deep
Hybrid DataCloud

DESIGNING AND ENABLING E-INFRASTRUCTURES FOR INTENSIVE PROCESSING IN A HYBRID DATACLOUD

www.deep-hybrid-datacloud.eu

DEEP Hybrid DataCloud is a 30 months research and innovation action financed by the European Commision under the H2020 Research and Innovation Programme in the topic of Platform-driven e-infrastructure innovation.

The key concept proposed in the DEEP Hybrid DataCloud project is the need to support intensive computing techniques that require specialized HPC hardware, like GPUs or low latency interconnects, to explore very large datasets. A Hybrid Cloud approach enables the access to such resources that are not easily reachable by the researchers at the scale needed in the current EU e-infrastructure. We also propose to deploy under the

common label of "DEEP as a Service" a set of building blocks that enable the easy development of applications requiring these techniques: deep learning using neural networks, parallel postprocessing of very large data, and analysis of massive online data streams. Several pilot applications exploiting very large datasets in Biology, Physics and Network Security are proposed, and further pilots for dissemination into other areas like Medicine, Earth Observation, Astrophysics, and Citizen Science will be supported in a testbed with significant HPC resources, including latest generation GPUs, to evaluate the performance and scalability of the solutions.

We are excited to present you the first issue of the DEEP Hybrid DataCloud Newsletter with the information about project activities and work done by the consortium during these months.

This issue contains the main objectives, a short introduction to our consortium team and project meetings, apresentation of the DEEP Open Catalog and additional articles regarding Alien4Cloud, Udocker, Software Quality. Finally, we will present the list of future events and related contacts.

All *Newsletters* will be available in the project website:

www.deep-hybrid-datacloud.eu

Don't forget to follow us on social media channels and stay always tuned for new updates and news!

@DEEP_eu

DEEP Hybrid Data Cloud

The final goal of the project is to prepare a new generation of e-infrastructures that harness latest generation technologies, supporting deep learni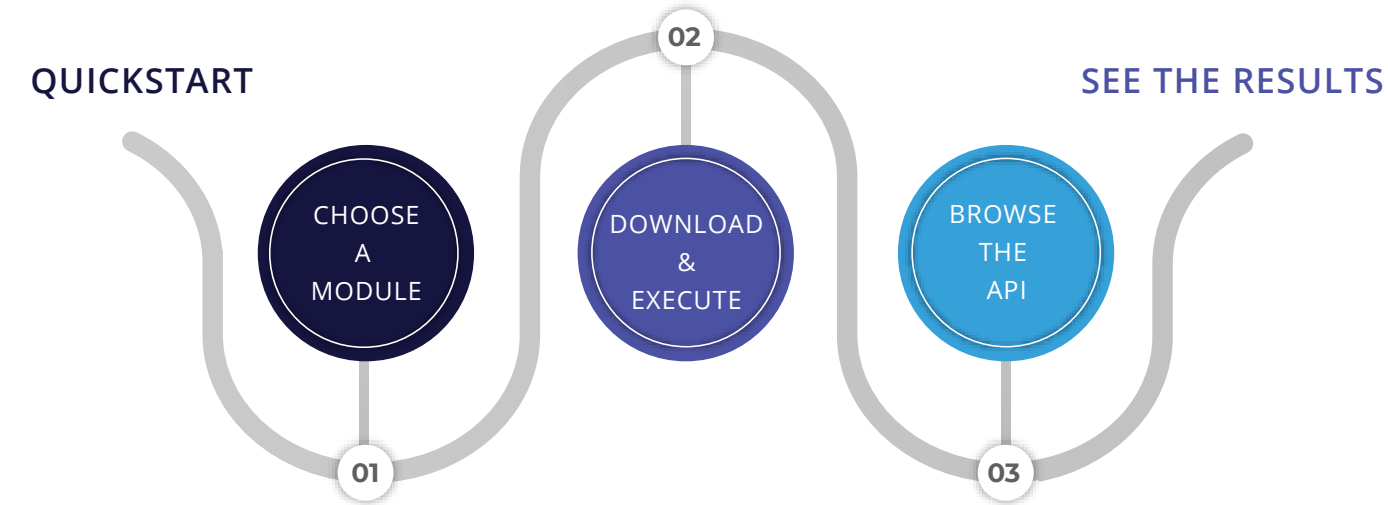ng and other intensive computing techniques to exploit very large data sources. The project will provide the corresponding services to lower the adoption barriers for new communities and users, satisfying the needs of both research, education communities and citizen science.

