Performance Optimization of Memory Intensive Applications on FPGA Accelerator

Doctoral Dissertation
Doctoral Program in Electronics Enginering (31st cycle)

Performance Optimization of Memory Intensive Applications on FPGA Accelerator

Arslan Arif
* * * * * *

Supervisor
Prof. Luciano Lavagno, Supervisor

Doctoral Examination Committee:
Prof. Jordi Cortadella Fortuny, Referee, Universitat Politecnica de Catalunya, Spain
Prof. Frederic Petrot, Referee, Universite Grenoble Alpes, France
Prof. Paolo F. M. Ienne Lopez, Ecole Polytechnique Federale de Lausanne, Switzerland
Prof. Mihai T. Lazarescu, Politecnico di Torino, Italy
Prof. Alex Yokovlev, Newcastle University, England

Politecnico di Torino
February 28, 2019

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Arslan Arif

Turin, February 28, 2019

Summary

Hardware accelerators are a fundamental part of modern high performance computing (HPC) systems due to their performance capabilities. The two most commonly used accelerators are GPUs and FPGAs. Despite the easier programmability and better memory performance of GPUs, generally FPGAs perform equally well for computationally challenging applications while dramatically reducing the energy consumption. Furthermore, with the availability of high level synthesis (HLS), the use of FPGAs has become easier. This makes them an excellent candidate for modern HPC systems. This dissertation describes my research work done in the field of electronic design automation with the major focus on optimizing memory intensive applications modeled using high level language for FPGAs. This work can be split into two parts, one dealing with manual memory optimization while other advocates the use of automated algorithms to select and optimize the best application-specific cache layout.

The first part covers the manual optimization of a realistic smart city application. The application implements two image processing algorithms in OpenCL language which computes velocity and density of vehicles on urban streets in real time. Several different implementations of these memory hungry algorithms are considered. The results show that using suitable optimizations and HLS optimization directives, FPGAs can produce results with performance similar to a GPU with an order of magnitude less energy consumption.

The second part of the dissertation starts by observing that custom data caches implemented on FPGAs are only useful if their layout is in accordance to their data access pattern. In this work, we present a tool, PEDAL (Pattern Evinced Determination of Appropriate Layout), that can automatically tune the custom data caches based on analyzing address traces. PEDAL uses artificial intelligence algorithms to detect the pattern of each array and then design the optimal cache for that pattern. The comparison of the results of PEDAL with the exhaustive search of cache configurations and cache designed through a state-of-the-art algorithm from the literature proves that it can produce better configurations in less time.

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.