Computers and Electrical Engineering 000 (2018) 1–17

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Managing high-performance computing applications as an on-demand service on federated clouds☆

Zhengxiong Hou[a,*], Yunlan Wang[a], Yulei Sui[b], Jianhua Gu[a], Tianhai Zhao[a], Xingshe Zhou[a]

[a] Center for High-Performance Computing, School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[b] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

ARTICLE INFO

Article history:
Received 28 June 2017
Revised 23 February 2018
Accepted 23 February 2018
Available online xxx

Keywords:
High-Performance Computing
Federated Clouds
On-demand
SaaS
Pay-per-use
Software Licensing

ABSTRACT

There are several challenges (e.g., imbalance between supply and demand of hardware resources and software licenses, and usability) under modern High-Performance Computing (HPC) environment. As a means of providing an on-demand service for end users, we propose a Software-as-a-Service (SaaS) approach for managing commercial HPC applications as a Web-based service deployed on top of federated clouds. Some inter-trusted private or public clouds are federated to create a unified service platform with a large amount of hardware resources. In addition, an on-demand, pay-per-use model for Web-service-enabled HPC applications is proposed. Further, we provide an economic analysis of the proposed approach from the perspective of end users, cloud service providers, and Independent Software Vendors (ISVs). We conduct a simulation using two HPC application services on three federated clouds. A combined Quality of Service (QoS) and economic evaluation demonstrates a better effect of the proposed approach comparing with existing HPC platforms.

1. Introduction

High-Performance Computing (HPC) centers usually provide in-house clusters or supercomputers for end users for high-performance scientific and engineering computing. There are many kinds of application software related to diverse domains. Almost all Independent Software Vendors (ISVs) providing software for commercial applications use software licenses for the protection of copyright and control of usage. Running HPC jobs needs the effective utilization of both hardware and software resources, which are usually shared by numerous end users from diverse application domains. The usage mode is often based on command lines and job scripts. The end users have to develop some basic knowledge about the Linux operating system, including scripts for a batch scheduling system, such as PBS, Condor, or IBM Platform LSF. Hardware resources and software licenses are limited, there are fluctuations in demand from end users, and the HPC service is inconvenient.

Recently, there has been growing interest in the use of cloud computing to enable on-demand HPC services on a pay-per-use basis. For service-oriented cloud computing [1–4], Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS) have all been offered by cloud service providers. Hardware services have already been

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Hong Shen.

* Corresponding author.
E-mail address: houzhx@nwpu.edu.cn (Z. Hou).

https://doi.org/10.1016/j.compeleceng.2018.02.036
0045-7906/

implemented on a pay-per-use basis. The cost for the end users is calculated based on their use of resources and services, such as CPU cycling, memory, storage resources, and network bandwidth. Payment is from the end users to the public cloud service providers. However, HPC applications are still less flexible than the hardware component in the HPC cloud computing environment. Our key insights regarding the limitations of managing HPC applications in the existing HPC platforms are listed as follows.

(1) Imbalance between supply and demand of hardware resources
Despite the emergence of public clouds, users still do not want to discard legacy in-house clusters or supercomputers, which are built for private clouds. Single private or public clouds have limitations on hardware resources. User demand fluctuates, resulting in dedicated resources that are either underutilized or overloaded. In many cases, jobs may need to wait in a batch queue (probably for a long time) before they can be processed. To alleviate this issue, some inter-trusted private or public clouds can be utilized to build a federated infrastructure.
(2) Imbalance between supply and demand of software licenses
Following the traditional "pay-per-license" model for perpetual licenses, cloud service providers buy the application software along with a given number of licenses. In reality, the usage of application software licenses is dynamic and elastic. The existing licensing mechanism is over-costly for many cases. There may be some excess licenses. Thus, sometimes there is a license wastage problem. On the contrary, in some cases, especially during peak request times, the users' requirements cannot be satisfied, i.e., there is a license shortage problem. Thus, when attempting to provide an on-demand service for commercial HPC applications, software licensing is a bottleneck. Consequently, pay-per-use-based SaaS can be used to provide on-demand services for HPC applications and software licenses.
(3) Complexity of access to the software
End users face a mixed scenario in terms of HPC resources, including different software versions. There are traditional clusters with batch schedulers, grids, and clouds (e.g., Amazon EC2) providing virtual clusters. Most end users are scientists, engineers, or graduate students, whose experience and expertise varies a lot in diverse application domains, and some are unfamiliar with setting up the necessary execution environments. Therefore, it is not convenient for them to alternate between different computing systems or software versions. A unified service portal for federated clouds is a promising solution to this problem.

To address the above challenges and provide an on-demand HPC service for end users, we present a pay-per-use-based SaaS approach for managing commercial HPC applications as an on-demand service on federated clouds. The proposed pay-per-use model serves end users via a Web portal by offering a large amount of hardware resources (usually virtualized) and unlimited software licenses for Web-service-enabled HPC applications on federated clouds. Similar to grid computing [5] and sky computing [6], several inter-trusted private or public clouds are federated to create a unified cloud computing platform for end users. Thus, the varying requirements in terms of software licenses can be satisfied anytime and anywhere. However, in our proposed approach, users need to pay for the use of software services and licenses, as well as hardware services. The on-demand services are based on a pay-per-use approach to software licensing. Therefore, we provide an economic analysis of the proposed approach. Nevertheless, there is still a limitation in terms of hardware resources on the federated clouds, so it is only possible to provide a quasi-on-demand HPC application service. The queuing time for the available resources is considerably shortened, and the service is easy to use.

Some new issues in relation to the proposed platform, such as workload distribution and data transfer, need further study. However, dealing with these other issues is beyond the scope of this work. The main contributions of this paper are as follows:

(1) We propose a pay-per-use-based SaaS approach to managing commercial HPC applications as an on-demand service on federated clouds, which can provide quasi-on-demand hardware services and on-demand software services with a unified, friendly Web portal for HPC end users.
(2) We provide a theoretical economic analysis of the proposed approach.
(3) A case study involving simulation of SaaS for two HPC applications on three federated clouds is presented.
(4) We present a comprehensive Quality of Service (QoS) and economic evaluation, including availability and scalability, average job response time, reliability, billing to end users and cost to cloud service providers, as well as profits for software vendors.

The rest of this paper is organized as follows. Section 2 introduces some related work. Section 3 describes our methodology and a detailed SaaS design for HPC applications on federated clouds. Section 4 presents an economic analysis from the perspective of the end users, cloud service providers, and software vendors. Section 5 presents a case study involving SaaS simulation for HPC applications on federated clouds. In Section 6, we evaluate the proposed model from different perspectives. Finally, Section 7 concludes the paper.

2. Related work

In this section, we discuss related work that is relevant to our solution, including software licensing and SaaS for HPC applications, and existing research on federated clouds and economic analysis.

Please cite this article as: Z. Hou et al., Managing high-performance computing applications as an on-demand service on federated clouds, Computers and Electrical Engineering (2018), https://doi.org/10.1016/j.compeleceng.2018.02.036