The Terracorrelator: a shared memory HPC facility for real-time seismological cross-correlation analyses

Malcolm Atkinson , Andrew Bell , Andrew Curtis , Elizabeth Entwistle , Rosa Filguiera , Amrey Krause, Ian Main , Giovani Meles , Mike Mineter , Sjoerd de Ridder, and Youqian Zhao

Roadmap

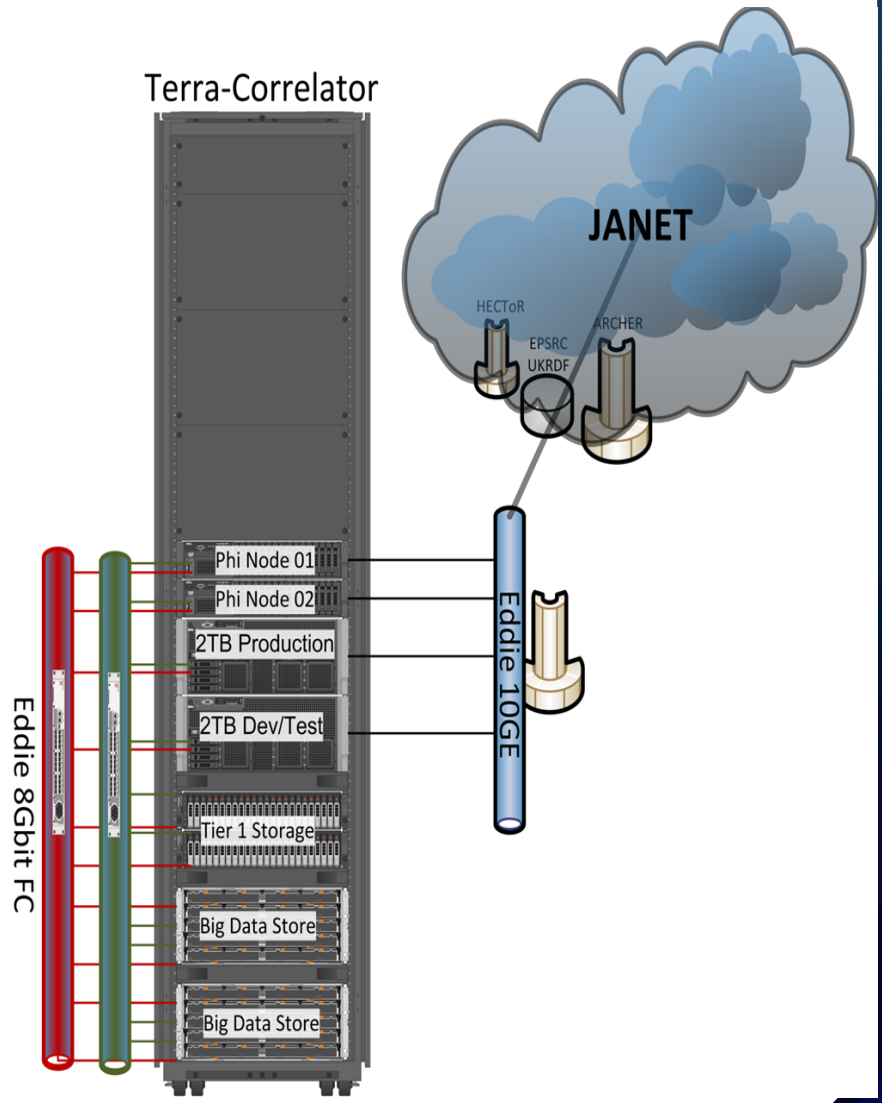
- Introduction -Motivation
- Terracorrelator Machine
- Seismic interferometry case study
- dispel4py
- Seismic ambient noise cross-correlation
- Performance Experiments
- Near real-time continuous correlations of two stations
- Conclusions and future work

Introduction - Motivation

- The development of reliable risk assessment methods for hazards events requires real-time analysis of seismic data.
- Earthquake ``repeater'' analysis, require a large number of waveform cross-correlations:
  - Computationally intensive
  - Challenging in real-time.
- We need:
  - Hardware and Software capable to accept those challenges

Terracorrelator Machine (TC)

- New HPC facility
- Funded by NERC and deployed for research in the GeoSciences during March 2014
- Designed for real-time cross-correlational analyses.
- 2 nodes with 2TB shared memory and 32 cores.
  - cross-correlation
  - post-processing
- 2 Intel Xeon Phi nodes
  - pre-processing



The University of Edinburgh
GeoSciences
www.geos.ed.ac.uk/masters

Seismic interferometry case study

- Seismic ambient noise cross-correlation
  - Goal: Preprocess + Cross-correlate 1000 stations in hourly intervals by using TC machine.
- Libraries used:
  - **ObsPy** for downloading data, seismic operations and processing
  - **Dispel4Py** for writing and executing the workflow in parallel.

