DEEP-Hybrid DataCloud project: Hybrid services for distributed e-infrastructures

A. Costantini[1], D.C. Duma[1], D. Michelotto[1], A. Falabella[1], E. Ronchieri[1], A. Ceccanti[1], M. Caberletti[1], D.Salomoni[1]

[1] INFN-CNAF, Bologna, Italy

E-mail: alessandro.costantini@cnaf.infn.it

Abstract. DEEP Hybrid DataCloud is an Horizon 2020 project that addresses the need to support intensive computing techniques that require specialized HPC hardware, like GPUs or low-latency interconnects, to explore very large datasets. Launched in November 2017 the H2020 DEEP Hybrid-DataCloud - DEEP-HDC is lasting for 30 months and is combining the expertise of 10 large European research organisations. The project proposes 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 post-processing of very large data, and analysis of massive online data streams. This contribution will introduce the project, presents the foreseen overall architecture and the contribution to the activities that are being carried on by INFN-CNAF personnel to achieve the project goals and objectives.

1. Introduction

Machine learning 'as-a-service' can clearly be seen as one of the main user requirements being asked for when thinking about using large-scale computing infrastructure. With ever more data being available, the wish to exploit this data is omnipresent. While machine learning pipelines for small-scale data sets are available in the form of several software libraries, large-scale learning tasks provide another level of challenge. With the need for a proper design of the learning task at hand, additional tasks result in the need to organize large-scale data, the provision of necessary computing power and storage capacity and, since large-scale data is commonly distributed, the orchestration of various infrastructure components at different places. It is obvious that such learning tasks cannot be managed by a user with domain knowledge in the field of application, only. Therefore, support by the infrastructure layer must break down the complexity of the task and allow the user to focus on what she/he is skilled on, i.e., modelling of the problem, evaluating and interpreting the results of the machine learning algorithms. As a consequence, infrastructure providers have to understand the needs of their user communities and help them to combine their services in a way that encapsulates technical details the end user does not have to deal with. Indeed, this is the goal of the DEEP-HDC project [1] aimed at deploying 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 post-processing of very large data, and analysis of massive online data streams.

The targeted platforms for the released products are the already existing and the next generation e-Infrastructures deployed in Europe, such as the European Open Science Cloud

(EOSC) [2], the European Grid Infrastructure (EGI) [3] and the computing infrastructures that will be funded by the upcoming H2020 EINFRA-12 call. DEEP-HDC is funded by the H2020 EINFRA-21-2017 Research and Innovation action under the topic Platform-driven e-Infrastructure innovation [4]. It is carried on by a Consortium that brings together technology providers with a proven long-standing experience in software development and large research communities belonging to diverse disciplines: Biological and Medical Science, Computing Security, Physical Sciences, Citizen Science and Earth Observation. DEEP-HDC started on 1st November 2017 and will run for 30 months until April 2020. The EU contribution for the project is 2.98 million euros.

2. Project Objectives

The DEEP-Hybrid-DataCloud project started with the global objective of promoting the usage of intensive computing services by different research communities and areas, and their support by the corresponding e-Infrastructure providers and open source projects. Other objectives followed by the project are:

- Focus on intensive computing techniques for the analysis of very large datasets considering highly demanding use cases.
- Evolve up to production level, intensive computing services exploiting specialized hardware.
- Integrate intensive computing services under a hybrid cloud approach.
- Define a "DEEP as a Service" solution to offer an adequate integration path to developers of final applications.
- Analyse the complementarity with other ongoing projects targeting added value services for the cloud.

The DEEP-Hybrid-DataCloud project aims to provide a bridge towards a more flexible exploitation of intensive computing resources by the research community, enabling access to the latest technologies that require also last generation hardware and the scalability to be able to explore large datasets. It is structured into six different work packages, covering Networking Activities (NA) devoted to the coordination, communication and community liaison; Service Activities (SA) focused on the provisioning of services and resources for the execution of the data analysis challenges; and Joint Research Activities (JRAs), dealing with the development of new components and technologies to support data analysis.

In order to achieve these objectives, we propose to evolve existing cloud services, taking into account the

following design principles:

- Evolve the required services from TRL6 to TRL8 under an open framework and considering existing standards for interoperability.
- Re-use if possible existing cloud services in production, and in particular those being adopted for proposed e-infrastructure of the European Open Science Cloud.
- Consider the integration of existing specialized resources into cloud services having in mind the point of view of the current daily management of those resources, like for example current HPC data centres.
- Ensure that the resulting framework will have a low learning curve for the developers of the solutions, by delivering a DEEP catalogue that can be directly exploited by users to build their applications.
- Assure the scalability and performance of the solution developed, which is key to guarantee the interest both of resource providers and users.

