Future Generation Computer Systems 34 (2014) 47–65

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

CLOUDRB: A framework for scheduling and managing High-Performance Computing (HPC) applications in science cloud

Thamarai Selvi Somasundaram, Kannan Govindarajan *

Madras Institute of Technology, Anna University, Chennai, India

HIGHLIGHTS

- We propose scheduling framework for managing HPC applications in Science Cloud (SC).
- Introduce Particle Swarm Optimization based scheduling for allocation of resources.
- Designed Discrete Event Simulator in Matlab to simulate the proposed research work.
- Embarrassingly parallel and Iterative parallel application execution in SC.
- Minimize makespan, cost, job rejection ratio and maximize jobs meeting deadline.

ARTICLE INFO

Article history:
Received 9 January 2013
Received in revised form
12 October 2013
Accepted 3 December 2013
Available online 22 December 2013

Keywords:
Cloud computing
High Performance Computing (HPC)
CLOUD Resource Broker (CLOUDRB)
Resource allocation
Job scheduling
Particle Swarm Optimization (PSO)
Science cloud

ABSTRACT

In recent years, the Cloud environment has played a major role in running High-Performance Computing (HPC) applications, which are computationally intensive and data intensive in nature. The High-Performance Computing Cloud (HPCC) or Science Cloud (SC) provides the resources to these types of applications in an on demand and scalable manner. Scheduling of jobs or applications in a Cloud environment is NP-Complete and complex in nature due to the dynamicity of resources and on demand user application requirements. The main motivation behind this research study is to design and develop a CLOUD Resource Broker (CLOUDRB) for efficiently managing cloud resources and completing jobs for scientific applications within a user-specified deadline. It is implemented and integrated with a Deadline-based Job Scheduling and Particle Swarm Optimization (PSO)-based Resource Allocation mechanism. Our proposed approach intends to achieve the objectives of minimizing both execution time and cost based on the defined fitness function. It is simulated by modeling the HPC jobs and Cloud resources using the Matlab programming environment. The simulation results prove the effectiveness of the proposed research work by minimizing the completion time, cost and job rejection ratio and maximizing the number of jobs completing their applications within a deadline and meeting the user's satisfaction. The proposed work has been tested in our Eucalyptus-based cloud environments by submitting real-world HPC applications and observed the improvements in performance.

1. Introduction

Cloud Computing provides on demand and scalable delivery models to the users. The Cloud service delivery models are categorized into three major types viz., Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The IaaS service delivery model plays a major role in hosting SaaS, PaaS and Scientific applications or web applications in Cloud data centers. Furthermore, it provides the computational and storage resources to the scientific and business community using the open-source Cloud middleware tools such as Nimbus [1], Eucalyptus [2], OpenNebula [3] and Hadoop [4] in a dynamic manner. Cloud Computing has been extended to HPC applications as well. HPC applications are mainly focused on scientific applications that may fall into any one of the categories, specifically parallel, iterative parallel, embarrassingly parallel, data intensive and workflow applications. The HPC applications are complex in nature; they require large amounts of computational cycles that also involves vast amount of data processing. In this research work, our focus is mainly on parallel iterative and embarrassingly parallel (parallel sweep) types of jobs. Originally, Cluster Computing has evolved to solve those scientific application problems in terms of High Performance Computing (HPC). The introduction of Grid Computing allowed a collaborative problem solving technique to solve the complex problems using the resources which are globally distributed. However, the

* Corresponding author.
E-mail addresses: stselvi@annauniv.edu (T.S. Somasundaram), kannan.gridlab@gmail.com (K. Govindarajan).
URL: http://www.annauniv.edu/care (K. Govindarajan).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.12.024

