HPC-as-a-Service for Life Sciences

Extended Abstract

Václav Svatoň
and Jan Martinovič
IT4Innovations
Ostrava, Czech Republic
name.surname@vsb.cz

Nina Jeliazkova
IDEAconsult Ltd.
Bulgaria
jeliazkova.nina@gmail.com

Vladimir Chupakhin
Janssen Pharmaceutika NV
Belgium
vchupakh@its.jnj.com

Pavel Tomančák
Max Planck Institute of Molecular
Cell Biology and Genetics
Germany
tomancak@mpi-cbg.de

Petr Vojta
IMTM, Palacky University Olomouc
Olomouc, Czech Republic
petr.vojta@gmail.com

1 INTRODUCTION

HPC as a Service is a well-known term in the area of high performance computing. It enables users to access an HPC infrastructure without a need to buy and manage their own infrastructure. Through this service academia and industry can take advantage of the technology without an upfront investment in the hardware. This approach further lowers the entry barrier for users who are interested in utilizing massive parallel computers but often do not have the necessary level of expertise in the area of parallel computing.

To provide this simple and intuitive access to the supercomputing infrastructure an in-house application framework called High-End Application Execution Middleware[3] (HEAppE) has been developed. HEAppE's universally designed software architecture enables unified access to different HPC systems through a simple object-oriented API. Thus providing HPC capabilities to the users but without the necessity to manage the running jobs form the command-line interface of the HPC scheduler directly on the cluster.

DHI Group, nation-wide company developing hydrologic software MIKE powered by DHI, was a collaborator during the design and implementation phase of the HEAppE[5] (formerly known as a HPC as a Service Middleware). At the start of the project the HEAppE was used in the area of hydrological modelling in a decision support system for crisis management. Since then the HEAppE was successfuly used in a number of projects from a different thematic domains.

2 HEAPPE ON SUPERCOMPUTING INFRASTRUCTURE

The IT4Innovations national supercomputing center operates supercomputers Salomon (2 PFLOP/s) and Anselm (94 TFLOP/s). The supercomputers are available to academic community within the Czech Republic and Europe and industrial community worldwide. Both supercomputers are available to users via HEAppE Middleware.

For security purposes HEAppE enables the users to run only pre-prepared set of so-called Command Templates. Each template defines arbitrary script or executable file that will be executed on the cluster, any dependencies or third-party software it might require and the type of queue that should be used for the processing. The template also contains the set of input parameters that will be passed to the executable script during run-time. The actual value of each parameter can be changed by the user for each job submission.

Main features:

- Providing HPC capabilities as a service to client applications and their users Unified interface for different operating systems and schedulers
- Authentication and authorization to provided functions
- Monitoring and reporting of executed jobs and their progress
- Current information about the state of the clusters
- Job accounting and job reporting for user or user group Secure data migration between different jobs
- Prepared job templates for domain specific tools Dedicated GUI for each domain specific use case

3 HEAPPE FOR LIFE SCIENCES

HEAppE has been successfully used in a number of research and commercial projects where there is a need for a remote access to an HPC infrastructure. This section describes the three use cases of HEAppE utilization in Life Sciences applications.

3.1 Machine Learning for Drug Discovery

Real-world pharma industry applications often encompass end-to-end data processing pipelines composed of a large number of interconnected tasks of various granularity. Most of the common tasks in the prediction of activity and toxicity of chemical compounds consist of several typical steps, such as compiling, cleaning and combining datasets, feature calculation, feature selection, model training and validation and applying models to predict properties of new compounds. Building and executing such pipelines on HPC systems can be challenging tasks for domain specialists who do not have sufficient level of experience in distributed computing. Therefore, a drug discovery web platform was developed that enables large-scale machine learning applications being executed on supercomputing facilities via specialized HEAppE middleware.

Pharma companies collected significant amount of protein-ligand interactions forming so-called chemogenomics matrix: interactions between compounds and proteins, but this matrix is very sparse, less than 1% of this matrix is filled. Predictive modelling can help to fill this matrix using classification or regression model, predictions in turn are used to speed-up drug design and development process, can help to cut cost and also reduce animal use. While machine learning is widely used on every step of the drug design and discovery process it is still a hurdle to use it on big data, taking into account all modelling steps needed: hyper parameter search, model and predictions.

HyperLoom[4] is an open-source platform developed by IT4-Innovations for an efficient definition and execution of scientific pipelines in distributed environments. HyperLoom enables to chain large number of computational tasks into a complex end-to-end data processing pipelines using a simple Python interface. The platform also utilizes the ExCAPE-DB[2].

