Adapting Reproducible Research Capabilities to Resilient Distributed Calculations

Manuel Rodríguez-Pascual, CIEMAT, Madrid, Spain
Christos Kanellopoulos, GRNET, Athens, Greece
Antonio Juan Rubio-Montero, CIEMAT, Madrid, Spain
Diego Darriba, Universidade da Coruña, A Coruña, Spain
Ognjen Prnjat, GRNET, Athens, Greece
David Posada, Universidade de Vigo, Vigo, Spain
Rafael Mayo-García, CIEMAT, Madrid, Spain

ABSTRACT

Nowadays, computing calculations are becoming more and more demanding due to the huge pool of resources available. This demand must be satisfied in terms of computational efficiency and resilience, which is compromised in distributed and heterogeneous platforms. Not only this, data obtained are often either reused by other researchers or recalculated. In this work, a set of tools to overcome the problem of creating and executing fault tolerant distributed applications on dynamic environments is presented. Such a set also ensures the reproducibility of the performed experiments providing a portable, unattended and resilient framework that encapsulates the infrastructure-dependent operations away from the application developers and users, allowing performing experiments based on Open Access data repositories. In this way, users can seamlessly search and lately access datasets that can be automatically retrieved as input data into a code already integrated in the proposed workflow. Such a search is based on metadata standards and relies on Persistent Identifiers (PID) to assign specific repositories. The applications profit from Distributed Toolbox, a framework devoted to the creation and execution of distributed applications and includes tools for unattended cluster and grid execution, where a total fault tolerance is provided. By decoupling the definition of the remote tasks from its execution and control, the development, execution and maintenance of distributed applications is significantly simplified with respect to previous solutions, increasing their robustness and allowing running them on different computational platforms with little effort. The integration with Open Access databases and employment of PIDs for long-lasting references ensures that the data related to the experiments will persist, closing a complete research circle of data access/processing/storage/dissemination of results.

KEYWORDS
Distributed Calculations, Fault Tolerance, Persistent Identifiers, Resilience

DOI: 10.4018/IJGHPC.2016010105

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
1. INTRODUCTION

Among the different distributed platforms, cluster and grid infrastructures have emerged as powerful options to face ambitious problems. However, the implementation and execution of distributed applications that can run both on cluster and Grid infrastructures is far from being trivial. Although the execution model of distributed applications on both platforms is similar, their particularities make the application requirements completely different. First, while the cluster is usually managed by someone from the same organization as the user, in the grid the resources belong to different organizations, each one with different hardware, software, usage and security policies (Foster, 2002). Also, while in local clusters the physical resources do not evolve in time, grid infrastructures are highly dynamic. Moreover, while local clusters are considered as reliable, execution of tasks on the grid can be problematic (Botón-Fernández et al., 2015), so the application must be designed to overcome failures.

One of the recent developments in distributed computing has been the use of Open Access databases. Data analysis can now be performed on different resources, possibly belonging to different institutions, and employing high performance networks to communicate. This introduces a new set of challenges, including how to refer to remote data and to ensure that the reference will remain stable for a reasonable amount of time. This is of utmost importance in the so-called data curation, this is, the active management of data over its life-cycle of interest, establishing long term repositories for current and future use.

The work presented here aims to solve both problems with a unified cross-disciplinary approach. The first objective is to provide developers with an efficient instrument to create and port distributed applications. For this sake, DistributedToolbox (Rodríguez-Pascual and Mayo-García, 2013) encapsulates a set of tools that enable the development and execution of highly portable distributed applications on clusters and grids. It ensures a correct task completion, so the application developer is released from the low level operations of task management and control.

DistributedToolbox is articulated around RemoteAPI, a very simple API designed to define the tasks to execute on the distributed infrastructure. Then, one of the devoted tools included on the toolbox takes care of the task execution. In this sense, the proposed solution does not compete with existing ones; instead, it embraces the different alternatives.

After dealing with the creation and execution of portable applications, this work tackles with the problem of reproducible research. Persistent IDentifiers (PIDs) are long- lasting references to digital objects (single files or set of files) (Hakala, 2010). In scientific computing, PIDs can reference primary and secondary scientific data with a unique and timeless manner, very similar to how DOI numbers are used to identify articles.

The objective is to ensure reproducibility, this is, the ability of an entire experiment or study to be reproduced, either by the researcher or by someone else and to create new work based on the research. For this sake, open on-line databases are a basic tool, making both raw and processed data freely available. Of course, with the commitment of making portable, long lasting distributed applications, the input data employed by these applications should persist. PIDs represent a powerful tool for this purpose, ensuring that future changes on URIs or internal organization of Open Access databases will be transparent to the user willing to repeat a given experiment.
Related Content

Placement and Scheduling over Grid Warehouses
[www.igi-global.com/chapter/placement-scheduling-over-grid-warehouses/19340?camid=4v1a](www.igi-global.com/chapter/placement-scheduling-over-grid-warehouses/19340?camid=4v1a)

A Grid and Cloud Based System for Data Grouping Computation and Online Service
[www.igi-global.com/article/grid-cloud-based-system-data/60254?camid=4v1a](www.igi-global.com/article/grid-cloud-based-system-data/60254?camid=4v1a)
A Survey of Scheduling and Management Techniques for Data-Intensive Application Workflows

A Generic Reference Architecture for Collaboratory Scientific Virtual Laboratory