A Hybrid Resource Reservation Method for Workflows in Clouds

Tyng-Yeu Liang, Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

Fu-Chun Lu, Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

Jun-Yao Chiu, Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

ABSTRACT

QoS and energy consumption are two important issues for Cloud computing. In this paper, the authors propose a hybrid resource reservation method to address these two issues for scientific workflows in the high-performance computing Clouds built on hybrid CPU/GPU architecture. As named, this method reserves proper CPU or GPU for executing different jobs in the same workflow based on the profile of execution time and energy consumption of each resource-to-program pair. They have implemented the proposed resource reservation method on a real service-oriented system. The experimental results show that the proposed resource reservation method can effectively maintain the QoS of workflows while simultaneously minimizing the energy consumption of executing the workflows.

Keywords: Cloud Quality of Service, Energy Consumption, Hybrid CPU/GPU, Hybrid Resource Reservation, Workflow

1. INTRODUCTION

Recently, General Purpose Graphic Process Unit (GPGPU, simply called GPU) (Owens, Luebke, Govindaraju, Harris, Krüger, Lefohn, & Purcell, 2007) has successfully become an alternative resource for Clouds (Voorsluys, Broberg, & Buyya, 2011) to provide users with high-performance computing services. For example, Amazon recently has proposed that they can provide their customers with the rental service of GPU clusters for high-performance computing. Since GPU has higher density of computation cores and lower EPI (energy per instruction) than CPU, using GPU clusters instead of CPU ones is helpful for Clouds to minimize the energy consumption, financial cost and building-space demand of offering high performance computing (HPC) services. From the viewpoint of users, this GPU solution also implies that they have a chance to reduce their pays for using Cloud resources, and to speed up the completion of their work.

However, GPU has some drawbacks in the hardware architecture (Owens, Houston, Luebke, Green, Stone, & Phillips, 2008). For

DOI: 10.4018/jghpc.2012100101
example, the performance of data synchronization in GPU is not good enough. The reason is that shared memory is allowed only for the threads in the same thread block, so the threads of a thread block must use global memory to communicate with the ones of another thread block while the access cost of global memory is more expensive than that of shared memory. Moreover, GPU is not efficient for executing branch instructions because it has no mechanism of branch prediction, and it cannot allow the threads in the same thread block to simultaneously jump to different targets. Consequently, GPU is not always better than CPU for reducing the execution time and the energy consumption of user programs because of these problems. Therefore, it is necessary to carefully select proper CPUs or GPUs for user programs in order for maintaining the QoS of HPC Clouds.

On the other hand, workflow is a mean of schedule that is originally used for organizing processes and resource planning in manufactories. Recently, it is also commonly applied for managing the processes of business and scientific analysis in Clouds. Basically, a scientific workflow is composed of a number of jobs and each job is assigned to execute a program that is developed by cloud-service consumers or cloud-service providers. Since the programs invoked in a workflow may be computation-bound or branch/synchronization bound, they must be allocated onto proper resources for execution in order to maintain the QoS of executing workflows. Although many past researches has proposed different resource allocation and reservation methods (Yu, Buyya, & Tham, 2005; Brandic, Benkner, Engelbrecht, & Schmidt, 2008; Khanli & Analoui, 2007; Afzal, Darlington, & McGough, 2006; Blythe, Jain, Deelman, Gil, Vahi, Mandal, & Kennedy, 2005; Amalarethinam, & Selvi, 2012) for resolving the QoS problem of workflows in grids, these methods focus only on the CPU resources, and usually considers only the time constrains or the budget limits of workflows but ignores the energy consumption of workflows. Therefore, it is necessary to develop an advanced resource reservation method for resolving the QoS and energy consumption problem of scientific workflows in the HPC Clouds, which provide both of CPU and GPU resources.

As previously discussed, we propose a hybrid resource reservation method for scientific workflows in the HPC Clouds built on a hybrid CPU/GPU architecture. The basic idea of the proposed method is to predict the execution time and energy consumption of resources to programs based on the execution history of resource-to-program pairs, and then select a proper CPU or GPU from the candidate resources according to the prediction result, and the requirements of QoS and energy consumption. We have implemented the proposed resource reservation method in a service-oriented system called Teamster-G. Moreover, we have done a series of experiments for evaluating the prediction precision and the effectiveness of the proposed method. Our experimental results show that the proposed method is indeed effective for meeting the QoS requirement of workflows while simultaneously minimizing the energy consumption of the workflows.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces the proposed resource reservation method. Section 4 describes the brief implementation job of the proposed method in Teamster-G. Section 5 discusses our experimental results of evaluating the performance of the proposed method. Finally, Section 6 gives the conclusions of this paper and our future work.

2. RELATED WORK

In recent years, several researchers have proposed the methods of resource allocation for optimizing the performance or achieving the QoS requirements of user applications in the hybrid resource architecture. For example, Gregg, Brantley, and Hazelwood (2010) proposed a method that dynamically allocates user programs on CPU and GPU for execution according to the load of resources and the predicted execution time of user programs. They defined a metadata of describing the execution history
Managing Inconsistencies in Data Grid Environments: A Practical Approach
www.igi-global.com/article/managing-inconsistencies-data-grid-environments/47211?camid=4v1a

Event Processing in Web Service Runtime Environments
www.igi-global.com/chapter/event-processing-web-service-runtime/44404?camid=4v1a