D-NA2.2: First report on validation and evaluation of enabled applications deployment and use cases

01/10/2012

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-NA2.2.pdf
Author(s):	A. Michelini
Location:	http://www.verce.eu/Repository/Deliverables/RP2/
Type of document:	Deliverable
Dissemination level:	Public
Status:	Final
Due date of delivery:	01/10/ 2012
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)	27/09/2012	Initial draft for comments
2	A. Michelini (INGV)	30/09/2012	Revised after reviewer's comments

Copyright notice

VERCE, *Virtual Earthquake and seismology Research Community e-science environment in Europe*, is a project co-funded by the European Commission as an Integrated Infrastructure Initiative within the 7th Framework Programme. VERCE began in October 2011 and will run for 4 years.

Contents

List of Figures

Executive Summary

The WP2/NA2 addresses various goals. It is responsible for selecting the existing pilot data-intensive applications and design sound use case scenarios and together with WP8/JRA1, WP5/SA1, WP9/JRA2 and WP7/SA3 analyse and define a priority strategy through the project. Secondly, WP2/NA2 supports and evaluates the "productising" transition of the methods and their implementation performed by WP8/JRA1. It also supports WP5/SA1 and WP9/JRA2 with application requirements for the definition of the workbenches and functionalities. Of major importance for WP2/NA2 is the support and evaluation of the deployment and of the efficiency of the pilot applications and their use case scenarios on the VERCE platform. In collaboration with WP3/NA3, the WP defines and provides improved documentation, best practice guides and tailored training session material and, together with WP4/NA4, defines and provides demonstrators and dissemination material while providing requirement and support to WP7/SA3 and WP9/JRA2 for tailored interfaces of the scientific gateways targeted to the developers and the users.

During this RP and according to the activities and deliverables provided in the DoW, NA2 started the validation and evaluation of the enabled applications deployment and use cases. The activities involved interaction with WP8/JRA1 to refine the workflow of the cpu-intensive "forward modelling and inversion" use case and the initial testing of some components of the data-intensive workflow enacted through the DISPEL language developed in WP7/SA3. In order to obtain a complete picture of the work done and given the close collaboration existing between the WPs, this deliverable should be "read together" with the WP8/JRA1and the WP7/ SA3 reports. This report is thus general on the activities of NA2 carried out during the RP whereas the companion D-NA2.2.1 report focuses solely on the metrics to be adopted for user evaluation of the platform and use cases.

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. These use cases are now being implemented and efforts have been made during this reporting period in the various work packages to set up both the VERCE platform with its associated services and the set up of the workflow framework upon which the use cases will be enacted. During this reporting period, 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 a first implementation of the use case using the workflow language DISPEL. 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*

The initial workflow presented in the first report has been refined and detailed during this second period. Figure 1 and figure 2 (cf. the D-JRA1.2 for additional technical details) provide the principal stages of the workflow of this use case. More specifically, the text in black in the flow-diagram representation provides information on the sought services or requirements whereas in red are shown the technical solutions that have been tested (or are being planned) and are now under implementation. The contribution of NA2

