Virtual Earthquake and seismology Research Community e-science environment in Europe
Project 283543 – FP7-INFRASTRUCTURES-2011-2 – www.verce.eu – info@verce.eu

SEVENTH FRAMEWORK PROGRAMME

VERCE
« As a « Data Management Use Case

Horst Schwichtenberg

EGI Technical Forum

Prag, 2012

VERCE

Content

- VERCE project
- VERCE platform for data intensive applications
- Seismology :
  - Data center
  - Use Cases / Applications
- Open Questions

VERCE : Virtual Earthquake and Seismology Research Community e-science environment in Europe

1	**CNRS-INSU**	Centre National de la Recherche Scientifique, France
2	**UEDIN**	University of Edinburgh, Scotland, United Kingdom
3	**KNMI-ORFEUS**	Royal Netherlands Meteorological Institute, Netherlands
4	**EMSC**	European-Mediterranean Seismological Centre , France
5	**INGV**	Istituto Nazionale di Geofisica e Vulcanologia, Italy
6	**LMU**	Ludwig-Maximilians-Universität , Germany
7	**ULIV**	University of Liverpool , England, United Kingdom
8	**BADW-LRZ**	Bayerische Akademie der Wissenschaften, Germany
9	**SCAI**	Fraunhofer-Gesellschaft e.V., Germany
10	**CINECA**	Centro di Calcolo Interuniversitario, Italy

Towards an e-Science environment for seismology and EPOS

- Provide a data intensive service-oriented e-Science environment to the EPOS community
- Lay the basis for transformative data-intensive research in the solid earth sciences
- Build trust and collaborative models for sharing of data , methods and tools
- Engage a new generation of researchers and experts in solid earth data intensive research

European and International domain context

Integrated European distributed Data Archives (EIDA), part of the international FDSN

A number of coordinated European projects in seismology: NERA, SHARE, GEM, ERC WHISPER, ITN QUEST…

The European Plate Boundary Observation System (EPOS): the ESFRI-PP project

Active collaborations within the FDSN with US (IRIS-DMC), and Japan (JAMSTEX, NIED)

European e-Science context

Fast evolution of seismology services and applications

RapidSeis Portal for Accessing & Processing FDSN archives

European initiatives: EPOS, ENVRI and EUDAT

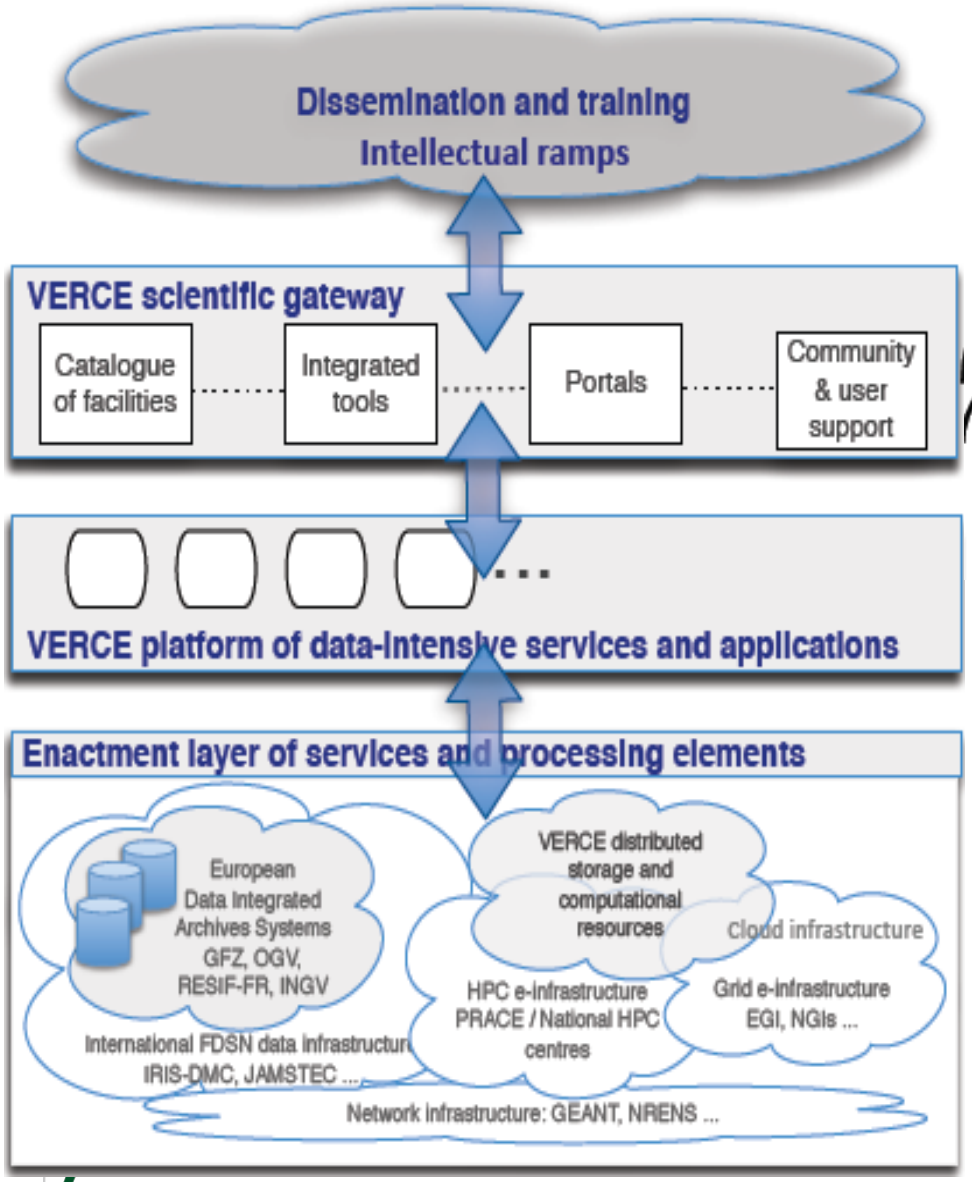
Converging e-Infrastructure ecosystem: EGI/NGIs, PRACE/NHPCs, GÉANT

