ABSTRACT

ALGORITHM-ARCHITECTURE-HARDWARE CO-DESIGN IN COMPUTING SYSTEMS: FROM CHIP MULTICORE TO THE CLOUD

Wenkai Guan, B.S., M.S.
Marquette University

The computational demands for training deep learning models doubled every three months recently. However, according to Moore's Law, the computational power available only doubled every two years. To bridge this demand-supply gap while optimizing *energy consumption* and *carbon emission*, through my dissertation, we propose a novel algorithm-architecture-hardware co-design cross-layer approach for computing systems: from chip multicore to the cloud. In this approach, we focused on multi-objective scheduling algorithms.

At the Chip Multicore Level: How can we design high performance network-on-chip based multiprocessors that are reliable and robust to uncertainty in design parameters? This dissertation seeks to answer this question by 1) laying the foundation for uncertainty modeling and robust multi-objective optimization for embedded systems design and 2) providing computer-aided design (CAD) automation tools, which incorporate a novel design method to achieve this multi-level goal. More specifically, we significantly jump from conventional approaches by directly dealing with variability and including reliability as a design concern. We developed probabilistic (Monte Carlo Simulation) and non-probabilistic (Information Gap Theory) approaches to capture uncertainty in design parameters. Chapter 3 proposed the first uncertainty aware reliability model for NoC based chip multicore; it integrated uncertainty models as a new design methodology constructed with Monte Carlo Simulation and evolutionary algorithms. Subsequently, Chapter 4 attempted for the first time to apply the info-gap theory to uncertainty modeling in the context of embedded systems design. We developed uncertainty-aware and reliability-oriented HW/SW co-synthesis tools that can effectively explore the solution space to identify the most robust design solutions that compose the 3D Pareto frontier. **This has not been done before.** We demonstrated that significant differences between actual values and estimations of design attributes exist when uncertainty in design parameters is considered.

At the Server and Cluster Levels: How should we build generic and effective machine learning models to improve datacenter scheduling algorithms? This dissertation attempts to answer this question by proposing the **first work** of using deep learning models within a unified hierarchical approach for scheduling that combines cluster and node levels scheduling while modeling interference and heterogeneity and considering performance and energy usage as design objectives. Chapter 5 combines a unified approach cluster and node level scheduling algorithms, and it can consider specific optimization objectives including job completion time, energy usage, and energy delay product (EDP). Its novelty lies in the unified approach and in modeling interference and heterogeneity. Experimental

results demonstrated that this approach outperforms state-of-the-art schedulers from industry and academia by 41.98% in energy delay product (EDP), 38.65% in energy usage, and 10.2% in job completion time. Subsequently, Chapter 6 harnesses additional external knowledge about applications and servers to develop AI-assisted datacenter scheduling. Exploiting simplicity also, we shift existing AI-assisted datacenter scheduling approaches that rely only on internal knowledge for DNNs to a hybrid approach that relies on both internal and external knowledge for improving DNN performance.

This is only the first step towards the way we must rethink and redesign energy-efficient and carbon-free computing systems - from chip multicore to the cloud - to support emerging Artificial Intelligence (AI) and Machine Learning (ML) applications.