ACCELERATING DNA SEQUENCE ALIGNMENT BASED ON SMITH WATERMAN ALGORITHM USING RECURSIVE VARIABLE EXPANSION

MUHAMAD FAIZ BIN ISMAIL

FACULTY OF ELECTRICAL ENGINEERING
UNIVERSITI TEKNOLOGI MARA
MALAYSIA

ACKNOWLEDGEMENT

In the name of Allah, the Beneficent and the Merciful. It is with the deepest sense of gratitude to Allah who has given the strength and ability to complete this project and the thesis as it today.

As a token of appreciation, I wish to express my deepest gratitude to my supervisor, En. Syed Abdul Mutalib Al Junid. En. Syed Abdul Mutalib Al Junid highly skilled guidance, patronizing critics, valuable comments, stimulating discussion and generous support during the execution of this project have been very valuable. I started the project with little knowledge of how computer architecture is build. At the end of the project, I managed to build the whole computer architecture and successfully putting the algorithm in operation.

I also want to extend my gratitude and appreciation to both honourable panels, Datin Dr. Fuziah and Madam Zaiton for their comments, valuable suggestions and outstanding advisement to improve the project during the project presentation.

Most importantly, I would like to express my deepest gratitude to my beloved parents and sisters who have supported me throughout my years at the university.

ABSTRACT

Reconfigurable computing, in which general purpose processor (GPP) is increasingly used for high performance computing where massive fine-grain parallelism can be exploited. A challenge is to exploit such massive parallelism on LabVIEW and more specifically how to map an application on the heterogeneous underlying platform. Similar to hardware compilers, software compilers can use loops to exploit such parallelism. The existence of a dependence between data is one the constraints that limits parallelism in a program. In this dissertation, we propose a transformation called Recursive Variable Expansion (RVE), which can be applied to an important category of loops. It removes all the data dependences by expanding the variable with its dependence expression until the expression becomes only a function of known variables. We classify two types of expressions, one which expands polynomially, and other which expands exponentially on the number of input variables. Irrespective of the type of expression, when we map an expression on LabVIEW, the area (LUT) required is proportional to the number of terms in the expression.

The exponentially expanding version is applicable to the category of dynamic programming (DP) problems for which RVE is combined with dataflow. We demonstrated better performance than dataflow only, which is the best technique known so far for such problems. We generalize the approach by proposing a framework such that the technique can be applied to a large range of DP problems.
Finally, we validate the proposed DP framework using the Smith-Waterman (SW) algorithm, which is a widely used, computation and data intensive application in bioinformatics. We show that our implementation yields a two times speedup at the cost of almost three times more area as compared to the conventional dataflow traditional array implementation.

Keywords: LabVIEW, Recursive Variable Expansion, dynamic programming, S-W algorithm

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

There are many computer applications from various fields whose computational demands exceed conventional processor's capability. A few examples include applications in the domain of financial analytics, bioinformatics, data mining, medical imaging and scientific computations. Even though all these applications have different program requirements, performance is a common objective.

Over last few years, we have seen a shift towards heterogeneous systems for high performance computing (HPC). In heterogeneous systems, a general purpose processor is augmented with application specific hardware or processors. This heterogeneous system of processors can be on multiple boards or one board connected with high bandwidth interconnections or can be on a single chip. The application specific hardware gives better performance/area and performance/power for specific applications as compared to a homogeneous system of processors; therefore overall, heterogeneous systems reduce the area and power requirements.

Computational methods in the field of biology have become a key factor since the advent of the human genome project in 1990. Since then many other genomes have been sequenced, generating a wide variety of sequence analysis problems. The sequencing of the human genome have impacted the study of human disease in significant ways and enabled many genome-wide association studies that aim to explain the genetic component of complex diseases. The recent introduction of instruments capable of producing