Title: "What am I waiting for? Energy and Performance Optimization on big.LITTLE Architectures: A Memory-latency Aware Approach"

The energy demands of modern mobile devices have driven a trend towards heterogeneous multi-core systems which include various types of core tuned for performance or energy efficiency, offering a rich optimization space for software. The downside of including powerful processors for performance is heat generation. Due to the small factor design of mobile devices, it is generally not possible to include an active cooling system. As such, thermal management relies on a mix between frequency scaling mechanisms and scheduling approaches to limit heat generation. In the context of heterogeneous multi-core systems where processor performance varies, scheduling decisions become tremendously difficult. In addition to this, data coherency between heterogeneous processors is automatically ensured by an interconnect; performance of this interconnect, and by extension of the entire multi-core system, is highly dependent on the software's memory access characteristics and on the set of frequencies of each processor, adding more complexity to both scheduling and frequency scaling decisions.

Existing frequency scaling and scheduling approaches do not consider memory interaction between the different processors through the interconnect and its associated latency, and so fail to achieve a holistically good decision in such a versatile environment.

Our research provides a deep insight into the memory subsystem hierarchy of a heterogeneous multi-core system and its timing behavior relative to application characteristics, scheduling, and frequency scaling. We propose a new adaptive frequency scaling governor that does consider the memory subsystem latency; our solution uses a simple trained hardware model of cache interconnect characteristics, along with real-time hardware monitors, to continually adjust core frequencies to maximize system performance. Evaluation of our governor on the Exynos5422 SoC, as used in the Samsung Galaxy S5, demonstrates that our approach achieves a speedup of more than 40%, and a 70% energy saving on a real-world application. Representative applications of general mobile device use are also tested using web-browsing and video decoding benchmarks.