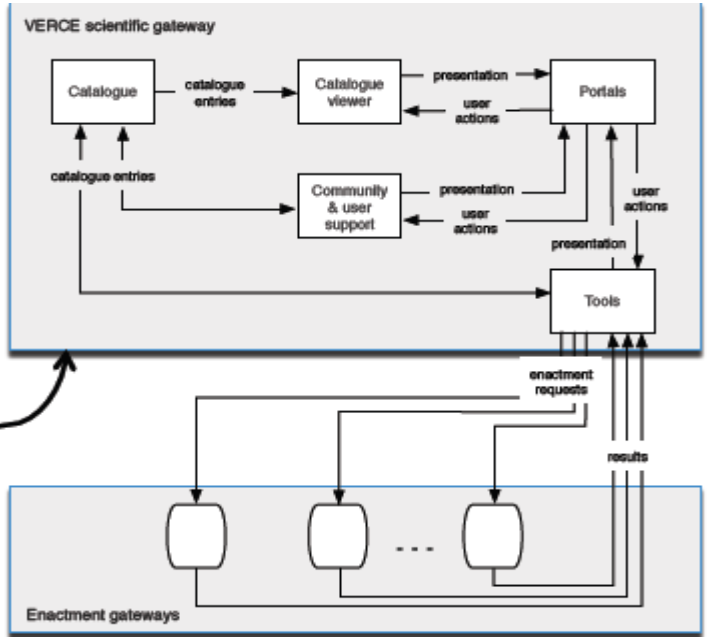
Emerging new data base management system and data centric architecture

Emergent data access and single-sign on protocols

A seismology architecture for data intensive applications: data analysis, mining and modelisation

Sharing with other disciplines: Astronomy & Astrophysics, Particle Physics, Biology


Dissemination and training
Intellectual ramps
VERCE scientific gateway
Catalogue of facilities
Integrated tools
Portals
Community & user support
VERCE platform of data-intensive services and applications
Enactment layer of services and processing elements
European Data Integrated Archives Systems GFZ, OGV, RESIF-FR, INGV
International FDSN data infrastructure IRIS-DMC, JAMSTEC ...
VERCE distributed storage and computational resources
HPC e-infrastructure PRACE / National HPC centres
Cloud infrastructure
Grid e-infrastructure EGI, NGIs ...
Network infrastructure: GEANT, NRENS ...


VERCE scientific gateway
Catalogue
catalogue entries
Catalogue viewer
presentation
user actions
Portals
catalogue entries
Community & user support
presentation
user actions
user actions
presentation
Tools
enactment requests
results
Enactment gateways

www.verce.eu

Initial Resources

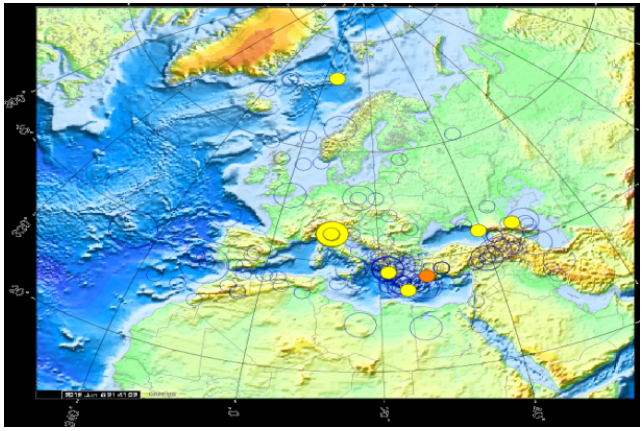
- Compute

  Public:

  - PRACE (HPC) sites: LRZ, CINECA
  - EGI-Infrastructure (GRID): ESR VO in EGI-Inspire, VERCE VO

  Private:

  - Department resources: UEDIN, ULIV, IPGP, SCAI
- Data Center:
  - KNMI/ORFEUS
  - IPGP
  - INGV
- Storage:
  - UEDIN
  - ULIV



Orfeus: Seismic monitor

Initial Software Components

- Components for Secure Access to resources:
  - *Different access methods in use*: from standard (gsi)SSH to EUGridPMA X.509 Certificate

    Challenge: No federated identity management available across the European e-infrastructures

- VERCE relevant data management tools:
  - different data management tools and
  - different technologies/protocols

    E.g. OGSA-DAI (see ADMIRE), IRODS, SRM, Arclink, GridFTP

- Job Management tools on public and private resources:
  - E.g. LSF, Torque on Clusters; gLite CREAM/WMS on Grid