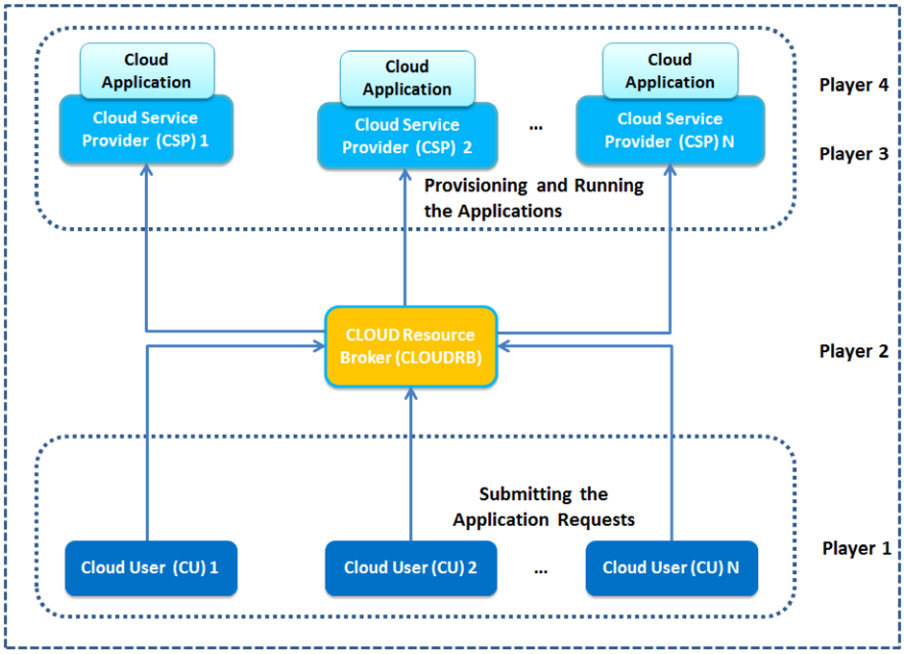

Fig. 1. Four major players of IaaS.

Grid Computing was unable to provision the resources as per the requirement of jobs in need of the software environment to execute the jobs. The Virtualization concept helped in provisioning Infrastructure from the Cloud resources in the form of IaaS. Thus, the emerging technology of Cloud Computing provides Cloud resources such as High Performance Computing Clouds (HPCCs) or Science Clouds (SCs). HPCCs provide the customized execution environment for scientific applications by employing the Virtualization Technology and Service Oriented Architecture (SOA) concepts in a dynamic manner.

1.1. Four major players in Cloud

The four major players in the Cloud environment are: (1) Cloud Users (CUs) (2) Cloud Service Providers (CSPs) (3) Cloud Applications (CAs) and (4) CLOUD Resource Brokers (CLOUDRBs). The interaction between the four major players is shown in Fig. 1. The CUs submit the jobs specifying the requirements of Software, Hardware and Quality of Service (QoS) parameters. The requirements vary in terms of Hardware (Processor Speed, RAM Memory, Bandwidth, etc.), Software (Mpich-1.2.7, Charm++ 3. X, FFTW-3. X, etc.) and QoS (Deadline, Throughput, etc.). The CSPs are responsible for managing the physical heterogeneous Cloud resources to host the Cloud applications which require computation, storage and network resources. The CAs may be of different types varying from simple web applications to complex high-performance computing applications. The Cloud Users are particular about completing their application's execution within a specified deadline and cost. To meet the objectives of CAs, CUs and CSPs we have proposed CLOUDRB. The Proposed CLOUDRB helps the CUs by selecting appropriate Cloud resources and enables the CSPs to deliver more profit with maximum resource utilization. To select the suitable Cloud resource a scheduling algorithm has been proposed and integrated with our proposed CLOUDRB to successfully complete the execution of tasks within the stipulated deadline with minimal time and cost.

Scheduling in Cloud Computing environments is considered to be an NP-complete problem [5]. The biologically inspired techniques are classified [6] into Evolutionary Technique, Swarm Intelligence (SI) and Ecology Technique. In recent years, the biologically inspired techniques are receiving greater attention to solve such types of NP-complete problems. Several researchers have investigated the SI-based biologically inspired algorithms such as Genetic Algorithm [7], Ant Colony Optimization [8] and Particle Swarm Optimization (PSO) [9] for solving optimization problems. PSO is one of the most popular biologically inspired optimization algorithms which come under the family of Swarm Intelligence (SI). It has been developed by Eberhart and Kennedy [10] in 1995. It uses the principle of social behavior of birds flocking or fish schooling. Moreover, it achieves a faster convergence rate [11,12] and global optimum solution [13] within minimal time as compared to ACO and GA. Some of the researchers [14–16] have investigated earlier for solving the job scheduling problem in grids and clouds using PSO because of the advantages over other algorithms mentioned above. However, most of the research papers are primarily focused on minimizing time [17], cost [15], energy [15,18] or time and cost [12] for workflow types of applications. Nevertheless, our proposed approach is mainly targeted to minimize the makespan and cost together for embarrassingly parallel and parameter sweep applications. Also, we have looked at the deadline as one of the important factors in our proposed job scheduling mechanism. Therefore, in this research work, we have made use of the advantages of Particle Swarm Optimization (PSO) for scheduling of jobs to the cloud resources for getting a near optimal solution. The jobs are first prioritized based on their deadline and resource selection is carried out by employing the population-based PSO algorithm. The proposed algorithm generates the possible schedule, and it builds all possible combinations of job requests with available cloud resources with the objectives of minimizing makespan and cost within the user specified deadline. We first formulate the scheduling problem in a cloud environment as an NP-complete problem; subsequently we have described the cloud-based architecture design of CLOUDRB. The proposed approach is simulated to test the efficacy of various performance metrics. Furthermore, we have tested the same with real-world HPC scientific applications in our cloud testbed. In brief, the contributions of the research study are summarized below:

- We have designed and developed a Deadline-based Job Scheduling algorithm for prioritizing the user job requests. (A)