Gesing, S., Carretero, J., García Blás, J., Montagnat, J. (2017). Boosting analyses in the life sciences via clusters, grids and clouds. *Future Generation Computer Systems*, 67, pp. 325-328.

DOI: 10.1016/j.future.2016.11.001

Boosting Analyses in the Life Sciences via Clusters, Grids and Clouds

Sandra Gesing[1,*], Johan Montagnat[b], Jesus Carretero[c], Javier Garcia Blas[c]

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

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 transnational research, clinical intervention, and decision-making.

Keywords: bioinformatics, biomedicine, and health, genomics, Life Sciences, cloud computing, workflows

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. 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 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 new situation demands appropriate IT-infrastructures, where bioinformatic and medical data can be processed within an acceptable timespan - reaching from minutes in

*Corresponding author
Email addresses: sandra.gesing@nd.edu (Sandra Gesing), johan.montagnat@cnrs.fr (Johan Montagnat), jesus.carretero@uc3m.es (Jesus Carretero), fjblas@inf.uc3m.es (Javier Garcia Blas)

health-care applications to days in large-scale research projects. Large-scale distributed IT-systems such as Grids, Clouds and Big-Data-Environments are promising to address research, clinical and medical research community requirements. 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 multi-centric collaborations. However, specific challenges in the employment of such systems for bioinformatic applications such as security, reliability and user-friendliness, often impede straightforward adoption of existing solutions from other application domains.

This special issue aims at bringing together developers of bioinformatic and medical applications and researchers in the field of distributed IT systems. On the one hand, it addresses 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. On the other hand, 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 Contents

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 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, being 13 of them were selected for publication after going through the Future Generation Computer Systems Journal peer review process. A brief presentation of contents of this special issue is shown below.

In "Building an open source cloud environment with auto-scaling resources for executing bioinformatics and biomedical workflows" Thomas et al. discuss how a full cloud stack ranging from Infrastructure-as-a-Service (IaaS) via Platform-as-a-Service (PaaS) to Software-as-a-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 organisation.

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. and they present "InfraPhenoGrid: A scientific workflow infrastructure for Plant Phenomics on the Grid". 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 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", 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" 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 analyse the GATK to expose opportunities for resource elasticity, demonstrate that it can be practically deployed at 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" 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, proposing also 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" 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 autotuning capabilities and optimisation strategies for the current tumour model. Development of such a platform could advance the development of novel therapeutic approaches for the treatment of CRC which can also be applied to other solid tumours.

Hamie et al. show in "Scaling Machine Learning for Target Prediction in Drug Discovery using Apache Spark" 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 and a reduction in the number of intermediate files while allowing 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. 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", by Duarte-Gomes et al., introduces a geometric method for detecting cavities on the molecular surface based on the theory of critical points. The method, called CriticalFinder, differs from other surface-based methods found in the literature because it directly takes advantage of the curvature of the scalar field (or function), which represents the molecular surface, instead of evaluating the curvature of the Connolly function over the molecular surface. To evaluate the accuracy of CriticalFinder, the authors compare it with other seven geometric methods, carrying out a performance analysis of the GPU implementation of CriticalFinder in terms of time consumption and memory space occupancy.

In "Parallel SuperFineA Tool for Fast and Accurate Supertree Estimation: Features and Limitations", the authors present Parallel SuperFinea, a tool that aims the fast and accurate supertree estimation and its features. Parallel SuperFine was derived from SuperFinea state-of-the-art supertree (meta)method. The authors describe an extension made to SuperFine, which permits a significant improvement of its performance, and how the EPIC framework is used to boost the overall performance of Parallel SuperFine. Additionally, the authors cope with the current limitations that impair to attain (even) a better performance. These studies reveal that Parallel SuperFine allows to significantly reduce the time required to perform supertree estimation. Moreover, the experiments show that Parallel SuperFine exhibits good scalability, even in the presence of asymmetric biological data sets. Furthermore, the achieved results enable to conclude that the radical improvement in performance does not

impair tree accuracy, which is a key issue in phylogenetic inference.

Energy efficiency of software tools for BioComputing is not a usual topic of research. In the paper "Energy Efficiency of Sequence Alignment Tools - Software and Hardware Perspectives", Kierzynka et al. compare the energy efficiency of the most established software tools performing exact pairwise sequence alignment on various computational architectures: CPU, GPU, and Intel Xeon Phi. The results show that the energy consumption may differ as much as nearly 5 times. Substantial differences are reported even for different implementations running on the same hardware. Moreover, they present an FPGA implementation of one of the tested tools – G-DNA, and show how it outperforms all the others on the energy efficiency front. Finally, they present the special RECS—Box servers, a hardware designed and manufactured with the special purpose to deliver highly heterogeneous computational environment supporting energy efficiency and green ICT.