- Seismic and seismological software
  - E.g. ObsPy, rdseed, seissol, soc2D, specfem3d,axisem

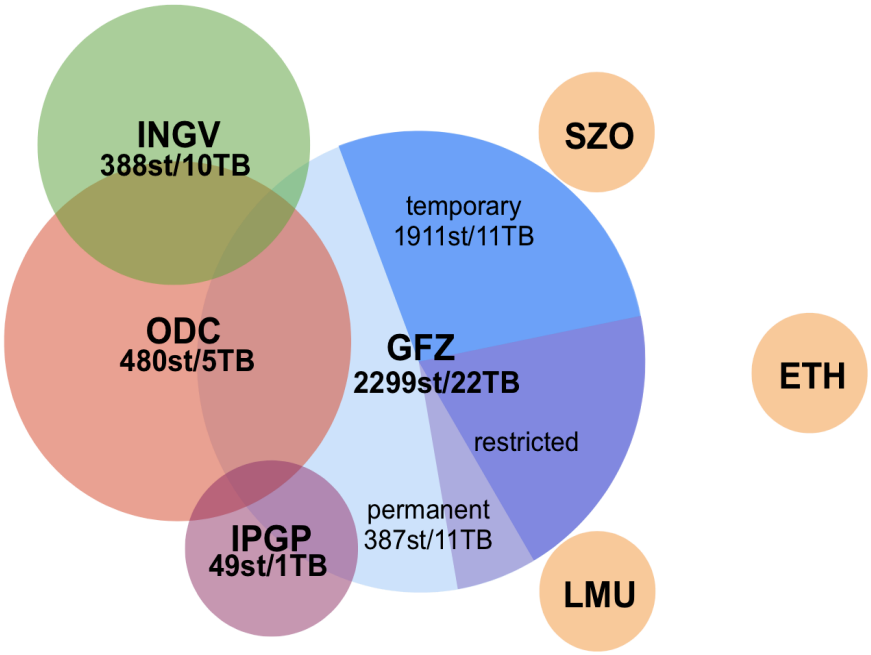
- First Components of the initial VERCE data intensive architecture
  - E.g. OGSA-DAI, ADMIRE/VERCE-DISPEL Workflow

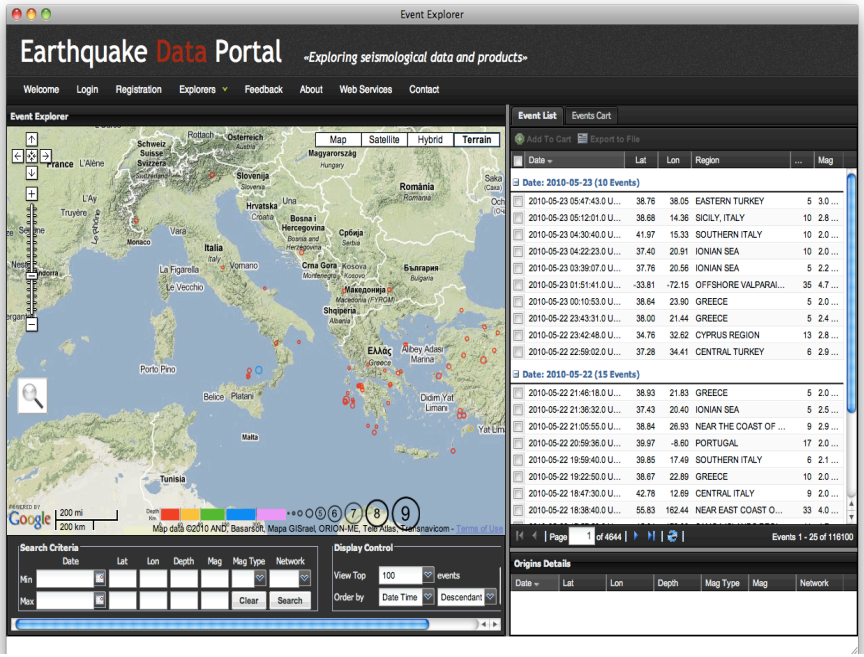
Seismological data archived at INGV

- Time series acquired at seismic stations
- Each station features a three component seismometer (3C) – vertical, N-S, E-W and a data logger having an A/D converter
- Digital acquisition at 100 samples per second using a 24 bit (3 Byte) per channel
- The data are usually compressed in SEED format - after compression and for data acquired in quite periods each sample is about 1.3 Byte. If an earthquake is recorded the compression is less to much less
- Thus for each station
  - 100 * 1.3 * (60*60 * 24) = 11.232 MB per day per channel
- INGV network consists of **~300 3C stations**
  - 11.232 * 300 * 3 = 10.11 GB per day
  - 10.11 * 365 = **~3.7 TB per year**
- Proper archiving started at INGV in 2005-2006 and right now the data set archived is **~26 TB**