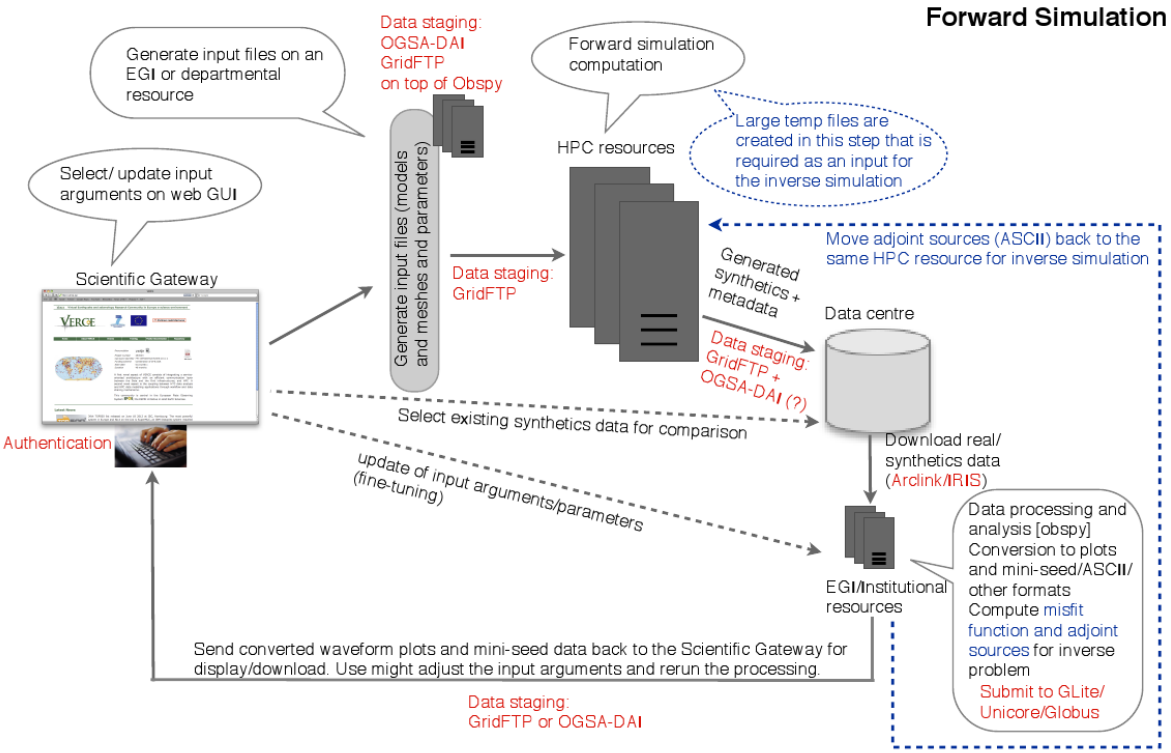

Figure 1 – Forward Modelling Workflow

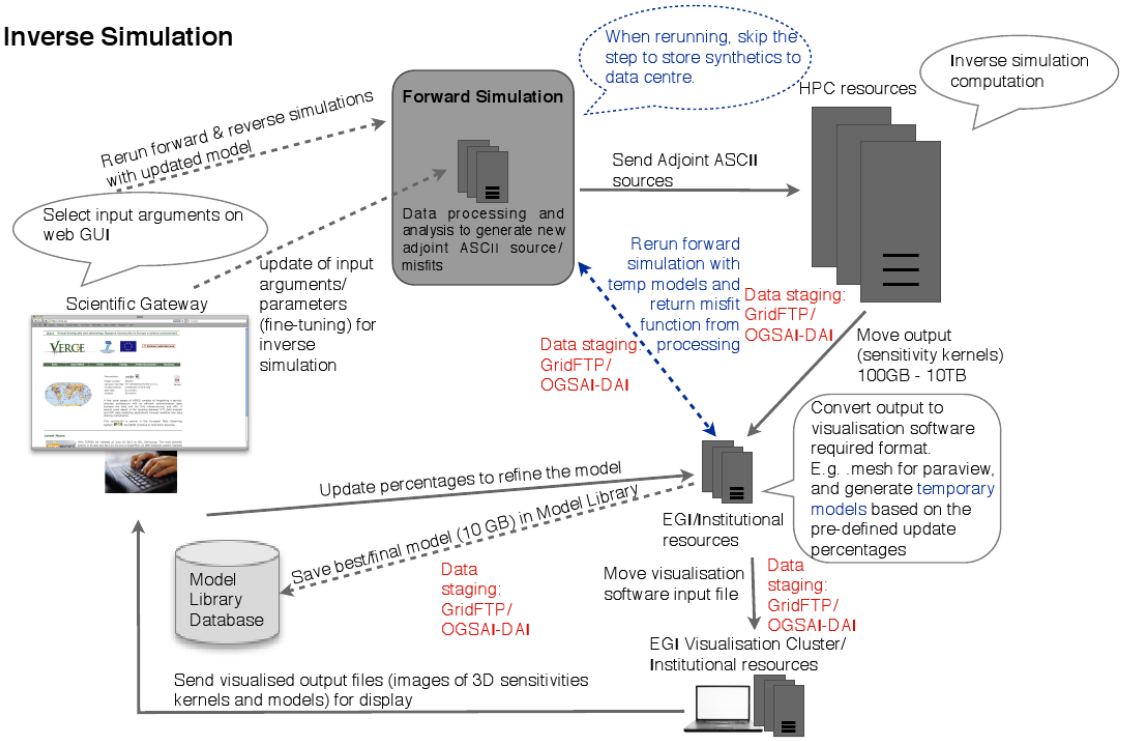

Figure 2 – Inverse Modelling Workflow

has been basically centred on refining the steps of the workflow by providing users' expertise.

3 Data-intensive—*Ambient Noise analysis*

In this use case, the role of NA2 has been of providing the seismological feedback during the developments of the workflow on the OGSA-DAI platform using the DISPEL language for the data-intensive workflow. The prototype developed by WP7/SA3 demonstrates how a workflow defining a preprocessing pipeline of seismic traces can be implemented with the DISPEL language and how it translates into the execution of the scientists' analysis code within a distributed deployment of the platform on the DEP-UEDIN-01, the data Intensive machine in Edinburgh. This represents a first prototypal module of the data-intensive workflow enacted upon the VERCE platform and it became of particular importance since it fostered discussion between the seismologists and the ITs during the Project meeting held in Liverpool at the beginning of September.

4 Contribution to WP3/NA3

The NA2 has contributed by lecturing on seismology to the IT experts during the WP3/NA3 training meeting held in Liverpool. This contribution appeared to have been of relevance since it attempted to provide to the non-seismologists the motivations for developing the VERCE platform for data- and compute-intensive computations.

5 Participation to other activities

