D-NA2.4: Third report on validation and evaluation of enabled applications deployment and use cases

29/09/2014

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-NA2.4.pdf
Author(s):	E. Casarotti and A. Michelini
Location:	http://www.verce.eu/Repository/Deliverables/RP4/
Type of document:	Deliverable
Dissemination level:	Public
Status:	Final
Due date of delivery:	03/10/ 2014
Keywords:	data-intensive, cpu-intensive, HPC, earthquake, seismology, data infrastructure, forward modeling, inversion

---

Version	*Author*	*Date*	*Comments*
1	E. Casarotti and A. Michelini (INGV)	27/09/2014	Initial draft for comments
2	A. Rietbrock (ULIV)	08/10/2014	Revision with suggestions
3	A. Michelini (INGV)	22/11/2014	Revised after adding suggestions
3	F. Magnoni and E. Casarotti (INGV)	09/11/2015	Review

Copyright notice

Contents

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 "productizing" 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 that are under construction are all well developed and already have their own line of implementation. This also implies that they have already been chosen their dissemination strategies through tutorials, web portals, etc.

The activities involved interactions with WP7/SA3 and JRA1 in the realization of the VERCE gateway front-end, in particular seismologists provided guidelines, suggestions, testing efforts and models/meshes to implement the current stage of the forward simulation portal. Technical details of the portal are in SA3 reporting deliverable.

1 Note on the modified NA2 data-intensive objectives

Two seismological use cases were chosen by NA2/WP2 in order to test the VERCE platform. The first consists of enacting 3D forward modelling wave simulation codes such as SPECFEM3D. The second is the "ambient noise cross-correlation" use case which is strictly a data-intensive use case. In Summer 2013, because of a delay accumulated by the project during the first two years, the Steering Committee, in agreement with the review panel, considered to push forward the 3D forward modelling wave simulation use case with the unavoidable accumulation of additional delay for the data-intensive use case. This change of priorities had an impact onto the overall objectives originally planned by NA2. Also, the amassed delay made it difficult to match the original roadmap laid out in the DoW. Nevertheless, focusing onto the forward modelling use case allowed to make a true step forward and this has been demonstrated during the Garching review meeting held at the beginning of June 2014. Moreover, the slow down of the data-intensive use case allowed also to reconsider it more thoroughly.

In general and in simple words, any data-intensive use case entails the ingestion of a large amounts of data and the execution in a pipeline of a number of processing analysess upon them. Thus, regardless of the selected ambient noise cross-correlation use case, attention is to be paid to the basic ingestion/pipeline analyses just mentioned. Dispel4py — a Python library used to describe abstract workflows for distributed data-intensive applications, also developed in the framework of the VERCE (e.g., WP7/SA3) — project provides these features and in July 22-23 2014 a workshop has been organized at INGV dedicated to these issues.

During the workshop emphasis has been put onto the following topics

1. Assessment of the status of the Data-Intensive (D-I) methods and technology
2. Better understanding of the scope and potential of D-I methods for seismology
3. commit ourself to immediate plans within VERCE
4. Develop a longer-term view about data-intensive applications and technology

The meeting involved bothsenior scientists of VERCE (Malcolm Atkinson, Alberto Michelini and, in Skype connection the first afternoon, Jean-Pierre Vilotte) and junior scientists —- the effective users and developers — who had the chance to test the dispel4py software directly on their laptop. The overall participation was around 10-15 people including scientists and technologists from University of Edinburgh, CNRS/IPGP, KNMI and INGV. A particular characteristics of the dispel4py is that abstract dataflows described in dispel4py can be executed in numerous environments, for example using a Storm cluster or as an MPI job besides own laptops where it can be developed and tested. Thus dispel4py allows to construct workflows without particular knowledge of the specific context in which they are to be executed, granting them greater generic applicability.

The workshop had informal sessions with intensive discussions. Result of the workshop in the framework of the objectives to be achieved by VERCE and discussed within the Steering Committee, is that the ambient noise implementation as originally planned is not achievable within the timeline of the project. There are different factors that contribute to this. First, the data-intensive community is more fragmented and there is currently no single authoritative development available in contrast with HPC where SPECFEM3D is definitely the authoritative code used by many scientists. It follows that this fragmentation would weaken the selection of a particular development, eventually. A second reason was mentioned above and it refers to the intrinsinc nature of the data-intensive use cases because, involving and relying on repetitive I/O, it results that only some specific parts of the workflow need to be re-engineered to gain computational speed (i.e., HTC). Highlighting of this aspect is important since it has re-directed the data-intensive development toward much more basic aspects than the

specific use case. To this regard, the development of the dispel4py library in Python is fundamental since it can be used by (computational) scientists who are already accustomed to this language. (In October a webinar courseon dispel4py has been held as part of the training activities of VERCE, http://www.verce.eu/Training/UseVERCE.php#october2014). Finally, dispel4py is now being benchmarked and the results will be presented in the next reporting.

2 Steps identified in the last report (D-NA2.3 10/13):

In order to achieve a beta-version of the portal, during the previous reporting period, NA2 suggested (i) to focus on SPECFEM3D, (ii) to expand the library of tomographic models and meshes, (iii) to allow the submission of models and meshes by users, (iv) to design a more flexible workflow submission control.

Furthermore, the collected feedbacks suggest to add more specific steps for the second part of this reporting period and for the upcoming one. In particular, NA2 considers: 1) automatising the meshing check in order to provide the user with a simpler and quicker automatic submission, 2) improving the visualization of synthetics, 3) collecting the feedbacks from seismological community in order to improve the forward modeling portal, 4) adding multiple features likes Shakemovies creation and visualization, 5) creating the workflow for comparison of data and synthetics, leading the path to the inverse problem use case.

3 ACTIVITIES during the third reporting period

A deep collaboration matured between NA2, SA3 e JRA1 and allowed to implement in the Forward Simulations Portal the requested features during last period. Face-to-Face (hackathon) meetings between components of the WPs involved in the development of the VERCE Science Gateway have been productive and well structured. While technical aspects have been address by SA3 and JRA1, NA2 provided fundamental seismological points of view and tests. In particular:

- Forward Modelling implementation for the VERCE portal
  - definition of the workflow steps to implement FW modeling in the gateway;
  - definition of the GUI UX that translates the file-centric user interface of SPECFEM3D in the structure of the gateway;
  - creation and validation of the preliminary database of meshes and and corresponding tomographic velocity models;
  - creation of visualisation scripts (wave propagation animation in 2D and 3D environment, syntethics seismograms, Google Earth KML);
  - creation of script for automatic check and validation of meshes;
- Test and Feedback reporting on:
  - functionalities of the gateway for 3D simulation tasks;
  - access to the gateway through certificates;
- PE development: collaboration to define the PE for:
  - creation of input files of stations, seismic events and simulation parameters;
  - visualisation;
  - misfit data and synthetics (in progress);
  - automatic meshing check (in progress);