Paper ID #220: Simulation and Exploration for Multi-Chiplet Systems using Open-Source Tools and Heuristic Algorithm

Luming Wang, Fangli Liu, Zichao Ling, Zheqin Cao, Yixin Xuan, Jianwang Zhai†, Kang Zhao
Beijing University of Posts and Telecommunications, China

Introduction

As Moore's Law plateaus, multi-chiplet systems have become pivotal for cost-effective SoC design[1],[2], yet their inherent heterogeneity—combining diverse chiplets (e.g., AMD's 3D V-Cache, Apple's unified memory) and complex networks-on-package—poses critical simulation challenges[3].

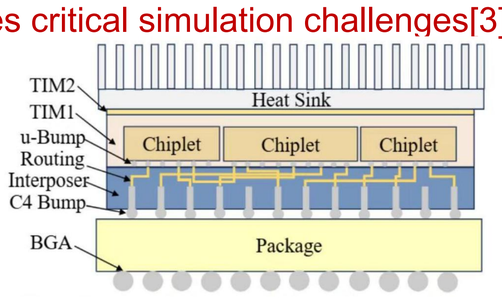


Fig.1 The architecture of chiplet-based 2.5D IC.

To address this, we enhance LegoSim[4] with three innovations: (1) A heterogeneous modeling framework integrating DRAMSim3 for cycle-accurate memory timing/bandwidth analysis; (2) A genetic algorithm-driven co-optimization of IPC and power across 15+ parameters. Experimental validation shows 8.96% CPU and 8.57% GPU performance gains while reducing design space exploration time by 4.3× versus manual tuning, establishing a new benchmark for scalable multi-chiplet simulation.

Design Space Exploration

The optimization process commences with population initialization where P0 = {X1, X2, . . . , XN } denotes the initial solution set. Each Xi encodes a parameter configuration within design space boundaries defined by simulator constraints and experimental specifications.

The fitness function f (X) considers both IPC (instructions per cycle) and power:

$$f(X) = \alpha_{CPU} \cdot IPC_{CPU} + \alpha_{GPU} \cdot IPC_{GPU} - \beta_{CPU} \cdot P_{CPU} - \beta_{GPU} \cdot P_{GPU},$$

where α_{CPU} and α_{GPU} are weights for IPC, and β_{CPU} and β_{GPU} are weights for power dissipation.

Table.1 Parameter Space of GPU

Parameter	Value/Range
GPU Clusters	range(1, 20)
Cores per Cluster	range(1, 5)
Shader Registers per SM	[16384, 32768, 49152, 65536]
Shared Memory Size	[16384, 32768, 49152, 65536]
L1 Data Cache Sets	[16, 32, 48, 64]
L2 Cache Sets	[32, 64, 96, 128]
Single-Precision Units	range(1, 5)
Special Function Units	range(1, 5)
Core Frequency (MHz)	range(500, 1000, 100)

Table.2 Parameter Space of CPU

Parameter	Value/Range
Main Frequency (GHz)	[2.6, 2.7, ..., 3.6]
Logical CPUs (SMT Threads)	[1, 2, 4]
L1 ICache Size (KB)	[32, 64, 128, 256]
L1 DCache Size (KB)	[32, 64, 128, 256]
L2 Cache Size (KB)	[256, 512, 1024, 2048]
L3 Cache Size (MB)	[8, 16, 32, 64]
Memory Bandwidth (GB/s)	[7.6, 15.2, 30.4, 60.8]
L2 Shared Cores	[1, 2, 4]
L3 Shared Cores	[1, 2, 4, 8, 16]

Standard Migration

The rigid coupling between memory protocols and system access patterns complicates multi-standard integration. We design an abstract wrapper bridging DRAMSim3 and LegoSim, enabling parameterized configuration of heterogeneous memory technologies through event-driven synchronization and memory-mapped state preservation. This eliminates redundant reconfiguration while maintaining cross-tool data integrity.