- During the 3^{rd} workshop of the training network QUEST VERCE was introduced to an audience of approx. 100 international seismologists. A discussion followed on how the VERCE platform could best help the community.
- During the 1^{st} EPOS-ORFEUS Coordination Meeting "Global challenges for seismological data analysis" that was held in Erice (Italy) from May 25 to May 30, 2012, VERCE was one of the organisers of the event and provided funding. The project itself was presented in its various parts (i.e., two seminars - the first on the project and the second on the VERCE architecture) to an audience of approx. 80 international scientists from seismology and ICT. A discussion followed on how the VERCE platform could best help the community and be part of the EPOS e-infrastructure.
- A two-day meeting was held in Rome in April (11-12) and had the specific goals of complementing the efforts (and the desiderata) of the seismologists and the IT by prioritising the implementation of the use cases on the VERCE platform on the implementation of the use cases.
- A two-day meeting with a specific focus on the implementation of the use cases prioritized by NA2 was held in Munich, June 24th and 25th, 2012. The meeting led to the definition of two task forces designed to optimize the implementation of cpu-intensive and data-intensive use cases.

6 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

- the initial implementation of the entire data-intensive workflow by exploiting the OGSA-DAI 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. This process is expected to amalgam further as the implementation of the use cases on the VERCE platform stregthens in the incoming months.

