D-NA2.3: Second report on validation and evaluation of enabled applications deployment and use cases

30/10/2013

Project acronym:	VERCE
Project n°:	283543
Funding Scheme:	Combination of CP & CSA
Call Identifier:	FP7-INFRASTRUCTURES-2011-2
WP:	WP2/NA2, Pilot applications and use cases
Filename:	D-NA2.3.pdf
Author(s):	A. Michelini
Location:	http://www.verce.eu/Repository/Deliverables/RP2/
Type of document:	Deliverable
Dissemination level:	Public
Status:	Draft
Due date of delivery:	30/10/ 2013
Reviewer:	J.-P. Vilotte
Keywords:	data-intensive, cpu-intensive, HPC, earthquake, seismology, data infrastructure, forward modeling, inversion

---

Version	*Author*	*Date*	*Comments*
1	A. Michelini (INGV)	27/10/2013	Initial draft for comments
2	A. Michelini (INGV)	30/10/2012	Revised after adding contributions
3	J-P Vilotte (IPGP)	30/10/2012	Revision and comments on introduction and conclusions

Copyright notice

Contents

List of Figures

Executive Summary

The main objectives of WP2/NA2 are: (1) select existing pilot data-intensive applications and design sound use case scenarios; (2) analyze and define a use case implementation strategy during the project with WP8, WP7 and WP9; (3) support and evaluate the "productising" transition of the methods and their implementation performed by WP8; (4) support and evaluate the deployment and the efficiency of the pilot applications and their use case scenarios on the VERCE platform; (5) define in collaboration with NA3 documentation and tailored training session material; (6) provide requirements and support to WP7 and WP9 for tailored interfaces of the scientific gateways targeted to the developers and the users.

VERCE's primary objective consists of "enabling" existing data- and HPC-intensive software applications through the development of processing elements (PEs) within dedicated workflows. It follows that the applications to be enabled or under enablement are all well developed and already have their own line of development. This also implies that they have already chosen their dissemination strategies through tutorials, web portals, etc.

During this RP and according to the activities and deliverables provided in the DoW, NA2 continued the validation and evaluation of the enabled applications deployment and use cases reported in the D-NA2.1. The activities involved interactions with WP8/JRA1 supporting the creation of the parameter file generator for multiple solvers and with WP7/SA3 providing the point of view of seismologists to the realisation of the VERCE gateway front- end. Technical details of these activities are covered in the JRA1 and SA3 reporting deliverables.

1 Introduction

WP2/NA2 activity is based on the seismological community of practice—the main stakeholder of VERCE. In the first reporting period, the main effort of NA2 has been devolved on the identification and selection of VERCE use cases. Two use cases were identified and first prioritised—one each for the compute- and data-intensive types of application. For the compute-intensive, it has been selected "forward modeling and inversion" use case whereas for the data-intensive, the ambient noise cross-correlation analysis.

The development and the implementation of the use cases on the VERCE platform have been modified to take into account the results of the review meeting held in April 2013, in Paris. Specifically, the Steering Committee, SC, during a meeting which took place at Charles De Gaulle Airport on 30 July 2013 decided to give precedence to the HPC compute-intensive use case to the goal of providing by March 2014 a working and ready-to-be-tested beta-version. This use case entails the development of waveform simulation of earthquake events and the analysis of the misfit between the simulated waveforms and the observed waveforms at a given set of stations to evaluate the quality of the earth models. The second decision of the SC regarded the data-intensive use case which needs for the moment much attention on the following main barriers identified by the seismologists — the ingestion of the raw data sets with appropriate data structure organisation and formats, and the data management layer during the end-to-end workflow.

Following what agreed by the SC in July, the use cases are now being implemented with different priorities and for the D-I special efforts are put into building a proper data federation platform through the iRods data management software that insures transparency and scalability across the different nodes.

During the reporting period object of this document, NA2 has contributed to improve defining the workflow of the compute-intensive use case. This activity has been carried out in close collaboration (symbiosis) with WP8/JRA1. For the data-intensive use case, NA2 has interacted closely with WP9/JRA2 and WP7/SA3 which have provided some initial developments of the data intensive architecture and much effort has been put by seismologists into the use of tools like Python they are most accustomed to. To this end, Python tools were explored to facilitate this development and prototyping approach, among others SAGA-Python, Disco and PyRods. In the following sections, we present also a summary of other activities in which NA2 has participated or given seismological feedback from the user community perspective.

2 Compute-intensive—*Forward modelling and Inversion*

During this reporting term, NA2 continued the validation and evaluation of the enabled applications deployment and use cases. The activities involved interactions with WP8/JRA1 supporting the creation of the parameter file generator for multiple solvers and with WP7/SA3 providing the point of view of seismologists to the realisation of the VERCE gateway front- end. Technical details of these activities are covered in the JRA1 and SA3 reporting deliverables.

Focusing on the activities concerning the Gateway Service Implementation for the forward modelling use case, NA2 provided feedback to SA3 shaping the user experience of the Forward Modelling GUI. We expect that the main end-users of VERCE gateway will be geoscientist and we designed a beta release that exposes the waveform computation as a service for the needs of the seismological community.

The VERCE Forward Modelling GUI is composed of two sections: an interactive map integrating geological overlayers and a tabbed windows for the definition of the workflow parameters and the submission of the simulation. In particular users can (i) select the solver (from a list of available codes already deployed within the VERCE computational resources), (ii) specify the input parameters, (iii) include the mesh that they want to adopt for the simulations, (iv) identify seismic stations and earthquake that are sought for simulation. Users can graphically select the mesh and the associated velocity model, choosing

from a database that will be created in the next reporting term. For the beta version, NA2 provided the mesh of Abruzzo region (Italy) and a corresponding tomography. Furthermore, the possibility to upload userÕs configuration files and input datasets is a relevant, useful and appealing feature.

