Chapter 14
Accelerating Large-Scale Genome-Wide Association Studies with Graphics Processors

Mian Lu
*Institute of High Performance Computing, A*STAR, Singapore*

Qiong Luo
Hong Kong University of Science and Technology, Hong Kong

ABSTRACT

Large-scale Genome-Wide Association Studies (GWAS) are a Big Data application due to the great amount of data to process and high computation intensity. Furthermore, numerical issues (e.g., floating point underflow) limit the data scale in some applications. Graphics Processors (GPUs) have been used to accelerate genomic data analytics, such as sequence alignment, single-Nucleotide Polymorphism (SNP) detection, and Minor Allele Frequency (MAF) computation. As MAF computation is the most time-consuming task in GWAS, the authors discuss in detail their techniques of accelerating this task using the GPU. They first present a reduction-based algorithm that better matches the GPU's data-parallelism feature than the original algorithm implemented in the CPU-based tool. Then they implement this algorithm on the GPU efficiently by carefully optimizing local memory utilization and avoiding user-level synchronization. As the MAF computation suffers from floating point underflow, the authors transform the computation to logarithm space. In addition to the MAF computation, they briefly introduce the GPU-accelerated sequence alignment and SNP detection. The experimental results show that the GPU-based GWAS implementations can accelerate state-of-the-art CPU-based tools by up to an order of magnitude.

DOI: 10.4018/978-1-4666-4699-5.ch014

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Nowadays, with the rapid progress of DNA sequencing techniques, large-scale genome-wide association studies (GWAS) have become practical. The research focuses on investigating DNA variation among a group of individuals to identify causes of complex traits (Johnson & O'Donnell, 2009). GWAS covers various applications for different uses. In this chapter, we focus on three fundamental tasks, which are sequence alignment, single-nucleotide polymorphism (SNP) detection, and minor allele frequency (MAF) computation.

To perform GWAS, a large number of short *reads* (sequence fragments generated from a Next-Generation Sequencing device) are first matched against a reference sequence (sequence alignment). Then the SNP information is calculated for every individual using a Bayesian model (SNP detection). Finally, based on the multiple individuals' SNP detection results, MAF is calculated by a probabilistic model (MAF computation). The results of MAF can be used to study the association between genes and traits, such as diseases (Manolio, 2010).

With the high throughput of modern DNA sequencing devices, a great amount of sequence data is generated on a daily basis at sequencing centers. For example, in BGI-Shenzhen, which is the largest sequencing center in the world, TBs of sequence data are generated per day. In this chapter, we identify two major challenges for large-scale GWAS.

- First, due to the large amount of data to process and high intensity of computation, the genome data analysis usually takes an excessively long running time using existing tools. For example, it takes several days for a commodity CPU server to perform sequence alignment and SNP detection on the data of a single individual. If MAF computation is applied to the data set of thousands of individuals, it will take months or even years to complete the analysis. Therefore, developing high-speed, scalable analysis tools is crucial for large-scale GWAS.
- Second, the computation may suffer from numerical issues due to the limit of floating point number representation in machines. In this chapter, we will show a floating point underflow issue in the MAF computation. It occurs when applying multiplications to a large number of small probability values from different individuals. The program will crash when the floating point underflow occurs.

As a result, due to the high computation intensity and high precision requirement, so far the state-of-the-art MAF results reported by genomists are based on the data set of up to hundreds of individuals (Kim et al., 2011; Li et al., 2010; Yi et al., 2010). To enable the study for data sets of up to thousands of individuals, these two challenges must be addressed.

Nowadays, the GPU has become a mainstream many-core hardware architecture to accelerate various scientific applications (Owens et al., 2007). However, applications can fully utilize the powerful GPU hardware resource only when their algorithm features match the GPU's architecture features, e.g., massive data parallelism and the coalesced memory access pattern. In practice, many algorithms do not expose sufficient data parallelism and have random memory accesses, including many algorithms in GWAS. Therefore, in this chapter, we show that careful design and optimization techniques are necessary in order to achieve the efficiency on the GPU. On the other hand, GPUs employ a similar numerical system

30 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/accelerating-large-scale-genome-wide-association-studies-with-graphics-processors/85463

Related Content

On the Use of Object-Role Modeling for Modeling Active Domains
Patrick van Bommel, Stijn Hoppenbrouwers, Erik Properand Theo van der Weide (2007). *Research Issues in Systems Analysis and Design, Databases and Software Development (pp. 123-145).*
www.irma-international.org/chapter/use-object-role-modeling-modeling/28435

Empirical Comparison of 3-D Virtual World and Face-to-Face Classroom for Higher Education
Xiaofeng Chen, Keng Siauand Fiona Fui-Hoon Nah (2012). *Journal of Database Management (pp. 30-49).*
www.irma-international.org/article/empirical-comparison-virtual-world-face/74702

Repositioning Data Management Near Data Acquisition
Paolo Diviacco, Jordi Sorribas, Karien De Cauwer, Jean Marc Sinquin, Raquel Casas, Alessandro Busato, Yvan Stoyanovand Serge Scory (2017). *Oceanographic and Marine Cross-Domain Data Management for Sustainable Development (pp. 178-199).*
www.irma-international.org/chapter/repositioning-data-management-near-data-acquisition/166841

INTECoM: An Integrated Approach to the Specification and Design of Information Requirements
Clare Atkins (2001). *Developing Quality Complex Database Systems: Practices, Techniques and Technologies (pp. 240-260).*
www.irma-international.org/chapter/intecom-integrated-approach-specification-design/8278

Knowledge Extraction From National Standards for Natural Resources: A Method for Multi-Domain Texts
Taiyu Ban, Xiangyu Wang, Xin Wang, Jiarun Zhu, Lvzhou Chenand Yizhan Fan (2023). *Journal of Database Management (pp. 1-23).*
www.irma-international.org/article/knowledge-extraction-from-national-standards-for-natural-resources/318456