D-JRA1.2.3: Enabling pilot applications: third report and validation of the VERCE architecture

23/11/2015

Project acronym:	VERCE
Project n°:	283543
Funding Scheme:	Combination of CP & CSA
Call Identifier:	FP7-INFRASTRUCTURES-2011-2
WP: WP:	WP8/JRA1, Harnessing data-intensive applications
Filename:	D-JRA1.2.3.pdf
Author(s):	H. Igel, L. Krischer
Location:	http://www.verce.eu/Repository/Deliverables/RP4
Type of document:	Deliverable
Dissemination level:	Restricted
Status:	Final
Due date of delivery:	30.10.2014
Keywords :	HPC, seismology, e-science environment, service-oriented architecture, Grid, earthquake, data infrastructure, seismology, report

---

Version	*Author*	*Date*	*Comments*
1	H. Igel, Lion Krischer LMU	10/10/2014	Initial draft for comments
2	H. Igel, Lion Krischer LMU	20/10/2014	Implementation of comments by E. Casarotti, A. Rietbrock, F. Magnoni
3	H. Igel, Lion Krischer LMU	28/10/2014	Implemented comments by internal reviewers (M. Atkinson, J.P. Vilotte)
4	H. Igel, Lion Krischer LMU	30/10/2014	Final Version

Copyright notice

Contents

Executive Summary

The main objective of JRA1 is to analyze the use cases proposed by the scientific domain experts of NA2 and map the components onto the VERCE platform, by means of implementation and development. In the initial stage of the project JRA1 prioritized and analyzed a list of applications and use cases and the decision for the pursued use cases was taken accordingly. To this means, the experts of JRA1 are responsible for providing the applications and software implementations with a computational harness to enable their use within the VERCE architecture proposed by JRA2. In this reporting period the focus was on 1) enabling the complete forward modelling workflow (data storage, job submission, data storage), 2) developing and extending processing elements for data-intensive and cpu-intensive use cases, 3) developing new scalable data and metadata formats for seismology.

1 Work Progress in this Reporting Period

In the last report JRA1.2 the following steps for this reporting period were identified. Other tasks were initiated in response to further comments by reviewers and technological requirements:

- Complete the mapping of the forward modelling use case on the VERCE platform
- Build mesh library to extend and allow high-level use of the VERCE platform for 3D simulation tasks
- Finish the definition of the new waveform file format and develop a library dealing with it.
- Development of a visualization PE for wave propagation
- Exploring the full potential of iRODS rule engine and micro services and use this functionality to obtain setups which allow efficient data movement and management
- Continuing the discussion on meta data structure and how to integrate it with into the data management layer
- Use of Python scripting to obtain a development and prototyping platform
- Testing concepts and extending the functionality of the current data intensive workflow
- Improving the efficiency of the current implementation of selected processing elements
- Integrating functionality with a more general purpose into the ObsPy toolkit

1.1 Enabling CPU-intensive and data-intensive use cases

In this section we report on progress developing the underlying scientific software for the implementation of the use cases.

1.1.1 ObsPy

ObsPy [1] is a central library used for both, the CPU-intensive and the data-intensive, use cases . It is a community-driven, open-source project dedicated to building a bridge for seismology into the scientific Python ecosystem. It offers read and write supports for all commonly used file formats in seismology, a comprehensive signal processing toolbox, integrated access to large data centers and web services, and convenient wrappers to legacy codes. It enjoys a large world-wide rate of adoption in the community and continues to see many successful applications. The following developments were made (steered and headed by the LMU):

- Released a new major and a new minor version (0.9.0 and 0.9.2, [2]) with new features and bug fixes. The new headline features are support for the FDSN StationXML standard [4] and a generic FDSN webservice [7] client. Further new features are support for the Nanometrics Y file format, the CSS waveform data format, the NEIC PDE bulletin event catalog, and a new client to access data from the NEIC CWB QueryServer. A very large number of bug fixes and stability improvements were also implemented. Please refer to the release notes for more details.
- The latest development version now supports Python 3.3 & 3.4 in addition to 2.6 & 2.7. The non-backwards compatible changes in Python 3 make this a tedious but necessary task essential for future prospects.

