CAREER: Holistic and Practical Performance Prediction for Diversely Heterogeneous Compute Systems

Overview

Computing systems are becoming more complex by scaling simultaneously in two directions: (a) the servers in data centers and supercomputers are now fostering more processor diversity, and (b) embedded systems are becoming more connected to each other and to the cloud. For example, the mobile and edge devices in emerging application domains, such as federated learning, autonomous vehicles and virtual reality, need to communicate with each other and also with cloud servers to exchange data and offload demanding jobs as such as machine learning, motion planning and 3D rendering. A key step in building future's such complex *diversely scaled systems* is to model the application behavior and hardware characteristics to predict the performance and other resource usage. Analytical modeling, simulation, and prototyping are the most common techniques used by industry and research. However, state-of-the-art approaches are long overdue for a major overhaul: (i) They are not generalizable and flexible to represent and *holistically* model diversely heterogeneous large-scale systems. (ii) Using such models during the HW/SW co-design stage or integrating them into the runtime is often ad hoc and tailored to specific application scenarios or hardware setup. (iii) Interference sources that adversely affect performance prediction accuracy, such as shared memory contention and processor sharing (*i.e.*, *multi-tenancy*), are mostly handled per processor basis. Overall, the existing performance models are far from *comprehensively* considering all these factors for the entire system.

In this project, we propose TASVIR, a new approach to performance modeling that can holistically address the current and future scalability challenges originating from computational diversity. The primary objective of our approach is to make SW/HW co-design and operation of diversely heterogeneous large-scale systems *easier*, *generalized*, *well-structured*, and more *performance efficient*. Our proposed approach spans the following:

- We propose *novel graph-based hardware representation scheme* that will enable flexible and scalable abstractions of diversely scaled systems.
- We create *slowdown models to predict interference due to shared memory usage and multi-tenancy*.
- We design a set of tools to *automate creation and integration of new performance models* from existing applications and hardware.

Intellectual Merit

The core contributions of this proposal are: (1) The new graph-based hardware representation enables the rapid development of complex and diverse heterogeneous systems that will allow system builders to quickly explore the performance of different configurations against the targeted workloads. (2) Taking memory and multi-tenancy related interference in diversely scaled heterogeneous architectures into the account will enable a more accurate prediction of the performance and resource usage in highly connected edge, autonomous, and mobile platforms. (3) The proposed automated model creation and profiling tools and the query-based performance prediction interface will further ease the integration of our proposed framework so that future diversely scaled systems can be built and operated more efficiently.

Broader Impacts

The framework developed in this project, TASVIR, will let engineers and researchers from diverse disciplines design & run complex computing systems *faster* and *more efficiently*, with more precise performance prediction capabilities. Efficient computing systems will be able to do more work with existing or less hardware, hence resulting in innovations and systems that are environmentally more sustainable.

TASVIR will also help develop the course curriculum and projects that will improve students' learning efficiency by integrating performance-oriented metrics into the assignment rubrics. This will enable a pedagogically inclusive college curriculum. In sum, TASVIR will benefit many engineering fields, promote sustainable computing, and provide enriching educational avenues.