7 Glossary and Links

Definition	Description
ADMIRE	Architectures for Data Intensive Research - http://www.admire-project.eu/
ArcLink	A protocol for data transfer from geographically distribute data archives based on time windows - http://www.seiscomp3.org/wiki/doc/applications/arclink
AXISEM	A parallel spectral-element method - http://www.seg.ethz.ch/software/axisem
CINECA	Consorzio Inter-universitario Cineca
compute-intensive	see cpu-intensive
cpu-intensive applications	Compute-intensive applications are those that devote most of their execution time to computational requirements and typically require small volumes of data although they can produce very large to huge data volumes. Compute-intensive is a term that applies to any computer application that demands a lot of computation, such as forward modeling programs for seismic wave propagation or other scientific applications.
cross-correlation	In signal processing, cross-correlation is a measure of similarity of two waveforms as a function of a time-lag applied to one of them.
data archive	The long-term storage of scientific data and methods.
data mining	The process of automatically extracting patterns from data using techniques such as classification, association rule mining and clustering.
data-intensive	An adjectival phrase that denotes that the item to which it is applied requires attention to the properties of data and to the ways in which data are handled.
DoW	Description of Work
e-Infrastructure	The ICT element of a research infrastructure, i.e. a distributed collection of data, storage and compute resources, interconnected by digital communications and organised to serve a common research purpose. It includes the hardware, software, middleware, staff, operational procedures and policies needed to make it operate for that purpose, and requires maintenance to function in the evolving digital environment and to meet the changing needs of its user communities.
Earth Model	Assumed one to three dimensional parameter sets of the earthÕs interior on which a simulation is based.
EIDA	European Integrated Data Archives infrastructure - http://www.verce.eu//ITCoordinationMeetingFebruary2012/EIDA-Overview.pdf

EPOS	"“European Plate Observing System” is an ESFRI approved infrastructure currently in its preparatory phase and funded by the EC (http://www.epos-eu.org). "
Forward Simulation	Simulation of seismic wave-propagation, results in synthetic seismograms.
Full-Waveform Inversion	Tomographic inversion of the real seismograms (or differences between real and synthetic seismograms) to determine the underlying earth model.
gateway	A software subsystem, typically at the middleware level, that accepts requests for computational and data-handling tasks. It vets those requests to establish whether they are valid, e.g. are syntactically and semantically consistent, and are authorised. Requests that are not validated are rejected. Requests that are accepted are passed to other software systems, at the same or other locations, for execution. The gateway may partition and translate requests in order to combine heterogeneous services.
GPU	Graphics Processing Unit
grid	A system that is concerned with the integration, virtualisation, and management of services and resources in a distributed, heterogeneous environment that supports collections of users and resources (virtual organisations) across traditional administrative and organisational domains (real organisations).
GridFTP	Grid File Transfer Protocol, an extension of the standard FTP for use with grid computing.
INGV	Istituto Nazionale di Geofisica e Vulcanologia
IRIS	Incorporated Research Institutions for Seismology (Data-Center)
globalCMT	Global Centroid-Moment-Tensor Project
metadata	Data that describes data. Metadata may include references to schemas, provenance, and information quality. In Seismology, metadata may also refer to data required in order to sanitise a seismograph's response.
miniSEED	The miniSEED format is a subformat of the commonly used SEED data format used for archiving seismological data.
NA	Network activities
NERA	Network of European RI for Earthquake Risk Assessmet and Mitigation. EC I3 project, www.nera-eu.org
NERIES	Network of RI for European seismology. EC I3 project ended 2010 www.neries-eu.org
ObsPy	A Python framework for processing seismological data. http://obspy.org/
OGSA	Open Grid Services Architecture supported by Globus. - http://www.globus.org/ogsa