Focus the interest of intensive computing techniques on the analysis of very large datasets, considering demanding cases from different research communities, in the context of the future generations of e-infrastructure.

Evolve, up to production level, intensive computing services exploiting specialized hardware components, like GPUs, low-latency interconnects, and others usually accessed as "bare metal" resources. The services, based on open source software, will follow existing standards to guarantee their deployment and orchestration on different platforms.

Integrate the intensive computing services under a Hybrid Cloud approach, assuring interoperability with the existing EOSC platforms and their services.

Define a "DEEP as a Service" solution to offer an easy integration path to the developers of final applications.

Integrate the intensive computing services under a Hybrid Cloud approach, assuring interoperability with the existing EOSC platforms and their services.

DEEP-Hybrid-DataCloud is a project that aims to deliver a framework to easily develop Machine Learning and Deep Learning modules on top of e-Infrastructures. In the DEEP Open Catalog, you can find ready to use modules in a variety of domains. These modules can be executed on your local laptop, on a production server or on top of e-Infrastrucutres supporting the DEEP-Hybrid-DataCloud stack. Alien4Cloud, Udocker, Software Quality.

QUICKSTART

SEE THE RESULTS

02

CHOOSE A MODULE

DOWNLOAD & EXECUTE

BROWSE THE API

01

03

CONSORTIUM MEETING

Poznan, Poland
11-13 of September 2019

In September our team had a consortium meeting which took place in Poland. During those three days, our team had the opportunity to share and discuss their progress in different work packages. A special focus has been placed on the refinement of technical details and definition of the workplan for the final phase of the project. These kinds of face-to-face discussions ensure that all partners are aligned with the DEEP Hybrid Data Cloud's objectives and help synchronize activities.

CSIC

Consejo Superior de Investigaciones Científicas
Spain

Consejo Superior de Investigaciones Científicas (CSIC) is the largest public multidisciplinary research organisation in Spain. CSIC collaborates with national and International universities, public RTD organisations, SMEs, companies and other institutions of a scientific/technological nature. It has a staff of more than 13,500 employees, among these about 3,500 are permanent researchers and more than 4,000 are pre and post-doctoral researchers.

LIP

Laboratorio De Instrumentacao e Fisica Experimental De Particulas
Portugal

Laboratório de Instrumentação e Física Experimental de partículas (LIP) is a Portuguese scientific organization, doing research in experimental particle physics and related detector technology. LIP was created in 1986 as the Portuguese Laboratory for collaboration with the European Organisation for Particle Physics (CERN). LIP develops its R&D activities mainly in the framework of experiments at CERN, but also with the European Space Agency (ESA). LIP also uses and shares other international scientific facilities such as GSI, SNOLAB, SURF (LUX) or the Pierre Auger Observatory. The LIP scientific activity is organized in research groups focused in three main areas: experimental particle physics and astroparticle physics, development of new instruments and methods (detectors for nuclear and particle physics, medical applications, radiation environment studies and applications for space missions), advanced computing.

