Abstract

Modern computing systems like laptops, smartphones, and Advanced Driver Assistance Systems (ADAS) use heterogeneous multicore CPUs, GPUs, FPGAs, and NPUs, leading to significant thermal challenges due to increased power dissipation. Traditional cooling methods like fans and heat sinks often fall short for embedded devices due to space constraints. Dynamic Thermal Management (DTM), including clock gating and Dynamic Frequency/Voltage Scaling (DVFS), provides runtime temperature control but can compromise Quality of Service (QoS) by prioritizing thermal considerations over performance. This makes on-chip temperature control while maintaining application timing a complex research problem, aggravated further by platform heterogeneity, performance demands, and the dynamic nature of workloads. Our research offers novel software-based online solutions to tackle this research problem. We introduce a series of adaptive thermal management approaches, spanning from heuristics to different learning-based techniques, including control theory and reinforcement learning, for efficient thermal-aware resource allocation on heterogeneous computing platforms. In particular, the contributions of the thesis are as follows:

- We propose a light-weight, adaptive, thermal-aware resource manager for CPU-GPU heterogeneous embedded systems. The framework mitigates dynamic thermal violations by adaptively modifying task mapping parameters using novel heuristics, with the eventual control objective of satisfying both platform-level thermal constraints and task-level deadline constraints.

- We present a Model Predictive Control (MPC) based thermal-aware scheduling framework for CPU-GPU heterogeneous embedded platforms. The framework aims to achieve smoother control of scheduling actions that account for future tasks' behaviour, and minimize the peak platform temperature while satisfying deadline constraints for task sets that can vary dynamically.

- We propose a portable, self-learning resource manager using Reinforcement Learning with Deep Q-network and Gaussian Process Regression-based models. This framework effectively manages core frequencies and task allocations across CPUs and GPUs in embedded heterogeneous computing platforms, balancing thermal management and task deadline constraints.

- Lastly, we devise an online learning-based thermal-aware resource manager using Reinforcement Learning with Proximal Policy Optimization technique for efficient thread-to-core mapping and frequency scaling in multicore heterogeneous CPU systems. The proposed solution minimizes peak and average temperatures while maintaining performance constraints, bypassing the need for exclusive system models and extensive task profiling.

Keywords:*Thermal management, Adaptive resource allocation, GPGPU, Heterogeneous computing platform*