MANTECH Publications

Journal of Research in VLSI Design Tools and Technology

Volume 1 Issue 3, September-December 2016

Chiplet and Heterogeneous 3D System Design

Priya Rangan[1], Avshek Chaudhary[2], Vikash Sharma[3]

Assistant professor, Students

Department of VLSI Packaging and Interconnects

Maharaja's College, Ernakulam, India

Email: *priyaranganjc@gmail.com[1], chaudharyavshek3333@yahoo.com[2], vikash_drsharma@rediffmail.com[3]*

Abstract

The rapid advancement in semiconductor technology has led to increasing demand for high-performance, energy-efficient, and cost-effective integrated circuits. Traditional monolithic system-on-chip (SoC) designs are facing scalability, yield, and thermal challenges as device dimensions continue to shrink. Chiplet-based and heterogeneous 3D system design approaches provide a promising alternative by integrating multiple smaller, function-specific dies into a single package. This paper reviews the fundamentals, architectures, design methodologies, and challenges associated with chiplet and heterogeneous 3D systems. We discuss the benefits of modular design, heterogeneous integration, advanced interconnect techniques, thermal management, and reliability considerations. Emerging trends, including co-packaged memory, advanced TSVs, silicon interposers, and AI-driven design tools, are highlighted. Tables and figures illustrate comparative performance metrics, integration strategies, and design workflows. The paper concludes with an outlook on future research directions, emphasizing the role of heterogeneous 3D systems in high-performance computing, AI accelerators, and next-generation consumer electronics.

Keywords: *Chiplet, 3D integration, heterogeneous system design, TSV, silicon interposer, advanced packaging, modular SoC, thermal management*

Introduction

The semiconductor industry has historically relied on scaling down transistor sizes to achieve higher performance and lower power consumption. However, the end of Moore's Law scaling and challenges in monolithic SoC fabrication—such as yield degradation, increased design complexity, and high manufacturing costs—have prompted exploration of alternative system integration approaches.

Chiplet-based and heterogeneous 3D integration techniques divide the functionality of a system across multiple smaller dies, each optimized for a specific purpose, and integrate them into a single package. This modular approach improves yield, allows reuse of IP cores, and supports heterogeneous integration of different technology nodes (logic, memory, analog, photonics).

Motivation

- **Yield improvement:** Smaller dies are easier to fabricate with higher yields.
- **Heterogeneous integration:** Allows mixing different process technologies (e.g., high-performance logic with low-power memory).
- **Scalability:** Enables incremental upgrades without redesigning the entire SoC.
- **Thermal management:** Easier heat dissipation with distributed dies.

Scope of the Paper

This paper explores the following aspects of chiplet and heterogeneous 3D system design:

1. Chiplet architectures and design methodologies.
2. 3D system integration techniques, including through-silicon vias (TSVs) and micro-bumps.
3. Heterogeneous integration of diverse technologies.
4. Interconnect technologies and communication protocols.
5. Thermal management and reliability challenges.
6. Future trends and research directions.

Fundamentals of Chiplet Design

Chiplet design represents a paradigm shift in modern semiconductor engineering. Unlike traditional monolithic system-on-chip (SoC) designs, where all functional blocks are integrated into a single die, chiplet-based designs break a system into multiple smaller, functionally complete

dies—called **chiplets**—that can be independently fabricated and integrated into a single package. This modularity addresses challenges in yield, cost, scalability, and heterogeneous integration, making it particularly suitable for high-performance computing, AI accelerators, and complex consumer electronics.

Chiplet Concept

A **chiplet** is a small, self-contained die that performs a specific function or a set of related functions within a larger system. Examples include a logic chiplet for CPU cores, a memory interface chiplet for high-bandwidth memory, or a specialized AI processing chiplet. These chiplets can be integrated with others in a single package using advanced packaging technologies to form a complete system.