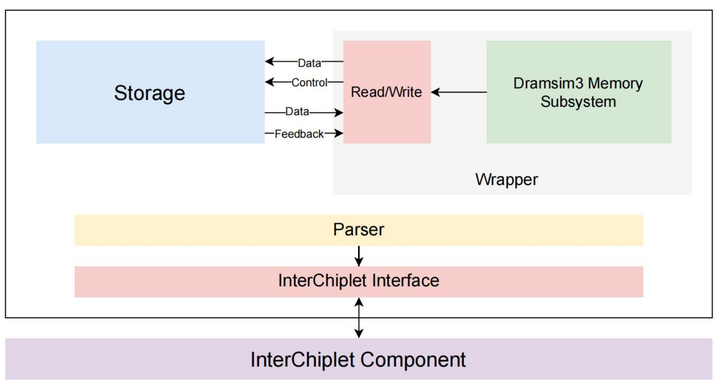


Fig.2 The memory chiplet design in our framework.

Timing Synchronization

Divergent clock domains across chiplets introduce synchronization challenges. By rearchitecting DRAMSim3 as an always-active server within LegoSim's simulation flow, our framework ensures real-time command processing and cycle-accurate alignment of memory accesses, even for tightly coupled CPU-GPU workflows with mixed clock frequencies.

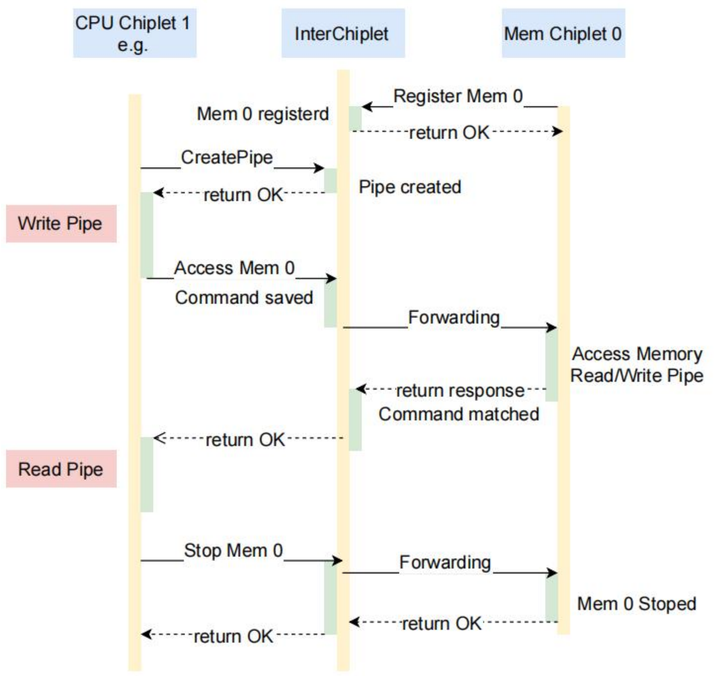


Fig.3 The complete system sequence diagram for accessing memory chiplets.

Experimental Setting

The simulation program was executed within a Docker container running Ubuntu 18.04 LTS with Intel® Xeon® Platinum 8383C CPUs (2.70 GHz, 80 cores, 160 threads) and 1024 GB of memory. Benchmark tasks include matrix multiplication (300×100 and 100×300 matrices) and a strided access microbenchmark (1GB working set).

Simulation of Configurable Memory

Fig 5 and 6 validate DRAMSim3 integration through emulation of real-world memory access patterns, while simultaneously stress-testing multi-memory-type simulation (DDR4/HBM/GDDR6) under programmable address mapping and timing constraints.

Fig.5 Latency distribution.