- Read support for the NDK file format used by the GCMT earthquake catalog [10]. Useful as input for the forward simulation use case as most other webservices do not serve moment tensor solutions which are an essential parameter for most waveform solvers. This has been used to generate per-mesh earthquake data sets for the forward simulation use case.
- Further features currently in the development version of ObsPy include, amongst others:
  - A new design for the documentation which is a major reason for the success of ObsPy. Now also properly displays on mobile devices.
  - More generic automatic tracking of processing information within the ObsPy framework easing provenance extraction.
  - Support for version 1.1 of the FDSN web services.
  - Built-in plotting of instrument responses.
  - Support for the SACPZ, ZMAP, DYNA/ITACA, and Kinemetric EVT data formats
  - A new easy to use SeedLink interface.
  - Bugfixes and stability improvements.

1.1.2 LASIF

LASIF [3] is an implementation of the complete full waveform inversion work flow using adjoint techniques that is intended to be gradually mapped on the VERCE platform. It performs most steps necessary to carry out an inversion in an automatic but user-controllable manner. These steps include the domain definition, automatic data acquisition from a wide range of sources, managing of various kinds of data, data verification tools, provenance tracking, misfit and adjoint source calculations, input file generation for a number of solvers, and other things.

Developments include (LMU):

- LASIF is now in a fully working state.
- Improvements with regards to parallel processing and iteration provenance tracking.
- Documentation and tutorial completed.
- Can now interface with SPECFEM and SES3D.
- New and better internal structure to facilitate future developments.
- Handling of some new file formats to keep up with the changing seismological data ecosystem.
- Developed a web interface to explore the data inside a LASIF project.

1.1.3 Waveform Solver Input File Generator

The wfs_input_generator project [11] is a data-driven generator producing input files for numerical waveform solvers from a variety of different sources. This is an important task when implementing workflows interfacing with numerical codes. It is used in the VERCE portal.

Recent developments include (LMU):

- More detailed event descriptions.
- Support for SPECFEM GLOBE and SES3D 4.1
- Bugfixes, tests, and full documentation

1.1.4 WHISPER

WHISPER is a library with functionalities for the data-intensive use case developed by the University at Grenoble. CNRS investigated the WHISPER library to see which functions or classes are appropriate to adapt to VERCE. Some of the processing functions of traces, such as whitening or time alignment were adapted to VERCE as PEs using dispel4py.

1.1.5 Other Software

VERCE – despite its focus now on the SPECFEM software for forward modelling – is designed to allow use of a variety of mature solvers developed for the seismic forward problem. The following software adaptations were carried out by the LRZ:

- SES3D-NT (spectral element solver on regular grids): coaching of Ses3d-NT main developer with respect to language/implementation specific questions, coarse code-analysis of Ses3d-NT, using small-scale test-runs one-on-one with main developer, identification of performance bottlenecks inside the code, and discussion of alternative implementations, identification of portability issues, mainly due to use of gcc-non-standard extensions, as well as use of non-mature but standardized features commonly still unsupported by most compiler vendors.
- SeisSol (discontinuous Galerkin method for unstructured tetrahedral grids): discussions on programming style and guidelines for SeisSol code-developments, partial code replacement of non-free or non-open-source numerical routines, partial code refactoring for improved encapsulation, modularity and reduction of unsafe programming pattern (reduction of possible points of failure)

1.2 Data and Meta-data formats

A project like VERCE requires the definition of appropriate standards for meta data and actual data that are scalable towatrds the use of large data volumes. No such standards existed at the beginning of the VERCE project and VERCE partners are beginning to play a leading role – in direct collaboration with the global seismological community – in proposing and establishing new standards. Recent developments include (LMU, CNRS):

- A new scalable, parallel seismic data format dubbed ASDF, the adaptable seismic data format [12]. Developments include (LMU in cooperation with Princeton):
  - A full description of the potential use cases and advantages of the new format.
  - Finished a first version of the description of the format.
  - Fully working Python implementation including a convenient interface to ObsPy. It includes a way to interactively edit and discover data, as well as and easy-to-use way to perform fully parallel I/O.

