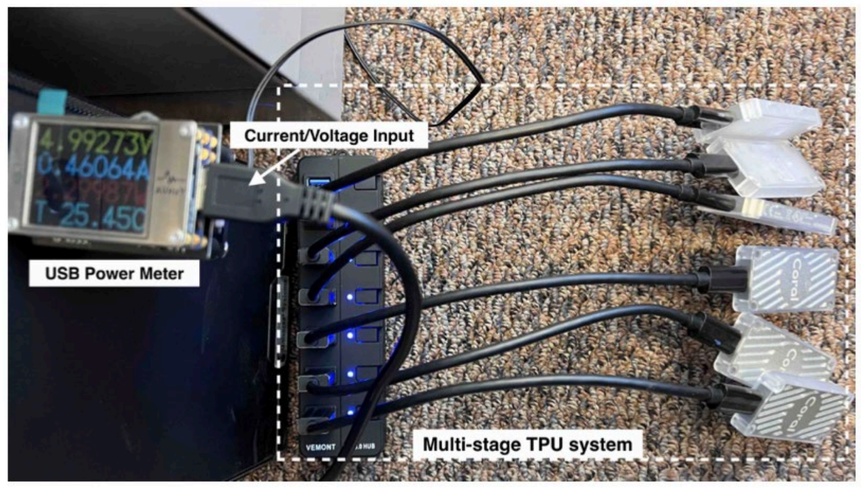
NSF-2349461/2019306 Collaborative Research: FMitF: Track I: DeepSmith: Scheduling with Quality Guarantees for Efficient DNN Model Execution


Cunxi Yu (University of Maryland - College Park), Zhiru Zhang (Cornell University)

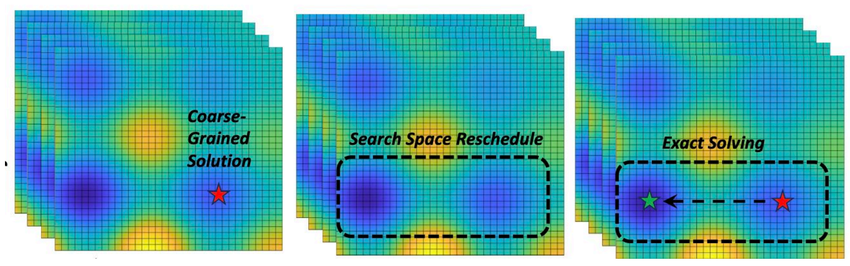
https://github.com/Yu-Maryland/FMiTF

Overview: Industries and research communities continue to face substantial computational and memory challenges, especially with the recent surge in ML/AI and cryptographic mechanisms, such as fully homomorphic encryption (FHE). This project aims to explore formal methods for generating optimal or near-optimal hardware compilation solutions, specifically by examining semi-formal approaches, i.e., combining formal methods with heuristics or machine learning.



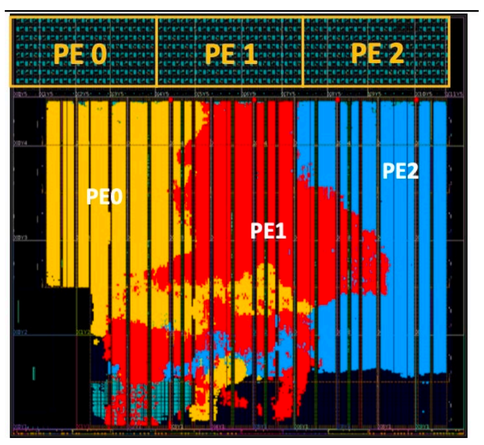
Major Challenges:

- Computational efficiency and performance are highly sensitive to resource allocation.
- Formal methods can ensure optimality in hardware compilation but struggle with runtime and scalability.
- Previous formal method-based optimizations lack flexibility due to domain-specific encoding.
- Hardware community faces a domain-knowledge barrier with formal methods.



