Future Generation Computer Systems 67 (2017) 325–328

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Preface

Boosting analyses in the life sciences via clusters, grids and clouds

Sandra Gesing [a,*], Jesus Carretero [b], Javier Garcia Blas [b], Johan Montagnat [c]

[a] University of Notre Dame, USA
[b] Universidad Carlos III de Madrid, Spain
[c] CNRS, France

ARTICLE INFO

Article history:

Keywords:
Bioinformatics
Biomedicine
And health
Genomics
Life Sciences
Cloud computing
Workflows

ABSTRACT

In the last 20 years, computational methods have become an important part of developing emerging technologies for the field of bioinformatics and biomedicine. Those methods rely heavily on large scale computational resources as they need to manage Tbytes or Pbytes of data with large-scale structural and functional relationships, TFlops or PFlops of computing power for simulating highly complex models, or many-task processes and workflows for processing and analyzing data. This special issue contains papers showing existing solutions and latest developments in Life Sciences and Computing Sciences to collaboratively explore new ideas and approaches to successfully apply distributed IT-systems in translational research, clinical intervention, and decision-making.

1. Presentation

In the last 20 years, computational methods have become an important part of developing emerging technologies for the field of bioinformatics and biomedicine [1–4]. Research areas such as biomodelling, molecular dynamics, genomics, neuroscience, cancer models, evolutionary biology, medical biology, biochemistry, biophysics, biotechnology, cell biology, nanobiotechnology, biological engineering, pharmacology, genetics therapy, or automatic diagnosis, rely heavily on complex computational resources as they need to manage Tbytes or Pbytes of data with large-scale structural and functional relationships, TFlops or PFlops of computing power for simulating highly complex models, or many-task processes and workflows for processing and analyzing data.

This new situation demands appropriate IT infrastructures, where biomedical data can be processed within an acceptable timespan reaching from minutes in health-care applications to days in large-scale research projects. Large-scale distributed IT systems such as Grids [5], Clouds [6] and Big-Data-Environments [7] are promising to address research, clinical and medical research community requirements [8]. They allow for significant reduction of computational time for running large experiments, for speeding-up the development time for new algorithms, for increasing the availability of new methods for the research community, and for supporting large-scale multidisciplinary collaborations. However, specific challenges in the employment of such systems for biomedical applications such as security, reliability and user-friendliness, often impede straightforward adoption of existing solutions from other application domains.

This special issue contains submissions from developers of bioinformatic and medical applications and researchers in the field of distributed IT systems. It targets diverse groups of audiences: Firstly, researchers who are already employing distributed infrastructure techniques in bioinformatic applications, in particular scientists developing data and compute intensive bioinformatic and medical applications that include multi-data studies, large-scale parameter scans or complex analysis pipelines; Secondly, it addresses computer scientists working in the field of distributed systems interested in bringing new developments into bioinformatic and medical applications. The special issue further intends to identify common requirements to lead future developments in collaboration between Life Sciences and Computing Sciences, and to collaboratively explore new ideas and approaches to successfully apply distributed IT systems in Life Sciences.

2. Special issue content

This special issue of Future Generation Computer Systems Journal contains papers selected from a set of invited papers extracted from the papers presented in International Workshop on Clusters, Clouds and Grids for Life Sciences (CCGrid-Life 2015) held

* Corresponding editor.
E-mail addresses: sandra.gesing@nd.edu (S. Gesing), jesus.carretero@uc3m.es (J. Carretero), fjblas@inf.uc3m.es (J.G. Blas), johan.montagnat@cnrs.fr (J. Montagnat).

http://dx.doi.org/10.1016/j.future.2016.11.001
0167-739X/

together with CCGrid 2015, held in Shenzhen, Guangdong, China, May 4–17, 2015, but it also covers papers coming from an open call. The objective of CCGrid-Life 2015 was to exchange and discuss existing solutions and latest developments in both fields, and to gather an overview of challenges (technologies, achievements, gaps, roadblocks).

The special issue received 19 papers, 13 of which were selected for publication after going through the Future Generation Computer Systems Journal peer review process. The submissions reflect three major trends in the Life Sciences and biomedical applications: Firstly, the use of novel concepts and cutting-edge software in Cloud Computing to increase the efficiency of applications under consideration of important aspects such as cost and energy efficiency. Established solutions such as workflows are ported to novel infrastructures and address the representation of scientific methods and enhancing the efficiency and reproducibility of computational tasks. They are further improved to meet the needs of the user communities in regard to usability and efficiency of applications. The efficiency of computation is also addressed via the second trend in the manuscripts, which elucidate algorithms exploiting novel accelerated hardware architectures such as GPUs. The third trend is concerned with the management of data, which is a challenge especially in the Life Sciences because of the vast amount of data and its diversity. The following briefly introduces to the manuscripts of the special issue and their focus areas.

