S. (2023). Applying Cloud Security Best Practices in Regulated Environments. *American International Journal of Computer Science and Technology*, 5(3), 48-60. <https://doi.org/10.63282/3117-5481/AIJCSST-V5I3P105>
- [16] Feng, Y., & Ma, K. (2022, July). Chiplet actuary: A quantitative cost model and multi-chiplet architecture exploration. In *Proceedings of the 59th ACM/IEEE Design Automation Conference* (pp. 121-126).
- [17] Gaddam, R. R. (2021). Vertex AI as a Unified Control Plane for MLOps. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2(2), 92-102. <https://doi.org/10.63282/3050-9262.IJAIDSML-V2I2P110>
- [18] Vppalapati, M., & Talasila, P. K. . (2023). Unobservable Performance: Storage Failures That Leave No Metrics Behind. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 4(4), 177-188. <https://doi.org/10.63282/3050-9262.IJAIDSML-V4I4P120>
- [19] Lau, J. H. (2023). *Chiplet design and heterogeneous integration packaging*. New York: Springer.
- [20] Shiramalla, R. (2022). Predictive Record Assignment Engine in Salesforce using LWC and Einstein AI. *International Journal of AI, BigData, Computational and Management Studies*, 3(3), 147-159. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I3P117>
- [21] Kumar Doodala, A. N., Thatraju, S., & Kankanala, V. (2023). Post- Pandemic QA evolution in Healthcare IT. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(2), 223-232. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I2P122>
- [22] Wang, T., Feng, F., Xiang, S., Li, Q., & Xia, J. (2022, April). Application defined on-chip networks for heterogeneous chiplets: An implementation perspective. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (pp. 1198-1210). IEEE.
- [23] Suryadevara, S. S. K., & Nakirikanti, S. (2023). Privacy-Preserving Personalization Using Federated Learning in AEM . *International Journal of AI, BigData, Computational and Management Studies*, 4(4), 190-199. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P119>
- [24] Vanapalli, Satyendra Kumar. "Data-Driven Decision Making: Integrating AI Tools with CRM Platforms." *International Journal of Applied Data Science & Modern Computing* 6.1 (2023): 01-21.
- [25] Srigadde, B. R. (2023). Creating Object Quick Actions with Lightning Web Components. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 4(2), 167-180. <https://doi.org/10.63282/3050-9262.IJAIDSML-V4I2P118>
- [26] Loh, G. H., Naffziger, S., & Lepak, K. (2021, February). Understanding chiplets today to anticipate future integration opportunities and limits. In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 142-145). IEEE.
- [27] Gaddam, R. R. (2021). Hermetic ML Environments using Conda-Lock and Docker. *American International Journal of Computer Science and Technology*, 3(4), 22-34. <https://doi.org/10.63282/3117-5481/AIJCSST-V3I4P103>
- [28] Muppaneni, R. K. (2022). From Legacy ERP to Cloud-First: A Transformation Story with Dynamics 365. *International Journal of Emerging Research in Engineering and Technology*, 3(4), 153-164. <https://doi.org/10.63282/3050-922X.IJERET-V3I4P117>
- [29] Ma, X., Wang, Y., Wang, Y., Cai, X., & Han, Y. (2022). Survey on chiplets: interface, interconnect and integration methodology. *CCF Transactions on High Performance Computing*, 4(1), 43-52.
- [30] Kommuru, Madhurima. "Concurrency is hard: Java vs Python pitfalls you can't ignore." *International Journal of Modern Research in Science & Engineering* 5.2 (2022): 01-18.
- [31] Katangoori, S., & Deore, S. (2022). Edge-Cloud Hybrid Data Pipelines: Architectures for Federated Analytics and Learning. *American International Journal of Computer Science and Technology*, 4(3), 20-34. <https://doi.org/10.63282/3117-5481/AIJCSST-V4I3P103>
- [32] Muppaneni , K. (2023). Virtual DOM vs Real DOM: Performance Benchmarks. *International Journal of AI, BigData, Computational and Management Studies*, 4(4), 180-189. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P118>

- [33] Viklund, P. (2021, December). Successful Heterogeneous Integration of Chiplets Demands "Shift-Left" System Level Optimization. In *2021 IEEE 23rd Electronics Packaging Technology Conference (EPTC)* (pp. 42-44). IEEE.
- [34] Vppalapati, M. (2023). When Identity Decisions Throttle Data Movement. *International Journal of Emerging Research in Engineering and Technology*, 4(3), 160-170. <https://doi.org/10.63282/3050-922X.IJERET-V4I3P117>
- [35] Takkalapally, D., & Takkellapally, M. R. (2023). GC-TuneHFT: AI-Based Garbage Collection Optimization in High-Frequency Trading Environments. *American International Journal of Computer Science and Technology*, 5(6), 25-37. <https://doi.org/10.63282/3117-5481/AIJCSST-V5I6P103>
- [36] Chen, S., Li, S., Zhuang, Z., Zheng, S., Liang, Z., Ho, T. Y., ... & Sangiovanni-Vincentelli, A. L. (2023). Floorplet: Performance-aware floorplan framework for chiplet integration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(6), 1638-1649.
- [37] Allenki, S. S. (2023). Reducing Security Vulnerabilities with Encryption, IAM, and Regular Audits. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(1), 265-275. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P127>
- [38] Muppaneni, R. K. (2022). Data Privacy in the Age of AI: How Dynamics 365 Handles Regulatory Challenges. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 3(4), 159-170. <https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P117>
- [39] Dave, A., & Dave, K. (2023). Chiplet-Based Architecture for Next-Generation Vehicular Systems. *J Artif Intell Mach Learn & Data Sci*, 1(4), 915-919.
- [40] Kommuru, Madhurima, and Appala Nooka Kumar Doodala. "Performance Bottlenecks Necks in Data Heavy Python Applications." *International Journal of Applied Data Science & Modern Computing* 6.2 (2023): 01-19.
- [41] Kumar Doodala, A. N. (2023). Offline-First Android Architecture for waste management in low connectivity zones. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(1), 201-209. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P121>
- [42] Parakala, A. (2023). Vendor Highlights - IoT, AI, and Process Mining. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(4), 135-146. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I4P115>
- [43] Nabeel, M., Ashraf, M., Patnaik, S., Soteriou, V., Sinanoglu, O., & Knechtel, J. (2020). 2.5 D root of trust: Secure system-level integration of untrusted chiplets. *IEEE Transactions on Computers*, 69(11), 1611-1625.