Key Features of Chiplets

1. **Functional Modularity** – Each chiplet encapsulates specific functionality, enabling engineers to design, test, and optimize individual blocks without affecting the entire system. For instance, a memory chiplet can be designed in a mature 22nm node while logic chiplets use a cutting-edge 5nm process.
2. **Known-Good-Die (KGD) Methodology** – Before integration, each chiplet is fully tested and verified as a "known-good-die," ensuring that only functional dies are assembled into the package. This significantly improves overall yield and reduces the risk of system failure, which is especially important in 3D stacking where a defect in one layer can compromise the entire chip.
3. **Heterogeneous Integration** – Chiplets enable mixing of different technologies within a single package. Logic, memory, analog, RF, and even photonics chiplets can be combined, overcoming limitations of monolithic SoCs. For example:
    - AI accelerators can integrate high-speed logic chiplets with HBM memory chiplets.
    - RF transceivers can be fabricated in specialized nodes while remaining tightly integrated with digital processing chiplets.
4. **Scalability and Reuse** – Modular chiplets allow reuse of intellectual property (IP) across multiple products. A CPU core chiplet designed for a desktop processor can be reused in a data

center processor with additional memory or AI chiplets. This reduces time-to-market and design costs.

5. **Thermal and Power Optimization** – Smaller chiplets generate lower local heat and can be placed strategically within the package to improve thermal dissipation, reducing hotspots and enabling more aggressive power management.

Chiplet Architectures

Chiplets can be integrated using multiple architectures, each with different advantages in performance, interconnect density, cost, and thermal characteristics.

2.2.1 2D Package-on-Package (PoP)

In PoP designs, chiplets are stacked vertically, typically using **micro-bumps** for electrical interconnects. PoP allows a compact vertical integration and is commonly used in mobile devices, combining logic and memory chiplets.

Advantages:

- Simple and cost-effective stacking.
- Short interconnects between stacked dies improve signal integrity compared to separate 2D placement.

Limitations:

- Interconnect density is limited by micro-bump pitch.
- Thermal challenges increase with higher power chiplets due to vertical stacking.

Example: Mobile SoCs often stack DRAM chiplets on top of the main CPU/GPU die using PoP to save PCB space.

2.5D Interposer

A **2.5D interposer** integrates multiple chiplets side-by-side on a high-density silicon or organic substrate. The interposer provides dense interconnects through a **redistribution layer (RDL)**, enabling high-bandwidth communication between chiplets.

Advantages:

- High-bandwidth, low-latency communication between chiplets.
- Allows heterogeneous integration of different technologies and nodes.

- Easier thermal management compared to stacked 3D designs.

Limitations:

- Cost of silicon interposers can be high.
- Requires precise alignment of multiple chiplets during assembly.

Example: Modern GPUs and AI accelerators often use 2.5D interposers to connect multiple memory chiplets with compute chiplets, achieving terabyte-per-second memory bandwidth.

3D Die-Stacking

3D stacking involves vertically stacking multiple chiplets or dies and connecting them using **through-silicon vias (TSVs)**, which are vertical electrical conduits passing through the silicon die. This approach enables ultra-high-density integration and very short interconnects.

Advantages:

- Extremely high bandwidth and low latency due to vertical TSV connections.
- Reduces footprint of the overall system.
- Enables tight integration of memory and logic for high-performance computing.

Limitations:

- Thermal management is complex due to heat accumulation in stacked dies.
- Mechanical stress and reliability issues are more pronounced with TSVs.
- Manufacturing cost is higher due to complex stacking and testing procedures.

Example: High-bandwidth memory (HBM) stacks multiple DRAM dies over a logic die in 3D to maximize memory density and bandwidth.

Table 1: Comparison of Chiplet Integration Architectures

Architecture	**Interconnect**	**Bandwidth**	**Thermal Complexity**	**Cost**
2D PoP	Micro-bumps	Medium	Low	Low
2.5D Interposer	Redistribution Layer (RDL)	High	Medium	Medium
3D Stacking	TSVs	Very High	High	High

Design Methodology

Designing a chiplet-based system requires a systematic methodology that considers both the individual chiplets and their integration into a complete system. Unlike traditional monolithic SoCs, chiplet design introduces additional complexity in terms of functional partitioning, interconnect standardization, package co-design, and verification. A well-defined methodology ensures high performance, yield, and reliability.

