Accepted Manuscript

Structure aware resource estimation for effective scheduling and execution of data intensive workflows in cloud

K. Kanagaraj, S. Swamynathan

PII: S0167-739X(16)30811-1
DOI: http://dx.doi.org/10.1016/j.future.2017.09.001
Reference: FUTURE 3656

To appear in: *Future Generation Computer Systems*

Received date : 19 December 2016
Revised date : 29 June 2017
Accepted date : 2 September 2017

Please cite this article as: K. Kanagaraj, S. Swamynathan, Structure aware resource estimation for effective scheduling and execution of data intensive workflows in cloud, *Future Generation Computer Systems* (2017), http://dx.doi.org/10.1016/j.future.2017.09.001

Structure Aware Resource Estimation for Effective Scheduling and Execution of Data Intensive Workflows in Cloud

Mr. K. Kanagaraj[a] and Dr. S. Swamynathan[b]

[a] Assistant Professor, MEPCO Schlenk Engg. College, Sivakasi, kanagaraj@mepcoeng.ac.in
[b] Associate Professor, CEG, Anna University, Chennai, swamyns@gmail.com

Abstract

A set of interdependent tasks used to automate a business or scientific process can be modelled as a workflow and represented in the form of a Directed Acyclic Graph (DAG) or Directed Acyclic Graph in XML (DAX). Cloud computing is the current popular technology that provides hardware and software resources that are accessible from anywhere and at any time. As the cloud users are relieved of the difficulties of managing hardware and software resources, it is the most convenient and suitable environment to execute workflows. Workflows that accept and process a large amount of data are termed as data intensive workflows. The execution cost of such workflows in the cloud depends not only on the configuration of the Virtual Machines (VMs) but also the cost of data transfer between the tasks. Due to the highly dynamic arrangement of tasks in the workflow, deciding the optimum configuration and exact number of VMs is a big challenge for researchers today. Hence, in this paper, an effective resource provisioning and scheduling mechanism based on the structure of the workflow is proposed. The significance of this work is to identify the required number of VMs and their configuration, based on the structure of the workflow and optimizing data transfer between the tasks. Popular workflows like Montage, Cyber Shake, Epigenomics and Inspiral are used to analyze the quality of this work, and the obtained results confirm that the proposed workflow scheduler is able to provide a notable reduction in execution cost without compromising the execution time.

Keywords: Workflow; Structure Aware; Data Intensive; Scheduling; Optimization

1. Introduction

Workflows are described as a set of tasks that have dependencies between them. A task would start its execution only when all its predecessors have completed their execution. Typically a workflow is composed with hundreds of tasks that need to be executed in a coordinated way and represented as DAG. Every node in the workflow is assigned a certain weight. The weight of a node is the execution time of the task. Most workflows will have a single entry and a single exit tasks. The elapsed time between the entry and exit tasks is called as the makespan (execution time) of the workflow. As workflows can represent any set of activities irrespective of their domain, it became a popular model to represent any business or scientific process. Workflow can be defined to handle a purchase order fulfilment, travel itinerary, health care recommendation and scientific applications which involve multiple tasks connected in a predefined order. The nature of workflow could be highly process intensive or data intensive depending on the application. The complexity of execution of tasks could range from the sequential execution to highly parallel execution with multiple inputs from various other tasks. The computing resource requirements for managing complex workflows requires high-end systems with huge processing and data storage capabilities. The obvious choice for executing complex workflows is the cloud, which provides Iaas or DaaS as services.

Cloud computing has emerged as a standard resource provisioning mechanism that can be used for running large scale scientific and critical applications. It also provides rapid access to computing resources at lower cost with flexible configurations. Without much investment in procuring and managing hardware and software resources, cloud resources can optimally be used to handle complex workflow systems. The cloud users can procure the right type and size of computing and storage resources and also can have access to as many resources as required. These resources can be obtained instantaneously and payment may be made for what is being used. Cloud