AgenticDSE: A Multi-Agent Design Space Exploration Framework with Multi-Phase Bayesian Optimization for Chiplet Accelerators

Zhantong Zhu, Zhuolin Li, Kangbo Bai, Hongou Li, and Tianyu Jia*
School of Integrated Circuits, Peking University, Beijing, China
*E-Mail: tianyuj@pku.edu.cn

***Abstract*—Chiplet accelerators offer a scalable solution for LLM inference with reduced manufacturing costs. However, the design space exploration (DSE) of chiplet accelerators is challenging due to the complex design space. Prior black-box Bayesian optimization (BO) solutions lack domain knowledge, limiting their effectiveness. In this work, we propose *AgenticDSE*[1], a multi-agent DSE framework that incorporates the collaboration of three LLM agents, i.e., exploration orchestrator, architecture analyst, and optimization engineer. This multi-agent framework enhances exploration efficiency through analysis-driven design space refinement and phase-wise design exploration using ensemble surrogate modeling. Over the same number of explorations, *AgenticDSE* achieves up to a 36.9% reduction of average distance to the real Pareto front, and a 66% increase in diversity of explored Pareto front compared to state-of-the-art DSE solutions.**

I. Introduction

GPUs [5], [14] and dedicated TPU accelerators [8], [9], [13] are currently mainstream hardware solutions for large language models (LLMs). Compared to costly GPUs, chiplet integration has emerged as a promising approach to meet the demand for increased computational resources by AI workloads. However, designing chiplet-based hardware presents significant challenges due to the vast and complex design space. Prior studies have explored architectural and dataflow co-optimizations for chiplet designs [3], [12], [18].

To automate the architectural design space exploration (DSE) process, prior works employed black-box optimizations, which treat the design parameters as a fixed design space and employ Bayesian optimization (BO) to iteratively evaluate promising design configurations [2], [6], [7], [15], [16], [19].

As shown in Fig. 1(a), simply applying conventional BO-based DSE for chiplet accelerators will pose new challenges. *First*, chiplet architectures introduce new design dimensions at the package level, together with diverse parallelism choices. Since LLM deployment introduces diverse parallelism strategies across multiple chiplets, it significantly complicates the chiplet design space. However, conventional black-box optimization solutions only treat design parameters as a fixed design space, which may contain infeasible parameter combinations and lead to inefficient exploration. *Second*, prior DSE processes focus solely on optimizing objective values, e.g., performance and cost, overlooking micro-architectural relationships between parameters and missing potential optimization opportunities. Consequently, an efficient DSE framework for chiplet architectures remains to be developed.

[1]This work has been accepted by DAC'26.

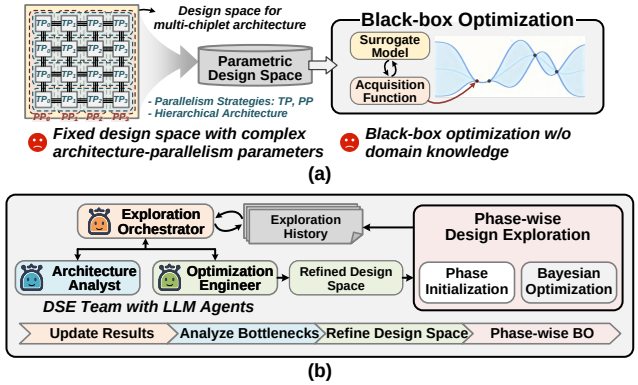

Fig. 1. (a) Limitations of black-box optimization, and (b) proposed multi-agent design space exploration framework.

In recent years, LLMs have demonstrated remarkable capabilities in understanding complex problems and generating creative solutions.In the domain of architectural DSE, LLMs have been proposed as surrogate models to suggest new design points [4], [11]. However, existing works have not evaluated LLMs for the DSE of large-scale designs, such as chiplet-based accelerators for LLM inference.