Beier et al. introduce in the paper "Multicenter Data Sharing for Collaboration in Sleep Medicine" a virtual research platform that supports inter-institutional data sharing and processing. The infrastructure is based on XNAT, a free and open source neuroimaging research platform, a loosely coupled service oriented architecture, and scalable virtualization in the back-end. The system is capable of local pseudonymization of biosignal data, mapping to a standardized set of parameters and automatic quality assessment derived from the "Manual for the Scoring of Sleep and Associated Events" of the American Academy of Sleep Medicine (AASM).

Up-to-date meta-databases are vital for the analysis of biological data. Pedersen et al. present in the paper "Large-scale Biological Meta-Database Management" a solution for biological meta-database management. It provides efficient storage and runtime generation of specific meta-database versions, and efficient incremental updates for biological data analysis tools. The approach is transparent to the tools, and it also provides a framework that makes it easy to integrate GeStore with biological data analysis frameworks. An evaluation of the performance characteristics of the system plus an evaluation of the benefits for a biological data analysis workflow are presented.

3. Acknowledgements

We would like to thank all the authors, reviewers and editors involved in the elaboration of this special issue, including also the reviewers that were involved in the CCGrid 2015 conference, where short versions of the papers were previously selected. We are especially grateful to Peter Sloot, editor in chief of the The International Journal of eScience (Future Generation Computer Systems), for approving this special issue and for his help along the process of its preparation.

Authors' Biographies

Sandra Gesing

Sandra Gesing is a research assistant professor at the Department of Computer Science and Engineering and a computational scientist at the Center for Research Computing at

the University of Notre Dame. Sandra Gesing research interests include science gateways, bioinformatic applications, grid and cloud computing, parallel programming, and GPU programming. In this context, She also works on disease modeling and analysis frameworks for modeling and simulations (e.g., Projects VectorBase and VecNet). She collaborate closely with the groups of Douglas Thain, Scott Emrich, Gregory Madey and Frank Collins. Prior to the position at Notre Dame, She was a research associate in the Data-Intensive Research Group at the University of Edinburgh, UK, in the area of data-intensive workflows and in the Applied Bioinformatics Group at the University of Tbingen, Germany, in the area of science gateways and grid computing. Additionally, she has perennial experience as a project manager and system developer in industry. As head of a system programmer group, she have led long-term software projects (e.g. infrastructure on web-based applications) for a major insurance company. She received my German diploma in computer science from extramural studies at the FernUniversitt Hagen and her PhD in computer science from the University of Tbingen, Germany.

Johan Montagnat

Johan Montagnat is a Computer Science researcher of the French National Center for Scientific Research in the area of large scale distributed systems, I3S laboratory (CNRS UMR 7271), and leader of the MODALIS team. He achieved his HRD thesis at Univeristy of Nice-Sophia Antipolis, (France), on Computer science processing and analyzing large medical image sets. His research interest includes large-scale distributed systems, distributed information systems, 3D and 4D image segmentation, anatomical structures modeling, and medical image processing. As a result of his research activity, Dr. Montagnat has led several national and international projects related to biomedicine and he has co-authored many international research journals papers and has participated in more that 60 international conferences. Hi is also co-author of the MOTEUR workflow engine, a tool to make flexible and efficient workflow deployment of data-intensive applications on grids.

Jesus Carretero

Jesus Carretero is a Full Professor of Computer Architecture and Technology at Universidad Carlos III de Madrid (Spain), since 2002. His research activity is centered on high-performance computing systems, large-scale distributed systems and real-time systems. He is Action Chair of the IC1305 COST Action "Network for Sustainable Ultrascale Computing Systems (NESUS)", and he is also currently involved in three other EU projects. Prof. Carretero is Associated Editor of the journal Computer and Electrical Engineering and International Journal of Distributed Sensor Networks and has been guest editor for special issues of journals as International Journal of Parallel Processing, Cluster Computing, Computers and Electrical Engineering, and New Generation Computing. He has been General chair of HPCC 2011 and MUE 2012, and Program Chair of ISPA 2012, EuroMPI 2013, C4Bio 2014, and ESAA 2014. Prof. Carretero is a senior member of the IEEE Computer Society and member of the ACM.

Javier Garcia Blas

Javier Garcia Blas has been a visiting assistant of the University Carlos III of Madrid since 2005. He has cooperated in several projects from various high performance research institutions including HLRS, DKRZ, and Argonne National Laboratory. He is currently involved in three projects funded by European projects. Additionally, he is currently involved in various projects on topics including parallel I/O, cloud computing, and accelerators for high-performance platforms. He has participated in many conference organization committees, and in the last three years he has been Program Chair of EuroMPI 2013, C4Bio 2014, ESAA 2014, CCGrid-Life 2015, and IASDS 2015. He counts with 35 research publications in international journals and conferences.