What is dispel4py ?

- New open source user-friendly tool
- Develop scientific methods and applications on local machines
- Run them at scale on a wide range of computing resources without making changes

Documentation
http://dispel4py.org
Source code
https://github.com/dispel4py/dispel4py
Installation
pip install dispel4py

The University of Edinburgh GeoSciences www.geos.ed.ac.uk/masters

edinburgh data-intensive research

epcc

Main dispel4py features

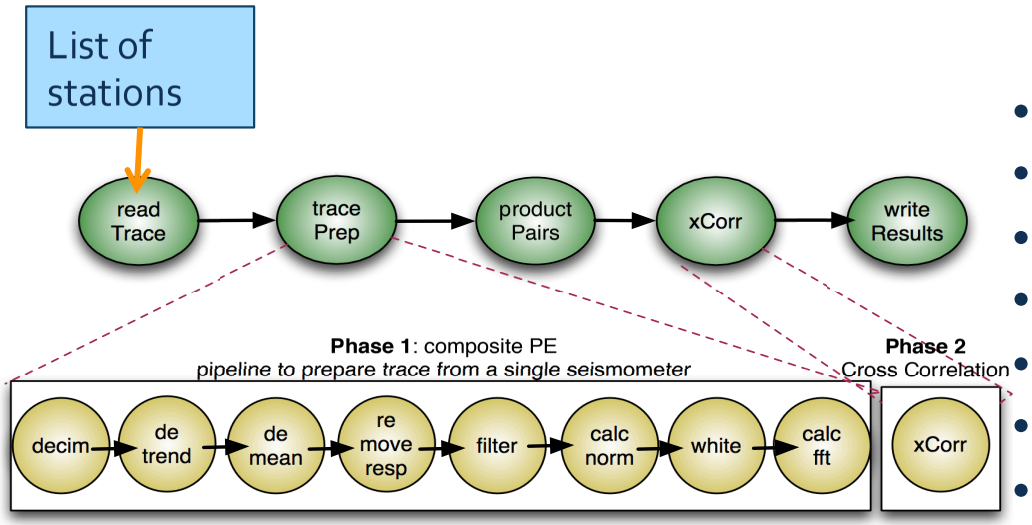
- **Stream-based**
  - Tasks are connected by streams and not by intermediate files
  - Multiple streams in & out
  - Optimization based on avoiding IO
- **Maps workflows dynamically** onto multiple enactment systems:
  - Automatic parallelization
  - Without cost to users
- **Python** language for describing tasks and connections

Main dispel4py features

- **Sequential**
  - Sequential mapping for local testing
  - Ideal for local resources: Laptops and Desktops
- **Multiprocessing**
  - Python's multiprocessing library
  - Ideal for shared memory resources
- **MPI**
  - Distributed Memory, message-passing parallel programming model
  - Ideal for HPC clusters
- **STORM**
  - Distributed Real-Time computation System
  - Fault-tolerant and scalable

Seismic ambient noise cross-correlation – dispel4py workflow + TC + multi mapping

- Phase 1- Preprocess: Time series data (traces) from seismic stations are preprocessed in parallel
- Phase 2: Cross-Correlation: Pairs all of the stations and calculates the cross-correlation for each pair (complexity O(n2)).



- Select vertical component (channel BHZ)
- Decimation to 4 Hz sampling rate
- Remove mean & trend
- Remove station response
- Bandpass filter for 0.01 Hz - 1.00 Hz
- Moving average re-normalization
- Spectral whitening

The University of Edinburgh GeoSciences www.geos.ed.ac.uk/masters

edinburgh data-intensive research

epcc

Performance Experiments - 1

- Stream every hour the available data form IRIS services:
  - All Operating USArray TA Stations (_US-TA) . 836 stations.
  - **Number of stations with data available**: Only **394 stations.**
- **Preprocess**: Stream data by requesting 10 stations each time:
  - Stream time **10 stations**: **12 seconds**.
  - Streaming time **40 sets of 10 stations**: **480 seconds**.
  - Workflow requests and preprocesses data in parallel
  - **Preprocess time**: 491 seconds (**aprox 8 minutes**)
- **Xcorr time**: 180 seconds (**aprox 3 min**).
- **Number xcorr = 77421**
- **Total time = Aprox 11 minutes**