European Integrated Data Archive

Present EIDA Architecture



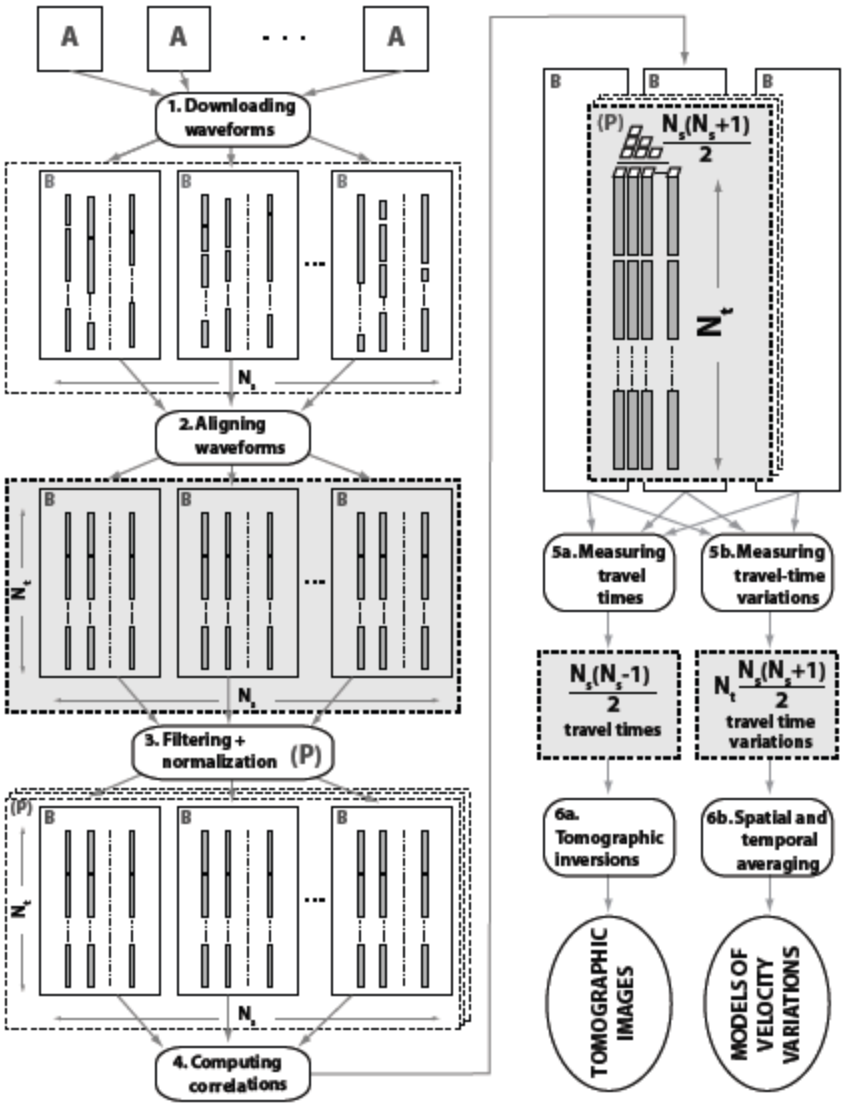


www.seismicportal.eu

Main Use Cases/Applications

1. Forward Modelling and Inversion (LMU).
2. Xspect: cross-spectrum analysis on noise cross-correlations (INGV).
3. High Resolution Tomography from 3D full waveform inversion in Italy (INGV).
4. TsuMaps: near real-time forecasting of tsunami wave height (INGV).
5. Automatic detection and High Resolution Location of Italian Seismicity (INGV, EOST).
6. L'Aquila 2009 quake: crustal velocity variation by means of seismic noise cross correlation (INGV).
7. Noise cross-correlations at the Valhall field (IPGP).
8. Automatic high resolution location of Maule aftershocks (ULIV).
9. Velocity and velocity changes of Japan: the Namazu project (IPGP/ISTerre).

Data Intensive analysis



Distributed Data Mining:

1. **Distributed Mining** of Data
2. Mining of **Distributed Data**

Data Storage and I/O bandwidth:

