eFlows4HPC

Enabling dynamic and Intelligent workflows in the future EuroHPC ecosystem

D1.2 eFlows4HPC interfaces and Iteration 1 software stack release

Version 1.0

Documentation Information

Contract Number	9555558
Project Website	www.eFlows4HPC.eu
Contractual Deadline	28.02.2022
Dissemination Level	PU
Nature	OTHER
Author	Domenico Talia (DtoK)
Contributors	Jorge Ejarque (BSC), Yolanda Becerra (BSC), Anna Queralt (BSC), Loic Albertin (Atos), Jedrzej Rybicki (FZJ), Alessandro D'Anca (CMCC), Donatello Elia (CMCC), José Flich (UPV), Nihad Mammadli (BSC), Salvatore Giampà (DtoK)
Reviewer	Rosa M. Badia (BSC)
Keywords	Software stack, Architecture, Programming interface

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955558. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Germany, France, Italy, Poland, Switzerland, Norway.

Change Log

Version	Description Change
V0.1	Proposed table of contents
V0.2	Preliminary version with links to Readthedocs
V0.3	Revised version after review
V0.4	Final comments added
V1.0	Document formatted for submission

Table of Contents

1. Executive Summary

This deliverable provides the first version of the design and implementation of the eFlows4HPC software stack made available in the public project repository. The main documentation is provided through the Read the Docs documentation framework. The link to the online documentation is https://eflows4hpc.readthedocs.io/ and the documentation repository is https://github.com/eflows4hpc/documentation.git. The modality adopted for this deliverable allows providing its contents online and evolving them as a live document as the project development proceeds. Releases to the document will be performed for each specific milestone which can be directly seen in the Read the Docs document. The version corresponding to this deliverable is version 1.0.

2. Introduction

WP1 encompasses integration and development activities for providing programming interfaces to support workflows that integrate HPC, high performance data analytics and machine learning. In particular Task 1.3 focuses on the design of the identification of the necessary programming interfaces for the integration of HPC, data analytics and machine learning in a single workflow. Given the architecture defined in Task 1.2, Task 1.3, together with Task 1.4 and Task 1.5, considers the necessary data-driven and control-driven interactions between the different layers and components and defines the interfaces required for a smooth execution of the workflows providing a complete software stack integration.

This deliverable releases the first version of the results obtained in Tasks 1.2, 1.3, and 1.4, which composes the Iteration 1 software stack release.

3. Release infrastructure

The eFlows4HPC Software Stack is released by using a set of tools, services and cloud resources to facilitate the open-source distribution, installation and usage documentation as well as the deployment of the Software Stack components. The following paragraphs describe the different components of the release infrastructure.

3.1. Github Organization

To manage the source code of the implemented components, documentation and other code examples, we have set a Github organization available on the following link.

https://github.com/eflows4hpc/

It provides a public space to manage different git repositories to store open-source code, package repositories to store binaries and container images, and the *github* teams working in the different implementations of the project. In this organization, we have stored the open-source releases of the components fully implemented in eFlows4HPC as well as forks of components which are modifications done in the eFlows4HPC project but not merged in the master branch of the original

component. These repositories are pinned to appear at the upper part of the main organization page.

Apart of component repositories, it also contains a repository to store the sources and build scripts for the on-line documentation of the eFlows4HPC Software Stack which is published in the *Read the Docs* platform[1].

3.2. Read the Docs documentation

Read the Docs is a platform to simplify software documentation by automating building, versioning, and free hosting of open source software. We have created a documentation project for eFlows4HPC in *Read the Docs*, with the goal of providing the users with the most updated version of the documentation in accordance with the modification done during the implementation of the project. This eFlows4HPC documentation project is linked to the documentation repository stored in the eFlows4HPC github organization. This repository can be found in the following link:

https://github.com/eflows4hpc/documentation

For the different version branches created in the repository, a new version is automatically created in the documentation. The master branch of the repository refers to the latest version of the documentation. Every time a push or merge is performed in one of these branches, the corresponding version is automatically updated. The version available for this deliverable is version 1.0.

The generated on-line documentation and the different versions can be found in the following link:

https://eflows4hpc.readthedocs.io/

The documentation git repository contains the *doc* section where the content material is stored and the *Readme* file together with the setup files. All the document sections are in the *doc/sources/Sections* folder which stores the source files of the documentation following the reStructuredText format[2].

The repository and its generated on-line documentation is structured in five main sections. The *eFlows4HPC* section provides an overview of the project. The *Software Stack* section provides an overview of the architecture and short description of the component together with useful information on how to find the source code, the component installation and usage guidelines; Afterwards, the *Programming Interfaces for integrating HPC and DA/ML workflows* introduces the workflow programming interfaces available to integrate different software in a single workflow. Then, the *HPCWaaS Methodology* introduces the HPC Workflows as a Service methodology proposed by the eFlows4HPC project. Finally, the *Usage Example* section provides an example on how to use the different components together to currently develop and execute a workflow using the eFlows4HPC components.