3. Project structure

The DEEP Hybrid DataCloud project is structured into six different work packages, covering Networking Activities (NA) devoted to the coordination, communication and community liaison; Service Activities (SA) focused on the provisioning of services and resources for the execution of the data analysis challenges; and Joint Research Activities (JRAs), dealing with the development of new components and technologies to support data analysis. Figure **??** describes the interaction between the different work packages.

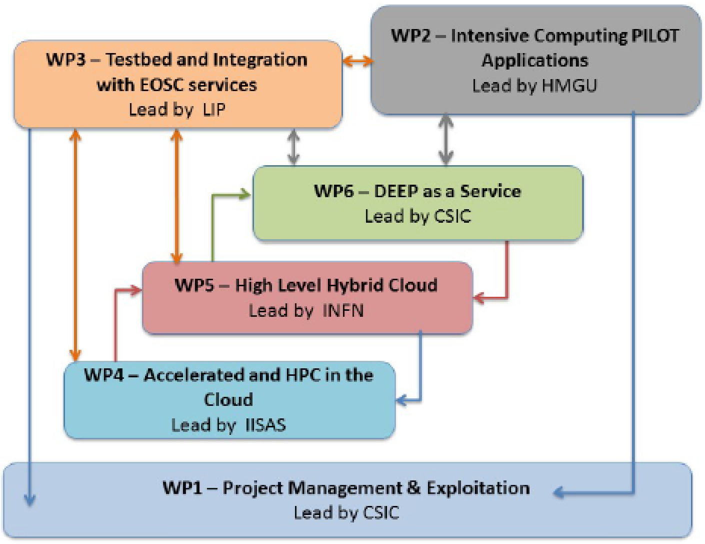

Figure 1. Diagram showing WPs interrelation.

It is important to remark the key role of WP2/NA2 and WP3/SA1 to define and channel towards the JRA work packages, WP4, WP5, WP6, the requirements for the solutions to be developed, and then provide feedback. The direct interaction between WP2 and WP6 will promote an agile interaction on the design and DEEP Hybrid DataCloud implementation of the services for final users.

Work Package 1, WP1 (NA1) - Project Management and Exploitation. This work package will perform the global oversight of the activities carried out within the project, ensuring that they are aligned with the DEEP Hybrid DataCloud work programme. WP1 will also coordinate the consortium management through the governance structure, including the promotion of an adequate interaction between the WP through the steering committee.

Work Package 2, WP2 (NA2) - Intensive Computing Pilot Applications. This work package is responsible for the definition and correct understanding of the pilot usage scenarios regarding the project's technical architecture and will propose an architecture that is applicable for the identified applications. Moreover, NA2 will interact with SA1 and JRA3 to ensure that the delivered outcomes are aligned with the expectations of the user communities, are compliant with the proposed scenarios and validated against the user applications.

Work Package 3, WP3 (SA1) - Testbed and integration with EOSC services. The service activities of the project will be supported by WP3, that will guarantee that the project pilot testbeds are correctly integrated with other state of the art and e-Infrastructures and services from the European Open Science Cloud (EOSC), so that the project can exploit their services in an easy way. Moreover, this work package will supervise the software development within the project, providing a continuous software improvement process that will involve quality assurance activities, software release management, maintenance and support. In particular, INFN-CNAF is coordinating Task 3.2 - Software quality assurance, release, maintenance and support. Expressed in terms of Software Quality Assurance and Software Release and Maintenance, CNAF is coordinating the management of those software products that became officially part of the first DEEP releases, codenamed Genesis [5], foreseen for late 2018 and effectivley released in January 2019. INFN CNAF is also coordinating the implementation of the continuous software improvement process, following a DevOps approach, through the definition and realization of an innovative Continuous Integration (CI) and Delivery (CD) system. INFN-CNAF is contributing also to Task 3.1 - Pilot testbeds and integration with EOSC platform and their services - by providing and maintainig the testbeds dedicated to developers, software integration and software preview. In particular, the activities were focused in implementing the services needed to support the software development and release management and included among others the source code repository, and continuous integration system.

Work Package 4, WP4 (JRA1) - Accelerated and High Performance Computing in the Cloud. This key research activity will be carried out close to the hardware and infrastructure, addressing the gaps that currently exist in the support of accelerators (like GPU), specialized hardware (such as low-latency interconnects) and HPC systems in general. In particular, INFN-CNAF contributes to Task 4.3 - Interaction with HPC resources with PaaS approach - by making available its experience on virtualization technologies and cloud middleware frameworks to develop advanced solutions that enable the delivery of bare-metal like performance and the resource sharing in multi-tenancy environments.