At each step, the GUI grants the users the option to download the configuration files in the current format of the chosen solver.

Thanks to these possibilities, the seismological community considers the GUI as a critical improvement of the current process of preparing the simulation.

Several fruitful feedback iterations with SA3 have been necessary to design the workflow. The possibility to access and play with implemented features in the beta release provides the possibility to individuate needs and request of the seismological community. Collaborative evaluation and design of functionalities of the VERCE science gateway represent the most important path in order to accomplish the expected results. Therefore, it is very important that the communication flow among the partners will be constantly updated and enriched by an active participation, as in this case.

3 Data-intensive—*Ambient Noise analysis*

In[1] previous reporting periods, most (of the) relevant PEs were developed, and a working prototype DISPEL workflow was setup and tested. Testing also revealed several issues, which needed to be addressed and detail feedback was provided to JRA2, SA1 and SA3:

The preliminary implementation of the portal did not yet allow the participating scientists to interact actively or reconfigure the workflow. The workflow was preconfigured by the developers, which limited the possibility of scientists to actively participate in the development and fully appreciate the advantages of a DISPEL based workflow.

The workflow uses a data stream abstraction, where data is passed directly from one PE to the next, the details of the implementation are hidden from the user. Currently, these data streams are based on the seismic waveform streams in the standardised MiniSEED format. This format fits the pre-processing phase quite well, but is not sufficiently versatile for a complete representation of the processing results of the subsequent phases, i.e. cross-correlograms, cross-correlograms integrated over time periods and post-analysis results. This hinders the possibility to proceed with the implementation of more complex and scientifically relevant time-integration (sliding time windows) and post-analysis PEs. During the preprocessing phase, optional interpolation of data with small gaps would require priori information on the gap length, which is not available in advance in the data stream, making it difficult to implement such functionality.

The identified issues were addressed in regular calls in the Task Force calls and a detailed feedback was provided to JRA2, SA1 and SA3. This should have helped to set the priorities and to shape further developments of the platform during the current and next reporting period. In particular, significant progress on the scientific gateway and the data management layer could be observed. See for also D-JRA-2.1.2 for further details.

In order to address the need of seismologists to pursuit their scientific objective and advance the development of their analysis while the VERCE platform is still under development it was agreed, that it should become possible for seismologist to compose their workflows by creating script entirely in Python to tide together the functionality of the Python processing elements. This would also allow for fast prototyping and comparison of new approaches, before their adaptation to the VERCE platform. A set of available Python tools were explored to facilitate this development and prototyping approach, among others SAGA-Python, Disco and PyRods.

[1] Part of text describing the D-I activities appears also in the D-JRA1.2.2. This is done in order to maintain self-consistency in the reporting.

The data management layer is of fundamental relevance for the data intensive use case. It is needed to manage persistent data both input data and processing results, but also for intermediate data with longer lifetime, which should not be purged immediately after processing. VERCE is opting for iRODS (integrated Rule-based Data System), a choice shared with the EUDAT project. Various options to organise data in the iRODS based archive, and to access data were explored. JRA1 also participates on the discussion and evaluation of the necessary meta-data structure for the archive and approaches on how to organise the interface between the data archive and the DISPEL workflow. The definition of the data formats of the intermediate results is an ongoing effort, and first ideas were developed. For efficient access to storage media standard container formats and in particular HDF5 are a promising choice. The HDF5 file format allows to collect and organise data in relatively large file, which are best suited for modern storage technology, i.e RAIDs and parallel file systems, but also provides efficient direct access to subsets of the stored data. However, direct access to portions of data needs to be reconciled, with the requirements of a streaming approach, which calls for a serialisation of data.

4 Participation to other activities

- The VERCE project was presented at the *41st Workshop of the International School of Geophysics* held at the 'Ettore Majorana' Center in Erice on August 26-31, 2013. The topic of the school was *A Roadmap for Earth Science in Europe: The next generation of Geophysical Research Infrastructures*. The project itself was presented to an audience of approx. 100 international scientists from the solid Earth Sciences and ICT. A discussion followed on how the VERCE platform could best help the community and be part of the EPOS e-infrastructure.

5 Conclusions and next steps

NA2 is actively involved in the developments of VERCE by providing feedback from the user side of the seismologists. During the next reporting term it is expected

- Use of Python scripting to obtain a development and prototyping platform;
- Testing concepts and extending the functionality of the current data intensive workflow;
- the initial implementation of the entire data-intensive workflow by exploiting the OGSA-DAI and STORM functionalities through the DISPEL workflow enactment;
- testing and evaluation of the data-intensive use case;
- the testing and evaluation of the "forward modelling and inversion" use case.
- provision seismologists' expertise to to JRA1, NA3, JRA2 and SA2

In conclusion, the activities of NA2 are blending progressively with those of the other work packages through the feedback provided. Following the reviewers' comments, the SC has decided to modify the VERCE program and pay particular attention to data ingestion and data management issues that are core to all the data-intensive developments. On the hand, the HPC (Compute-intensive) use case is expected to benefit of the ongoing middleware and job submission software being implemented. This process is expected to amalgam further as the implementation of the use cases on the VERCE platform strengthens in the incoming months.