In this work, we propose *AgenticDSE*, a multi-agent DSE framework that leverages multiple collaborating LLM agents to navigate the complex design space of chiplet accelerators, as illustrated in Fig. 1 (b). By leveraging the extensive knowledge of chiplet architecture gained during LLM pre-training, along with the reasoning and collaboration between multiple LLM agents, we aim to enhance the existing BO-based approach.

II. Multi-Agent DSE Framework

In this work, we present *AgenticDSE*, a multi-agent DSE framework as outlined in Fig. 2. *AgenticDSE* features three collaborating LLM agents, i.e., **exploration orchestrator**, **architecture analyst**, and **optimization engineer**. The overall DSE flow is performed in a multi-phase manner. Initially, an LLM samples design points across the design space, provided with overall design space, along with architecture and workload information (❶).

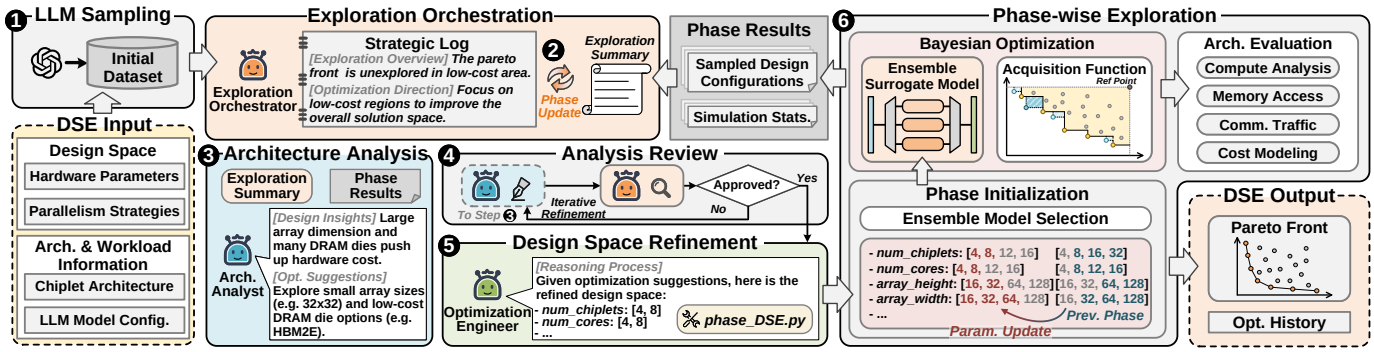

Fig. 2. Overview of our *AgenticDSE* framework, which is the first agentic DSE framework for chiplet accelerators.

During the phase-iterative DSE process, *AgenticDSE* invokes multi-agent with tight collaborations for efficient DSE. The exploration orchestrator first updates its strategic log based on the latest phase results (❷). The orchestrator queries the architecture analyst to evaluate the previous exploration phase, utilizing high-level summaries from the orchestrator and detailed phase results, i.e., design configurations and simulation statistics of the sampled design points (❸). The exploration analysis and suggestions are then sent to the orchestrator for review. If the analyst's response diverges from the orchestrator's exploration trajectory, an iterative refinement process is initiated to ensure alignment (❹). Once validated, the analysis and suggestions are sent to the optimization engineer, who translates the analyst's insights into a refined design space for the next exploration phase, facilitating dynamic adjustments to the design space and optimization direction (❺). After that, the phase-wise BO exploration is initiated by updating the candidate values for each design parameter and selecting ensemble surrogate models.

The phase-wise exploration involves iterative BO over the updated design space, incorporating ensemble learning for surrogate model construction and architectural evaluation providing optimization objectives, i.e., LLM inference latency and hardware cost, along with detailed simulation statistics to enable phase-wise evaluations by LLM agents (❻). After repeating steps ❷ - ❻ for several iterations, *AgenticDSE* terminates and reports the DSE results.

