Future Generation Computer Systems 37 (2014) 46–59

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Enabling scientific workflow sharing through coarse-grained interoperability

Gabor Terstyanszky [a,*], Tamas Kukla [a], Tamas Kiss [a], Peter Kacsuk [a,b], Akos Balasko [b], Zoltan Farkas [b]

[a] Centre for Parallel Computing, University of Westminster, London, United Kingdom

[b] MTA-SZTAKI, Budapest, Hungary

HIGHLIGHTS

- Coarse-Grained Interoperability concept developed to support workflow sharing.
- Data structure elaborated to enable workflow management in a repository.
- Created a simulation platform to allow the execution of meta workflows.

ARTICLE INFO

Article history:
Received 21 December 2012
Received in revised form
18 December 2013
Accepted 17 February 2014
Available online 19 March 2014

Keywords:
Workflow
Workflow interoperability
Workflow repository
Simulation platform
Distributed Computing Infrastructure

ABSTRACT

E-scientists want to run their scientific experiments on Distributed Computing Infrastructures (DCI) to be able to access large pools of resources and services. To run experiments on these infrastructures requires specific expertise that e-scientists may not have. Workflows can hide resources and services as a virtualization layer providing a user interface that e-scientists can use. There are many workflow systems used by research communities but they are not interoperable. To learn a workflow system and create workflows in this workflow system may require significant efforts from e-scientists. Considering these efforts it is not reasonable to expect that research communities will learn new workflow systems if they want to run workflows developed in other workflow systems. The solution is to create workflow interoperability solutions to allow workflow sharing. The FP7 Sharing Interoperable Workflow for Large-Scale Scientific Simulation on Available DCIs (SHIWA) project developed two interoperability solutions to support workflow sharing: Coarse-Grained Interoperability (CGI) and Fine-Grained Interoperability (FGI). The project created the SHIWA Simulation Platform (SSP) to implement the Coarse-Grained Interoperability approach as a production-level service for research communities. The paper describes the CGI approach and how it enables sharing and combining existing workflows into complex applications and run them on Distributed Computing Infrastructures. The paper also outlines the architecture, components and usage scenarios of the simulation platform.

1. Introduction

E-scientists have put large efforts in the exploitation of Distributed Computing Infrastructures (DCIs) for their ability to run compute- and data intensive applications. Many DCIs with large user communities have emerged during the last decade, such as the Distributed European Infrastructure for Supercomputing Applications (DEISA) [1], Enabling Grids for e-Science Grid (EGEE) [2], the German D-Grid initiative (D-Grid) [3], UK National Grid Service (NGS) [4] and the North American TeraGrid (TG) [5] etc. They are based on different middleware stacks that provide an abstraction layer between computer resources and applications. For example NGS and TeraGrid are built on the Globus Toolkit [6], EGEE on gLite [7], DEISA relies on both the Globus Toolkit and Unicore [8], while D-Grid supports gLite, the Globus Toolkit and Unicore. In Europe the European Grid Infrastructure (EGI) federates all major European organizations through National Grid Infrastructures (NGIs).

Production DCIs are commonly built on a large number of components, such as data resources, metadata catalogues, authentication and authorization methods, and repositories. As a result, managing the execution of applications on DCIs is a complex task.

* Correspondence to: University of Westminster, 115 New Cavendish Street, W1W 6UW London, United Kingdom. Tel.: +44 20 79115000; fax: +44 20 79115089.
E-mail addresses: terstyg@wmin.ac.uk, terstyg@gmail.com (G. Terstyanszky).

