Future Generation Computer Systems 71 (2017) 113–128

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Cost minimization for deadline-constrained bag-of-tasks applications in federated hybrid clouds

Somayeh Abdi [a], Latif PourKarimi [b], Mahmood Ahmadi [c,*], Farzad Zargari [d]

[a] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[b] Department of Mathematics, Razi University, Kermanshah, Iran
[c] Computer Engineering and Information Technology Department, Razi University, Kermanshah, Iran
[d] Information Technology Faculty, Iran Telecom Research Center, Tehran, Iran

HIGHLIGHTS

- The cost of bag-of-tasks applications in a federated cloud is minimized.
- Cost minimization is modeled as a binary linear programming problem.
- The proposed model is solved with the CPLEX solver.
- The sensitivity of optimal solutions to input data is investigated.

ARTICLE INFO

Article history:
Received 2 September 2016
Received in revised form 2 January 2017
Accepted 30 January 2017
Available online 3 February 2017

Keywords:
Federated hybrid clouds
Scheduling
Bag-of-tasks applications
Binary linear programming
Cost minimization
Deadline constrained

ABSTRACT

A mathematical programming model is proposed for a resource allocation problem in federated clouds, where bag-of-tasks (BoT) applications are assigned to instance types with different costs and performance levels. The proposed model is a binary linear programming problem containing deadline and resource constraints in the cloud federations and by the objective of minimizing the total cost of applications. These constraints and objective are explicitly expressed using mathematical functions, and the model is solved with the CPLEX solver. This paper also discusses a post-optimality analysis that deals with stability in assignment problems. Numerical results show that the optimal cost and optimal solutions in the cloud federations are lower and more stable, respectively, than those presented by single-provider clouds. In contrast to optimality in single-provider clouds, that in the cloud federations is less sensitive to input data.

1. Introduction

Cloud computing has received considerable attention in recent years because of its affordances, such as the provision of pay-per-use services [1,2]. The emergence of cloud computing prompted the development of interconnected clouds, called inter-clouds, to achieve improved quality of service (QoS) and provide scalable services. Inter-clouds are a reflection of the evolution of cloud computing. Two fundamental inter-cloud models that have been proposed are cloud federations and multi-clouds, which are still in their infancy and are confronted with many challenges related to portability and interoperability between clouds [3]. Despite these challenges, however, the inter-cloud concept has created many opportunities for cloud providers and customers [3–6].

A federation of clouds refers to multiple interconnected and interoperable cloud providers, whose services help costumers avoid vendor lock-in and access highly available and cost-effective services. Cloud federations also bring new opportunities for cloud providers to share their infrastructure, maximize the capacity of perishable resources (e.g. CPU cycles), save energy, address unexpected loads, and avoid the extra cost incurred from overprovisioning in data centers [3,7–9]. A federated hybrid cloud, which consists of a private cloud and a federation of public clouds, has been put forward as a cloud deployment model [10]. This innovation enables research centers to execute compute-intensive scientific applications on multiple public IaaS[1] platforms.

Scheduling of scientific applications with cost minimization objective in inter-cloud models has received attention in recent years [11–13]. Cost-efficient scheduling is possible in a cloud federation because of the diversity in price and performance of the

* Corresponding author.
E-mail addresses: s.abdi@srbiau.ac.ir (S. Abdi), l.pourkarimi@razi.ac.ir (L. PourKarimi), m.ahmadi@razi.ac.ir (M. Ahmadi), zargari@itrc.ac.ir (F. Zargari).

[1] Infrastructure as a Service (IaaS).

http://dx.doi.org/10.1016/j.future.2017.01.036
0167-739X/© 2017 Published by Elsevier B.V.

instance types provided by multiple IaaS providers. Commercial clouds limit the maximum number of VMs that a customer can run simultaneously. For example, Amazon Elastic Compute Cloud (EC2)[2] allows a customer to concurrently run a maximum of 20 VMs, albeit the customer can issue a request for a limit increase to Amazon. In spite of such limits, scientists can simultaneously run numerous VMs in a cloud federation. Compute-intensive bag-of-tasks (BoT) applications have an appropriate structure that enables execution on multiple providers because these applications include independent homogeneous tasks [14].

This study uses a mathematical programming model to address a resource allocation problem for BoT applications in federated hybrid clouds. The mathematical model, objective, and constraints of the problem are explicitly expressed using mathematical functions and expressions. Three tasks are carried out: a mathematical model is created, the derived model is solved, and a post-optimality or sensitivity analysis is conducted. The derived model is a binary linear programming (BLP) problem that can be solved in a reasonable time. Given that some efficient algorithms for solving BLP problems are available, the model obtained in this work can be of great interest from the perspective of computation.