Work Package 5, WP5 (JRA2) - High Level Hybrid Cloud solutions. Lead by INFN, WP5 will take care of the provisioning of the platform exploiting the outcomes from JRA1 in a hybrid approach, delivering an execution platform for JRA2, ensuring that applications can be spawned in across several cloud infrastructures. In particular, INFN-CNAF contributes to Task 5.1 - PaaS-level Orchestration Supporting Multi-IaaS Hybrid Infrastructures - by contributing both to the development of solutions aimed at provision and orchestrate resources among multiple IaaS and to the test of the new TOSCA Templates that are able to deploy clusters of services.

Work Package 6, WP6 (JRA3) - DEEP as a Service. This activity focuses on bridging the outcomes of NA2, JRA1 and JRA2 so as to deliver the final solution to the users in the form of a DEEP as a Service solution. This service will ensure that scientists have an easy way to deploy and execute their intensive compute applications based on containers (from NA2) that will be executed in an hybrid cloud platform (JRA2), exploiting the specialized hardware that their application requires (JRA1). INFN-CNAF contributes to the JRA3 activities and task by supporting integration and testing activities aimed at composing a set of defined building blocks that will model the user application and deploying these applications as services that can be offered to final users, as a way to deliver scientific results to a wider scope of stakeholders. In particular, INFN-CNAF is testing the DEEP-Alien4Cloud [6] plugin, elected as the tool able to provide an easy to use and intuitive application composition to deliver the DEEPssS solutions.

4. DEEP Overall architecture

The DEEP PaaS layer is based on the components developed and integrated in the INDIGO-DataCloud project [7]. The architecture is depicted in Figure 2 and the main components are briefly described hereafter:

The PaaS Orchestrator is the core component of the PaaS layer. It receives high-level deployment requests and coordinates the deployment process over the IaaS platforms. The Identity and Access Management (IAM) Service provides a layer where identities, enrolment, group membership, attributes and policies to access distributed resources and services can be managed in an homogeneous and interoperable way.

The Monitoring Service is in charge of collecting monitoring data from the targeted clouds,

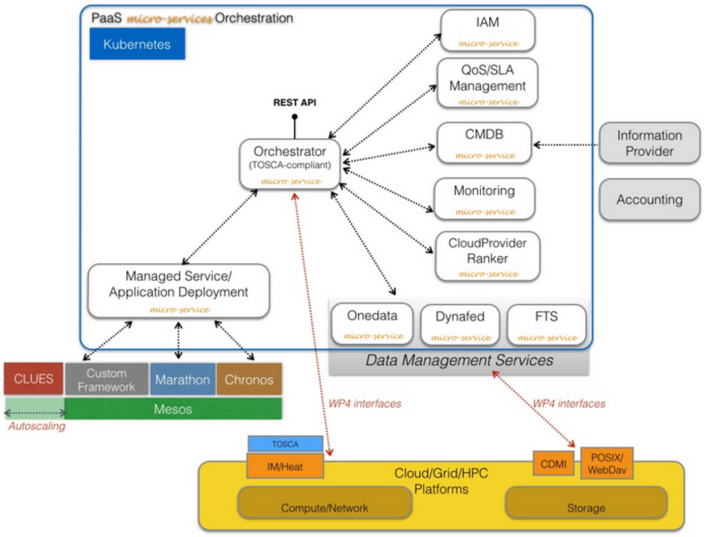

Figure 2. The architecture of the DEEP PaaS layer is based on the building blocks provided by the INDIGO-DataCloud project.

analysing and transforming them into information to be consumed by the Orchestrator. The Cloud Provider Ranker (CPR) is a rule-based engine that allows to rank cloud providers in order to help the Orchestrator to select the best one for the requested deployment. The ranking algorithm can take into account preferences specified by the user and other information like SLAs and monitoring data. The SLA Management (SLAM) Service allows the handshake between users and a site on a given SLA. The Managed Service/Application (MSA) Deployment Service is in charge of scheduling, spawning, executing and monitoring applications and services on a distributed infrastructure; the core of this component consists of an elastic Mesos cluster with slave nodes dynamically provisioned and distributed on the IaaS sites. The Infrastructure Manager (IM) deploys complex and customized virtual infrastructures on a IaaS site providing an abstraction layer to define and provision resources in different clouds and virtualization platforms. The Data Management Services is a collection of services that provide an abstraction layer for accessing the data storage in a unified and federated way. The Information Provider and Accounting System collects detailed information from an IaaS provider about the current status of the resources from the amount of resources of CPU, RAM or storage to the availability of a service.