INFN

Instituto Nazionale Di Fisica Nucleare
Italy

The Italian National Institute of Nuclear and Particle Physics (INFN) was founded in 1951 and is a governmental research organization with 20 divisions, 4 national laboratories and three national centers spread throughout Italy. Its mission is to promote, coordinate and fund nuclear, particle and high-energy physics research in Italy. Since its creation, INFN has been developing open ICT innovative solutions for its own advanced needs of distributed computing and software applications. It has a remarkable expertise on Grid and Cloud technologies, having fostered and participated, with leadership roles, to many of the large projects financed by the EC that eventually led to the realization of the European Grid Infrastructure (EGI). The Grid middleware, based on open standards and protocols and on high-level security technologies, provides researchers with an integrated access to computing and storage systems, heterogeneous and owned by various institutions, establishing a new scalable paradigm, replicable in other sectors of the society (e.g. eGovernment).

PSNC

Instytut Chemii Bioorganicznej Polskiej Akademii Nauk
Poland

Poznań Supercomputing and Networking Center (PSNC) is affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences. PSNC employs about 250+ people divided into four departments. It is an HPC Center, Systems and Network Security Center as well as R&D Center of Future Internet, e-Infrastructure, Digital Content, Portals and Mobile Applications. PSNC is also the operator of Polish National Research and Education Network PIONIER, which is connected to the GEANT2 network, Operator of Poznań Metropolitan Area Network POZMAN. PSNC is a leading HPC center in Poland with its 300+ Tflops of computing power, storage and data management infrastructure, providing direct support for the scientific communities in Poland as well as in Europe (e.g. Nuclear Fusion, Astrophysics, Bioinformatics, Chemistry). It has been coordinating 6 international EU projects (e.g. DORII, RINGRID, GRIDLAB, Phosphorus, Porta Optica, CoolAmAll) and participated (or participating) in 80+ European and national projects (PRACE, EUDAT, IGE, NEXPRES, EGI_INSPIRE, EUFORIA, EUROFusion etc.). PSNC is the Microsoft Innovation Center and one of the founding members of PRACE AISBL, and a partner of Polish NGI. PSNC takes an active part in many international conferences and forums, including the Open Grid Forum, e-IRG.

KIT

Karlssruher Institut Fuer Technologie
Germany

Karlsruhe Institute of Technology (KIT) is the combination of a science and research institution with a University in Germany. Funded by the Federal Republic of Germany and the state of Baden-Württemberg KIT is leading in innovation, research and teaching with a staff of 10000 employees and 25000 students. KIT is a member of the Helmholtz Association of German Research Centres.
Steinbuch Centre for Computing (SCC) operates the central computing centre of KIT to support the IT-demands of the various research Programmes within KIT and within the Helmholtz-Association of German research centres. It has solid experience in both, distributed computing infrastructures (DCIs) as well as identity management and security in DCIs. SCC developed and runs the largest German university cloud storage for sync-and-share, providing access to more than 350,000 students and 100,000 researchers. SCC provides 22 PB disk (+22 PB tape) storage to a diverse range of scientific experiments hosted at more than 130 institutes on site, having fostered and participated, with leadership roles, to many of the large Projects financed by the EC that eventually led to the realization of the European Grid Infrastructure (EGI). The Grid middleware, based on open standards and protocols and on high-level security technologies, provides researchers with an integrated access to computing and storage systems, heterogeneous and owned by various institutions, establishing a new scalable paradigm, replicable in other sectors of the society (e.g. eGovernment).

FIRST PART | RUNNING A MODULE LOCALLY FOR PREDICTION

Deep Learning is nowadays at the forefront of Artificial Intelligence, shaping tools that are being used to achieve very high levels of accuracy in many different research fields. Training a Deep Learning model is a very complex and computationally intensive task requiring the user to have a full setup involving a certain hardware, the adequate drivers, dedicated software and enough memory and storage resources. Very often the Deep Learning practitioner is not a computing expert and want all of this technology as accessible and transparent as possible to be able to just focus on creating a new model or applying a prebuild one to some data.

With the DEEP-HybridDataCloud solutions you will be able to start working from the very first moment! The DEEP-HybridDataCloud project offers a framework for all users, and not just for a few experts, enabling the transparent training, sharing and serving of Deep Learning models both locally or on hybrid cloud system.