[1] https://readthedocs.org/
[2] https://docutils.sourceforge.io/rst.html

3.3. Deployment Environments

To facilitate the component integration and uptake of the Software Stack, we plan to use two cloud environments: the *HDF Cloud* at the Jülich Supercomputing Center and *nCloud* (https://ncloud.bsc.es) at Barcelona Supercomputing Center. Both Cloud services are managed by OpenStack, which allow developers to create VMs where they can deploy the implemented gateway services of the eFlows4HPC Software Stack. In the *HDF Cloud*, users can find deployed versions of the Data Catalogue, a prototype version of the Data Logistic Service, the Alien4Cloud, the Ysitia Orchestrator, and the HPCWaaS execution service. The end-points of the deployed services can be found in the Software Stack page of the documentation. We are working on replicating this environment in the *nCloud* infrastructure to allow developers to advance in new developments in one environment while Pillar's workflow developers can validate the implementations in the other environment.

4. Implementations

This section introduces the material provided in the Read the Docs repository about the implementation components of the eFlows4HPC software stack release together with the interface implementations. The project is delivering the eFlows4HPC software stack that integrates different components to provide an overall workflow management system. One of the core functionalities of the software stack is the definition of the complex workflows that combine HPC, HPDA and ML frameworks and the integration of large volumes of data from different sources and locations. Using this software stack, the project builds an HPC Workflow as a Service (HPCWaaS) platform to facilitate the reusability of these complex workflows in federated HPC infrastructure. The sections of the Read the Docs repository that describe the adopted solutions can be found here: https://eflows4hpc.readthedocs.io/. Apart from the introductory one (eFlows4HPC), they are as follows.

4.1. Software Stack

Here is presented an overview of the eFlows4HPC software stack release and deployment. The software components of the designed stack are listed and introduced. In particular the software stack is composed of three layers: On the top, the programming models used for the definition of the complex workflows that combine HPC, HPDA and ML frameworks and the integration of large volumes of data from different sources and locations. In the middle there are the components to facilitate the accessibility and reusability of workflows. Finally, in the bottom part of the stack are the different components for deployment, execution and data management.

(See: https://eflows4hpc.readthedocs.io/en/latest/Sections/01_Software_Stack.html).

4.2. Programming Interfaces for integrating HPC and DA/ML workflows

This section introduces the programming interfaces and explains how the eFlows4HPC programming interface has been designed to reduce the effort required to integrate different frameworks in a single workflow. The integration is divided in two parts: Software Invocation Management and Data Integration. In this first iteration, we have defined the software invocation descriptions and we have extended the PyCOMPSs programming model and runtime accordingly.

(See:https://eflows4hpc.readthedocs.io/en/latest/Sections/02_Programming_Interfaces.html)

4.3. HPCWaaS methodology

In this section the HPC Workflow as a Service (HPCWaaS) methodology is sketched that applies the usage of the Functions as a Service (FaaS) model in Cloud environments for workflows in HPC systems. This model defines different main roles. In particular, the function developer is in charge of developing and registering the function in the FaaS platform, and the final user executes the deployed function using a REST API. In the case of running workflows in HPC systems, we can find similar roles. In fact, the workflow developer is in charge of developing and deploying the workflow in the computing infrastructure, and the users' communities that execute the workflow and collect their results to advance in their scientific goals. This section in the Read the Docs repository includes two subsections describing the Development interface and the Execution API.

(See:https://eflows4hpc.readthedocs.io/en/latest/Sections/03_HPCWaaS_Methodology.html)

4.4. Usage example

This section introduces a usage example of a scientific workflow that integrates the main components (a data logistic pipeline and a PYCOMPSs workflow) of the first release of the eFlows4HPC software stack. In this first version of the workflow, the required software and the access credentials are already deployed in the infrastructure. Next versions will include how to do it with the eFlows4HPC Software Stack.

This section in the Read the Docs repository includes three subsections describing the implementation of the Data Logistics pipeline, the PyCOMPSs workflow and the integration of both data logistics and PyCOMPSs application in TOSCA.

(See: https://eflows4hpc.readthedocs.io/en/latest/Sections/04_Usage_Example.html)

5. Conclusions

The eFlows4HPC software stack integrates different components to provide an overall workflow management system. It is a key element for implementing a European workflow platform enabling the design of complex applications that integrate HPC processes, data analytics and artificial intelligence, making use of the HPC resources in an easy, efficient and responsible way as well as

enabling the accessibility and reusability of applications to reduce the time to solution. In this deliverable are presented all the software components and a first release of their integration.

In the provided Read the docs repository all the software stack components are described together with links and references to their code and user manuals to allow their deployment. The HPC Workflow as a Service (HPCWaaS) methodology is also presented and through a usage example is described how to implement, deploy and execute a real-world workflow using the eFlows4HPC Software Stack.

6. Acronyms and Abbreviations

- DA – Data Analysis
- DoA – Description of Action (Annex 1 of the Grant Agreement)
- EB – Executive Board
- EC – European Commission
- HPC – High Performance Computing
- ML - Machine Learning