Major Technical Solutions:

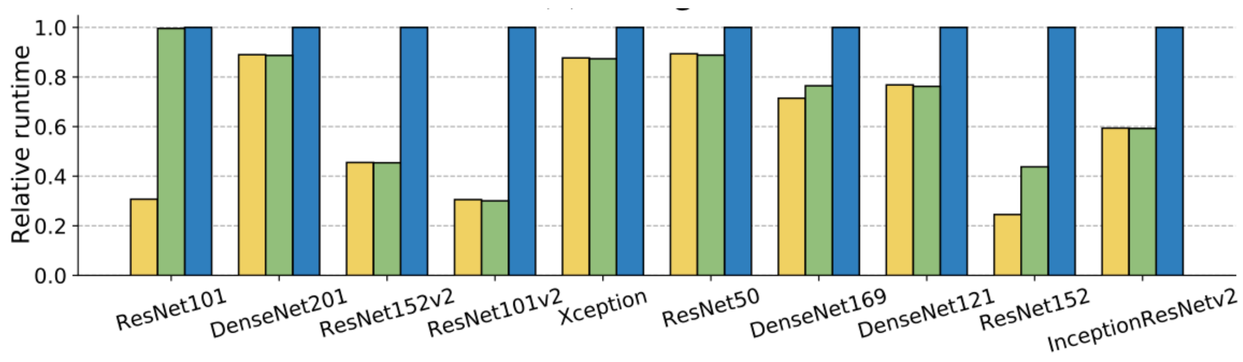
- Equality saturation graphs for compiler and hardware synthesis
- Integration of ML and formal solving to advance runtime and quality limits
- Accelerating semi-formal synthesis methodologies using differentiable programming on GPUs.

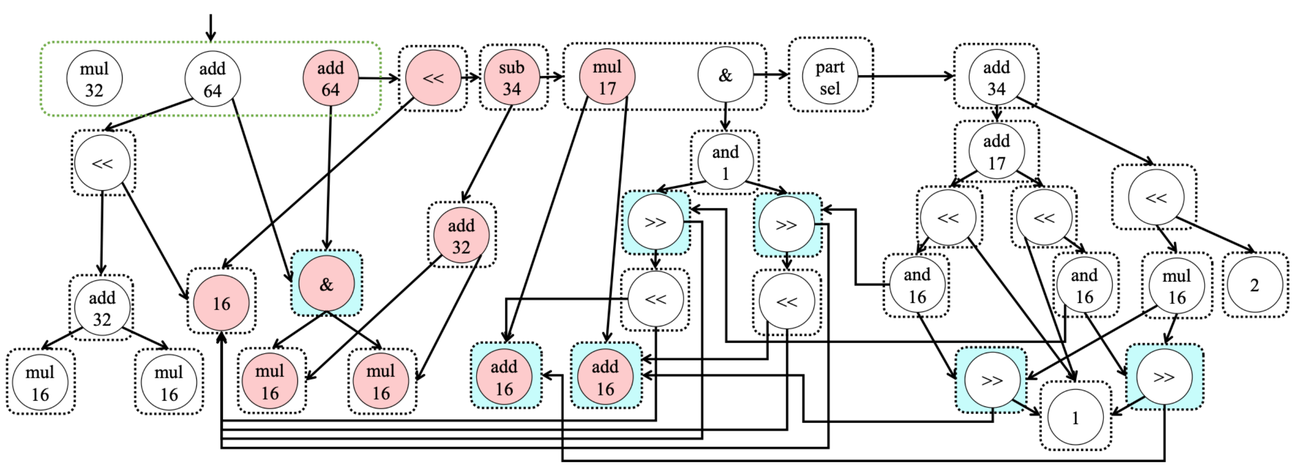


Scientific Impact

Our developed scheduling demonstrated in an end-to-end real-world edge ML system using Google EdgeTPU. The developed scheduler outperforms the commercial EdgeTPU compiler in both runtime and energy efficiency results.

Our proposed equality saturation graph based datapath optimization have demonstrated significant performance improvement in end-to-end FPGA design, which outperforms SOTA commercial tool Vitis.





Broader Impacts – Research Highlights

- Multiple outcomes released as open-source projects, including *IMPRESS, FlowTune,* and *AIM*, with over 300 stars.
- Successful industrial technology transfer in synthesis and formal reasoning using ML. Recipient of the Best Paper Award at the Design Automation Conference (DAC '23).

Broader Impacts - Education and Outreach Highlights

- Course developments in Utah, Maryland, and Cornell University, specifically in CAD of Digital Design and High-Level Synthesis.
- Research involvement with an underrepresented PhD student and an undergraduate REU student (who joined the MIT EECS PhD program).