The VERCE Architecture for Data-Intensive Seismology

Iraklis A. Klampanos

iraklis.klampanos@ed.ac.uk

http://www.verce.eu
Overall Objectives

• VERCE: Virtual Earthquake and Seismology Research Community in Europe

• Unified, Europe-wide computing infrastructure

• Support common data- and CPU-intensive tasks

• Exploit local and remote processing and storage resources

• Use-cases:
  • Ambient noise cross-correlation → Data-intensive
  • Earth model adjustment → CPU-intensive
Technical Goals

- Transparent parallelism and maximum use of local resources
  - Fine-grained Dispel streaming workflows
- Automated data handling
  - Preemptive and reactive movement/replication
  - Extensive use of science-specific metadata
- Customisation of workflows
  - Use of community oriented library tools, e.g. ObsPy
Data-Intensive Requirements

- Data-intensive:
  - Size, numbers of files, frequency of movement, etc.
- Data staging, pre-processing and cross-correlation
  - mainly on private or Grid resources
- Arbitrarily many continuous waveforms (typically 24h records from local, regional, global networks)
- Over arbitrarily long periods of time
- Unified view of data sources (files, online feeds)
  - Through metadata (e.g. station, channel, time, etc.)
Architectural Overview

- **Science Gateway**: extensible portal
- **Users, resources, profiles, collaboration**
- **Services Integration**
  - Dispel processing - interpreter, optimiser
  - Execution engine
  - Registry
  - Provenance

**Catalogues and stores**: external scientific and resource description catalogues and registries (e.g. UNICORE)
**Computing resources**: local Dispel/D-I/adhoc, local HPC, external PRACE/HPC, external EGI/Grid, (future) Cloud
Dispel: Fine-Grained Scientific Workflows

- Composition of arbitrarily fine-grained workflows through processing elements and functions/constructors

- Stream-oriented / online processing:
  - Decoupling from lower-level, technical details
  - Composable workflow patterns

- An example seismology processing element may have
  - High-level specification in Dispel
    - Typed I/O streams and parameters
  - Lower-level implementation in ObsPy, or other libraries
  - Tweaking parameters alters behaviour
The VERCE Registry

- Provides consistency across sites
  - w.r.t. Dispel components, resources, data
- Concurrent collaborative workflow development
- Interfaces with external registries and catalogues
- Major registry components
  - Dispel
  - Resources
- User and workload profiling information
Provenance

- Pioneering use of W3C PROV model
- Gathered from remote resources
- Mirrored workflow
- User access through the science gateway
Execution Engine

- OGSA-DAI
  - Distributed data access and management
  - Inherited from the ADMIRE project
- Storm
  - Well designed high-level optimisation defaults
  - Stress tested-used by companies and researchers
- Dispel-to-Storm interpreter
Research Challenge 1: Optimisation

- Dispel semantics - close to users’ cognitive level
- Optimisation levels
  - High-level: workflow partitioning and deployment
  - Low-level: on-site data handling, data and code movement, execution management
- Constant ‘learning’ model making use of information from Registry and Provenance
- Adjust to changes in the environmental context
Cross-Correlation Optimisation example

User’s view

Data Gathering → Pre → X-Corr. → Post → Data Storage

In: 3 stations time period
Out: [waveform]

In: waveform
Out: waveform

In: [waveform]
Out: waveform

In: waveform
Out: waveform

In: waveform
Out: N/A

High-level optimisation

Data Gathering → Pre → X-Corr. → Post → Data Storage

In: 3 stations time period
Out: [waveform]

In: waveform
Out: waveform

In: waveform
Out: waveform

In: waveform
Out: N/A
Research Challenge 2: Registry Design and Evaluation

- Registry’s role central for
  - informing and supporting users
  - being a good citizen in a community of services
  - enabling efficient operation
- Distributed Vs. Centralised Vs. Hybrid design
- Balance between replication and federation
- Prioritisation of features/metadata
- Models for data and information retrieval
- Multiple evaluation angles