Result: Custom web interface enabling the execution of a specialized drug discovery pipelines for model creation, prediction and statistics on HPC infrastructure via HEAppE Middleware.

3.2 Bioimage Informatics on HPC

Biomedical research is currently undergoing revolutionary transition caused by dramatic progress in microscopic imaging technologies. Using the state-of-the-art microscopes, it is possible to thoroughly examine the interior of cells and living systems and to study biological processes with unprecedented resolution in space and time. This leads to important discoveries in basic biological research and sub-sequently to improving detection and intervention of serious human diseases problems associated with nature.

The data are collected either systematically for a large amount of samples under certain biological conditions or individual biological systems are observed in 3D, by means of time-lapse microscopy for a long time. These two approaches can also be combined which leads to generation of big data. In terms of volume, these datasets are comparable with the data generated in the field of particle physics under international projects such as CERN.

State-of-the-art imaging devices, such as light sheet microscopes, produce datasets so large that they can only be effectively analyzed by employing methods of image processing on high-performance computing clusters. To address this issue, an HPC plugin for Fiji[1], one of the most popular open-source software tools for image processing, has been developed. The plugin enables end users to make use of HPC clusters to analyze large scale image data remotely and via the standard Fiji user interface.

Result: Fiji plugin utilizing HEAppE Middleware for remote execution of SPIM image processing pipeline on selected HPC infrastructure. The created framework will form a foundation for parallel deployment of any Fiji/ImageJ2 command on a remote HPC resource, greatly facilitating big data analysis.

3.3 Massive Parallel Sequencing

The methods of massive parallel sequencing (MPS) have started to play a key role in clinically oriented research and DNA diagnostics of molecular pathologies. Thus, the concept of personalized medicine replaces low-throughput classical approaches, which are often methodically time-consuming to cover long DNA regions. Whole exome sequencing (WES), amplicon sequencing of pooled PCR products of long genes, or genes panel sequencing, which are associated with specific diseases, are methods currently applied for demonstration of genetic risk factors. MPS methods, especially WES generate huge amount of data, which must be further processed.

Therefore the MPS processing platform for the next generation DNA sequencing (NGS) data processing in detection of hereditary and somatic DNA variants was created. Integrated pipeline is usable for genome/exome or gene panel DNA sequencing projects. The main goal of the platform is to provide easy and intuitive access to HPC computing resources to scientific researchers in the area of clinical MPS data via a specialized web-based interface.

Pipeline integrates custom developed annotation tool for DNA variants. The annotation program is designed for human genetic variants annotation, but the functionality was successfully tested on other types of genomes with the different ploidy. One of the advantages over existing annotation programs is the effective phenotypic prioritization of variants on the basis of ontological relationships allowing the effective annotation of genetic variants in the broad range of human diseases.

Result: Specialized web platform for the next generation DNA sequencing with custom annotation tool and a number of open-source bioinformatics software. Platform is deployed at I4Innovations and also at the Institute of Molecular and Translational Medicine. Both instances are utilizing HEAppE Middleware to access the local HPC infrastructure.

ACKNOWLEDGMENTS

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPS II) project "IT4Innovations excellence in science - LQ1602", by the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center - LM2015070", by the ExCAPE project - the European Union's Horizon 2020 research and innovation programme under grant agreement No. 671555, by the European Regional Development Fund in the IT4Innovations national supercomputing center - path to exascale project, project number CZ.02.1.01/0.0/0.0/16_013/0001791 within the Operational Programme Research, Development and Education and by TAČR grant TE02000058.

REFERENCES

[1] Fiji. 2018. Fiji. Retrieved August 7, 2018 from http://fiji.sc
[2] ExCAPE: Exascale Compound Activity Prediction H2020. 2018. ExCAPE-DB. Retrieved August 7, 2018 from https://zenodo.org/record/173258#.WO3q6FOGP0c
[3] IT4Innovations. 2018. HEAppE Middleware. Retrieved August 7, 2018 from http://heappe.eu
[4] IT4Innovations. 2018. HyperLoom. Retrieved August 7, 2018 from http://hyperloom.eu
[5] J. Martinovic, S. Kuchar, V. Svaton, V. Vondrak, L. Vojacek, M. Golasowski, A. Ronovsky, D. Bezdek, T. Bech, J. Hartnack, J. Carlson, and O. R. Sorensen. 2017. Hydrological Model Remote Execution and HPC as a Service. *Supercomputing in Science and Engineering* 1 (2017), 244.