Thesis Title: Energy Aware Synthesis of Accelerators on a Network of HyperCells

Abstract:

With supply voltage no longer scaling down at the same rate as transistor feature size, keeping power dissipation to practical levels while maximizing performance is becoming a challenge in future computing systems. Increasing performance per watt for target applications is critical. Heterogeneous computing systems which consist of General Purpose Processors (GPPs), Graphic Processing Units (GPUs) and application specific accelerators can provide improved performance while keeping power dissipation at a realistic level. Application specific accelerators give the best performance per watt for a given application, but their lack of flexibility prevents their applicability in case of any small modification in the application or for a closely related application. In such scenarios, Coarse Grained Reconfigurable arrays or CGRAs are drawing increasing attention due to their promise of providing more flexibility than application specific accelerators, but with better energy efficiency than GPPs.

One key feature of the majority of CGRAs is to naturally layout computational data paths in space, so as to avoid the hardware complexity associated with general purpose processor pipelines. This makes CGRAs more energy efficient when compared to GPPs. However, existing compilation frameworks for CGRAs are targeted towards maximizing performance for a given application kernel while neglecting power dissipation. While the very nature of CGRAs make these kernels run at lower power compared to the GPPs, existing techniques do not attempt to get the least power footprints for these kernels on the CGRA. With power dissipation becoming critical, CGRA compilation techniques should try to optimize the performance for a given kernel while simultaneously optimizing for power dissipation. Extracting parallelism inherent in kernels and exposing it efficiently to the CGRA is an effective way to achieve maximum performance at minimum power dissipation.

This thesis presents a CGRA targeted for realizing kernels specified as function compositions. Function composition is defined as applying one function to the results of another to form a new function. A functional style of programming is more effective in expressing parallelism compared to imperative style and is better suited for kernels targeting CGRAs. The proposed CGRA consists of a set of reconfigurable datapaths called HyperCells which can be stitched together to form a single datapath of required granularity as dictated by the targeted kernel. We call this CGRA, a Coarse Grained Composable Reconfigurable Array or CGCRA. We also propose a synthesis methodology for mapping kernels to the CGRA, for a given performance while minimizing power dissipation. A comprehensive throughput and power model for the CGCRA proposed here enables accurate estimation of performance and energy during synthesis.

An RTL prototype for the proposed CGRA has been developed and synthesized to gate level netlist using Cadence RTL Compiler with 40 nm LowK (RVT) standard cell library from Faraday Technology. A 5X9 array with 32 HyperCells has an area of 32.27 mm^2 and can operate at a maximum clock frequency of 275 MHz. This gives a theoretical peak performance of 220 GFLOPS. A few application kernels from signal processing, machine learning, and HPC domains have been mapped to the CGCRA using the proposed synthesis methodology. Estimated power efficiency for these kernels falls within a range of 9 to 19 GFLOPS/Watts with an average 13.8 GFLOPS/Watts. Higher performance is observed for kernels with significant data reuse with a maximum observed performance of 120 GFLOPS which is 55% of the theoretical peak.