Theme: Next Generation Computing

Sub-Theme: Application of Machine Learning Techniques for Accurate Power Prediction and Active Power/Thermal Control in Mobile Devices

Accurate application-level power prediction algorithms are important both during design and development (to track against power targets) as well as during runtime (to stay within power limits). Such power predictions are important both for short transient behavior (to stay within current supply limits or to drive scheduling decisions) as well as long duration averages (to stay within thermal constraints). Given increasing complexities of both applications and architectures, traditional simulation-based or analytical power estimation methods are either too slow or too inaccurate to drive design- or run-time decisions. Machine learning approaches hold the promise to provide such fast yet accurate power predictions simply based on tracking of workload behavior, e.g. using hardware counters. Such algorithms should use the minimum number of architecture counters to minimize runtime overhead and hardware complexity, and deliver accurate and stable phase power behavior across a wide set of relevant workloads and across a wide range of heterogeneous targets. Furthermore, they should allow integration into power management and runtime systems to drive throttling, scheduling and mapping decisions.

Given a set of applications for which power predictions are desired on a target processor, relevant research should

- Identify power phases that need to be modeled
- Determine architecture counters needed
- Develop training vectors/ directed tests and train model
- Validate model for application power prediction
- Extend model for temperature and leakage power prediction
- Use prediction models to identify power/thermal hotspots, and

- Develop optimum model-based runtime throttling, scheduling and power management policies

We are interested in the following research questions. These questions are not exhaustive but different research questions are open to discuss with research partners.

- What is the optimum number of architecture counters needed to predict power of the target workloads at desired accuracy?
- How robust is the model in terms of prediction accuracy for workloads outside the training set?
- Are different models needed for ‘short’ and ‘long’ time intervals, and if so, what algorithm should be used by the runtime engine to switch between the two?
- Are different models needed for ‘coarse’ and ‘fine’ spatial granularity (e.g., block vs. sub-block level) to trigger appropriate local micro-architectural throttling mechanisms to control local power and thermal hot-spots?

※ The topics are not limited to the above examples and the participants are encouraged to propose original idea.

※ Funding : Up to USD $100,000 per year