- Data life cycles and replication
- Fast sequential I/O

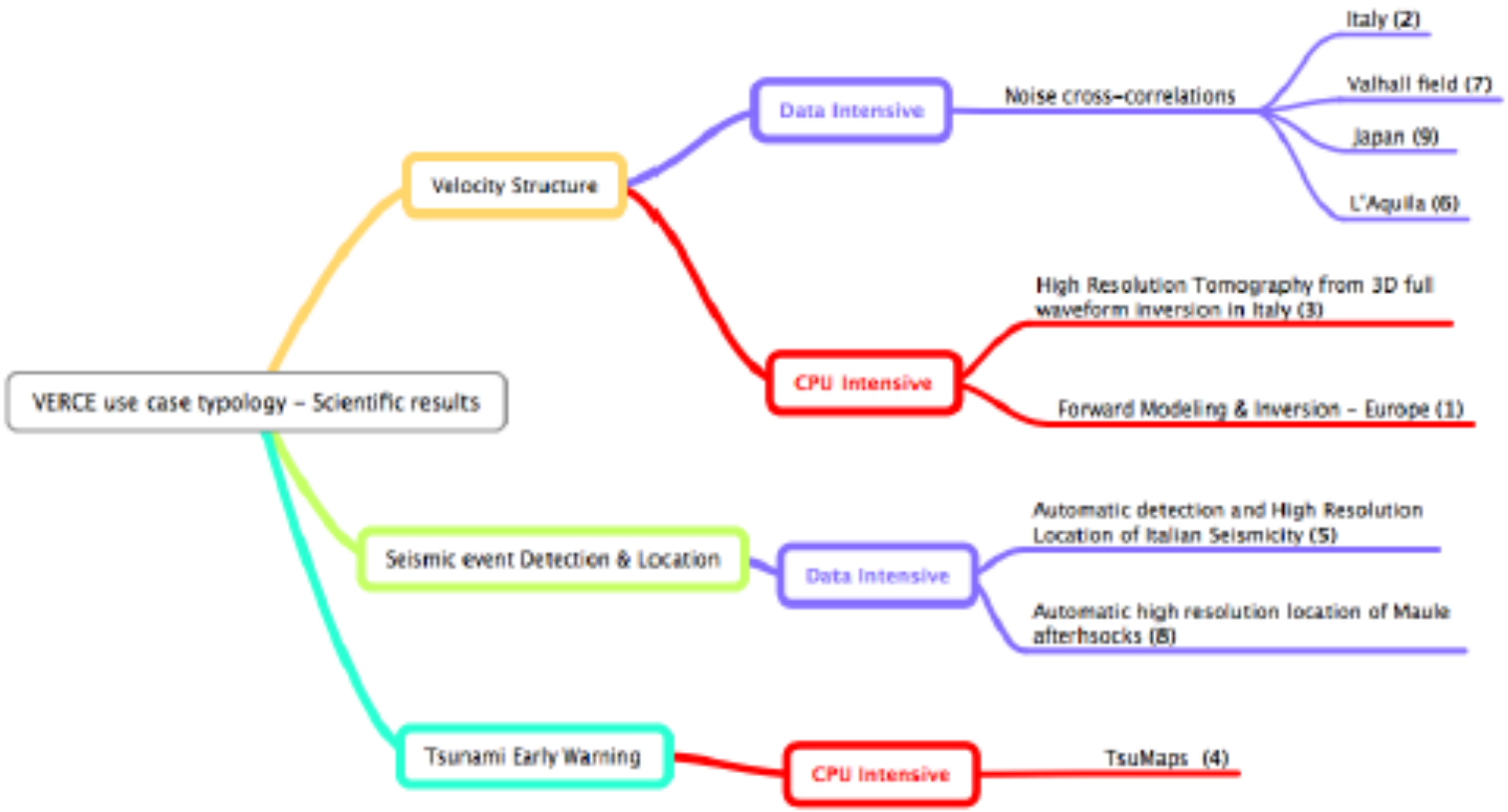
Software development:

1. **Reusable libraries**
2. Workflows: **Interactions/ Traceability**

Data and Infrastructures policies:

Explore new data-intensive paradigms enabled by several technologies (workflow engines, Map/Reduce, GPU)

VERCE USE CASES



Defining "Data-Intensive" by Malcolm Atkinson

- Generally
  - A computational task is data-intensive if you have to think hard about an aspect of data handling to make progress
    - distribution, permissions and rules of use, complexity, heterogeneity, rate of arrival, unstructured or changing structure, long tail of small and scattered instances, size of data, number of users
    - invariably in combination
- Quantitatively
  - The computation's Amdahl numbers are close to 1
    - CPU operations : bits transferred in or out of memory
    - 1000 CPU operations : 1 I/O operation
  - Total volumes expensive to store
  - Total requests/unit time hard to accommodate
  - Data transport too slow or expensive

Verce Platform

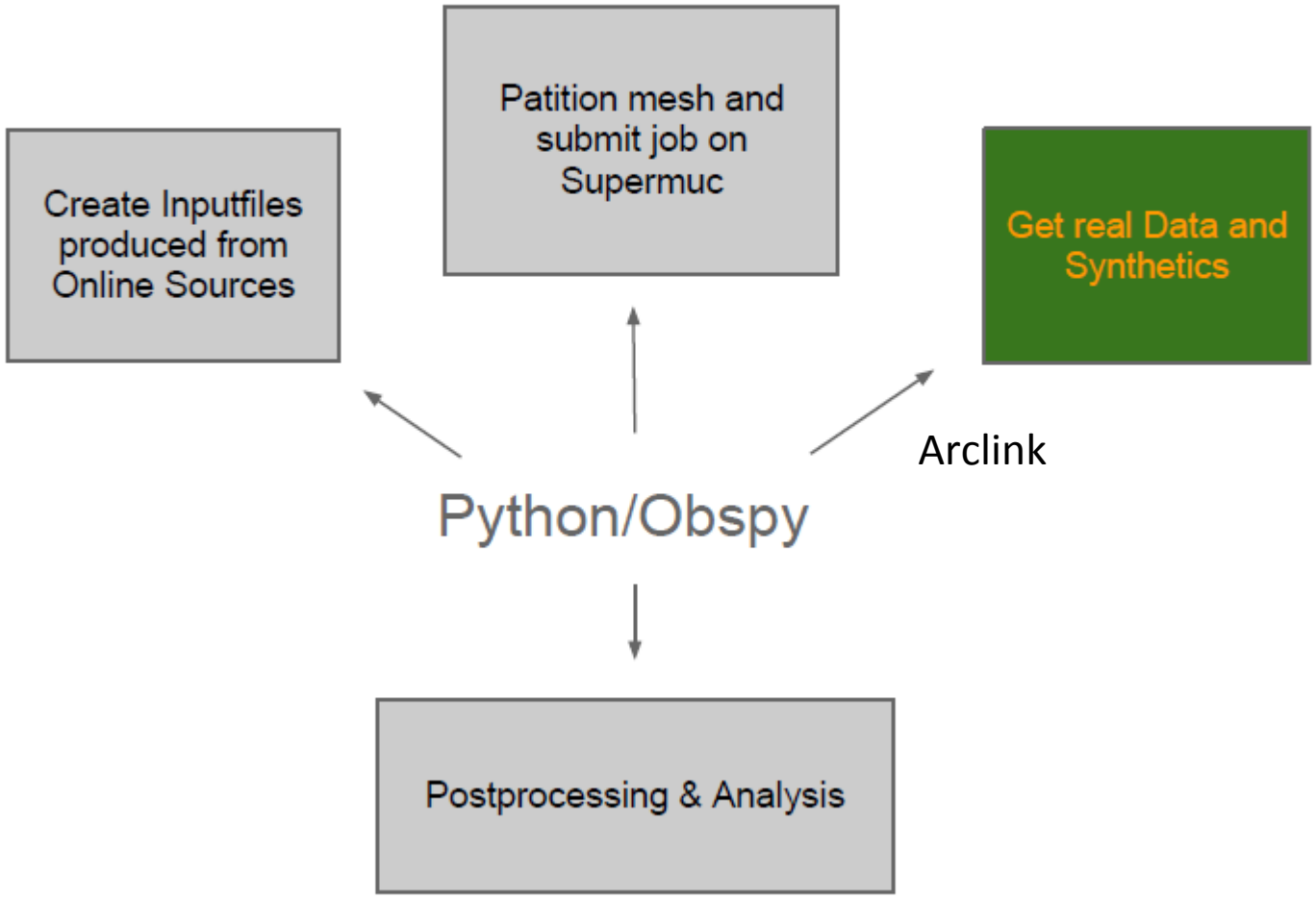
- Implement different seismological workflows on-top existing e-Infrastructures
- Different interfaces available to implement:
  - Simple workflows -> Python based Obspy
  - Distributed complex workflows -> DISPEL Gateways

