Special Issue for Emerging Computational Methods for the Life Sciences Workshop

Judy Qiu
Indiana University
Bloomington, Indiana
xqiu@cs.indiana.edu

Ian Foster
University of Chicago
Chicago, IL
foster@anl.gov

Ronald Taylor
Pacific Northwest National Laboratory
Richland, WA
ronald.taylor@pnl.gov

Abstract

This paper surveys the contents of the special issue on Emerging Computational Methods for the Life Sciences Workshop with six contributed papers. They cover a rich variety of topics on interface of life sciences and computation which in detail are parallelizing two popular micro array data analysis techniques using the Simple Parallel R Interface (SPRINT); parallelization of PEMer structural variation pipeline and the BWA alignment tool for execution on clusters, grids and clouds using the Weaver/Starch/Makeflow workflow stack; hierarchical MapReduce framework for utilizing computational resources from multiple clusters simultaneously to execute a MapReduce computation across them; framework which can utilize HPC, Grid and Cloud infrastructure through a unified framework to achieve task-level concurrency; port of the AutoDock molecular docking program to run within the open source Hadoop MapReduce framework; current research topics on computational methods of genomics, the complexity of biological applications and computational assays, and the increasing demands of improving algorithms and parallel systems;

TABLE OF CONTENTS

1. BACKGROUND

Computing systems are rapidly changing with multicore, GPUs, clusters, volunteer systems, clouds, and grids offering a confusing dazzling array of opportunities. New programming paradigms such as MapReduce and Many Task Computing have joined the traditional repertoire of workflow and parallel computing for the highest performance systems. Meanwhile the Life Sciences are continuing to expand in data generated with continuing improvement in the instruments for high throughput analysis. This “fourth paradigm” (observationally driven science) is joined by complex systems or biocomplexity

that can build phenomenological models of biological systems and processes. This special issue juxtaposes these trends seeking those computational methods that will enhance scientific discovery. Within this overall scope, this special issue encouraged researchers to submit and present original work related to the latest trends in parallel and distributed high performance systems applied to Life Science problems.

Relevant contributions have been provided by Mitchel et al. [Mitchell 2012], by Lanc et al. [Lanc 2012], by Luo et al. [Luo 2012], by Kim et al. [Kim 2012], by Ellingson et al. [Ellingson 2012], and by Schatz [Schatz 2012]. These contributions focus on:

- parallelizing two popular micro array data analysis techniques using the Simple Parallel R Interface (SPRINT);
- parallelization of PEMer structural variation pipeline and the BWA alignment tool for execution on clusters, grids and clouds using the Weaver/Starch/Makeflow workflow stack;
- a hierarchical MapReduce framework for utilizing computational resources from multiple clusters simultaneously to execute a MapReduce computation across them;
- presenting a framework which can utilize HPC, Grid and Cloud infrastructure through a unified framework to achieve task-level concurrency;
- porting the AutoDock molecular docking program to run within the open source Hadoop MapReduce framework;
- introducing the current research topics on computational methods of genomics, the complexity of biological applications and computational assays, and the increasing demands of improving algorithms and parallel systems;

2. SPECIAL ISSUE PAPERS

Mitchel et al. [Mitchell 2012] resents parallel implementations of two popular micro array data analysis techniques: exploratory clustering analyses using the random forest classifier; and feature selection through identification of differentially expressed genes using the rank product method. The authors have parallelized these two applications using the Simple Parallel R Interface (SPRINT), which is a library for R that aims to reduce the complexity of using HPC systems by providing biostatisticians with drop-in parallelised replacements of existing R functions. The paper demonstrates how one can parallelize R routines with minimum changes to the existing codes with the help of SPRINT, speeding up serialized and time-consuming analysis procedures written in R. Authors also implemented a tree-reduction algorithm for parallel combining of the results, which showed surprisingly large effect on the overall performance. The paper also provides experimental results achieving 40 times speed-up over serialized codes by using 128 processes.

Lanc et al. [Lanc 2012] describes the adaptation and parallelization process of PEMer structural variation pipeline and the BWA alignment tool for execution on clusters, grids and clouds using the Weaver/Starch/Makeflow workflow stack. Authors describe the application of previous obtained lessons to a new workflow with and without shared file storage to tract the intractable sequential running times of these applications on large datasets. Authors present lessons and results for refactoring bioinformatics tools for elastic scaling on personal clouds and describe the various challenges faced when constructing such a workflow, from dealing failure detection, to managing dependencies, to handling the quirks of the underlying operating systems. Authors scale the workflows on hundreds of processors reducing the run times of the two workflows to hours from days with high speedup. The lessons and the experiences presented in this paper can lower the barrier to

scalable execution of workflows, allowing users to better harness the power of heterogeneous distributed systems for their own tools.

