Automating Scientific Computations: from the User's Desktop to World-Class Supercomputers

Rafael Ferreira da Silva[1], Rosa Filgueira[2], Ewa Deelman[1]
[1]University of Southern California, Information Sciences Institute, Marina del Rey, USA
[2]School of Informatics, University of Edinburgh, Edinburgh, UK

Modern science often requires the processing and analysis of vast amounts of data in search of postulated phenomena, and the validation of core principles through the simulation of complex system behaviors and interactions. This is the case in fields such as seismology, astronomy, bioinformatics, physics, climate and ocean modeling. In order to support the computational and data needs of today's science, new knowledge must be gained on how to deliver the growing capabilities of the national cyberinfrastructures and more recently commercial clouds to the scientist's desktop in an accessible, reliable, and scalable way. In over a decade of working with domain scientists, the Pegasus project has developed tools and techniques that automate the computational processes used in data- and compute-intensive research. Among them is the scientific workflow management system, *Pegasus* [1], which is being used by researchers to model seismic wave propagation, to discover new celestial objects, to study RNA critical to human brain development, and to investigate other important research questions. As part of its research program of earthquake system science, the Southern California Earthquake Center (SCEC) has developed *CyberShake*, a high-performance computing software platform that uses 3D waveform modeling to calculate physics-based probabilistic seismic hazard analysis (PSHA) estimates for populated areas of California. A CyberShake hazard curve computation can be divided into two phases. In the first phase, a 3D mesh of approximately 1.2 billion elements is constructed and populated with seismic velocity data. This mesh is then used in a pair of wave propagation simulations that calculates and outputs strain Green tensors (SGTs). The SGT simulations use parallel wave propagation codes and typically run on 4,000 processors. In the second phase, individual contributions from over 400,000 different earthquakes are calculated using the SGTs, then these hazard contributions are aggregated to determine the overall seismic hazard. These second phase calculations are loosely coupled, short-running serial tasks. To produce a hazard map for Southern California, over 100 million of these tasks must be executed. The extensive heterogeneous computational requirements and large numbers of high-throughput tasks necessitate a high degree of flexibility and automation; as a result, SCEC utilizes Pegasus workflows for execution [2]. Recently, the Pegasus team and the dispel4py [3] team have collaborated to enable automated processing of real-time seismic interferometry and earthquake "repeater" analysis using data collected from the IRIS database. The workflow (*Seismic Ambient Noise Cross-Correlation*) periodically reads data from the repository (about every hour), and performs waveform cross-correlations analyses through thousands of computational tasks. The workflow consists of two main phases: *Preprocess*–each continuous time series from a given seismic station (called a trace), is subject to a series of treatments. The processing of each trace is independent from other traces, making this phase "embarrassingly" parallel (complexity O(n), where n is the number of stations); and *Cross-Correlation*–pairs all of the stations and calculates the cross-correlation for each pair (complexity $O(n^2)$). This integration is freely available online as a service using Docker containers, which can be easily used in mostly cloud computing environments.

References

1. E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, "Pegasus, a Workflow Management System for Science Automation," Future Generation Computer Systems, 46:17-35, 2015.
2. S. Callaghan, E. Deelman, D. Gunter, G. Juve, P. Maechling, C. Brooks, K. Vahi, K. Milner, R. Graves, E. Field, D. Okaya, "Scaling up workflow-based applications," Journal of Computer and System Sciences, 76(6):428-46, 2010.
3. R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Spinuso, S. Sanchez-Exposito, "dispel4py: An agile framework for data-intensive escience," IEEE 11th International Conference on eScience, pp. 454-464, 2015.

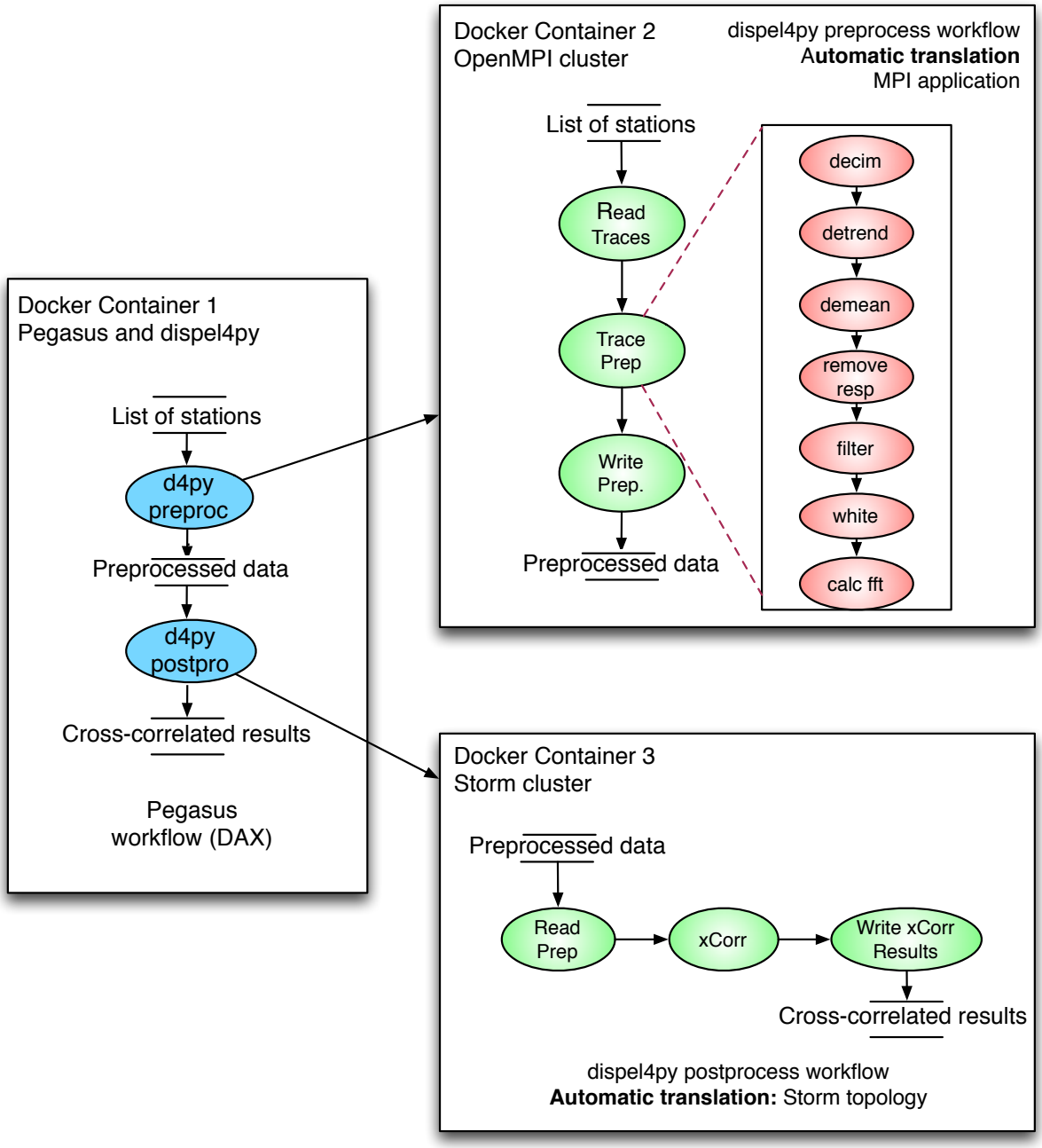

Figure 1. Overview of the Seismic Ambient Noise Cross-Correlation workflow using Pegasus and dispel4py to automate computation on distributed resources (OpenMPI and Storm clusters).