ENERGY OPTIMIZATIONS FOR SCRATCH PAD MEMORY BASED SIMD ARCHITECTURES

NAMITA SHARMA

AMAR NATH AND SHASHI KHOSLA SCHOOL OF IT
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
APRIL 2015

ENERGY OPTIMIZATIONS FOR SCRATCH PAD MEMORY BASED SIMD ARCHITECTURES

by

NAMITA SHARMA

Amar Nath and Shashi Khosla School of IT
Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of
Doctor of Philosophy

to the

Indian Institute of Technology Delhi
April 2015

Certificate

This is to certify that the thesis titled **Energy Optimizations for Scratch Pad Memory based SIMD Architectures** being submitted by **Ms. Namita Sharma** for the award of **Doctor of Philosophy** in Amar Nath and Shashi Khosla School of IT is a record of bona-fide work carried out by her under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Preeti Ranjan Panda
Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi
New Delhi - 110 016

Acknowledgments

The satisfaction that accompanies the successful completion of this thesis would be incomplete without making mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts. First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Preeti Ranjan Panda, Department of Computer Science and Engineering, IIT Delhi for providing me an opportunity to work under his precious guidance. It is my fortune to work with Prof. Panda who has been enthusiastically guiding me at every stage of this work and encouraging me to achieve the targets set well in time. Working with him was a great pleasure and vast learning experience. His critical comments during the discussions enabled me to overcome my weaknesses and make valuable improvements. Thanks for your belief in my abilities and to direct my passion towards the right direction.

It is my fortune to have my thesis work in collaboration with IMEC, Belgium. Thanks to Prof. Francky Catthoor for agreeing to serve as my co-supervisor at IMEC. I really admire him for his patience and long future insight he has. I would like to thank him for keeping timely meetings that made this distant collaboration fruitful. I would also like to extend my thanks to my colleagues – Tom Vander Aa, Praveen Raghavan, Eddy De Greef, Amir Amin and Min Li whose timely assistance greatly helped in developing a deep understanding of the project. Special thanks to my PhD friends Prashant Agrawal, Matthias Hartmann, Raf Appeltans, Robert Fasthuber, and Halil Kukner for friendly discussions and support that made my stay at IMEC memorable.

My sincere thanks to Professors Anshul Kumar, G.S. Visweswaran and Vinay Ribeiro for serving on my Student Research Committee. Thanks to my labmates – Vaibhav Jain, Sharat Varma, G. Krishnaiah, Swanti Satsangi, Rajshekhar Kalayappan, Sandeep Chandran, and Prathmesh Kallurkar for their great company. I would also like to thank S.D. Sharma and Vandana Ahluwalia for their timely assistance in lab related issues.

Finally, I dedicate this thesis to my beloved parents Gajendra Sharma and Uma Sharma for their blessings, support, and persistence which has always been a source of inspiration that kept me going to complete this venture. Special thanks to my mom and brother Hardik Sharma who always encouraged me and shared my tough and sweet moments without any hesitation.

Namita Sharma

Abstract

Mobile communication devices have evolved tremendously in the last few decades. These devices are not bound to an application but are useful for a wide range of functionalities. For example, the modern smartphones provide not only the calling facility but also multiple other functionalities such as 3D gaming, Internet access, Bluetooth, and so on. These evolutions require higher data rates, and low latencies. To achieve these features, techniques such as OFDM and MIMO are being introduced leading to co-existence of multiple wireless standards. Separate standards exist for supporting varying connectivity ranges. Relying on hardware implementation to accommodate these features on mobile devices leads to design challenges such as higher area for multiple hardware units, costlier implementation, longer time to market and so on. To overcome these challenges, having the entire baseband running on a programmable architecture is an attractive alternative. This is possible if we have the wireless physical layer functions software defined and run on a re-configurable architecture. This is popularly known as Software Defined Radio (SDR) implementation. With SDR architectures, we overcome the drawbacks of custom hardware implementations by bringing down the implementation cost, reducing the size as resource sharing is applicable, lowering the time-to-market as standard revisions simply require software upgradation rather than building the customized hardware. SDR thus proves to be an attractive alternative as it enables the reuse of design efforts across generations.

