Automatic Soft CGRA Overlay Customization for High-Productivity Nested Loop Acceleration on FPGAs

Cheng Liu, Hayden Kwok-Hay So
Department of Electrical and Electronic Engineering, The University of Hong Kong
{liucheng, hso}@eee.hku.hk

1. INTROUDCTION

Compiling high level compute intensive kernels to FPGAs via an abstract overlay architecture has been demonstrated to be an effective way to improve designers' productivity. However, achieving the desired performance and overhead constraints requires exploration in a complex design space involving multiple architectural parameters and counteracts the benefit of utilizing an overlay as a productivity enhancer.

In this work, a soft CGRA (SCGRA) which provides unique opportunity to improve the power-performance of the resulting accelerators is used an FPGA overlay. With the observation that the loop unrolling factor and SCGRA size typically have monotonic impact on the loop compute time and the loop performance benefit degrades with the increase of the two design parameters, we took a marginal performance revenue metric to prune the design space to a small feasible design space (FDS) and then performed an intensive customization on the FDS by using analytical models of various design metrics such as power and overhead.

2. PROPOSED CUSTOMIZATION FRAMEWORK

Figure 1 illustrates the proposed customization framework for nested loop acceleration using an SCGRA overlay. Centering the SCGRA size and loop unrolling factor, the revenue aware (RA) design space exploration (DSE) algorithm is applied to acquire a feasible sub design space. By taking advantage of the regularity of the SCGRA overlay, the design metrics such as power and energy of the FDS can be explored effectively and customization for various design goals can be obtained. Afterwards, the customized accelerator and communication interface can be generated. Then both the software running on host CPU and the generated accelerator can be compiled to the CPU-FPGA system rapidly.

3. EXPERIMENTS & RESULTS

We take four applications including Matrix Multiplication (MM), FIR, Kmean(KM) and Sobel Edge Detector (SE) as our benchmark. In order to evaluate the quality and efficiency of the proposed design framework, we have the benchmark implemented using both the proposed RA DSE and an exhaustive search (ES) based DSE. As shown in Figure 2 and Figure 3, when compared to the ES based DSE, the proposed customization framework reduces run time by 2 orders of magnitude on average while producing similar energy-performance Pareto-optimal curve and the same customized architecture that achieves minimum energy consumption.

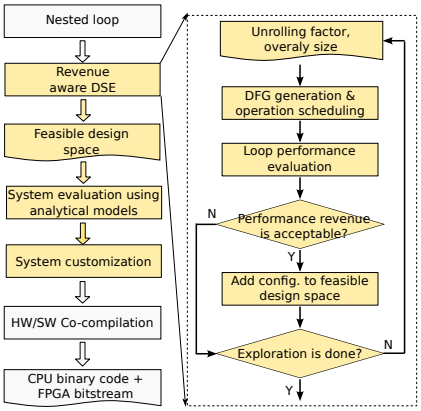

Figure 1: The proposed customization framework

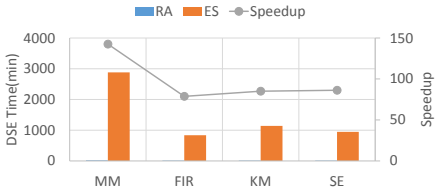

Figure 2: RA DSE Vs. ES DSE

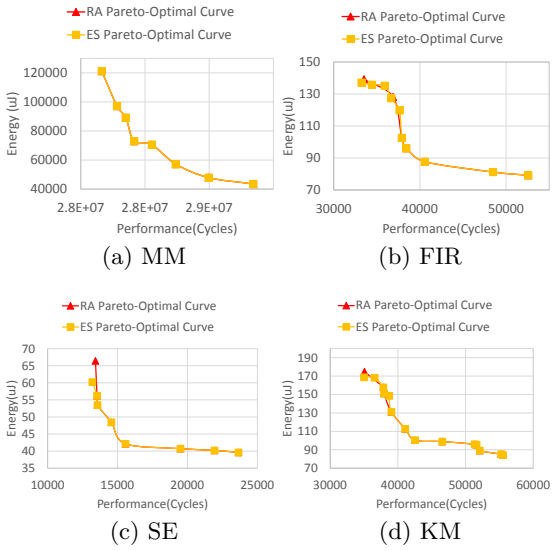

Figure 3: Performance-energy Pareto-optimal curve