OGSA-DAI	Open Grid Service Architecture Data Access and Integration, an open source product for distributed data access and management.
ontology	In computer science, a formal explicit specification of a shared conceptualisation.
ORFEUS	Observatories and Research Facilities for European Seismology.www.orfeus-eu.org
PID	Persistent Identifier : A persistent identifier is a permanent, location- independent and globally unique identifier for a resource. Persistent identifiers are generally assigned by agencies who undertake to provide reliable, long-term access to resources. Examples of persistent identifiers include Digital Object Identifiers, Uniform Resource Names, Handles and Archival Resource Keys.
Pilot application	main software routine within a use case (e.g., the cross-correlation analysis in the use case addressing the velocity variations of the Italian peninsula crust properties).
portal	In the context of knowledge discovery, a tool designed for a particular group of domain experts that can be used via their browsers; it enables them to establish their identity and rights, and to pursue conveniently a set of research tasks for which the portal is designed.
pre-processing	One or operations performed on the observed data to prepare the latter for the analysis an/or for performing quality control checks.
QUEST	QUantitative Estimation of Earth's Seismic Sources and Structure
RAPID	Rapid portals for Seismological Waveform Data - http://research.nesc.ac.uk/node/423
RapidSeis	Portal for interactively running C++ scripts on seismological waveform data Not yet ready for Python.
registry	A persistent store of definitions and descriptions of data or software components and their relationships accessed by tools and other elements of a distributed research environment. It is intended to facilitate discovery and use of the components.
repository	A store holding software definitions, other shared code and data, that supports distributed concurrent access, update and version management.
Research Infrastructure	The collection of equipment, resources, organisations, policies and community support that enables a particular discipline to conduct research. Normally, this refers to the advanced facilities that enable frontier research, such as the research infrastructures endorsed by ESFRI.
SAC	http://www.iris.edu/software/sac

science gateway	A consistently presented set of facilities designed to be a convenient working environment for researchers in a particular domain, in this case seismology. It should bring together access to all of the capabilities and resources such a researcher needs: including catalogues of available data and tools, established methods and arrangements for applying them with specified parameters to specified data.
SEED, mSEED, SAC	Standard seismic data formats
SeisSol	A simulation software based on the Discontinuous Galerkin Finite Element Method - http://www.geophysik.uni-muenchen.de/ kaeser/SeisSol/
SEM	Spectral Element Method wave propagation
Shibboleth	Standards based, open source software package for web single sign-on across or within organizational boundaries - http://www.shibboleth.net
SPECFEM3D	A simulation software code based on the spectral-element method for 3D seismic wave propagation - http://www.seg.ethz.ch/software/specfem3D
Synthetic Seismograms	Waveform(time series) calculated in a computer simulation (size of data depends on duration and sampling rate, also meta-data). It is dependent on the solver, the computational grid(mesh), the earth model, the event parameters, and the location of "observation".
UEDIN	The University of Edinburgh
Use case	In software and systems engineering, a use case is a list of steps, typically defining interactions between a role and a system, to achieve a goal. The actor can be a human or an external system (cf http://en.wikipedia.org/wiki/Use_case). In VERCE it is assumed to represent the entire scientific application (e.g., analysis of the noise cross-correlation of the Italian seismic networks for 6 years period to detect temporal variations of the Crust material properties)
VERCE architecture	A high-level and coherent design for the VERCE e-Infrastructure; it evolves as the seismological goals and digital environment evolve and become better understood. It should guide the development of successive VERCE platforms.
VERCE e-Infrastructure	An envisaged result of VERCE, as an integrated computational and data environment that presents a coherent virtual research environment in which to conduct seismology research and eventually research in other Earth sciences.

VERCE Platform	The current realisation of the VERCE e-Infrastructure at any time in the VERCE project. Initially this is not fully integrated and may only constitute a partial implementation. Nevertheless, it is sufficient both to pursue research identified as priority seismology use cases and to develop and test the design of the VERCE e-Infrastructure. The VERCE platform is an approximation to the VERCE e-Infrastructure. These approximations should converge on the VERCE e-Infrastructure by the end of the VERCE project.
W3C	World Wide Web Consortium, an international community of member organisations and the public that works to define and promote standards for web technologies.
web service	A software system designed to support interoperable machine- or application-oriented interaction over a network.
workflow	A process of composed data-handling tasks, computational tasks and human interactions intended to implement a research method or established working practice.
WP	Work Package
WP1	NA1
WP2	NA2
WP3	NA3
WP4	NA4
WP5	SA1
WP6	SA2
WP7	SA3
WP8	JRA1
WP9	JRA2
XML	Extensible Markup Language.
XSEDE	Extreme Science and Engineering Discovery Environment - https://www.xsede.org/
DISPEL	Verce Workflow Enactment Engine