The DEEP Open Catalog provides the universal point of entry to all services offered by DEEP. It's offers several options for users of all levels to get acquainted with DEEP:

- Basic Users can browse the DEEP Open Catalog, download a certain model and apply it to some local or remote data for inference/prediction.

- Intermediate Users can also browse the DEEP Open Catalog, download a model and do some training using their own data easily changing with the parameters of the training.

- Advanced Users can do all of the above. In addition, they will work on more complex tasks, that include larger amounts of data.

The DEEP-HybridDataCloud **solution is based on Docker containers** packaging already all the tools needed to deploy and run the Deep Learning models in the most transparent way. No need to worry about compatibility problems, everything has already been tested and encapsulated so that the user has a fully working model in just a few minutes.

To make things even easier, **we have developed an API allowing the user to interact with the model directly from the web browser**. It is possible to perform inference, train or check the model metadata just with a simple click! Visit our website and see how all this work:

https://deep-hybrid-datacloud.eu/2019/07/16/deep-as-a-service-deep-learning-for-everybody/

Alien4Cloud® is an advanced cloud management tool that aims at simplifying the composition (both visual and text), management (e.g. source storing/versioning/editing), handling, and deployment (deploy, undeploy, update) of cloud topologies described using the TOSCA specification language. Alien4Cloud® supports users (with different roles) who can utilize either basic TOSCA template components or full TOSCA template topologies to deploy their own cloud infrastructures.
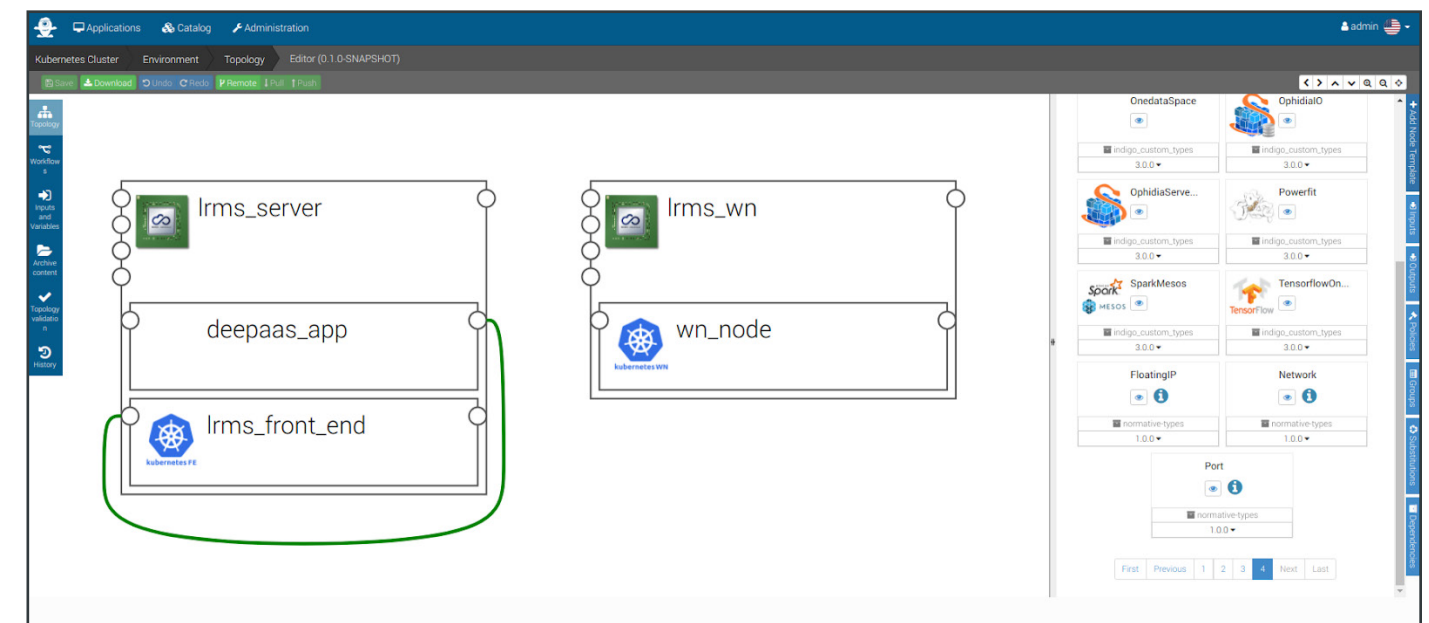

