PROCESS

A Survey of Interactive Execution Environments for Extreme Large-Scale Computations

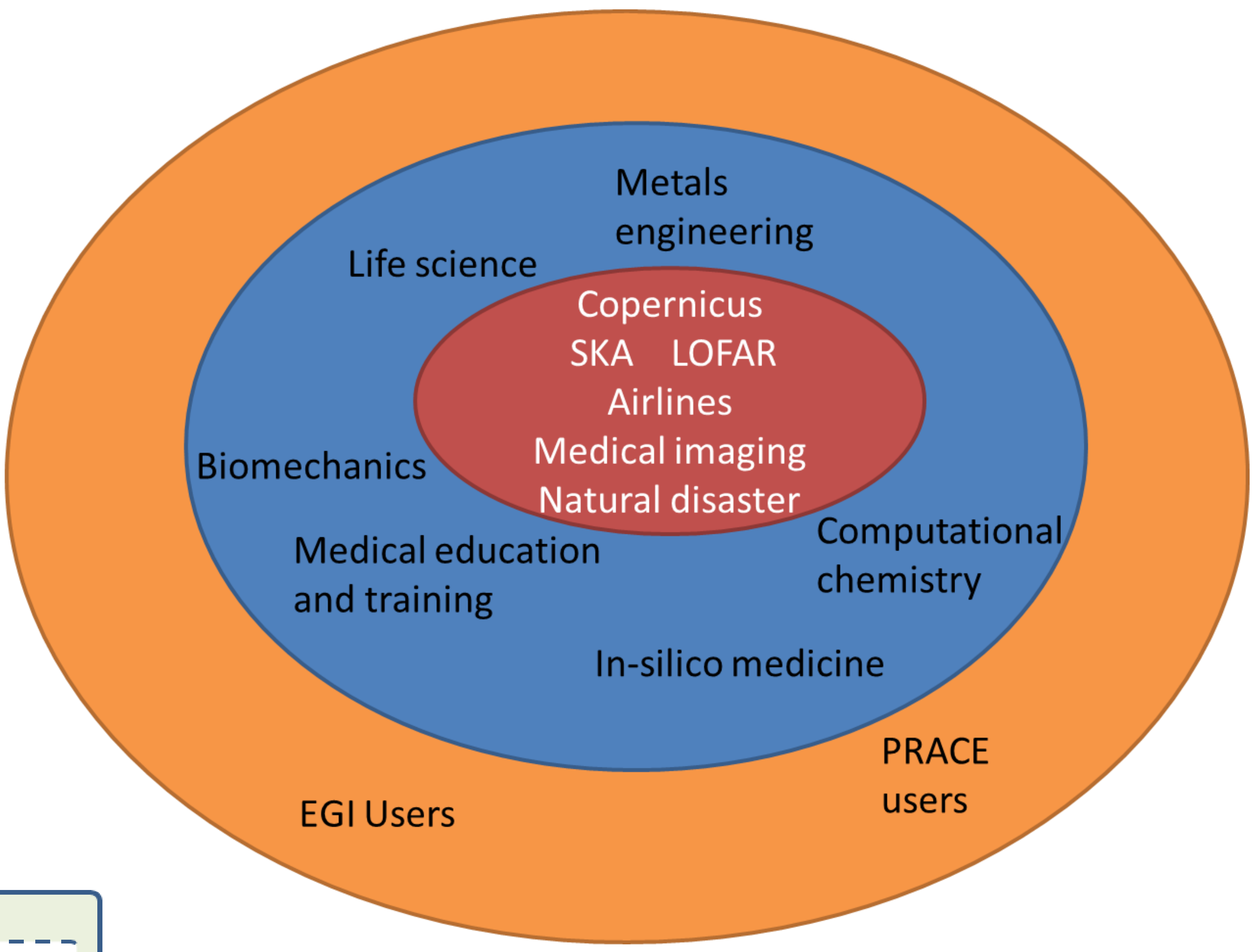
Katarzyna Rycerz[1], Piotr Nowakowski[2], Jan Meizner[2], Bartosz Wilk[2], Jakub Bujas[1], Łukasz Jarmocik[1], Michał Krok[1], Przemysław Kurc[1], Sebastian Lewicki[1], Mateusz Majcher[1], Piotr Ociepka[1], Lukasz Petka[1], Krzysztof Podsiadło[1], Patryk Skalski[1], Wojciech Zagrajczuk[1], Michał Zygmunt[1], and Marian Bubak[1,2]

[1]Department of Computer Science, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland

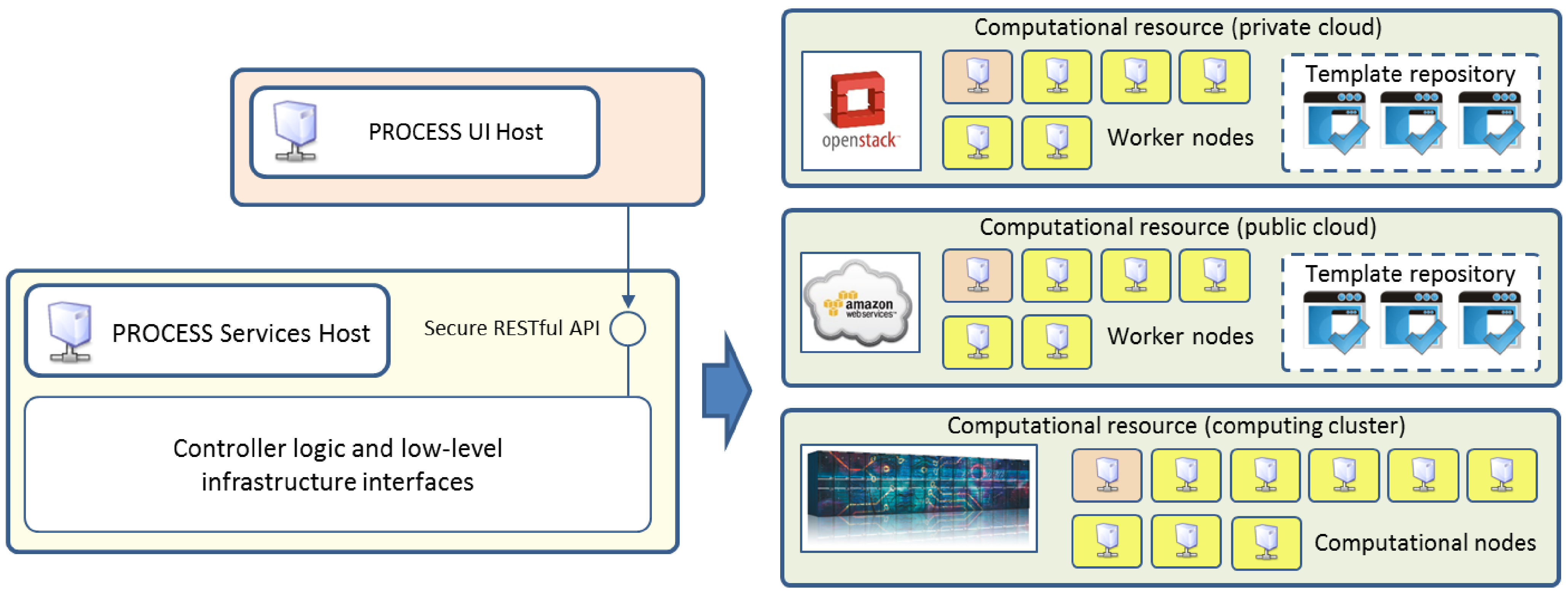
[2]Academic Computer Centre Cyfronet AGH, Nawojki 11, 30-950, Kraków, Poland

Goals

- To provide exascale ready computational and data services that will accelerate innovation
- To validate the services in real-world settings, both in scientific research and in industry pilot deployments:
  - Square Kilometre Array – a large radiotelescope project
  - medical informatics
  - airline revenue management
  - open data for global disaster risk reduction
  - agricultural analysis based on Copernicus data



Extreme Large Computing Services


PROCESS UI Host
PROCESS Services Host
Secure RESTful API
Controller logic and low-level infrastructure interfaces
Computational resource (private cloud)
openstack
Worker nodes
Template repository
Computational resource (public cloud)
amazon web services
Worker nodes
Template repository
Computational resource (computing cluster)
Computational nodes

- Based on *"focus on services and forget about infrastructures"* idea
- Support computational activities: analysis, data mining, pattern recognition, etc.
- Use heterogeneous research datasets (input and output data from modelling, simulation, visualization and other scientific applications stored in data centers and on storage systems available on European e-infrastructures)
- Support HPC and cloud based computations needed for various data analyses

Survey of interactive execution environments

- Focus on:
  - integration of scripting notebooks with HPC infrastructures to support building extreme large computing services
  - extension mechanisms required to add support specific to exascale processing of large data sets
  - ability to mix multiple languages in one document
  - integration with cloud infrastructures

Name	Large data set support	Integration with Cloud/HPC infrastructures	Extension mechanisms
R Notebook	using additional custom libraries (e.g. for Apache SPARK)	using custom libraries communicating with HPC queuing systems (e.g. SLURM)	It is possible to develop custom engines for languages which are not natively supported.
DataBricks	the whole platform is based on Apache SPARK	Available only on Amazon Web Services or Microsoft Azure	almost none
Beaker	using additional custom libraries	no specific support for HPC; Docker version available	Users can add Beaker support for unsupported languages via a dedicated API.
Jupyter	using additional custom libraries	no mature solution for HPC; Docker version available	Additional languages can be supported by writing a new Jupyter kernel.
Cloud Datalab	support for Google data services (e.g. BigQuery, Cloud Machine Learning Engine, etc.)	restricted to the Google Cloud platform	limited
Zeppelin	native support for Apache Spark	can be run on HPC using connection to the YARN cluster	support for additional languages can be added

Summary

- DataBricks and Cloud Datalab must be run on specific cloud resources
- Zeppelin and DataBricks are based on Apache SPARK, which potentially limits their usage to that platform
- R Notebooks seems promising; however, some important features are only available with a commercial version of Rstudio
- BeakerX (successor to Beaker) and Cloud Data are based on the Jupyter solution
- Jupyter seems to be a suitable base for developing extreme large computing environments

References

1. Ciepiela, E., Harężlak, D., Kasztelnik, M., Meizner, J., Dyk, G., Nowakowski, P. and Bubak, M., 2013. The collage authoring environment: From proof-of-concept prototype to pilot service. Procedia Computer Science, 18, pp.769-778.
2. Beaker Notebook webpage http://beakernotebook.com/features
3. Databricks webpage https://databricks.com/product/databricks
4. Datalab webpage https://cloud.google.com/datalab/
5. Jupyter webpage http://jupyter.org/
6. Rstudio webpage https://www.rstudio.com
7. Zeppelin web page https://zeppelin.apache.org/

AGH CYFRONET INFORMATYKA Akademia Górniczo-Hutnicza w Krakowie

http://dice.cyfronet.pl

HORIZON 2020

Funding by EU H2020 grant 777533.