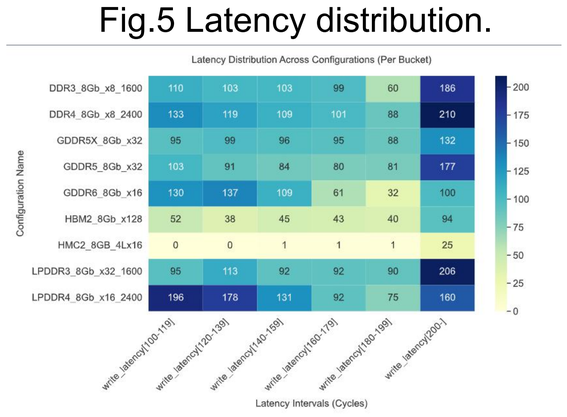
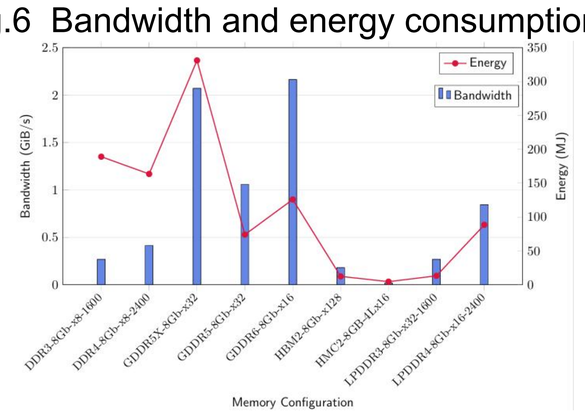


Fig.6 Bandwidth and energy consumption.



DSE Comparison with Baseline

As shown in Fig. 7, the fitness ratio increases from 1.062 in Generation 1 to 1.104 in Generation 26, showing a 10.4% improvement over the baseline. The speedup ratio rises from 1.156 in Generation 1 to 1.396 in Generation 26, as demonstrated in Fig. 8. It reflects a 28.38% reduction in execution time compared to the baseline. As demonstrated in Table.3, we have achieved notable IPC improvements (8.96% for CPUs and 8.57% for GPUs) without compromising power efficiency.

Fig.7 The fitness ratio over generations.

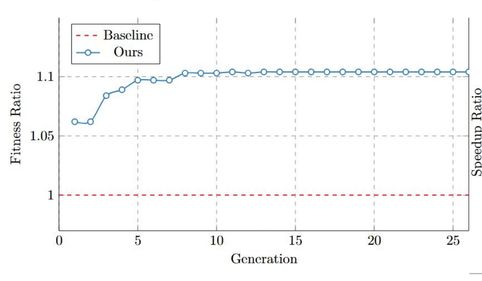


Fig.8 The speedup ratio over generations.

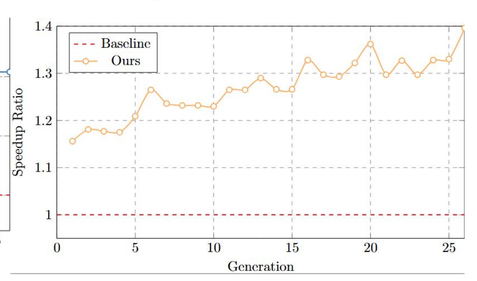


Table.3 Comparison of optimal configuration

Metric	Baseline	Ours	Change
CPU IPC	1.34	1.46	+8.96%
GPU IPC	0.0245	0.0266	+8.57%
CPU Power	21.80	22.36	+2.57%
GPU Power	33.91	33.89	-0.06%
Fitness	2.0709	2.2861	+10.40%
Time	53469	38293	-28.38%

Conclusion

Our work enhances the existing open-source simulator by integrating DRAMSim3 to enable memory chiplet simulation with timing, bandwidth, and power modeling. Importantly, we introduce a GA-based optimization framework to automate configuration exploration. These contributions advance Chiplet design exploration methodologies for practical applications.

References

[1] G. H. L. R. Swaminathan, "The next era for chiplet innovation," in IEEEInternational Conference on Chiplet Technology, 2023, pp. 1–10.
[2] S. Chen, H. Zhang, Z. Ling, J. Zhai, and B. Yu, "The Survey of 2.5D Integrated Architecture: An EDA Perspective," in Proceedings of IEEE/ACM Asian and South Pacific Design Automation Conference (ASP-DAC), 2025.
[3]P. Vivet, E. G. Y. Thonnart, G. Pillonnet et al., "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, vol. 56, no. 1, pp. 79–97, 2021
[4] H. Zhi, X. Xu, W. Han, Z. Gao, X. Wang, M. Palesi, A. K. Singh, and L. Huang, "A methodology for simulating multi-chiplet systems using open-source simulators," in Proceedings of the 8th Annual ACM International Conference on Nanoscale Computing and Communication, 2021, pp. 1–6.