Battery life limits the utility of handheld devices such as mobile phones. With more and more applications and features enabled in modern devices the energy consumption is increasing. Current smart phones suffer a major drawback of short battery life, with average lasting time of less than a day. This poses a need to lower the energy consumption so as to increase the battery life. The lower the energy consumption, the longer the battery life will be. Bigger batteries are not a good solution for this problem as the portability of handheld devices decreases with increasing device sizes and weights. Thus, there is a need to have energy efficient implementations for applications running on such architectures. Memory plays a dominating role in the system design leading to significant amount of research in data and memory optimizations. Optimizations related to memory accesses and data storage make a significant difference to the performance and energy of a wide range of data-intensive applications. Such strategies need to

evolve with modern SoC and processor architectures, which lead to new optimization opportunities. In this thesis, we propose some strategies that significantly reduce the memory accesses thereby reducing the overall implementation energy. The proposed strategies are targeted at embedded processor systems with features such as scratch pad memories (SPM), Single Instruction Multiple Data (SIMD) FUs, and vector register files with wide interfaces to both SPM and FUs. The main contributions of this thesis are:

- We study the inter- and intra-kernel data dependencies for the LTE MIMO wireless standard and propose and compare efficient data layouts for the application. Our focus is on the LTE downlink receiver as this is a data- and computation-intensive part of the LTE application with tight energy and latency constraints.

- Array Interleaving, the proposed data layout transformation for the LTE application, is further developed to make it a more broadly applicable compiler transformation strategy for SIMD architectures. Based on the estimations of the benefits and interleaving costs, we perform a global analysis of arrays accesses spanning multiple loops, and take interleaving decisions that are predicted to be globally optimal. We also incorporate the effect of interleaving granularity in our exploration.

- We propose an energy efficient data flow transformation for Givens Rotation based QR Decomposition, a widely used function for matrix inversion and triangularization, for mapping to SIMD architectures. The proposed sequencing strategy is compared with the conventional sequences for matrices of different sizes. We also explore different possible implementations for QRD of multiple matrices using the SIMD feature of the processor.

- We propose a strategy to select an energy optimal Register File (RF) size for FFT processing on input data having large percentage of zeros. FFT is widely used in signal processing and image processing domains to transform the inputs in the time domain to the frequency domain for simplifying the computations. The strategy, based on memory access and compute count estimates, results in an energy efficient RF configuration out of the large number of configurations possible with the variation in number of registers and the SIMD widths.

With the proposed data layout strategy – Array Interleaving, for LTE application, a reduction of 7-15% in memory energy consumption is obtained over a highly hand-optimized code. Through an exploration for inter- and intra- kernel data dependencies in the application we conclude that there exist no dependencies across the Physical Resource Blocks (PRBs) allowing the merging of some kernels in the data processing block, thereby reducing the overall memory access count by around 15%. Experimental evaluation of the proposed layout strategy on

several applications in the wireless communication, multimedia, and image processing domain showed a significant reduction (6% – 34%) in memory energy consumption.

Our proposed data flow transformation strategy for QR Decomposition results in up to 36% reduction in overall energy across different implementations. Up to 75% reduction in memory accesses is achieved over the conventional sequences. A comparison of these sequences with standard tiling transformation shows that tiling results in an energy efficient implementation with respect to conventional sequences but has higher energy consumption compared to the proposed sequencing.

Using our proposed exploration strategy for RF size selection for FFT, we obtain a reduction of up to 59% in data memory energy. This percentage variation is obtained when comparing over the range of possible RF configurations with varying number of registers and SIMD widths.

Contents

List of Figures

List of Tables