Luo et al. [Luo 2012] describes an enhanced MapReduce based programming model ”Map-Reduce-GlobalReduce”, where the computations are expressed as three functions: Map, Reduce, and GlobalReduce. The authors name this model as ‘Hierarchical MapReduce’. The hierarchical MapReduce framework divides the MapReduce computations and utilizes computation resources from multiple clusters simultaneously to execute MapReduce job across them. The design is a powerful extension to MapReduce, especially to provide additional processing power for very large computations. Two static prior-knowledge based scheduling algorithms are proposed, one that targets compute-intensive jobs and another data-intensive jobs, evaluated using a life science application, AutoDock, and a simple Grep. The authors demonstrate the utility of their design and the performance metrics by greatly accelerating the application AutoDock across three large clusters.

Kim et al. [Kim 2012] present a runtime-environment, Distributed Application Runtime Environment (DARE) that supports the scalable, flexible and extensible composition of capabilities exploring the interoperability among heterogeneous distributed computing environments for pleasingly parallel applications. DARE is a SAGA-BigJob based framework motivated by the next-generation sequencing (NGS) analysis and other similar data intensive applications. The proposed framework would enable NGS like applications to run automatically on different infrastructures. DARE can utilize HPC, Grid and Cloud infrastructure through a unified framework to achieve task-level concurrency. In this work, authors use BFAST as a representative standalone tool used for NGS data analysis and a ChIP-Seq pipeline as a representative pipeline based approach. This paper represents the initial steps in the design and development of a general-purpose, scalable and extensible infrastructure to support next-generation (gene) sequencing data analytics.

Ellingson et al. [Ellingson 2012] describe their experience porting the AutoDock molecular docking program to run within the open source Hadoop MapReduce framework. Virtual molecular docking is a task parallel computational method used in computer-aided drug discovery that calculates the binding affinity of a small molecule drug candidate to a target protein. Authors evaluate the performance of the Autodock Hadoop implementation on the 1088 core Kandinsky cluster located at Oak Ridge National Laboratory. In this environment, the authors were able to achieve an impressive 450-fold speedup over a serial execution, reducing >1 year of work to ~1 day.

Finally, **Schatz** [Schatz 2012] introduce the current research topics on computational methods of genomics, the complexity of biological applications and computational assays, and the increasing demands of improving algorithms and parallel systems. The challenges brought by the ever-increasing amount of data produced by advanced instruments are elaborated systematically in detail. Authors discuss how parallel computing and cloud computing have been used to run large-scale biological applications and lists the challenges of Cloud computing for digital genomics. Issues such as big data, data security and privacy, cost of cloud utility are discussed. The author also discusses the advantage of using hardware accelerators to empower the genomics analysis and speculate several future trends of digital demands of genomics that can potentially help researchers to reshape their thinking.

3. ACKNOWLEDGEMENTS

We would like to thank the authors for contributing papers on their research on latest trends in data intensive technologies and applications for this special issue, and thank all the reviewers for providing constructive reviews and in helping to shape this special issue. Finally we would like to thank the editors of Concurrency and Computation: Practice and Experience for providing us an opportunity to bring this special issue to the research community.

4. REFERENCES (typesetters: please add correct Concurrency and Computation: Practice and Experience citations)

[Mitchell 2012] Lawrence Mitchell, Terence M. Sloan, Muriel Mewissen, Peter Ghazal, Thorsten Forster, Michal Piotrowski and Arthur Trew. A parallel random forest classifier for R, Special Issue on Emerging Computational Methods for the Life Sciences.

[Lanc 2012] Irena Lanc, Peter Bui, Douglas Thain and Scott Emrich. Adapting Bioinformatics Applications for Heterogeneous Systems: a case study, Special Issue on Emerging Computational Methods for the Life Sciences.

[Luo 2012] Yuan Luo, Zhenhua Guo, Yiming Sun, Beth Plale, Judy Qiu and Wilfred Li. A Hierarchical Framework for Cross-Domain MapReduce Execution, Special Issue on Emerging Computational Methods for the Life Sciences.

[Kim 2012] Joohyun Kim, Sharath Maddineni and Shantenu Jha. Characterizing Deep Sequencing Analytics Using BFAST: Towards a Scalable Distributed Architecture for Next-Generation Sequencing Data, Special Issue on Emerging Computational Methods for the Life Sciences.

[Ellingson 2012] Sally Ellingson and Jerome Baudry. High-Throughput Virtual Molecular Docking: Hadoop Implementation of AutoDock4 on a Private Cloud, Special Issue on Emerging Computational Methods for the Life Sciences.

[Schatz 2012] Michael Schatz, Answering the demands of digital genomics, Special Issue on Emerging Computational Methods for the Life Sciences.