HPC-as-a-Service for Life Sciences

Václav Svatoň, Jan Martinovič
IT4Innovations, VŠB - Technical University of Ostrava
Czech Republic

Pavel Tomančák
Max Planck Institute of Molecular Cell Biology and Genetics
Germany

Petr Vojta
IMTM, Palacky University Olomouc
Czech Republic

Nina Jeliazkova
IDEAconsult Ltd.
Bulgaria

Vladimir Chupakhin
Janssen Pharmaceutika NV
Belgium

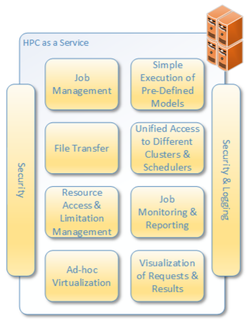
HEAppE: High-End Application Execution Middleware

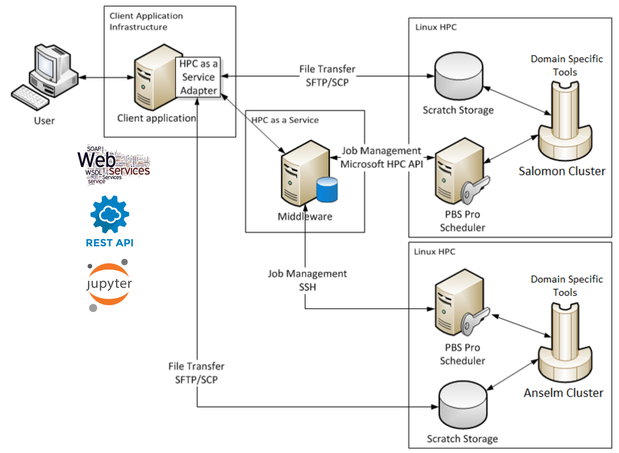
HPC as a Service is a well known term in the area of high performance computing. It enables users to access an HPC infrastructure without a need to buy and manage their own physical servers or data center infrastructure. Through this service **academia and industry** can take **advantage of the technology** 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 **HEAppE** has been developed. HEAppE's universally designed software architecture enables **unified access** to different HPC systems through a simple object-oriented **client-server interface** using standard **web services**, **REST API** or **Jupyter notebooks**. 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.

The **IT4Innovations national supercomputing center** operates two 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**.

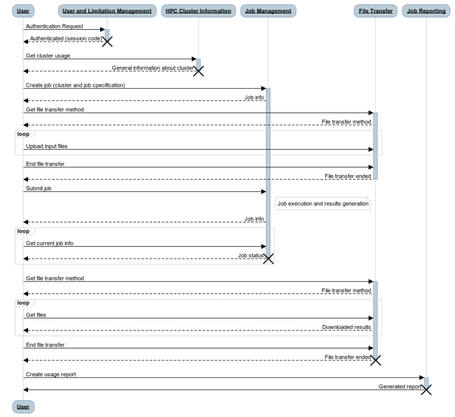
This software will soon be available as a open-source at ***http://www.heappe.eu***





HEAppE Middleware

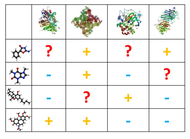
- Providing **HPC capabilities as a service** to client applications and their users
- **Unified interface** for different operating systems and schedulers
- **Authentication** and **authorization** to provided functions
- **Monitoring** and **reporting** of executed jobs and their progress
- Current information about the **state of the clusters**
- Job **accounting** and job **reporting** for user or user group
- Secure **data migration** between different jobs
- **Batch job** processing or **interactive mode**
- **Sandbox** processing via **Docker/Singularity** images
- Pre-prepared **job templates** for domain specific tools
- Number of different **APIs**
- **Dedicated GUI** for each domain specific use case



UC1 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.

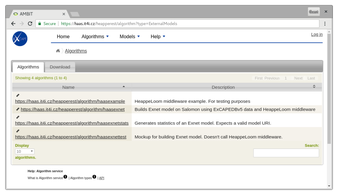
Pharma companies collected significant amount of **protein-ligand interactions** forming so-called **chemogenomics matrix**: interactions between compounds and proteins, but this matrix is very sparse, less than 1% of this matrix is filled. **Predictive modelling** can help to fill this matrix using **classification or regression model**, predictions in turn are used to **speed-up drug design and development process**.



Large scale deep learning modelling: **ExCAPEDBv5** : 955,386 compounds, 526 protein targets, chem2vec descriptors

- Chemical structure standardisation by AMBIT (http://ambit.sf.net)
- Nested cross-validation
- Fully connected deep net with two hidden layers (a flavour of binet by JKU-Linz)
- Hyper parameter search for best network architecture and learning rate

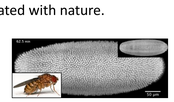
HyperLoom framework for distributed task execution was used for an **efficient definition** and **execution** of **drug discovery pipelines** in distributed environments.

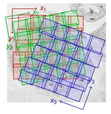


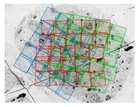
Drug discovery web platform enabling the execution of a specialized drug discovery pipelines for model creation, prediction and statistics on HPC infrastructure via HEAppE Middleware.

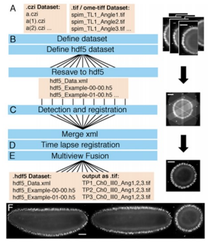
UC2 Bioimage Informatics on HPC

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 subsequently to **improving detection and intervention of serious human diseases** and other social problems associated with nature.





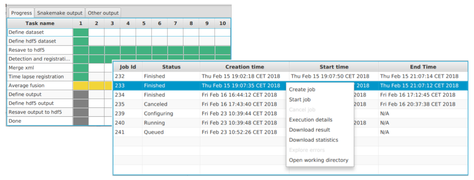




SPIM registration of Drosophila embryo

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.

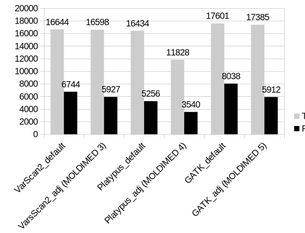
An HPC plugin for **Fiji** (Fiji Is Just ImageJ), one of the most popular **open-source** software tools for **image processing**, has been developed. This **plugin** enables end users to make use of HPC clusters to analyze **large scale image data** remotely and via the standard Fiji user interface.



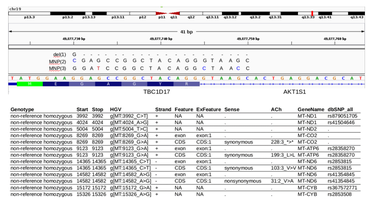
Fiji plugin utilizing HEAppE Middleware for remote execution of SPIM image processing pipeline on selected HPC infrastructure. The created framework will form a foundation for parallel deployment of any Fiji/ImageJ2 command on a remote HPC resource, greatly facilitating big data analysis.

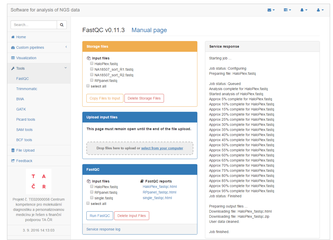
UC3 Massive Parallel Sequencing

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. 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) and data processing in **detection of hereditary and somatic DNA variants** was created.



- Custom **annotation tool** for DNA variants
- Designed for **human genetic variants** annotation
- Tested on other types of **genomes with the different ploidy**
- Effective **phenotypic prioritization** of variants
- Effective **annotation of genetic variants**
- Applicable for the **broad range of human diseases**





Specialized 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 to access the local HPC infrastructure.

Acknowledgements

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.