Workflow driven by Data

Verce Platform

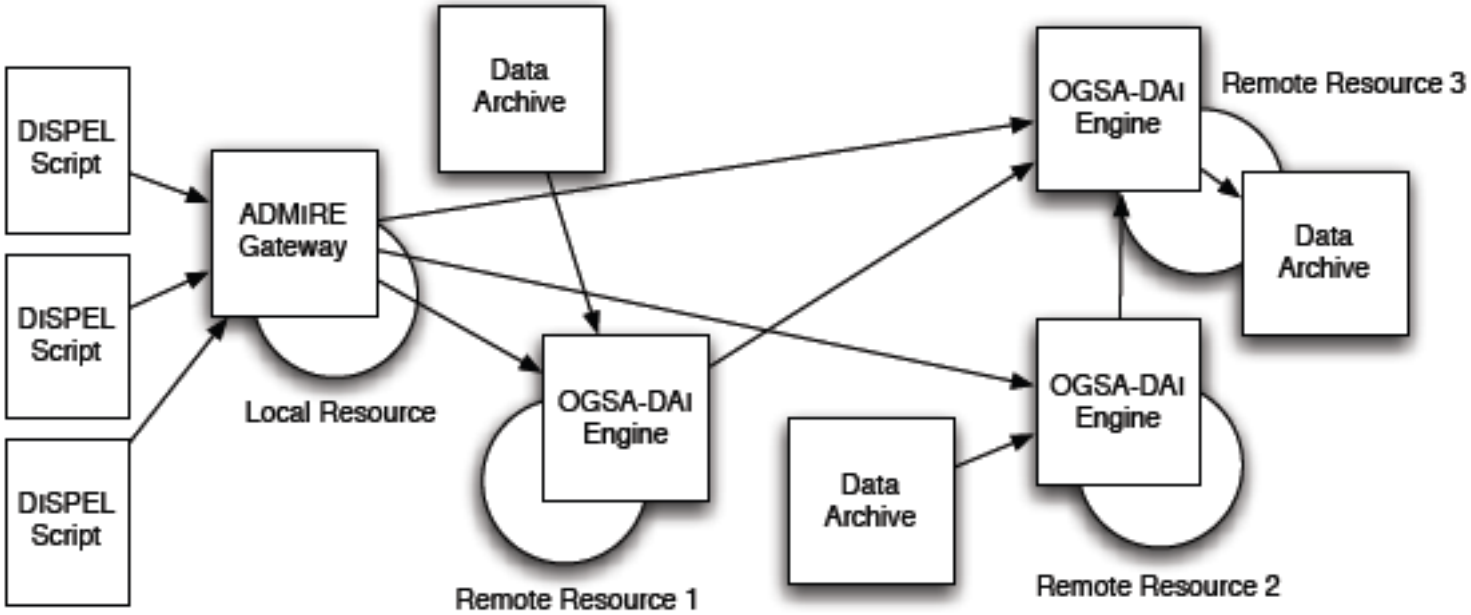
- VERCE Workbench :
  - Seismologists may use graphical Interface
  - Typical workflows or sequences stored in repositories
  - Data Management is completely hidden
  - Execution is recorded - Provenance

