Geophysical Research Abstracts
Vol. 15, EGU2013-4772, 2013
EGU General Assembly 2013
© Author(s) 2013. CC Attribution 3.0 License.

Supporting Large-Scale Data-Intensive Computation for Seismology in VERCE

Iraklis Klampanos (1), Alessandro Spinuso (2), Luca Trani (2), Amy Krause (1), Visakh Muraleedharan (3), Celine Hadziioannou (4), Gaia Soldati (5), Licia Faenza (5), Lucia Zaccarelli (5), Peter Danecek (5), Xavier Briand (6), Alberto Michelini (5), Malcolm Atkinson (1), and Jean-Pierre Vilotte (3)

(1) University of Edinburgh, UK, (2) Koninklijk Nederlands Meteorologisch Instituut, The Netherlands, (3) Institut de Physique du Globe de Paris, France, (4) Ludwig-Maximilians-Universitat Munchen. Germany, (5) Istituto Nazionale di Geofisica e Vulcanologia, Italy, (6) Universite Joseph Fourier, France

One of the objectives of the VERCE project (*V*irtual *E*arthquake and Seismology *R*esearch *C*ommunity in *E*urope – http://www.verce.eu/) is to provide scientists with a unified, Europe-wide, computing environment able to support data-intensive scientific computation. The term “data-intensive” is used to characterise computation that either requires or generates large volumes of data, or that its data access patterns are complex due to algorithmic or infrastructural reasons. In this work we will present our approach to designing and building the VERCE data-intensive infrastructure.

The work of the modern seismologist typically involves managing, storing and working against large datasets, typically distributed at remote sites, and accessible via different transfer protocols. Performing experiments and scientific analyses against these data generates more data, which in turn have to be managed, further analysed and frequently be made available and shared within or outside scientific communities. At the same time, scientists have access to increasingly powerful computing facilities, often managed away from their workplace, designed to address different needs and computing requirements. In the vast majority of cases, the laborious tasks of data management, transfer and execution of scientific codes is being handled manually by the scientist. This state of affairs is hardly manageable and, more importantly, it hinders seismologists from making full use of the data and tools they have at their disposal for scientific discovery.

In VERCE, we address the aforementioned issues by introducing levels of abstraction suitable to different kinds of experts and by modelling scientific procedures through the Dispel workflow language. Dispel workflows are based on the data-streaming model, allowing for efficient and transparent passing of data through computing blocks called *processing elements* (PEs). File- and metadata-management technologies we experimented with include IRODS (https://www.irods.org) as well as triplet stores. Based on metadata pertaining to raw and derivative data, available processing elements and available computing facilities, the VERCE enactment engine can resolve higher-level requests and direct them to appropriate resources. This, allows the scientist to concentrate on the logic and approach of the scientific analysis, communicate own results or replicate other experiments in a manageable, tractable way.

Dispel PEs can be parameterised and combined to form enactable graphs or composite PEs, etc. The set of PEs available to the scientists form the VERCE library, a distributed and remotely accessible ecosystem which can be broadly divided into *generic* and *domain-specific* elements. VERCE’s seismology PEs are provided by the scientists. VERCE allows for the specification of seismology PEs in languages other than Dispel, such as Python, making use of highly specialised scientific libraries, such as ObsPy (http://obspy.org). Scientists can, therefore, continue working using tools they are familiar and productive with, while at the same time getting access to a Europe-wide infrastructure, in a transparent manner.

In VERCE, the main data-intensive use-case involves accessing and pre-processing raw data from an increasing number of stations, before they use them to cross-correlate seismic noise. The architecture allows for incrementally improving seismology PEs, modifying various stages of the workflow, such as the sequence of pre-processing, the exact pre-processing steps and relevant parameters, as well as it allows scientists to create their own custom workflows in order to describe different experiments. We will show the results of the effort so far, as well as how this approach can be successfully sustained, contributing to community building and ultimately advancing science.