Automating FPGA Accelerator Design using Large Language Models

Vinamra Sharma[1], Xingjian Fu[1], Jude Haris[1], José Cano[1]

School of Computing Science, University of Glasgow, Scotland, UK

ABSTRACT

The design of FPGA accelerators requires navigating through the vast space of complex trade-offs between computation, memory, and resource utilization. This work presents SECDA-DSE, an LLM-guided framework for FPGA accelerator design space exploration (DSE) that combines structured exploration, retrieval-augmented generation, chain-of-thought reasoning, and iterative hardware feedback loop. Based on the SECDA ecosystem, SECDA-DSE aims to autonomously generate SECDA-compliant accelerator designs and refine them using evaluation data. The framework is validated through generation and execution of three major accelerators designs including element-wise vector multiplication, 2D convolution, and matrix transpose on FPGA hardware. The results demonstrate the feasibility of LLM-guided DSE and highlight its potential to reduce manual design effort and streamline hardware development without need of extensive hardware expertise.

KEYWORDS: FPGA Accelerators; Design Space Exploration; Large Language Models; Hardware-Software Co-Design; FPGA Design Automation

1 Introduction

Field-Programmable Gate Arrays (FPGAs) are widely used to accelerate artificial intelligence (AI) workloads due to their flexibility, reconfiguration ability, and energy efficiency. However, designing efficient FPGA accelerators remains a big challenge as it requires exploring a large design space involving compute parallelism, memory organization, dataflow strategies, and resource allocation. As design decisions impact the performance and resource utilization vastly, choosing the right combination makes accelerator development a time-consuming process that often requires deep hardware expertise.

SECDA simplifies FPGA accelerator development by providing a hardware-software co-design workflow based on SystemC simulation, high-level synthesis (HLS) and FPGA execution. Though SECDA reduces implementation effort, identifying optimized accelerator configurations still relies heavily on manual design space exploration (DSE) and iterative tuning. As accelerator complexity grows, exhaustive exploration becomes increasingly expensive and time consuming.

[1]E-mail: {2805343s,3065487f}@student.gla.ac.uk, {jude.haris, jose.canoreyes}@glasgow.ac.uk

While, recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in reasoning, code generation, and engineering assistance. These creates opportunities to automate FPGA DSE by enabling reasoning-guided exploration of architectural trade-offs. However, the application of LLMs to FPGA accelerator DSE remains relatively under-explored, particularly in workflows that combine structured exploration with hardware-aware feedback and on-board execution.

To address this challenge, we present SECDA-DSE, an LLM-guided framework for FPGA accelerator DSE. SECDA-DSE combines structured exploration with retrieval-augmented generation (RAG), chain-of-thought (CoT) reasoning, and an iterative feedback loop to generate and refine SECDA-compliant accelerator designs. The framework uses hardware evaluation data created for this development to guide exploration decisions and adapt generated architectures across different workloads.

Further this work evaluates SECDA-DSE by generating and executing three accelerator designs on FPGA hardware: element-wise vector multiplication, 2D convolution, and matrix transpose. The generated accelerators successfully pass HLS and final hardware validation while exhibiting workload-specific architectural characteristics. These results demonstrate the feasibility of LLM-guided FPGA accelerator exploration and highlight its potential to reduce manual design efforts and development time.

This work contributes an LLM-guided FPGA accelerator DSE framework that integrates RAG, CoT reasoning, and iterative hardware feedback within the SECDA ecosystem. The work also demonstrates the applicability of the frame SECDA-DSE framework through generation and FPGA implementation.

2 Related Work

SECDA [HGC+21] introduced a hardware-software co-design methodology for FPGA-based AI accelerators, enabling rapid accelerator development using SystemC simulation and FPGA execution workflows. SECDA-TFLite[HGC+23] extended the ecosystem through reusable architectural templates and TensorFlow Lite integration, further reducing SECDA integration effort. Despite these advances, accelerator configuration and DSE remain very manual processes that require adept hardware expertise.

Recent advances in LLMs have motivated their use in hardware design automation, including architecture exploration, hardware code generation, and optimization of HLS. These efforts form part of a broader trend towards AI-assisted electronic design automation (EDA), where machine learning (ML) techniques are used to support DSE and optimization. However, the integration of LLM-guided reasoning with FPGA accelerator co-design workflows and iterative hardware feedback remains relatively under-explored.

SECDA-DSE is an LLM-guided framework for FPGA accelerator design space exploration that combines structured exploration, RAG, reasoning, and iterative hardware feedback [SFHC26a, SFHC26b]. Prior work demonstrated SECDA-compliant accelerator generation through high-level synthesis and validated the approach via end-to-end FPGA execution. This work presents the SECDA-DSE framework and its core components, highlighting how LLM-guided reasoning and iterative feedback can be integrated into a unified workflow for FPGA accelerator design generation and refinement.

Compared with prior SECDA-DSE work, this paper emphasizes the integrated workflow, hardware feedback loop, and cross-workload behaviour observed in FPGA execution.