Step 1: Functional Partitioning

Functional partitioning is the process of breaking down a complex system into smaller, manageable chiplets, each implementing a specific function. This step is critical because it determines the overall system architecture and performance.

Key considerations:

- **Functionality Mapping:** Identify which components should be separate chiplets. For example, compute cores, memory interfaces, AI accelerators, and analog/RF blocks can be implemented as separate chiplets.
- **Technology Selection:** Assign appropriate semiconductor nodes to each chiplet based on performance, power, and cost. For example, high-performance logic can use 5nm or 3nm nodes, whereas DRAM or analog circuits may remain on mature 28nm or 65nm nodes.
- **Communication Requirements:** Determine the bandwidth and latency requirements between chiplets. Chiplets with high data exchange, like CPU cores and HBM memory, may require TSVs or high-speed interposers.

Example: In a GPU system, the functional partitioning may result in separate chiplets for:

- Compute cores (logic)
- High-bandwidth memory (HBM)
- Memory controllers and interconnect logic
- IO/PCIe interface

This partitioning allows modular design and scalability, enabling engineers to upgrade or replace chiplets without redesigning the entire system.

Step 2: Interface Standardization

Standardized interfaces are critical to ensure reliable and efficient communication between chiplets. Non-standard or proprietary interconnects increase design complexity, cost, and limit interoperability.

Key interconnect standards:

- **UCIe (Universal Chiplet Interconnect Express):** Provides a high-bandwidth, low-latency, scalable interface for heterogeneous chiplets. It supports multiple topologies, including 2D, 2.5D, and 3D stacking.
- **CXL (Compute Express Link):** Optimized for coherent memory sharing between chiplets, especially useful in CPU-memory and accelerator-memory integration.
- **AMBA AXI/ACE:** Widely used in embedded systems and SoCs for on-chip communication.
- **Design considerations:**
- Define the data width, clocking, and protocol requirements early in the design cycle.
- Minimize signal skew and latency through careful floorplanning and interconnect routing.
- Ensure power-efficient operation through low-swing signaling and adaptive clocking.

Example: In a multi-chip AI accelerator, UCIe may be used to connect compute and memory chiplets, while CXL handles coherent memory access across chiplets.

Step 3: Co-Design of Package and Die

Chiplet performance is not determined solely by the individual dies; the **package design plays an equally important role**. Co-design ensures that chiplets, interconnects, and package layers are optimized together for signal integrity, power delivery, and thermal management.

Key considerations:

- **Placement Optimization:** Position high-communication chiplets close together to reduce interconnect length and latency. For instance, memory chiplets should be near compute chiplets.
- **Thermal Management:** Distribute high-power chiplets strategically to avoid hotspots. Integrate heat spreaders, thermal TSVs, or microfluidic cooling as required.

- **Power Delivery:** Design robust power distribution networks to maintain voltage stability across stacked dies.
- **Signal Integrity:** Optimize routing and shielding to minimize crosstalk and reflection in high-speed interconnects.

Example: A 2.5D GPU package may co-design multiple compute chiplets and HBM memory on a silicon interposer, balancing bandwidth and thermal constraints to achieve optimal performance.

Step 4: Testing and Verification

Testing and verification are critical in chiplet-based design because defects in even a single chiplet can affect the entire system. A **Known-Good-Die (KGD) approach** is widely adopted to mitigate this risk.

Verification steps:

- **Individual Chiplet Testing:** Each chiplet undergoes functional, timing, and parametric tests to ensure it meets specifications.
- **Interface Verification:** Ensure communication protocols between chiplets are correctly implemented and signal integrity is maintained.
- **Package-Level Testing:** After integration, perform system-level tests, including thermal stress, voltage variation, and high-speed communication tests.
- **Design for Test (DFT) Features:** Incorporate embedded test circuits such as BIST (Built-In Self-Test) to facilitate post-assembly testing.

Heterogeneous 3D System Design

Concept

Heterogeneous 3D systems integrate dies from multiple technologies or nodes within a single package. For instance, a system could combine:

- Logic chiplets fabricated in 5nm for high performance.
- High-density DRAM or HBM memory in 22nm.
- Analog or RF blocks in mature nodes like 65nm.

