Automatic Soft CGRA 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

Introduction

Compiling high level compute kernels to FPGAs via an abstract overlay architecture is 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.

To address the above problem, a soft CGRA (SCGRA) with regular tiling structure is used an FPGA overlay. Because of the regularity of the SCGRA overlay, design metrics such as power consumption, implementation frequency are highly predictable and the system customization problem can be reduced to a simpler sub design space exploration (DSE) centering SCGRA scheduling and a straightforward customization with all the potential configurations well estimated using analytical models. With the proposed customization framework, we can achieve both high design productivity and performance.

SCGRA based accelerator

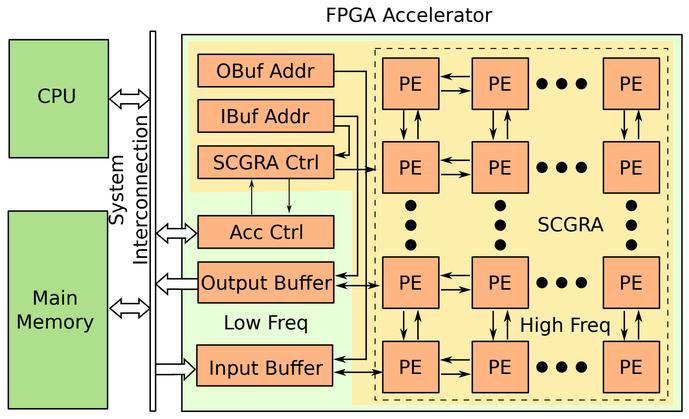


Figure 1: SCGRA based FPGA accelerator

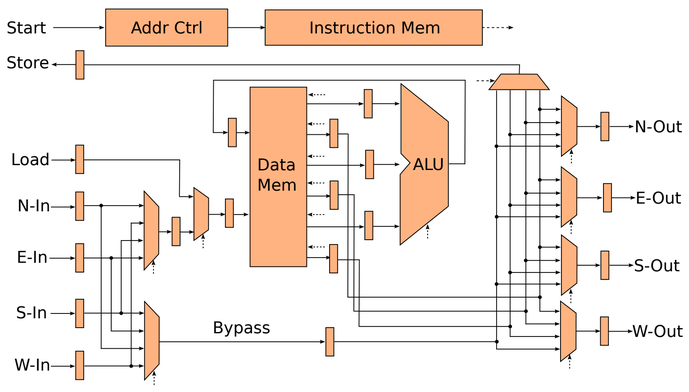


Figure 2: PE structure

Experiment setup

We took four applications including Matrix multiplication (MM), FIR, Kmean(KM), and Sobel edge detector (SE) as our benchmark. The detailed configuration of the benchmark is listed in the following table.

Benchmark	Parameters
MM	Matrix Size(128)
FIR	# of Input (1024) # of Taps+1 (64)
SE	# of Vertical Pixels (128) # of Horizontal Pixels (8)
KM	# of Nodes(1024) # of Centroids(4) # of Dimensions(2)

FPGA acceleration framework

QuickDough–a rapid FPGA acceleration design framework using soft CGRA overlay is presented in Figure 3.

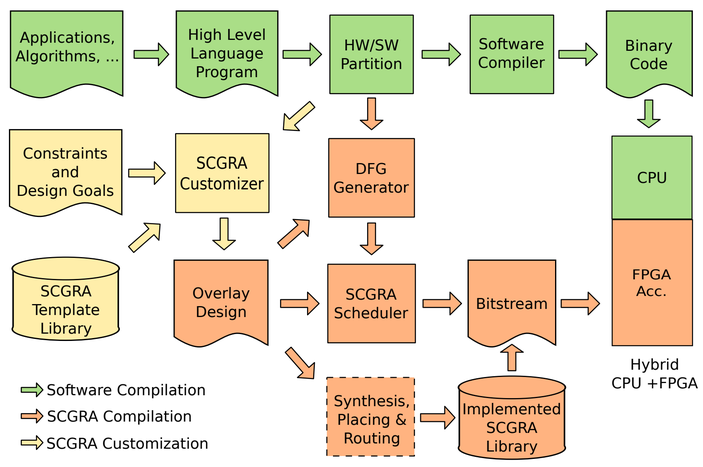


Figure 3: QuickDough: a rapid FPGA acceleration design framework

Customization framework

The customization is a critical part of QuickDough. By taking advantage of the regularity of the SCGRA overlay, the complex nested loop acceleration problem is greatly simplified and divided into two simpler steps as shown in Figure 4.