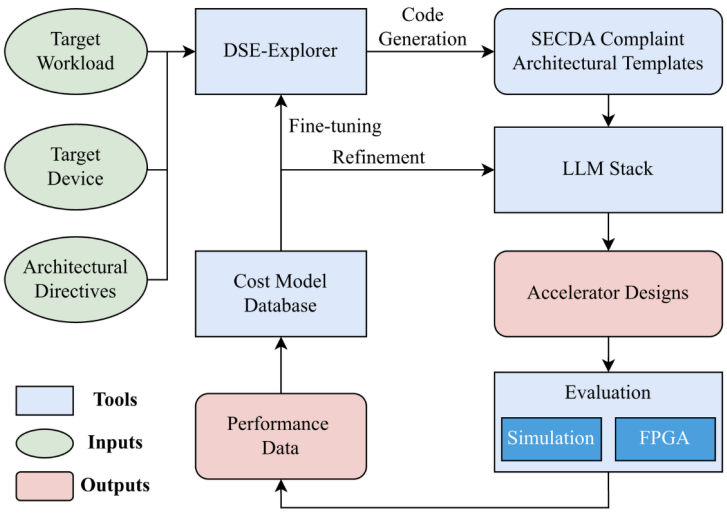


Figure 1: Overview of SECDA-DSE.

3 SECDA-DSE

SECDA-DSE is an LLM-guided framework for FPGA accelerator design space exploration built on top of the SECDA ecosystem. Figure 1 provides an overview of SECDA-DSE. SECDA-DSE takes a target workload, target FPGA device, and architectural directives as inputs to the DSE Explorer module, which generates candidate accelerator configurations. These configurations are combined with SECDA-compliant architectural templates and provided to the LLM Stack for accelerator design generation.

The generated designs are then evaluated through the SECDA toolchain using simulation, synthesis, and FPGA execution workflows. Performance metrics, including execution latency and FPGA resource utilization, are collected and stored as performance logs in a cost model database. These hardware datapoints are then reused to fine-tune for future design generation, creating an iterative feedback loop. The full LLM Stack pipeline is developed in a modular way using the open-source framework Ollama for on-device inference.

By combining retrieval-based reasoning with hardware-aware feedback and SECDA-native accelerator generation, SECDA-DSE enables more automated exploration of FPGA accelerator design spaces while reducing the manual effort traditionally required during accelerator development.

4 Results and Evaluation

SECDA-DSE was evaluated through three generated FPGA accelerators: element-wise vector multiplication, 2D convolution, and matrix transpose. All designs successfully passed synthesis and FPGA execution. Table 1 reports execution latency and resource utilization.

The generated accelerators show workload-dependent resource utilization patterns. The element-wise vector multiplication accelerator achieves the lowest execution latency (135 ms) but exhibits the highest DSP utilization (21.82%), reflecting its arithmetic-intensive nature. While, the transpose accelerator shows the highest latency (238 ms) and LUT utilization (8.85%), indicating a stronger dependence on data movement and memory reorganization.

Table 1: Xilinx Zynq-7000 FPGA execution and resource utilization of generated accelerators.

Metric	2D Conv	Vector Multiplication	Transpose
Validation	Pass	Pass	Pass
Latency (ms)	163	135	238
BRAM (%)	2.50	3.57	2.86
DSP (%)	1.36	21.82	5.45
LUT (%)	6.64	8.23	8.85
FF (%)	4.30	6.02	5.84

The convolution accelerator demonstrates a balanced resource profile while maintaining successful FPGA execution.

5 Conclusion

The SECDA-DSE framework enables LLM-guided FPGA accelerator design space exploration within the SECDA ecosystem while reducing the manual effort required during accelerator development. By integrating retrieval-augmented generation, hardware evaluation, and iterative feedback, the framework supports automated generation and refinement of SECDA-compliant accelerator designs. Evaluation with representative workloads demonstrated successful synthesis and FPGA execution of generated accelerators. Future work will focus on expanding the exploration capabilities, improving automation of the design loop, and evaluating additional workloads and FPGA platforms.

Acknowledgment

All authors gratefully acknowledge UK Research and Innovation (UKRI) and Engineering and Physical Sciences Research Council (EPSRC) funding for the AI Hub for Productive Research and Innovation in Electronics (APRIL) AI Hub [grant number EP/Y029763/1].

References

[HGC+21] J. Haris, P. Gibson, J. Cano, N. Bohm Agostini, and D. Kaeli. SECDA: Efficient Hardware/Software Co-Design of FPGA-based DNN Accelerators for Edge Inference. In *SBAC-PAD*, 2021.

[HGC+23] J. Haris, P. Gibson, J. Cano, N. Bohm Agostini, and D. Kaeli. SECDA-TFLite: A toolkit for efficient development of FPGA-based DNN accelerators for edge inference. *Journal of Parallel and Distributed Computing*, 2023.

[SFHC26a] V. Sharma, X. Fu, J. Haris, and J. Cano. LLM-Driven Design Space Exploration of FPGA-based Accelerators. In *InSyDe workshop @ EuroSys*, 2026.

[SFHC26b] V. Sharma, X. Fu, J. Haris, and J. Cano. Towards Autonomous Accelerator Design: FPGA Accelerator Generation with SECDA. In *MLArchSys workshop @ ISCA*, 2026.