Accepted Manuscript

A security and cost aware scheduling algorithm for heterogeneous tasks of scientific workflow in clouds

Zhongjin Li, Jidong Ge, Hongji Yang, Liguo Huang, Haiyang Hu, Hao Hu,
Bin Luo

PII: S0167-739X(15)00398-2
DOI: http://dx.doi.org/10.1016/j.future.2015.12.014
Reference: FUTURE 2928

To appear in: *Future Generation Computer Systems*

Received date: 31 May 2015
Revised date: 16 December 2015
Accepted date: 20 December 2015

Please cite this article as: Z. Li, J. Ge, H. Yang, L. Huang, H. Hu, H. Hu, B. Luo, A security and cost aware scheduling algorithm for heterogeneous tasks of scientific workflow in clouds, *Future Generation Computer Systems* (2016), http://dx.doi.org/10.1016/j.future.2015.12.014

A Security and Cost Aware Scheduling Algorithm for Heterogeneous Tasks of Scientific Workflow in Clouds

Zhongjin Li[a], Jidong Ge[a,b,*], Hongji Yang[c], Liguo Huang[d], Haiyang Hu[e], Hao Hu[a], Bin Luo[a]

[a] *State Key Laboratory for Novel Software Technology, Software Institute, Nanjing University, China, 210093*

[b] *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China*

[c] *Centre for Creative Computing (CCC), Bath Spa University, England, UK*

[d] *Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX, USA, 75275-0122*

[e] *School of Computer, Hangzhou Dianzi University, Hangzhou, China, 310018*

Abstract

Security is increasingly critical for various scientific workflows that are big data applications and typically take quite amount of time being executed on large-scale distributed infrastructures. Cloud computing platform is such an infrastructure that can enable dynamic resource scaling on demand. Nevertheless, based on pay-per-use and hourly-based pricing model, users should pay attention to the cost incurred by renting virtual machines (VMs) from cloud data centers. Meanwhile, workflow tasks are generally heterogeneous and require different instance series (i.e., computing optimized, memory optimized, storage optimized, etc.). In this paper, we propose a security and cost aware scheduling (SCAS) algorithm for heterogeneous tasks of scientific workflow in clouds. Our proposed algorithm is based on the meta-heuristic optimization technique, particle swarm optimization (PSO), the coding strategy of which is devised to minimize the total workflow execution cost while meeting the deadline and risk rate constraints. Extensive experiments using three real-world scientific workflow applications, as well as CloudSim simulation framework, demonstrate the effectiveness and practicality of our algorithm.

Keywords: scientific workflow scheduling; cloud computing; big data application; security awareness; particle swarm optimization (PSO); deadline constraint

1. Introduction

Workflow applications have emerged as an attractive paradigm for programming distributed computing infrastructures, which are extensively applied in diverse scientific computing areas such as astronomy, bioinformatics, and physics [1]. With the increasing complexity of scientific computing systems, such workflow applications have become big data applications, which demands large-scale infrastructures in order to be executed within in a reasonable amount of time [2].

Among such infrastructures, the cloud computing environment is one of the special interests [3]. Cloud computing is a large-scale distributed computing driven by the emerging needs of economical and efficient system operations, in which the infrastructures are available in a pay-per-use system and can provide dynamic scaling in response to the needs of workflow applications. Recently, the cloud and non-cloud storage systems are respectively deployed for biomedical scientists to conduct the controlled experiments on performance comparison. The results show that the cloud system outperforms the non-cloud system in terms of execution time, consistency, and efficiency improvement [4]. Furthermore, the type and number of VM resources for workflow execution can be provisioned on demand in clouds. Although an infinite amount of resources can be accessed in the context of clouds, users should pay special attention to the economic cost incurred by resource lease [5]. Commercial cloud providers typically charge users by an hourly-based pricing model. Therefore, cost is calculated based on the time unit model instead of the actual usage of resources, which meaning that users have to