This integration allows leveraging the best process for each function without the cost of a single monolithic die.

Integration Techniques

Through-Silicon Vias (TSVs)

TSVs are vertical electrical connections passing through silicon wafers. They allow:

- High-density interconnects.
- Reduced signal latency.
- Lower power consumption compared to long planar interconnects.

Micro-bump and RDL Interconnects

Micro-bumps and redistribution layers enable fine-pitch connections between stacked dies or chiplets on a silicon interposer.

Silicon and Organic Interposers

- **Silicon interposers** provide high-performance, high-density connections, suitable for high-bandwidth applications like GPUs and AI accelerators.
- **Organic interposers** are cost-effective but have lower density and performance.
-

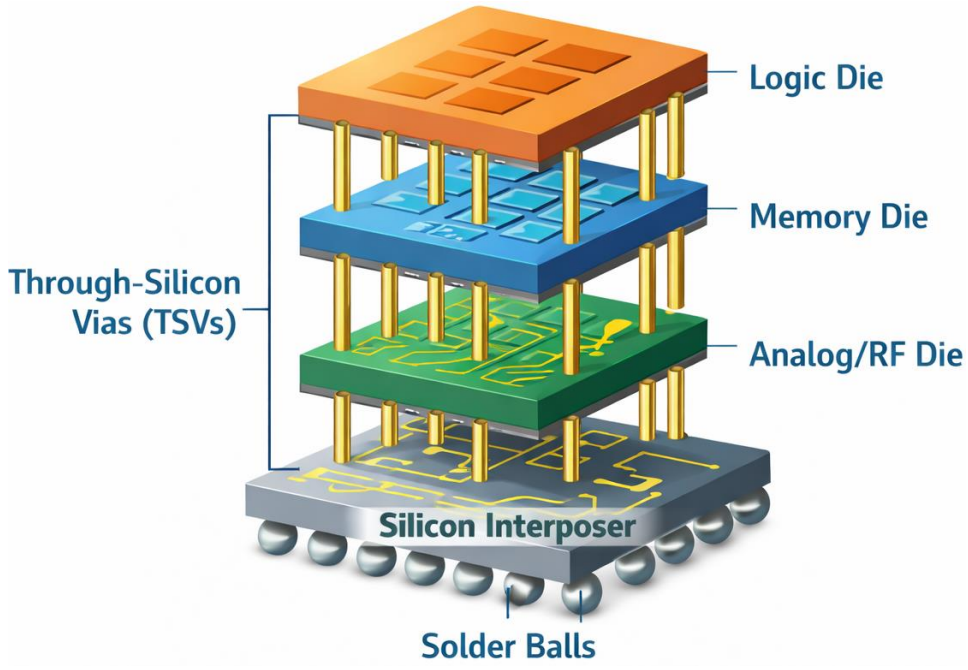

Figure 1: Heterogeneous 3D Integration Using TSVs and Silicon Interposers

Communication Protocols

High-performance chiplets require standardized communication interfaces:

- **UCIe (Universal Chiplet Interconnect Express):** Supports scalable, high-speed connectivity.
- **CXL (Compute Express Link):** Enables memory pooling and coherence.
- **AMBA AXI/ACE:** Widely used in embedded systems.

Thermal Management in 3D Systems

Thermal Challenges

3D stacking increases power density, causing hotspots and reduced reliability. Key issues include:

- Uneven heat distribution across layers.
- Thermal-induced stress affecting TSVs and interconnects.

Thermal Solutions

- **Microfluidic cooling:** Embedded channels for liquid cooling.
- **Thermal vias and heat spreaders:** Direct heat conduction paths.
- **Dynamic power management:** Adaptive voltage and frequency scaling.

Table 2: Thermal Management Techniques for 3D Systems

Technique	Pros	Cons	Use Case
Heat Spreaders	Simple, passive	Limited efficiency	General-purpose chips
Microfluidic Cooling	High efficiency	Complex, expensive	High-power AI chips
Thermal TSVs	Direct conduction	Adds design complexity	Dense 3D stacking

Reliability and Testing

Reliability Concerns

- **TSV stress:** Mechanical stress may cause cracks or failure.
- **Electromigration:** High current density in small interconnects.
- **Thermal cycling:** Repeated heating and cooling can delaminate interfaces.