http://dx.doi.org/10.1016/j.future.2014.02.016

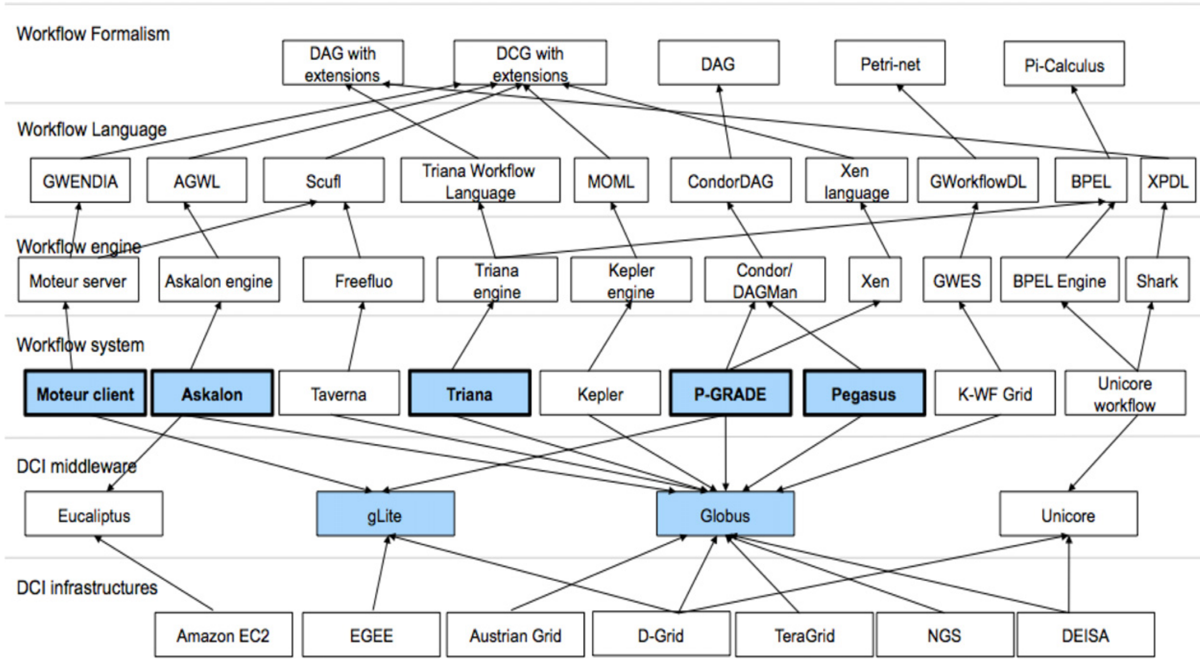

Fig. 1.1. Snapshot of heterogeneous technologies in current workflow systems.

Moreover, solutions developed for one DCI are difficult to port to other infrastructures. In order to hide this complexity from researchers workflow systems are widely used as a virtualization layer on top of the underlying infrastructures. Workflows have become essential to integrate expertise about both applications (user domain) and DCIs (infrastructure domain) in order to support research communities. Workflows have emerged as a new paradigm for e-scientists to formalize and structure complex scientific experiments to enable and accelerate many significant scientific discoveries. A workflow is a formal specification of a scientific process, which represents, streamlines and automates the analytical and computational steps that e-scientists need to go from data selection and integration, computation and analysis to final data presentation and visualization. A workflow system supports the specification, modification, execution, failure recovery and monitoring of a workflow using workflow engine to control its execution.

Research communities have developed different workflow systems and created large numbers of workflows to run their experiments. Fig. 1.1 presents some of the best known workflow systems and DCIs where workflows are executed. Although this picture is far from being complete, it demonstrates how heterogeneous the workflow technology is. These systems differ in workflow description languages, which are based on different formalisms, enactment strategies and middleware providing access to different infrastructures. This often has a profound impact on the resulting workflow performance, development effort, management and portability. It takes significant effort and time to learn how to use workflow systems, and requires specific expertise and skills to develop and maintain workflows. As a result, creating, running and maintaining workflows on DCIs require substantial efforts and expertise. E-scientists would like to share workflows (automatic porting of workflows across workflow systems and DCIs) to optimize their efforts. Currently, the major obstacle of workflow sharing is that workflow systems are not compatible. E-scientists hesitate to learn new workflow systems or port their experiments to other workflow systems, as this is a time-consuming and error prone process. Sometimes they are even constrained to a particular workflow system by DCIs they can use, or services operated by developers or system administrators. Without workflow interoperability e-scientists often depend on specific workflow systems and DCIs. As a result, they cannot develop and run their experiments on multiple workflow systems to unleash performance and share workflows.

Remark. The SHIWA project addressed the highlighted middleware and workflow systems.

Workflow sharing needs workflow systems interoperability for two reasons: first, being able to reuse workflows written in different languages; second, being able to execute workflows using different DCIs. The first challenge reflects the execution of workflows by different workflow systems to enable workflow sharing. The second one relates to the execution of compute and/or data intensive workflows on different DCIs to harvest available compute and data resources to improve the overall performance of workflows. It is difficult to address these two challenges not only due to the diversity and complexity of DCIs and workflow systems, but also because of the wide range of user requirements and the complex data/control flow dependencies of workflows.

Workflow interoperability enables the integrated execution of workflows of different workflow systems that may span multiple heterogeneous infrastructures. It can facilitate application migration due to infrastructure, services and workflow system evolution. Workflow interoperability allows workflow sharing to support and foster the adoption of common research methodologies, improve efficiency and reliability of research by reusing these common methodologies, increase the lifetime of workflows and reduction of development time for new workflows. Interoperability among workflow systems does not only permit the development and enactment of large-scale and comprehensive workflows, but also reduces the existing gap between different DCIs, and consequently promotes cooperation among research communities exploiting these DCIs.

The FP7 Sharing Interoperable Workflow for Large-Scale Scientific Simulation on Available DCIs (SHIWA) project [9] and [10] developed two workflow interoperability solutions: Coarse-Grained Interoperability (CGI) and Fine-Grained Interoperability (FGI). The project created a simulation platform: the SHIWA Simulation Platform (SSP) to support these two interoperability