III. Evaluations

Fig. 3 shows the explored Pareto front of various DSE methods, including five baseline methods and our *AgenticDSE*. We also evaluate two mainstream LLM acceleration hardware, i.e., DGX-H100 [14] and 8 TPUv4 [8]. Chiplet-based accelerators consistently outperform these architectures with higher D2D bandwidth compared to interconnects of monolithic chips.

Normalized average distance to reference set (ADRS), diversity indicator (DIR), and hypervolume (HV) of explored Pareto front of different DSE methods are listed in Table I. *AgenticDSE* achieves 10.4%-36.9% reduction in ADRS and 4.91%-10.07% higher HV than baseline solutions, indicating a higher overall quality of its explored Pareto front. *AgenticDSE* not only explores better Pareto-optimal designs in terms of objective values, it also provides a better diversification of solutions, achieving 28%-66% increase in DIR.

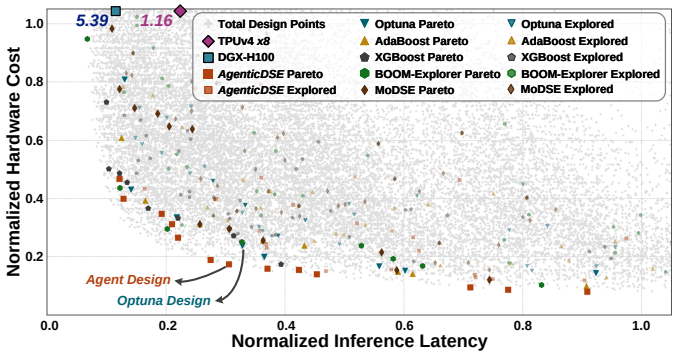

Fig. 3. Explored Pareto front from different DSE methods.

TABLE I
Comparison between different DSE methods.

Methodologies	ADRS ↓	HV (%) ↑	DIR ↑
Optuna [1]	0.086	88.06	0.42
AdaBoost [10]	0.095	88.12	0.41
XGBoost [17]	0.122	89.65	0.52
BOOM-Explorer [2]	0.087	89.96	0.41
MoDSE [16]	0.118	84.8	0.53
Agentic-AdaBoost	0.094	92.47	0.55
Agentic-XGBoost	0.100	90.03	0.58
Agentic-MoDSE	0.093	85.35	0.53
AgenticDSE	**0.077**	**94.87**	**0.68**

We also choose three baseline methods as the phase-wise exploration backend, denoted as *Agentic*-AdaBoost, *Agentic*-XGBoost, and *Agentic*-MoDSE, respectively. Incorporating multi-agent collaboration into existing DSE solutions yields up to a 21.2% reduction in ADRS, 4.35% improvement in HV, and 34.1% increase in DIR.

IV. Conclusions

This work introduces *AgenticDSE*, a novel multi-agent DSE framework for chiplet accelerators. *AgenticDSE* achieves an efficient exploration process characterized by LLM-driven design space refinement and phase-wise exploration utilizing ensemble surrogate modeling. *AgenticDSE* outperforms existing DSE solutions, achieving 10.4%-36.9% ADRS reduction and 28%-66% higher diversity in the explored Pareto front within the same number of iterations.

REFERENCES

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Jul. 2019, arXiv:1907.10902 [cs]. [Online]. Available: http://arxiv.org/abs/1907.10902

[2] C. Bai, Q. Sun, J. Zhai, Y. Ma, B. Yu, and M. D. Wong, "BOOM-Explorer: RISC-V BOOM Microarchitecture Design Space Exploration Framework," in *IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, Nov. 2021.