Furthermore, a system for provenance of seismological data is in preparations (LMU). It is called SEIS PROV [14] and is an extension to the W3C PROV provenance data model. It is intended as a provenance exchange format stored alongside the data with information about data origin and applied processing. A large number of provenance record types common in seismology have been described. We are currently working on an implementation within the ObsPy framework. To that end, we added PROV XML support to a Python package for provenance [13].

CNRS has initiated a discussion on metadata for seismic data towards the ”Definition of Data Structure for the Catalogue of seismic data” in particular in connection with the data-intensive use case (provenance

structure, processing history). In addition, python scripts were developed for an iRODS micro service to extract Catalogue metadata from mini seed structure of raw files at injection time in iRODS. The script can produce (output JSON) either the ds-Cat from document above or the structure defined in the metadata api.

1.3 Platform Developments

The progress reported here in this section refers to the actual technical enabling of the use cases on the VERCE science gateway.

INGV focused on the forward modelling implementation for the VERCE portal using SPECFEM3D (in cooperation with KNMI), in particular:

- Final definition of the workflow steps to implement forward modelling use case in the portal;
- Support for implementation of solver-specific (SPECFEM3D, https://github.com/geodynamics/specfem3d) requirements in the portal structure;
- Supply of 3D meshes and corresponding tomographic velocity models to build a preliminary archive for the portal;
- Tests and feedback reporting on the functionalities (including visualization of synthetic seismogram, wavefield propagation, ground shaking) of the portal for 3D simulation tasks, particularly using SPECFEM3D;
- Tests and feedback reporting on the preliminary steps to access to the portal through Grid and proxy certificates

Several processing elements (PE) were developed in cooperation with UEDIN.

- Finalization of a PE that interrogates available databases and produces suitable input files of stations and events to run SPECFEM3D through the portal (in collaboration with Alessandro Spinuso and Lion Krischer);
- Creation of a PE to visualize snapshots and movies of the wavefield propagation and the peak ground velocity by managing the output files of SPECFEM3D;
- (In progress) creation of a PE to visualize the mesh and the superimposed velocity model from the output files of SPECFEM3D;
- (In progress) creation of a PE to visualize movies of the wavefield propagation in three-dimensional volumes;
- (In progress) definition of the workflow to process observed and simulated seismograms, to compare them and calculate the corresponding misfit measurements, and creation of the required processing elements (in collaboration with Mario David).

2 Next Steps

The focus for the final project year is coordinated in connection with the roadmap presented elsewhere. JRA1 will support and/or deliver the following tasks identified in the roadmap (Table 1)

- Waveform solvers and forward simulation servers

- Metadata definition, formats, and collection during the workflows
- Creating the link between simulation and observation: misfit calculations
- mesh and model upload functonalities
- Further development and provision of ObsPy library
- Seismogical examples pipeline for workflow implementation
- Development of provenance functionalities for metadata

References

[1] http://obspy.org/

[2] https://github.com/obspy/obspy/releases

[3] http://krischer.github.io/LASIF

[4] http://www.fdsn.org/xml/station/

[5] https://quake.ethz.ch/quakeml/

[6] http://www.hdfgroup.org/

[7] http://www.fdsn.org/webservices/

[8] http://krischer.github.io/wfs_input_generator/

[9] https://www.irods.org/

[10] http://www.globalcmt.org/

[11] http://krischer.github.io/wfs_input_generator/

[12] http://asdf.readthedocs.org

[13] https://github.com/trungdong/prov

[14] http://asdf.readthedocs.org/en/latest/provenance.html