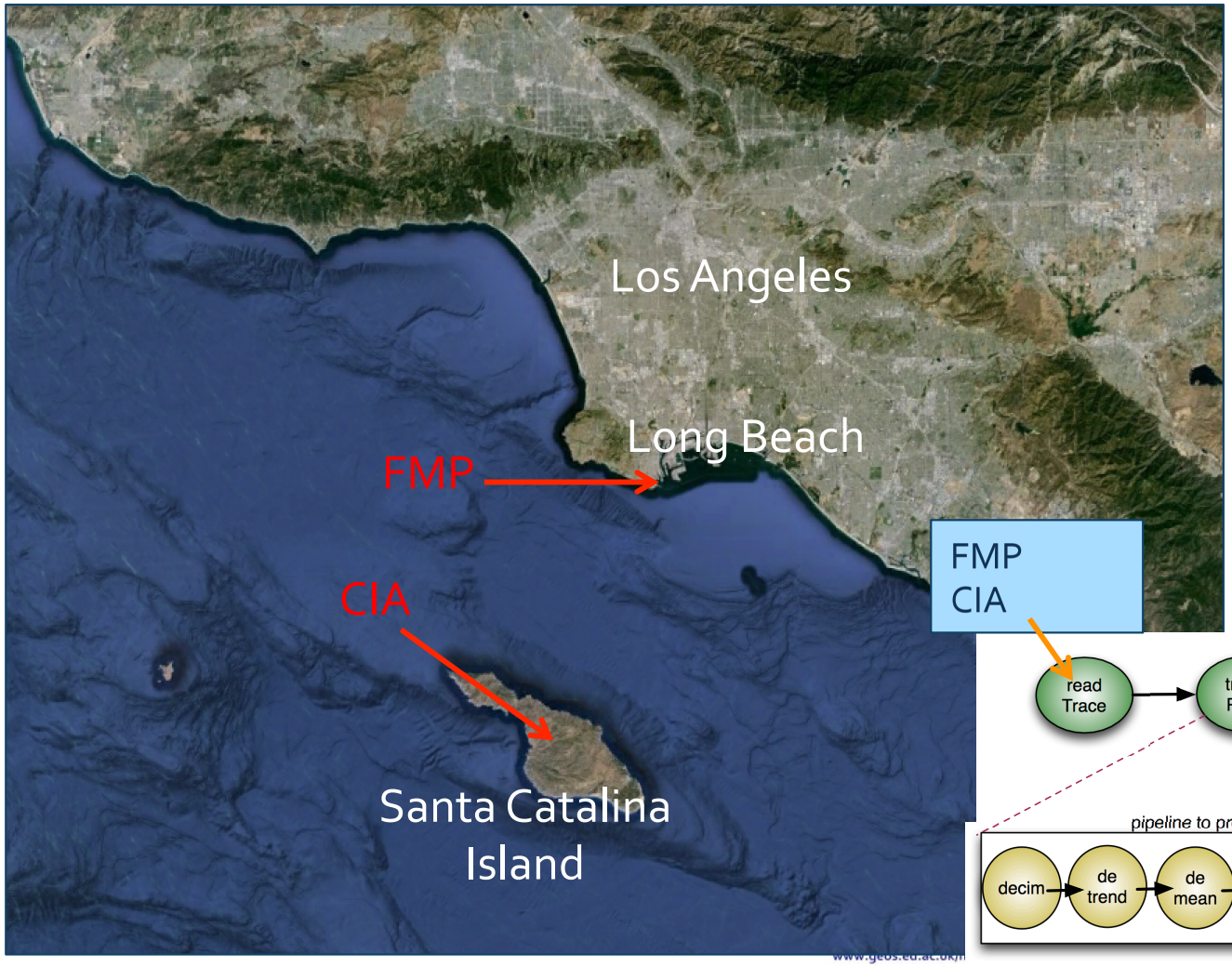
Performance Experiments - 2

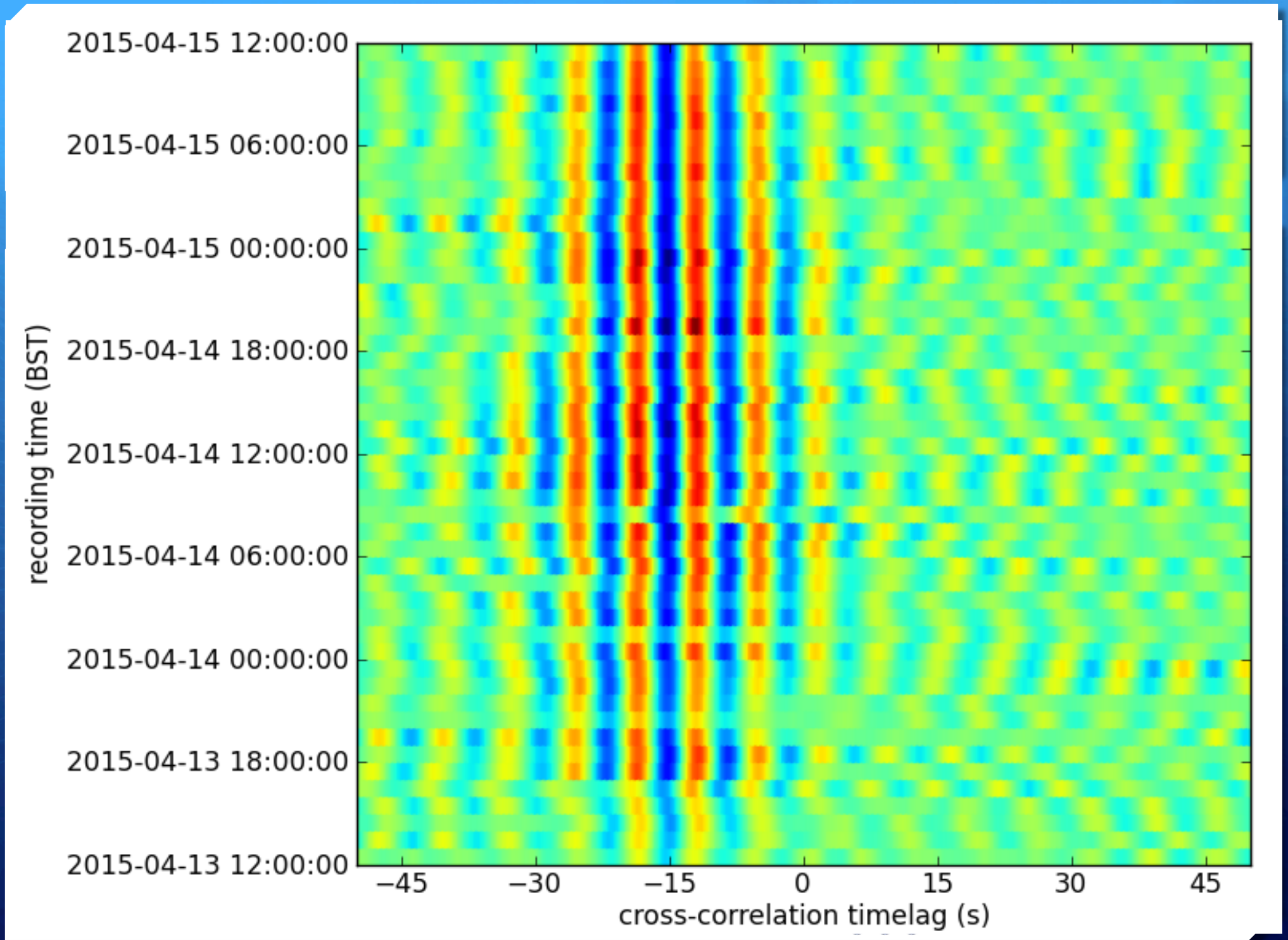
- Simulation of 1000 stations.
  - **Important:** We do not request data to IRIS services in real-time.
- **Number of stations: 1000**
- **Preprocess time: 24 seconds**
- **Xcorr time**: 1332 seconds (**Aprox 22 minutes**).
- **Number xcorr: 4999500**
- **Total time = Aprox 23 minutes**

Performance Experiments - 3

- List of 10 stations, which we know that data is available
- Request and stream in real-time 100 times, data from these 10 stations. Request + Preprocess in parallel
- **Number of stations: 1000**
- **Preprocess time: 1335 seconds (Aprox 22minutes).**
- **Xcorr time: 1332 seconds (Aprox 22 minutes).**
- **Number xcorr: 4999500**
- **Total Time = 44 minutes**

Near real-time continuous correlations of two stations: CI-FMP and CI-CIA from the Caltech Regional Seismic Network




2015-04-15 12:00:00
2015-04-15 06:00:00
2015-04-15 00:00:00
2015-04-14 18:00:00
2015-04-14 12:00:00
2015-04-14 06:00:00
2015-04-14 00:00:00
2015-04-13 18:00:00
2015-04-13 12:00:00
recording time (BST)
−45
−30
−15
0
15
30
45
cross-correlation timelag (s)

Conclusions and Future work

- Achievements:
  - Capacity to pre-process and cross-correlate 1000s of seismic stations in real-time
- Future work:
  - Use data stations from different services. E.g Orfeus
  - Cross-correlation based analyses of large numbers of seismic records
  - Identification of repeating events in large seismic datasets
  - Climate science: acceleration of research requiring the integration of climate model and satellite data.