[3] J. Cai *et al.*, "Gemini: Mapping and Architecture Co-exploration for Large-scale DNN Chiplet Accelerators," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Mar. 2024, pp. 156–171, iSSN: 2378-203X. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10476408

[4] G. Chen, K. Zhu, S. Kim, H. Zhu, Y. Lai, B. Yu, and D. Z. Pan, "LLM-Enhanced Bayesian Optimization for Efficient Analog Layout Constraint Generation," Dec. 2024, arXiv:2406.05250 [cs]. [Online]. Available: http://arxiv.org/abs/2406.05250

[5] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "NVIDIA A100 Tensor Core GPU: Performance and Innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, Mar. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9361255/?arnumber=9361255

[6] V. Fu, M. Benazouz, L. Zaourar, and A. Munier-Kordon, "High-Performance Computing Architecture Exploration with Stage-Enhanced Bayesian Optimization," in *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, Jun. 2025, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/11132525

[7] Y. Gao, D. Luo, C. Bai, B. Yu, H. Geng, Q. Sun, and C. Zhuo, "Is Vanilla Bayesian Optimization Enough for High-Dimensional Architecture Design Optimization?" in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*. ACM, Oct. 2024, pp. 1–9. [Online]. Available: https://dl.acm.org/doi/10.1145/3676536.3676746

[8] N. Jouppi *et al.*, "TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings," in *2023 ACM/IEEE 50th Annual International Symposium on Computer Architecture (ISCA)*, 2023.

[9] S. Kim, S. Kim, W. Jo, S. Kim, S. Hong, N. Lee, and H.-J. Yoo, "A Low-Power Large-Language-Model Processor with Big-Little Network and Implicit-Weight-Generation for On-Device AI," in *2024 IEEE Hot Chips 36 Symposium (HCS)*, 2024, pp. 1–1.

[10] D. Li, S. Yao, Y.-H. Liu, S. Wang, and X.-H. Sun, "Efficient Design Space Exploration via Statistical Sampling and AdaBoost Learning," in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, Jun. 2016, pp. 1–6.

[11] J. Li, J. Zhang, Y. Li, W. Yin, and L. Wang, "LEMOE: LLM-Enhanced Multi-Objective Bayesian Optimization for Microarchitecture Exploration," in *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, Jun. 2025, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/11132704/

[12] N. B. Mallya, B. Goel, and I. Sourdis, "MEMPLEX: A Memory System with Replication and Migration of Data for Multi-Chiplet NUMA Architectures," in *Proceedings of the 39th ACM International Conference on Supercomputing*, ser. ICS '25. Association for Computing Machinery, 2025, p. 1219–1233. [Online]. Available: https://doi.org/10.1145/3721145.3725776

[13] T. Norrie, N. Patil, D. H. Yoon, G. Kurian, S. Li, J. Laudon, C. Young, N. Jouppi, and D. Patterson, "The Design Process for Google's Training Chips: TPUv2 and TPUv3," *IEEE Micro*, vol. 41, no. 2, pp. 56–63, Mar. 2021, conference Name: IEEE Micro. [Online]. Available: https://ieeexplore.ieee.org/document/9351692/?arnumber=9351692

[14] NVIDIA, "Nvidia h100 tensor core gpu architecture," https://resources.nvidia.com/en-us-data-center-overview-mc/en-us-data-center-overview/gtc22-whitepaper-hopper, accessed: 2025-11-16.

[15] C. Sakhuja, Z. Shi, and C. Lin, "Leveraging Domain Information for the Efficient Automated Design of Deep Learning Accelerators," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Feb. 2023, pp. 287–301, iSSN: 2378-203X. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10071095

[16] D. Wang, M. Yan, X. Liu, and M. Zou, "A High-accurate Multi-objective Exploration Framework for Design Space of CPU," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*, Jul. 2023.

[17] Z. Xie *et al.*, "Fist: A feature-importance sampling and tree-based method for automatic design flow parameter tuning," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2020, pp. 19–25.

[18] Z. Xu *et al.*, "WSC-LLM: Efficient LLM Service and Architecture Co-exploration for Wafer-scale Chips," in *Proceedings of the 52nd Annual International Symposium on Computer Architecture*. ACM, Jun. 2025, pp. 1–17. [Online]. Available: https://dl.acm.org/doi/10.1145/3695053.3731101

[19] X. Yi *et al.*, "Graph Representation Learning for Microarchitecture Design Space Exploration," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*, 2023, pp. 1–6.