Data- and Compute Intensive

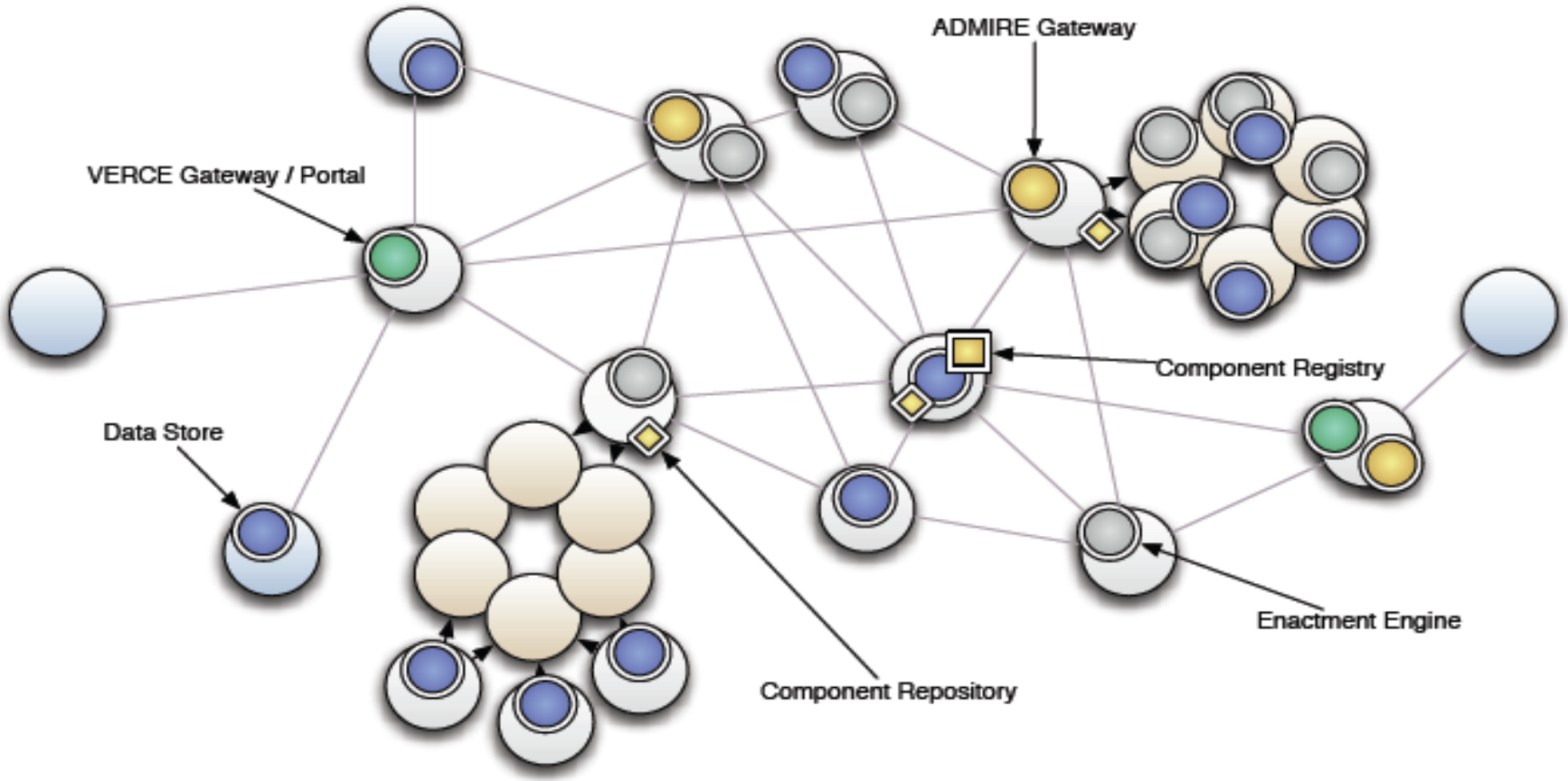


VERCE Workflow Architecture

Design Workflows by DISPEL
„Enactment Gateways are Service Provider (Expansion of Patterns)
Interpreting and Executing Workflows written in DISPEL“