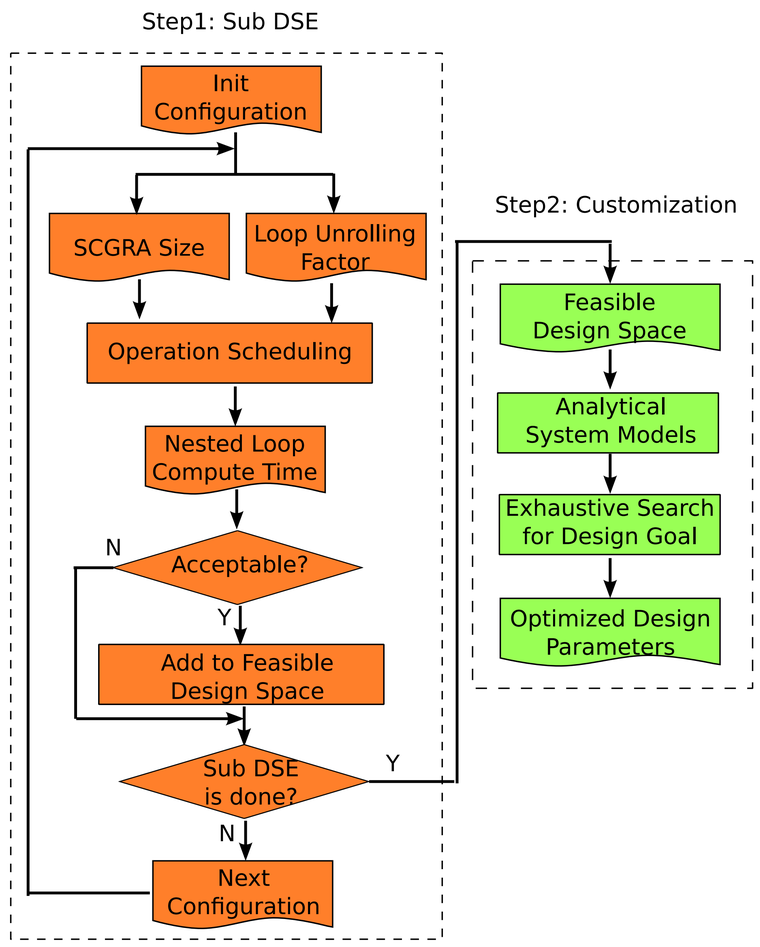


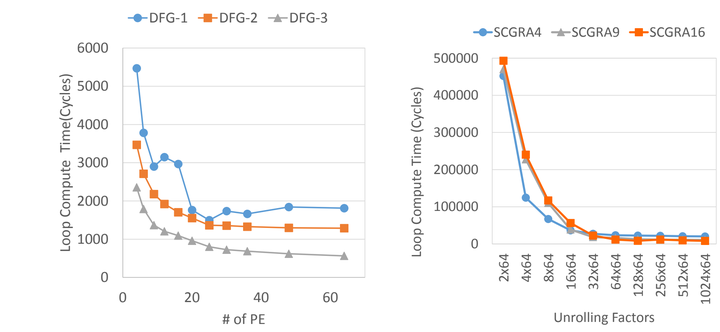
Figure 4: Customization framework for nested loop acceleration

Conclusion

In this work, we have presented an automatic nested loop acceleration framework based on a homogeneous SCGRA overlay built on top of off-the-shelf FPGA devices. Given high level user constraints and design goals, the framework performs an intensive system customization specifically to a nested loop. According to the experiments, it can produces quite similar energy-performance Pareto-optimal curve to that obtained from an exhaustive search. The customized implementations exhibit competitive performance to the optimized HLS implementations.

