Contribution ID: **47** Type: **not specified**

Allo: A Python-Embedded Programming Model for Composable Accelerator Design

Thursday 22 May 2025 10:00 (30 minutes)

Special-purpose hardware accelerators are critical for performance gains amid slowing technology scaling, but designers lack effective tools to build complex accelerators. Existing high-level synthesis (HLS) tools require intrusive source-level changes to attain high performance while most accelerator design languages excel only with simple kernels. In this talk, we present Allo [PLDI'24], a Python-embedded programming model for composable accelerator design. Allo decouples hardware customizations (compute, memory, communication, data types) from algorithms, and encapsulates them as primitives. By preserving program hierarchy, Allo integrates customizations bottom-up, enabling holistic optimizations across functions. Evaluated on HLS benchmarks and real-world applications, Allo outperforms state-of-the-art tools. Allo-generated accelerators for machine learning models achieve superior latency and energy efficiency compared to GPUs, showing scalability for complex, large-scale designs.

Talk's Q&A

End of talk

Will you be able to present in person?

Yes

Talk duration

20'+10'

Authors: CHEN, Hongzheng (Cornell University); ZHANG, Niansong (Cornell University); Prof. ZHANG, Zhiru (Cornell University)

Presenter: CHEN, Hongzheng (Cornell University)

Session Classification: HDL Development Tools

Track Classification: HDL development tools