Figure 1 - *Alien4Cloud®'s TOSCA visual editor*

The DEEP Hybrid DataCloud project integrates Alien4Cloud® as the main component that allows our users to interact with the rest of the infrastructure. Figure F1 shows a custom infrastructure loaded from an existing topology template file that represents a small Kubernetes® cluster running the DEEPPaaS application.

You can find more details about DEEP and the work being done at http://www.deep-hybrid-datacloud.eu. Figure 1 depicts the role of the cloud management tool in the project. Alien4Cloud® is already a usable part of the project, please be sure you check our public demos on Youtube to see it in action.
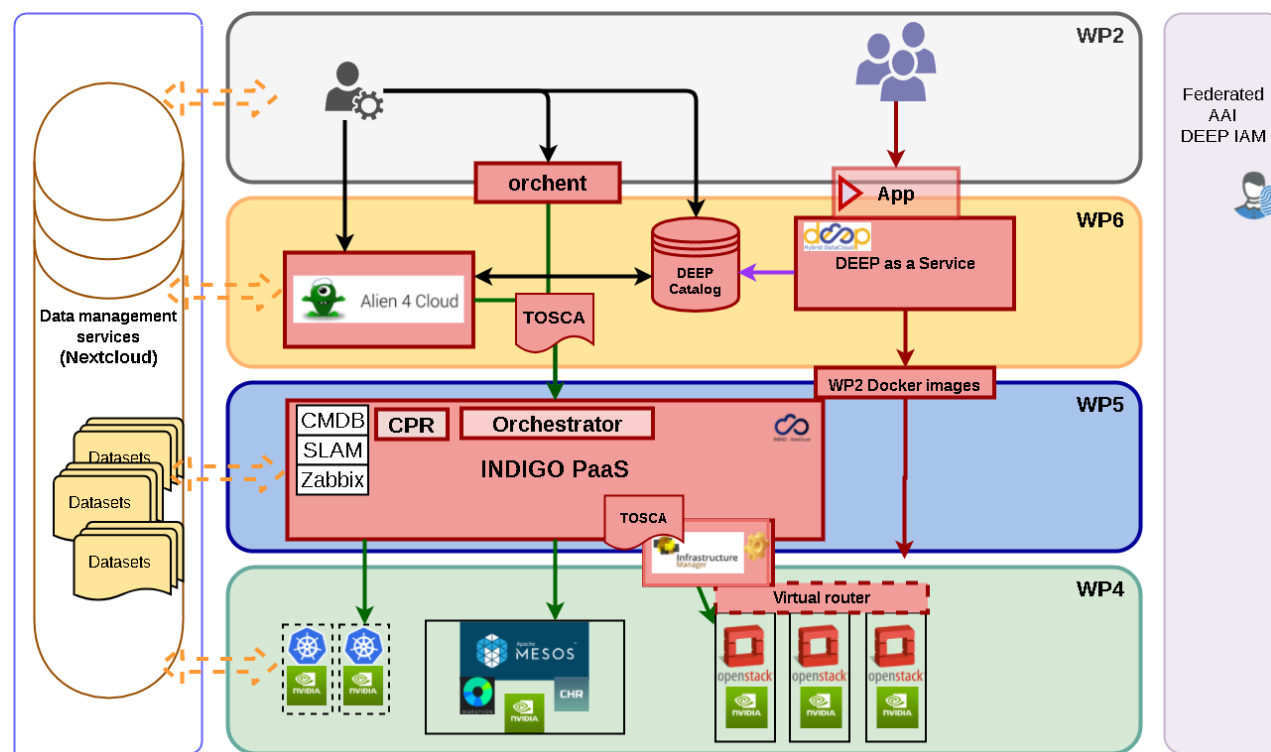
Figure 2 - *The Alien4Cloud® role in DEEP*

