Chapter 66
A Survey of Scheduling and Management Techniques for Data-Intensive Application Workflows

Suraj Pandey
The Commonwealth Scientific and Industrial Research Organisation, Australia

Rajkumar Buyya
The University of Melbourne, Australia

ABSTRACT

This chapter presents a comprehensive survey of algorithms, techniques, and frameworks used for scheduling and management of data-intensive application workflows. Many complex scientific experiments are expressed in the form of workflows for structured, repeatable, controlled, scalable, and automated executions. This chapter focuses on the type of workflows that have tasks processing huge amount of data, usually in the range from hundreds of mega-bytes to petabytes. Scientists are already using Grid systems that schedule these workflows onto globally distributed resources for optimizing various objectives: minimize total makespan of the workflow, minimize cost and usage of network bandwidth, minimize cost of computation and storage, meet the deadline of the application, and so forth. This chapter lists and describes techniques used in each of these systems for processing huge amount of data. A survey of workflow management techniques is useful for understanding the working of the Grid systems providing insights on performance optimization of scientific applications dealing with data-intensive workloads.

INTRODUCTION

Scientists and researchers around the world have been conducting simulations and experiments as a part of medium to ultra large-scale studies in high-energy physics, biomedicine, climate modeling, astronomy and so forth. They are always seeking cutting-edge technologies to transfer, store and process the data in a more systematic and controlled manner as the data requirements of these applications range from megabytes to petabytes. Thus, to help them manage the complexity of execution, transfer and storage of results of these large-scale applications, the use of a Workflow

DOI: 10.4018/978-1-4666-4153-2.ch066

Management Systems (WfMS) is in wide practice (Yu & Buyya, 2005).

Scheduling and managing computational tasks of a workflow were the main focus of WfMS in the past. With the emergence of globally distributed computing resources and increasing output data from scientific experiments, scientists began to realize the necessity of handling data in conjunction with computational tasks. Scientific workflows were then modeled taking into account the flow of data. However, even with a plethora of techniques and systems, many challenges remain in the area of data management related to workflow creation, execution, and result management (Deelman & Chervenak, 2008; Gil et al., 2007).

Some challenges for managing data-intensive application workflows are:

- High throughput data transfer mechanisms
- Massive, cheap, green and low latency storage solutions and their interfaces
- Composition of scientific applications as workflows
- Multi-core technology and workflow management systems
- Standards for Interoperability between workflow systems
- Globally distributed data and computation resources

In this chapter, we classify and survey techniques that have been used for managing and scheduling data-intensive application workflows to meet the challenges listed above. The classification is based on techniques that take into account data, storage, platform and application characteristics. We sub-divide each general heading into more specific techniques. We then list and describe several work under each sub-heading. Most systems use a combination of existing techniques to achieve the objectives of an application workflow.

The rest of the chapter is organized as follows. In next section, we present previous studies that focused more on systems side of Grid workflows and Data Grids along with their taxonomy. We then describe the terms and definitions used in this chapter followed by an abstract model of a WfMS and its component responsible for data and computation management. In the rest of the chapter, we present the survey. We finally conclude identifying some future trends in management of data-intensive application workflows.

RELATED WORK

Over the last few years, we can find much work being done on data-intensive environments and workflow management systems. We list taxonomies for Data Grid Systems and Workflow management Systems that present the grounds for our survey.

Venugopal, Buyya, & Ramamohanarao (2006) proposed a comprehensive taxonomy of data Grids for distributed data sharing, management and processing. They characterize, classify and describe various aspects of architecture, data transportation, data replication and resource allocation, and scheduling for Data Grids systems. They list the similarities and differences between Data Grids and other distributed data-intensive paradigms such as content delivery networks, peer-to-peer networks, and distributed databases.

Yu & Buyya (2005) proposed taxonomy of workflow management systems for Grid computing. They characterize and classify different approaches for building and executing workflows on Grids. They present a survey of representative Grid workflow systems highlighting their features and pointing out the differences. Their taxonomy focuses on workflow design, workflow scheduling, fault management and data movement.

Bahsi, Ceyhan & Kosar (2007) presented a survey and analysis on conditional workflow management. They studied workflow management systems and their support for conditional structures such as *if, switch and while*. With case studies on existing WfMS, they listed the differences in