Testing Strategies

- **Known-Good-Die (KGD) testing:** Individual chiplets tested before assembly.
- **Built-in Self-Test (BIST):** Embedded test logic in each chiplet.
- **Package-level stress testing:** Ensures performance under real-world thermal and electrical conditions.

Emerging Trends

Co-packaged Memory and Logic

Integration of high-bandwidth memory (HBM) close to logic reduces latency and power.

AI-Driven Chiplet Design

Machine learning tools assist in:

- Optimal floorplanning of chiplets.
- Predicting thermal hotspots.
- Designing interconnect topology.

Photonic Interconnects

Optical links for ultra-high-speed, low-latency communication between chiplets.

Applications

High-Performance Computing

Chiplet-based GPUs and CPUs benefit from modular upgrades and high memory bandwidth.

AI Accelerators

Heterogeneous integration allows combining logic, memory, and specialized AI cores efficiently.

Consumer Electronics

Enables modular smartphone SoCs with optimized battery life, performance, and cost.

Challenges and Future Directions

- **Standardization:** Lack of widely adopted chiplet interface standards.
- **Design Complexity:** Co-design of package, die, and interconnect requires advanced EDA tools.

- **Cost Trade-offs:** 3D stacking and interposers increase fabrication complexity.
- **Thermal and Reliability Management:** Continued research needed in cooling and stress mitigation.

Future Directions:

1. Development of universal chiplet interfaces (UCIe adoption).
2. Advanced AI-driven EDA for automated chiplet integration.
3. Photonic and wireless interconnects to overcome bandwidth limitations.
4. Co-packaged heterogeneous dies for next-gen exascale computing.

Conclusion

Chiplet and heterogeneous 3D system design represent a paradigm shift in semiconductor integration, addressing the limitations of monolithic SoCs. By enabling modular, scalable, and heterogeneous integration of logic, memory, and analog functions, these systems improve performance, yield, and flexibility. Advanced interconnect techniques, thermal management solutions, and AI-driven design tools further enhance the viability of these architectures. Despite challenges in standardization, reliability, and cost, chiplet-based heterogeneous systems are poised to dominate high-performance computing, AI accelerators, and next-generation consumer electronics. Continued research in interconnect technologies, thermal solutions, and automated design methodologies will drive widespread adoption and innovation.

References

1. Bhardwaj, A., & Kumar, P. (2022). *Chiplet-based heterogeneous integration: Opportunities and challenges*. IEEE Transactions on Components, Packaging, and Manufacturing Technology, 12(8), 1400–1415.
2. Liu, Y., et al. (2021). *3D integration technologies for high-performance computing.* Microelectronics Journal, 112, 105001.
3. Chen, H., & Wang, Z. (2020). *Thermal management in 3D stacked systems.* Journal of Electronic Packaging, 142(3), 031001.
4. Universal Chiplet Interconnect Express (UCIe) Consortium. (2023). *UCIe Specification v1.1*.

5. International Technology Roadmap for Semiconductors (ITRS). (2020). *3D Integration and Heterogeneous Systems.*
6. Kim, S., & Patel, A. (2021). *AI-assisted chiplet floorplanning and optimization.* ACM Transactions on Design Automation of Electronic Systems, 26(4), 50.
7. Zhang, X., et al. (2020). *Co-packaged memory and logic integration for high-bandwidth applications.* IEEE Micro, 40(5), 72–82.
8. Huang, L., & Tsai, M. (2019). *Reliability issues in TSV-based 3D ICs.* Microelectronics Reliability, 97–98, 38–49.
9. Choi, J., et al. (2021). *Photonic interconnects for heterogeneous 3D systems.* Optica, 8(6), 712–720.
10. Song, Y., & Lee, H. (2022). *Emerging trends in chiplet-based heterogeneous integration.* Journal of Semiconductor Technology, 37(2), 95–108.