Workflow enactment



VERCE WORKFLOW
DISPEL

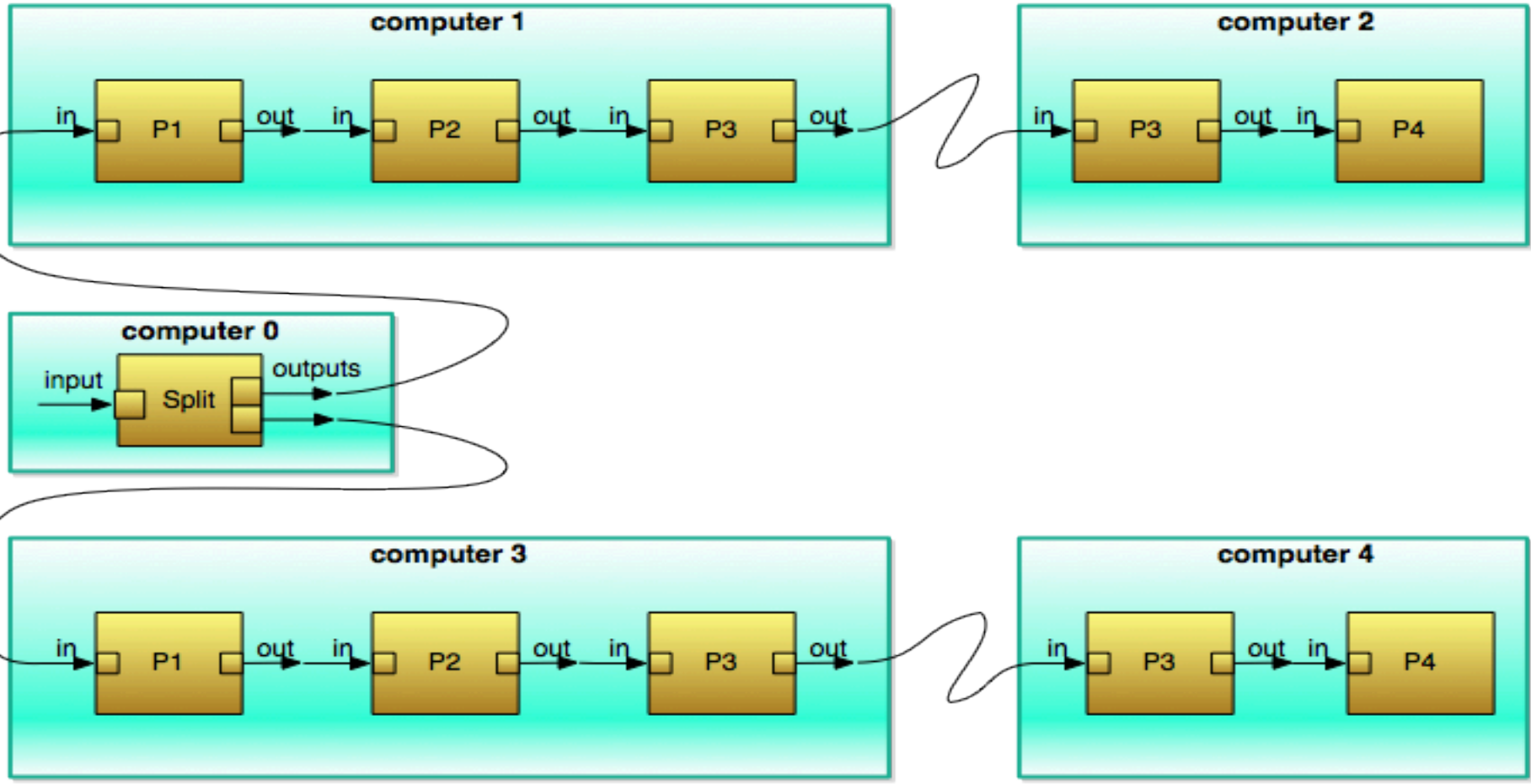
```
1   package tutorial.example {
2       // Import existing PE from the registry.
3       use dispel.db.SQLQuery;
4
5       // Define new PE type.
6       Type SQLToTupleList is PE( <Connection expression> => <Connection data> );
7
8       // Define new PE constructor.
9       PE<SQLToTupleList> lockSQLDataSource(String dataSource) {
10          SQLQuery query = new SQLQuery;
11          |-repeat enough of dataSource-| => query.resource;
12          return PE( <Connection expression = query.expression> =>
13                     <Connection data = query.data> );
14      }
15
16      // Create new PEs.
17      PE<SQLToTupleList> TutorialQuery = lockSQLDataSource("uk.org.UoE.dbA");
18      PE<SQLToTupleList> MirrorQuery   = lockSQLDataSource("uk.org.UoE.dbB");
19
20      // Register new entities.
21      register TutorialQuery, MirrorQuery;
22  }
```

Data-Intensive Process Engineering Language

- A language for constructing data-flow graphs
  - Nodes are processing elements
  - Arcs are data-flow paths
- A language for generating data-flow patterns
  - Functions hide detail of graphs
  - Functions generate graphs
- A language for discussing data-flow engineering
  - Designed to be read and written by humans
  - As well as by programs
  - Supports validation and optimisation

designed to encourage data-intensive thinking

Parallel Streams



VERCE Data Management
Old „open“ Questions

- How can we get data in and out of HPC resources
- How can we access seismological data
  from a HPC resource
- How to shuffle data between GRID (EGI) to HPC (PRACE)
- Large datasets may be transfered by „sneaker.net“
  But where to cache the data for short time
- Which are the best data transfer protocols or solutions?
- Is Globus Online a solution for us?
- Permanent storage for seismic and meta data?

Verce Data Management
Open Questions

- Integration of Data Center Policies
- Encouraging Data Centers to install anonymous GridFTP servers to replace their existing anonymous FTP servers

- How to manage „AA“ - still no European federated identity management available

- Workflow layers on top of e-Infrastructures:
  - No support today:

    Are „Event-Gateways“ possible close to resource providers (e.g. HPC centers)

Ultimate aim

- Enable European seismologists to work on the existing European e-Infrastructure seamlessly independent of whether it is a Grid, HPC or department resource!

Contacts

- "Jean-Pierre Vilotte" vilotte@ipgp.fr
  (coordinator)
- "Malcolm Atkinson" mpa@staffmail.ed.ac.uk

  VERCE Architecure (DISPEL, ADMIRE/VERCE Gateway)

- Transparencies by: Marek Simon (LMU), Malcolm Atkinson (UEDIN), Alberto Michelini (INGV), Siew Hoon Leong (LRZ)

horst.schwichtenberg@scai.fraunhofer.de

Thank you … more to come !