5. DEEP as a Service solution

The high level decomposition of the DEEP as a Service design is depicted in Figure 3 and consists on the following key components:

- The DEEP open Catalog where the users, communities, etc. can browse, store and download relevant modules for building up their applications (like ready to use machine learning frameworks, complex application topologies, etc.).
- An application modeler or composition tool, that will be used to build up complex application topologies in an easy way.
- A runtime engine, that will take the defined topology as input, provision the required computing resources and deploy the application.

- The DEEP PaaS layer, that will coordinate the overall workflow execution to choose the appropriate Cloud sites and manage the deployment of the applications to be executed.
- The DEEP as a Service solution, that will offer the application functionality to the user.
- The EINFRA/EOSC data services, to be integrated with the DEEP solutions in order to provide access to any of the data facilities existing in the European Open Science Cloud

The system is designed with extensibility in mind, taking great care in designing a framework which can be updated easily and where a component can be replaced with a new one in case it is needed. Many of the anticipated changes to our system in future phases will only require adding additional functionality on top of existing components, remaining backwards compatible with previous versions.

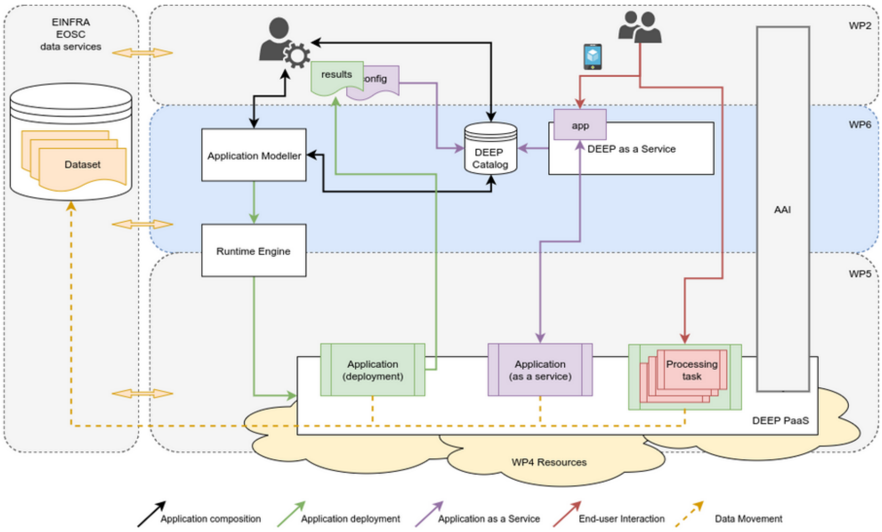

Figure 3. DEEPaaS high level architecture.

6. Conclusions

In the present contribution the DEEP-XDC project and its objectives have been presented and discussed. These objective, together with the related needs proper of the research communities involved in the project, are the real driver to develop innovative and reliable open source solutions able to fill up the technology gaps that currently prevent effective exploitation of distributed computing and storage resources by many scienti

c communities.

For the second part of project, the activities carried on at INFN-CNAF will continue to ensure the fulfilment of the project objectives. In particular, the already available softare solutions will be enriched by advanced functionalities (provided by JRAs) aimed at addressing the use case requirements provided by NA2. The implementation and related testing of those new solutions will be performed in the testbeds maintained by SA1. SA1 will also continue its activities aimed at further validate the software, its robustness and scalability and will follow the preparation of the second project release, DEEP-2 , foreseen for the second half of 2019.

Moreover, DEEP-HDC project can complement and integrate with other running projects and communities and with existing multi-national, multi-community infrastructures. As an example,DEEP-HDC is collaborating with the eXtreme-DataCloud (XDC) [8] project aimed at developing scalable technologies for federating storage resources and managing data in highly distributed computing environments.

7. References

[1] Web site: www.deep-hybrid-datacloud.eu
[2] Web site: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[3] Web site: https://www.egi.eu/
[4] Web site: http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/einfra-21-2017.html
[5] Web site: https://deep-hybrid-datacloud.eu/2019/01/18/deep-genesis-first-software-release-is-out/
[6] Web site: https://github.com/indigo-dc/alien4cloud-deep/
[7] Web site: https://www.indigo-datacloud.eu
[8] Web site: www.extreme-datacloud.eu