Heterogeneous Computing: Confronting the Challenges of Dark Silicon



Yatish Turakhia | **Siddharth Garg (U. Waterloo)** | **Diana Marculescu**

While transistor density continues to scale, a rapid growth in power density triggers power limitations curtailing the fraction of the chips that can be concurrently active for future designs. For performance scaling to continue, we would require exploration of heterogeneous architectures in which the "active" portions of the chip are utilized most efficiently.

To this end, we propose an iterative algorithm that generates a configuration from a vast design space that tries to maximize the performance within the area and power constraints for a given set of workloads. Our method involves comprehensive exploration design space varying different microarchitectural parameters to identify the Pareto optimal frontiers for power, area and performance for our core library. We first characterize the performance and power (Fig. 1) for these set of cores with varying areas from the given library for individual benchmarks. Our run-time aware optimization algorithm solves for a design vector (Fig. 2) that maximizes the overall performance for the given workloads within these power and area constraints. Our experiments using synthetic data involving six workloads and a library of ten cores provide an overall performance improvement of 10-15% (Fig. 3) over the best homogeneous design (Fig. 2) under the same power and area constraints.

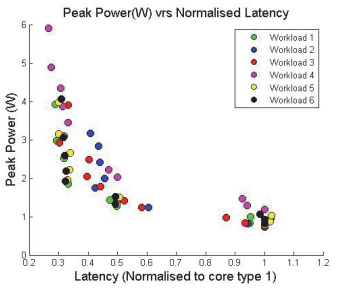

Fig. 1: Power vs. latency for workload set

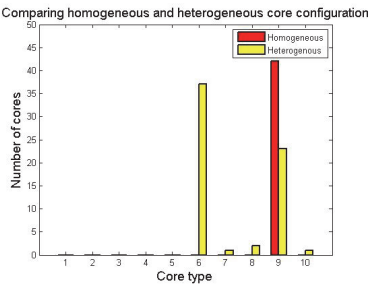

Fig.2: Homogeneous and heterogeneous configs.

Fig.3: Execution time for homogeneous vs. heterogeneous configurations