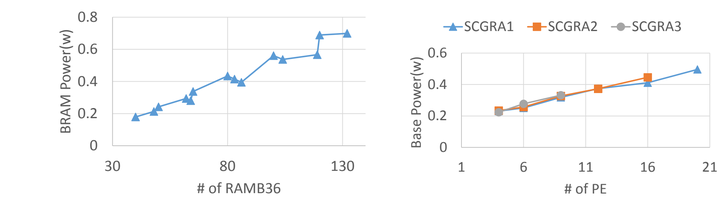
Experiment results

The customization problem is simplified with the observations in Figure 5 and Figure 6.



(a) Unrolling factor (b) SCGRA size

Figure 5: Design parameters influence on loop perforemance



(a) BRAM power (b) Power excluding BRAM power

Figure 6: Predictable power consumption of the SCGRA overlay based FPGA accelerators

The benchmark is implemented using both the proposed two-step customization (TS) and exhaustive search (ES) customization. Figure 7 shows the DSE time comparison and Figure 8 presents the Pareto-Optimal curve comparison. Figure 9 shows the benchmark performance of implementations using the proposed design framework and Vivado HLS.

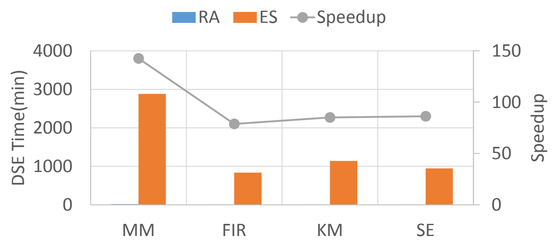
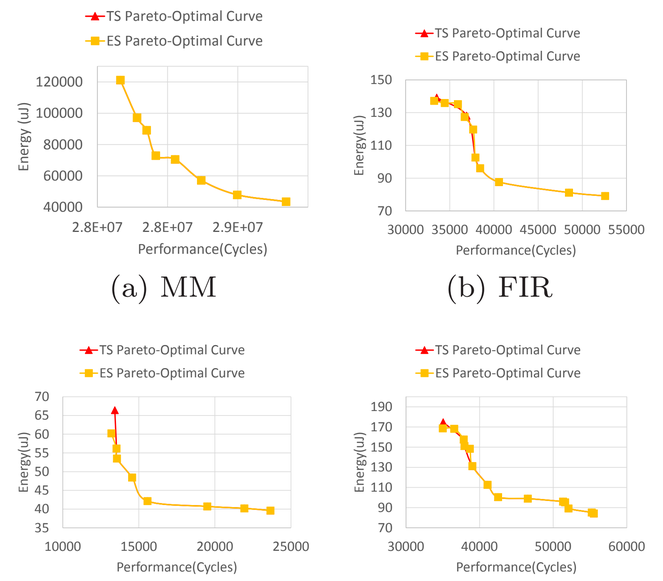
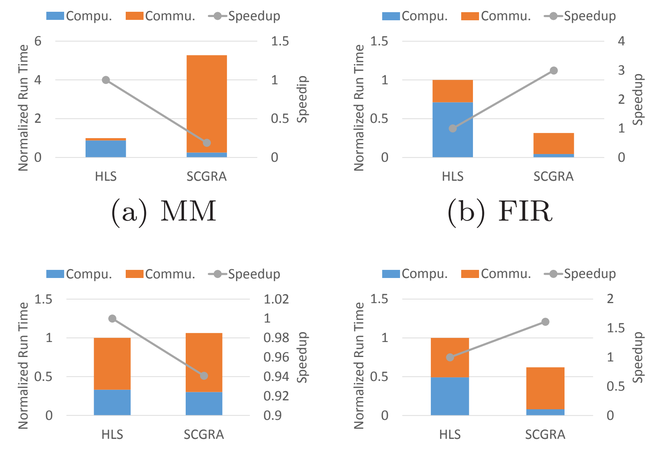


Figure 7: RA DSE time Vs. ES DSE time



(a) MM (b) FIR (c) SE (d) KM

Figure 8: Performance-energy Pareto-optimal curve comparison between RA DSE and ES DSE



(a) MM (b) FIR (c) SE (d) KM

Figure 9: Performance comparison between the customized design and optimized HLS design