In "Building an open source cloud environment with auto-scaling resources for executing bioinformatics and biomedical workflows" [9] the authors discuss how a full cloud stack ranging from Infrastructure-as-a-Service (IaaS) via Platform-as-a-Service (PaaS) to Software-asa-Service (SaaS) can be built on open source technologies. On the PaaS level, they present a scaling strategy for the Galaxy workflow platform with bioinformatics and biomedical use cases. Due to its open nature, it is able to run on any IaaS cloud platform, ranging from public commercial providers to private research/academic clouds, also allowing the easy (re-)construction of this platform for on-premise computing, which can be a requirement for processing sensitive data, which must not leave the security domain of an organization. Two distinct use-cases are presented to demonstrate the feasibility and performance of the solution.

A Grid-based science workflow infrastructure is presented by Cohen-Boulakia et al. in "InfraPhenoGrid: A scientific workflow infrastructure for Plant Phenomics on the Grid" [10]. The infrastructure was designed and deployed to efficiently manage data sets produced by the PhenoArch plant phenomics platform in the context of the French Phenome Project. The solution consists of deploying scientific workflows on a Grid using a middleware to pilot workflow executions and to provide a user-friendly environment in the sense that, despite the intrinsic complexity of the infrastructure, running scientific workflows and understanding results obtained (using provenance information) is kept as simple as possible for end-users.

Santana-Prez et al. propose in the paper "Reproducibility of Execution Environments in Computational Science Using Semantics and Clouds" [11], a novel approach based on semantic vocabularies that describes the execution environment of scientific workflows, so as to conserve it, together with a process for documenting the workflow application and its related management system, as well as their dependencies. Then, they apply this approach over three different real workflow applications running in three distinct scenarios, using public, private, and local Cloud platforms (one astronomy workflow and two life science workflows for genomic information analysis). Experimental results demonstrate that the approach presented can reproduce an equivalent execution environment of a predefined virtual machine image on all evaluated computing platforms.

The research work "A Cost-Effective Approach to Improving Performance of Big Genomic Data Analyses in Clouds" [12], presented by Xing et al. describes how the Genome Analysis Toolkit (GATK) can be deployed to an elastic cloud and defines policy to drive elastic scaling of the application. The authors extensively analyze the GATK to expose opportunities for resource elasticity, demonstrate that it can be practically deployed at large scale in a cloud environment, and demonstrate that applying elastic scaling improves the performance to cost trade-off achieved in a simulated environment.

Guzzeti et al. propose in "Platform and Algorithm Effects on Computational Fluid Dynamics Applications in Life Sciences" [13] methodologies and protocols to identify the optimal choice of computing platforms for hemodynamics computations, that will be increasingly needed in the future, and the optimal scheduling of the tasks across the selected resources. The authors focus on hemodynamics in patient-specific settings and present extensive results on different platforms, also proposing a way to measure and estimate performance and running time under realistic scenarios tailored to the utility function of the simulation. They discuss in detail the optimal (parallel) partitioning of the domain of a problem of interest with different mathematical approaches, showing that an overlapping splitting is generally advantageous and the detection of optimal overlapping has the potential to significantly reduce computational costs of the entire solution process and the communication volume across the platforms.

In "Colorectal Tumour Simulation using Agent Based Modelling and High Performance Computing" [14] the authors propose to apply advanced HPC techniques to achieve the efficient and realistic simulation of a virtual tissue model that mimics tumour growth or regression in space and time. These techniques combine extensions of the previously developed agent-based simulation software platform (FLAME) with auto-tuning capabilities and optimization strategies for the current tumour model. Development of such a platform could advance the development of novel therapeutic approaches for the treatment of CRC could also be applied to other solid tumours.

Hamie et al. show in "Scaling Machine Learning for Target Prediction in Drug Discovery using Apache Spark" [15] the implementation of the traditional pipeline for identification of candidate molecules that affect proteins associated with diseases in drug discovery by using Apache Spark, which enabled them to lift the existing programs to a multinode cluster without making changes to the predictors. Evaluations show almost linear speedup, due in part to the reduction in the number of intermediate files. It also allows easier checkpointing and monitoring.

The problem of "Finding Exact Hitting Set Solutions for Systems Biology Applications using Heterogeneous GPU Clusters" is faced out by Carastan-Santos et al. [16]. In the paper, the authors propose a novel algorithm for solving HSP instances with thousands of variables by using: (i) clause sorting, which enables the efficient discarding of non-solution candidates, (ii) parallel generation and evaluation of candidate solutions through the use of GPUs, and (iii) support for multiple GPUs. To permit the execution on heterogeneous clusters, the authors determine the minimum kernel size that does not incur extra overhead and distribute tasks among available GPUs on demand. The experimental results show that the combination of these techniques results in a speedup of 118.5, when using eight NVIDIA Tesla K20c in comparison with a ten-core Intel Xeon E5-2690 processor. Consequently, the presented algorithm can enable the usage of exact algorithms for solving the Hitting Set problem and applying it to real world problems.

The paper "Multi-GPU-Based Detection of Protein Cavities using Critical Points" [17], by Duarte et al., introduces a geometric method for detecting cavities on the molecular surface based on