In the resource allocation investigated in this work, we have a federation of multiple clouds that provide instance types with different prices and performance levels and some BoT applications that need to be executed within a predetermined deadline. The objective is the scheduling of the applications in the federations at minimum total cost. The proposed model fulfills constraints related to the application requirements, scheduling policy, and resource limits imposed by the federations. It is solved by the CPLEX solver and developed in OPL, which is an algebraic modeling language. Finally, analysis is conducted to determine the sensitivity of optimal cost to changes in input data under a given set of assumptions.

A cloud federation can apply different policies for establishing resource limits. We formulate three scenarios of resource limitation in a cloud federation:

(1) No limitation on resource provisioning (NLRP): A customer can simultaneously run its required VMs until demand no longer exceeds the available VMs in a federation. The federation does not constrain the maximum number of simultaneously run VMs.
(2) Cloud provider-level restriction (CPLR): Resources are allocated in accordance with the limits set by providers that participate in a federation.
(3) Federation-level restriction (FLR): Resource limits are applied in cloud federations to prevent the system saturation caused by substantial resource provisioning to a single user. In this scenario, a limit is imposed on the total number of simultaneously run VMs that can be allocated to a customer in a federation.

The scheduling policy implemented in a federated cloud environment can affect the total cost of applications. All tasks that are associated with an application can be submitted to a single participating cloud (i.e., non-distributed scenario) or distributed among multiple clouds (i.e., distributed scenario). Some recent studies [11,13,15] have focused on the distribution of application tasks across multiple cloud providers. In the current work, we formulate distributed and non-distributed scenarios and evaluate them with respect to their suitability for compute- and data-intensive applications.

The contributions of this study are summarized as follows:

- The problem of scheduling BoT applications in federated hybrid clouds is formulated as a BLP model underlain by the objective of minimizing total cost.
- Variations in optimal cost with respect to resource limits, data sizes, and deadline parameters are investigated.
- The optimal cost and sensitivity of the resource allocation problem under distributed and non-distributed scheduling scenarios are compared.
- A post-optimality analysis examines the sensitivity of optimal cost to changes in deadlines, resource limits, and data sizes.

The numerical results indicate that the optimal cost in the cloud federations is considerable lower than that presented by single-provider clouds in the presence of stringent limitations and rules (e.g., short deadlines).

The rest of this paper is organized as follows. In Section 2, we review the related works. Section 3 introduces the scheduling method, system model of federated hybrid clouds, BoTs applications and problem statement. Section 4 deals with the proposed mathematical model for the scheduling problem as well as input data and auxiliary parameters. Numerical results of the model are reported in Section 5 followed by concluding remarks on Section 6.

2. Related works

In cloud computing, the problem of resource allocation with profit-maximization [7,9,16–20] or cost-minimization [10–12,15,21–25] objectives has received attention in recent years, from the perspective of providers and customers, respectively. Scheduling of applications with cost minimization objective has received considerable attention in hybrid clouds [15,21,22,24] and inter-cloud models [10–13].

The problem of resource allocation for deadline-constrained BoT applications in hybrid IaaS clouds has been addressed in [22,24]. They consider computation and data transfer costs. These approaches, schedule each task of applications on one instance type which this kind of scheduling is not cost-efficient for tasks with short run-time, because VMs are charged in terms of dollar-per-hour. Our model schedules multiple tasks on one instance type to utilize allocated VMs. Also, they did not address the resource limits imposed by commercial clouds, which we consider it as an important factor in resource allocation.

M. Malawski et al. [21] addressed scheduling of BoT applications in hybrid IaaS clouds. They formulated the resource allocation problem as a mixed integer nonlinear programming model. They considered data transfer and computation costs. This approach considered one bag of identical tasks. We proposed a general model for scheduling of multiple BoT applications that each application may consist of multiple bags e.g. Astro workflow [26]. We formulated the resource allocation problem for BoT applications as a BLP problem. The derived BLP model can be of great interest from the perspective of computation.

The cost minimization for scientific workflows has been recently addressed in [13,15]. These approaches optimized cost and makespan in IaaS clouds using Pareto-based approach and game theoretic scheduling, respectively. These works use different models from ours. We optimize the total cost for deadline-constrained applications using mathematical programming technique.

A. Jaikar et al. [12] addressed resource allocation in federated cloud. They did not consider data transfer cost. Moreover, they considered one instance type in each cloud provider while requests are identical. Our model considers computation and data transfer costs where computational size of applications are different. We also assumed that VMs are charged in dollar-per-hour.

I. Moschakis et al. [11] proposed simulated annealing algorithm for scheduling of BoT applications on multi-clouds with the cost

[2] https://aws.amazon.com/ec2.