Abstract

Over the last couple of decades, Chip Multi-Processors (CMPs) – also called multicores, have been the leading architectural choice for computing systems ranging from battery-operated devices to high-end servers. Even though CMPs enhance performance through concurrent execution of application programs, the contention for shared resources makes the performance of individual application programs and associated energy consumption unpredictable. Modelling the inter-application resource contention is essential as it gives valuable information to achieve performance and energy optimization.

In this thesis, we first propose lightweight metrics that accurately capture the potential contention among concurrently running applications on CMP. We also show and capture the time varying behavior of applications and extract them using phase-wise profiling. With the help of these metrics, we propose and validate a methodology for optimizing performance through contention aware scheduling decisions. In addition to this, we also propose a learning model that predicts the performance of applications with partially shared cache.

On the other hand, energy consumption is a critical parameter as it constitutes the most significant operating cost for computing clouds. Analogous to this, limited use time before the need to recharge batteries, continues to be an essential user concern in mobile devices. To optimize on power consumption, modern CMP processors are designed with Dynamic Voltage and Frequency Scaling (DVFS) support at the individual core as well as for the uncore part. When applications slow down due to resource contention, typically their performance sensitivity to DVFS also reduces. Hence, the performance-energy trade-off curve of each application varies with its co-runners. In this thesis, we also model the performance-energy trade-off for applications running on a CMP platform with provision for both core and uncore DVFS. We use a learning algorithm to build the model. The proposed model is not statically tuned to support a specific DVFS policy. It builds a relationship between the DVFS steps based on the behavior of the chosen application and the co-runners represented through two simple metrics and performance. It is flexible enough to be part of any DVFS controller. We demonstrate the efficacy of our proposed model by considering various QoS policies where the user specifies the maximum permissible performance loss. The model predicts the lowest possible voltage/frequency step such that the QoS requirement is met and the energy saving is maximized.