1859
Politecnico
di Torino
ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Electrical Engineering (35thcycle)

Improving Quality of Results (QoR) for High-Level Synthesis (HLS) based FPGA designs

By

M. Usman Jamal

******

Supervisor(s):
Prof. Luciano Lavagno, Supervisor

Doctoral Examination Committee:
Prof. Roberto Passerone, Referee, Universita degli Studi di Trento
Prof. Mohammad Mozumdar, Referee, California State University at Long Beach
Prof. Mario R. Casu, Politecnico di Torino
Prof. Mihai T. Lazarescu, Politecnico di Torino
Dr. Osama B. Tariq, Newcastle University

Politecnico di Torino
2023

Declaration

I hereby declare that, the contents and organization 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.

M. Usman Jamal
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Improving Quality of Results (QoR) for High-Level Synthesis (HLS) based FPGA designs

M. Usman Jamal

High-level synthesis (HLS) is an Electronic design automation (EDA methodology that work towards designing complex digital systems using high level programming languages like C, C++ or SystemC and automatically transforms them into a hardware description language (HDL) in a relatively short time. This not only increases designer productivity but also helps in exploring different designs faster and trade offs between cost and performance. One of the open issues of HLS is the memory bandwidth bottleneck which limits the performance which is extremely important for the memory bound algorithms. Thus, designs implemented on Field-programmable gate array (FPGA) via HLS suffer from this bandwidth bottleneck and off chip memory latency. Current HLS tools are incapable of automatically exploiting the memory hierarchy on FPGAs and the only way to exploit the memory hierarchy is in a scratchpad fashion, but this requires considerable design effort and therefore, time-consuming. Secondly, the existing HLS tools currently exhibit a deficiency in providing dependable estimates of final Quality of results (QoR), thereby impeding designers ability to make well-informed decisions regarding the trade-offs between cost and performance.

This thesis explores and examines both these issues and addresses them by developing solutions in order to overcome aforesaid matters in question.

The first part of this thesis addresses the issue of off-chip memory latency and bandwidth bottlenecks in FPGA designs implemented via HLS. We propose an automated FPGA memory management approach using a fully-configurable source-level cache in Xilinx *Vitis HLS*. The primary objective of our cache implementation is to minimise the amount of design effort required while enabling the designer to focus on algorithmic optimizations, specifically for memory access patterns that are data-dependent or irregular in nature. Experimental results shows that our cache implementation improve the performance of different benchmarks by up to 60 times compared to the out-of-the-box HLS solution.

The second part pertains to the enhancement of QoR estimation in HLS. For this purpose, by taking advantage of the widespread use of Machine learning (ML), we propose Graph neural network (GNN)-based model that learn and predict post-implementation QoR using pre-schedule control data flow graphs (CDFGs) and HLS optimization directives. Experimental results show that our model can estimate the timing and resource usage of a previously unseen design (i.e, a completely new CDFG) within milliseconds with high accuracy, reducing prediction errors by up to 74 % compared to the estimate generated by the HLS tool.

Acronyms

EDA electronic design automation

FPGA field-programmable gate array

GNN graph neural network

HLS high-level synthesis

ML machine learning

QoR quality of results