Since Alien4Cloud® allows functionality extension by the means of plugins, we developed one freely available on Github that connects it with the INDIGO PaaS Orchestrator used in DEEP. The Github repository doesn't contain only the code for the orchestrator plugin, it also includes the Dockerfile necessary to create a Docker image (prebuilt at https://hub.docker.com/r/indigodatacloud/alien4cloud-deep) that includes all the components needed for deployment and use in DEEP:

- Alien4Cloud, as of now our home baked version.
- The INDIGO PaaS Orchestrator plugin.
- The normative TOSCA types adapted from Openstack's version.
- The custom TOSCA types created during the INDIGO-DataCloud project and being updated during DEEP.

If you find our project interesting, stay tuned for more. There are new and exciting features coming up. We are currently working on:

- Extending the TOSCA support in the Alien4Cloud® parser, we plan to support normative TOSCA 1.2
- Improve the user experience with the graphical TOSCA editor.
- Integrate the TOSCA parser from newer Alien4Cloud® versions into the INDIGO PaaS Orchestrator, with the normative support in both components, ensuring interoperability between the two components.

While you are here, you might want to take a look at a short video to help you get started with our Docker release. More detailed information you can find here:
https://youtu.be/mCxz8LQJWWA

UDOCKER: DOCKER FOR USERS WITH NO PRIVILEGES

UDOCKER is a user tool to execute simple docker containers in user space, no root privileges required.

Containers and specifically Docker containers have become pervasive in the IT world, either for development of applications, services or platforms, the way to deploy production services, support CI/CD DevOps pipelines and the list continues.

One of the main benefits of using containers is the provisioning of highly customized and isolated environments targeted to execute very specific applications or services.

When executing or instantiating Docker containers, the operators or system administrators need in general root privileges to the hosts where they are deploying those applications or services. Furthermore, those are in many cases long running processes publicly exposed.

In the world of science and academia, the above mentioned benefits of containers are perfectly applicable and desirable for scientific applications, where for several reasons users often need highly customized environments; operating systems, specific versions of scientific libraries and compilers for any given application.

Traditionally, those applications are executed in computing clusters, either High Performance Computing (HPC) or High Throughput Computing (HTC), that are managed by a Workload Manager.

The environment on those clusters is in general very homogeneous, static and conservative regarding operating system, libraries, compilers and even installed applications.

As such, the desire to run Docker containers in computing clusters, has arisen among the scientific community around the world but, the main issue to accomplish this has to do with the way Docker has to be deployed in the hosts, raising many questions regarding security, isolation between users and groups, accounting of resources, and control of resources by the Workload Manager among others.

To overcome the issues of executing Docker containers in such environments (clusters), several container framework solutions have been developed, among which is the udocker tool.

More information about **UDOCKER** you can find on our website.

📅 **EOSC-hub week**
Prague (April 2019)

📅 **5th Summer School on Intelligent Signal Processing for Frontier Research and Industry**

(Infieri), Wuhan, China

📅 **IoT Solutions World Congress**
29-31 October 2019   **Barcelona** (Spain)

📅 **The Premier Machine Learning Conference**
18-19 Novembre 2019   **Berlin** (Germany)

📅 **UCC 2019 International Conference on Utility & Cloud Computing**
02-05 December 2019   **New Zealand** (Australia)

www.deep-hybrid-datacloud.eu

 @DEEP_eu

 DEEP Hybrid Data Cloud

 deep-info@listas.csic.es