Ain Shams University
Faculty of Computer & Information Sciences
Computer Systems Department

Solving Computationally Intensive Problem Using GPUs- Solving Motif Finding Problem as a Case Study

A Thesis submitted to Computer Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University,
in partial fulfillment of the requirements for
the Ph.D. of Science in Computer and Information Sciences

By

Mirvat Mahmoud Ahmed Al-Qutt
M.Sc. in Computer and Information Sciences,
Teaching Assistant at Computer Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University.

Under Supervision of

Prof. Dr. Hossam El-Deen Mostafa Fahim
Prof. of Computer Systems,
Department of Computer Systems,
Faculty of Computer and Information Sciences, Ain Shams University.

Dr. Rania Abdel Rahman El Gohary
Associate Professor,
Department of Information Systems,
Faculty of Computer and Information Sciences, Ain Shams University.

Dr. Heba Khaled Ahmed Mahmoud
Assistant Professor,
Department of Computer Systems,
Faculty of Computer and Information Sciences, Ain Shams University.

Cairo 2018

Acknowledgement

First and foremost, thanks to Allah, the most merciful, the most graceful, for His great help throughout this work.

I sincerely acknowledge and express my appreciation to prof. Dr. Hossam Faheem for his valuable advice during the study and fulfillment of this thesis and for his support to get an access to High Performance Computing Center in King Abdulaziz University (Aziz Supercomputer) to run the work described in this thesis.

I owe my deepest gratitude and appreciation to my supervisors for their great care, valuable advices, and helpful guidance to carry out this work.

Thanks also are extended to my colleague in the Computer System department Mahmoud Fayez for his kind support and help.

My special gratitude is due to my beloved late father, my warmhearted mother, my all of my beloved family members who supported me throughout this work with all their love and faith, without their encouragement and understanding it would have been impossible for me to finish this work.

Mirvat Al-Qutt.

Abstract

With scientific problems getting larger and more complicated, level of parallelization needed to be increased to provide the needed amount of computational power, so more adequate, efficient and robust parallel environments became an essential need for scientific research. This thesis proposes hybrid parallel paradigms to solve computationally intensive problem, taking into consideration one of the most common problems in Bioinformatics, Motif Finding as a case study. Motifs are defined as the short patterns these consist of nucleotide; they are usually positioned close to the genes contained in Deoxyribo Nucleic Acid (DNA). They occur in a frequent manner within the sequence, these patterns are not identical but it comes with some mutations in several of their nucleotide positions. Motif Finding Problem aims to discover unknown motifs that are expected to be common in a set of sequences. Generally it can be viewed as a large-length sequence matching problem.

There are a numerous number of algorithms available to solve this problem; these can either be exact or approximate. Motif finding problem is categorized as one of the most computationally intensive problems in the field of bioinformatics and it requires a large amount of memory. It has been categorized as Nondeterministic Polynomial Time Order Problem. Both software and hardware accelerators have been proposed and deployed to accelerate Motif finding problem algorithms, Software based acceleration solutions are easier to implement and do not require hardware experience.

This research is kicked-off by exploring the existing algorithms and parallelization techniques for accelerating Motif Finding problem and identifying the main points that could represent a contribution.

In this research, SKIP brute-force algorithm is accelerated using different High Performance Computing implementations. The results showed that GPU significantly reduced the execution time and improved the performance better than using Multicore architecture. In Addition, it has been noticed that different sequence lengths affect the speedup and power consumption. These implementations are considered a step towards building a complete parallel architecture for solving computationally intensive problems of bioinformatics.

Consequently, machine learning techniques have been exploited to develop an intelligent recommender system that advices the optimal HPC architecture for specific Motif finding problem size and other parameters. A neural network–based multi-objective optimization approach is employed (Neural Network Inversion), which is used for direct problem approximation by mean of a neural network. The objective functions are maximizing the speedup ratio and minimizing the power consumption. The importance of this system is clear as it employs an automatic decision regarding optimal number of processors in terms the optimal hardware configuration that can efficiently solve computationally intensive problems taking into consideration the resources availability and a set of requirements and environment restrictions that might be conflicting sometimes. The proposed system achieved prediction accuracy reached over 89 % with “390” iterations on average for CPU based hardware configuration prediction system, and 87% with “306” iterations on average for GPU based hardware configuration prediction system.

Finally, considering the solution scalability, cloud platform services have been explored for storage and analytics purposes, taking into consideration the huge volume of data, tremendous cloud storage platforms that currently exist and provide a scalable, distributed storage

service. Therefore, the challenges of implementing Motif Finding based on employing the services provided by the cloud storage platform are addressed. In this research, Apache Hadoop is picked for Big Data processing which is an open source Platform that provides enormous number of clusters that is used for parallel implementations. MF solution was implemented using two different Big Data frameworks: MapReduce and Apache Spark, they have different implementation schemes and resources usage plans. The results are collected and analyzed against different parameters such as the speedup value and power consumption. From experiments Spark achieves much more speedup than MapReduce. This finding is expected due to MapReduce Input/Output operations overheads. The main purpose of this step is to accomplish effective parallelization and evaluates the performance of such an integrated solution.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural network
API	Application Program Interface
CPU	Central Processing Unit
CUDA	Computer Unified Device Architecture
DNA	Deoxyribo Nucleic Acid
FPGA	Field-Programmable Gate Array
GPU	Graphics Processing Unit
HDFS	Hadoop Distributed File System
HPC	High Performance Computing
HW	Hardware
IaaS	Infrastructure as a Service
IT	Information Technology
MEME	Multiple Expectation Maximization for Motif Elicitation
MFP	Motif Finding Problem
MIC	Intel Many Integrated Cores
MIMD	Multiple Instruction Multiple Data
ML	Machine Learning

MOOF	Multi-Objective Optimization Function
MPI	Message Passing Interface
MSE	Mean Square Error
MSP	Median String Problem
NBP	Nonamer-Binding Protein
NGS	Next Generation Sequencing
NP-Complete	Nondeterministic Polynomial Time Order Problem
NUMA	None Uniform Memory Access
OpenHMPP	Open Hybrid Multicore Parallel Programming
PaaS	Platform as a Service
PMP	Planted motif problem
PMSE	Power Mean Square Error
PSSM	Position Specific Scoring Matrix
RDBMS	Relational Database Management Systems
RDD	Resilient Distributed Datasets
RNA	RiboNucleic Acid
SaaS	Software as a Service
SKIP BF	SKIP Brute Force