

## Chapter 12

# Specification and Performance Characteristics of Scientific Grid Workflows

**Radu Prodan**

*Institute of Computer Science, University of Innsbruck, Austria*

### ABSTRACT

*Grid computing promises to enable a scalable, reliable, and easy-to-use computational infrastructure for e-Science. To materialize this promise, Grids need to provide full automation of the entire development and execution cycle starting with application modeling and specification, continuing with experiment design and management, and ending with the collection and analysis of results. Often, this automation relies on the execution of workflow processes. Not much is known much about Grid workflow characteristics, scalability, and workload, which hampers the development of new techniques and algorithms, and slows the tuning of existing ones. This chapter describes techniques developed in the ASKALON project for modeling and analyzing the executions of scientific workflows in Grid environments. The authors first outline the architecture, services, and tools developed by ASKALON and then introduce a new systematic scalability analysis technique to help scientists understand the most severe sources of performance losses that occur when executing scientific workflows in heterogeneous Grid environments. A method for analyzing workload traces is presented, focusing on the intrinsic and environment-related characteristics of scientific workflows. The authors illustrate concrete results that validate the methods for a variety of real-world applications modeled as scientific workflows and executed in the Austrian Grid environment.*

DOI: 10.4018/978-1-4666-0249-6.ch012

## INTRODUCTION

The Grid computing vision (Foster & Kesselman, 2004) promises an easy-to-use, reliable, and efficient computing infrastructure for e-Science. For this promise to become reality Grids must fully automate the application development and execution process that starts with application modeling and specification, continues with experiment design and management, and ends with analysis of results.

Workflow modeling is a well established area in computer science that has been strongly influenced and driven by business process modeling work (Workflow Management Coalition) (Fischer, 2007). Recently, the Grid community has generally acknowledged that orchestrating existing software applications implemented as Grid services in coarse-grain workflows represents an important class of applications that matches the loosely coupled Grid model and, therefore, can benefit from being executed in distributed Grid infrastructures (Yu & Buyya, 2005). Similarly, in order to efficiently harness the computational resources provided by the Grid, existing monolithic scientific applications are currently being re-engineered and decomposed in a set of atomic activities orchestrated in a loosely coupled scientific workflow.

There is currently a large amount of research in the Grid community devoted to the specification of scientific workflow applications that range from low-level scripting languages (The Condor project, n. d.; Deelman et al., 2005; Krishnan et al., 2001; Mayer, McGough, Furmento, Lee, Newhouse, & Darlington, 2003; Seidel, Allen, Merzky, & Nabrzyski, 2002), to high level abstract XML representations (Alves et al., 2007; Amin, Hategan, von Laszewski, Zaluzec, Hampton, & Rossi, 2004; Fahringer, Qin, & Hainzer, 2005; von Laszewski & Hategan, 2005), and user friendly graphical interfaces (Erwin, 2002; Ludascher et al., 2006; Oinn et al., 2004; Taylor, Shields, Wang, & Rana, 2003). Still, a common consensus on the

fundamental structural and runtime characteristics of scientific Grid workflows is missing. In the first part of this chapter we aim to complement these efforts by introducing the approach taken by the ASKALON Grid environment (Fahringer et al., 2006) to provide specification and transparent runtime support for their main distinctive features:

- Large number of activity instances (i.e., hundreds to thousands) which are difficult or impossible to express individually;
- Computationally intensive activities with long and often unpredictable execution times;
- Complex data dependencies of various sizes ranging from few bytes to several gigabytes;
- Sequential loops that transform directed acyclic graph (DAG)-based workflows into more complex and general directed graph-based structures;
- Dynamic control and data flow structure, often unknown before the execution, that may change at runtime depending on the input parameters or on the output results produced by the workflow activities;
- Unreliable execution resources that raise complex fault tolerant issues.

Alongside effective modeling and specification, execution of workflows is the step that mostly distinguishes scientific workflows from their business counterparts making them interesting for research. While several Grid workflow execution engines have recently emerged (Yu & Buyya, 2005), not much is known about their demand, impacting adversely the evolution of old and new workflow engines. One of the distinctive requirements of scientific workflows is performance, for which no clear scalability model exists yet. We propose in this chapter a simple method for computing the speedup and efficiency exhibited by a workflow in a Grid environment and validate it for three real-world applications.

25 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

[www.igi-global.com/chapter/specification-performance-characteristics-scientific-grid/64146](http://www.igi-global.com/chapter/specification-performance-characteristics-scientific-grid/64146)

## Related Content

---

### A Column Generation Procedure for the Split Delivery Vehicle Routing Problem Using a Route-Based Formulation

Joseph Hubert Wilck IV and Tom M. Cavalier (2014). *International Journal of Operations Research and Information Systems* (pp. 44-63).

[www.irma-international.org/article/a-column-generation-procedure-for-the-split-delivery-vehicle-routing-problem-using-a-route-based-formulation/120447](http://www.irma-international.org/article/a-column-generation-procedure-for-the-split-delivery-vehicle-routing-problem-using-a-route-based-formulation/120447)

### Project Management Guide and Project Management Maturity Models as Generic Tools Capable for Diverse Applications

Robert Chee Choong Gan and Christina M. M. Chin (2019). *Diverse Applications and Transferability of Maturity Models* (pp. 269-315).

[www.irma-international.org/chapter/project-management-guide-and-project-management-maturity-models-as-generic-tools-capable-for-diverse-applications/214793](http://www.irma-international.org/chapter/project-management-guide-and-project-management-maturity-models-as-generic-tools-capable-for-diverse-applications/214793)

### Baker Hughes IO and BEACON with a Focus on Downsizing Personnel Requirements at Rig-Site

Joanna Karin Grov Fraser, Jan Ove Dagestad and Barry L. Jones (2013). *Integrated Operations in the Oil and Gas Industry: Sustainability and Capability Development* (pp. 213-224).

[www.irma-international.org/chapter/baker-hughes-beacon-focus-downsizing/68718](http://www.irma-international.org/chapter/baker-hughes-beacon-focus-downsizing/68718)

### Protection of Critical Homeland Assets: Using a Proactive, Adaptive Security Management Driven Process

William J. Bailey (2018). *Global Business Expansion: Concepts, Methodologies, Tools, and Applications* (pp. 1705-1738).

[www.irma-international.org/chapter/protection-of-critical-homeland-assets/202292](http://www.irma-international.org/chapter/protection-of-critical-homeland-assets/202292)

### Fuzzy Based Parameter Adaptation in ACO for Solving VRP

Sandhya and Rajiv Goel (2019). *International Journal of Operations Research and Information Systems* (pp. 65-81).

[www.irma-international.org/article/fuzzy-based-parameter-adaptation-in-aco-for-solving-vrp/225815](http://www.irma-international.org/article/fuzzy-based-parameter-adaptation-in-aco-for-solving-vrp/225815)