Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Virtual machine placement for elastic infrastructures in overbooked cloud computing datacenters under uncertainty

Fabio López-Pires [a,b,*], Benjamín Barán [b], Leonardo Benítez [b], Saúl Zalimben [b], Augusto Amarilla [b]

[a] Information Technology and Communications Center, Itaipu Technological Park, Hernandarias, Paraguay
[b] Polytechnic School, National University of Asunción, San Lorenzo, Paraguay

HIGHLIGHTS

- A first proposal for a complex IaaS environment for VMP problems considering service elasticity, including both vertical and horizontal scaling of cloud services, as well as overbooking of physical resources, including server (CPU and RAM) as well as networking resources (Ortigoza et al., 2016).
- A two-phase optimization scheme for VMP problems, combining advantages of both online and offline VMP formulations in the proposed IaaS environment, introducing a prediction-based VMPr Triggering method to decide when to trigger a placement reconfiguration (*Research Question 1*) as well as an update-based VMPr Recovering method to decide what to do with VMs requested during placement recalculation times (*Research Question 2*).
- A first scenario-based uncertainty approach for modeling the following relevant uncertain parameters of the proposed complex IaaS environment: (i) virtual resources capacities (vertical elasticity), (ii) number of VMs that compose cloud services (horizontal elasticity), (iii) utilization of CPU and RAM memory virtual resources (relevant for overbooking) and (iv) utilization of networking virtual resources (also relevant for overbooking).
- A first formulation of a VMP problem considering the above mentioned contributions, for the optimization of the following four objective functions: (i) power consumption, (ii) economical revenue, (iii) resource utilization, as well as (iv) placement reconfiguration time.
- An experimental evaluation of the presented two-phase optimization scheme against state-of-the-art alternatives for VMP problems, considering 400 different scenarios.

ARTICLE INFO

Article history:
Received 26 February 2017
Received in revised form 25 August 2017
Accepted 7 September 2017
Available online xxxx

Keywords:
Virtual machine placement
Cloud computing
Overbooking
Elasticity
Uncertainty
Incremental VMP
VMP reconfiguration

ABSTRACT

Infrastructure as a Service (IaaS) providers must support requests for virtual resources in highly dynamic cloud computing environments. Due to the randomness of customer requests, Virtual Machine Placement (VMP) problems should be formulated under uncertainty. This work presents a novel two-phase optimization scheme for the resolution of VMP problems for cloud computing under uncertainty of several relevant parameters, combining advantages of online and offline formulations in dynamic environments considering service elasticity and overbooking of physical resources. In this context, a formulation of a VMP problem is presented, considering the optimization of the following four objective functions: (i) power consumption, (ii) economical revenue, (iii) resource utilization and (iv) reconfiguration time. The proposed two-phase optimization scheme includes novel methods to decide when to trigger a placement reconfiguration through migration of virtual machines (VMs) between physical machines (PMs) and what to do with VMs requested during the placement recalculation time. An experimental evaluation against state-of-the-art alternative approaches for VMP problems was performed considering 400 scenarios. Experimental results indicate that the proposed methods outperform other evaluated alternatives, improving the quality of solutions in a scenario-based uncertainty model considering the following evaluation criteria: (i) average, (ii) maximum and (iii) minimum objective function costs.

1. Introduction

Achieving an efficient resource management in cloud computing datacenters presents several research challenges, including relevant topics in resource allocation [1]. This work focuses on

[*] Corresponding author at: Information Technology and Communications Center, Itaipu Technological Park, Hernandarias, Paraguay.

E-mail address: fabio.lopez@pti.org.py (F. López-Pires).

https://doi.org/10.1016/j.future.2017.09.021
0167-739X/© 2017 Elsevier B.V. All rights reserved.

one of the most studied problems for resource allocation in cloud computing datacenters: the process of selecting which requested virtual machines (VMs) should be hosted at each available physical machine (PM) of a cloud computing infrastructure, commonly known as Virtual Machine Placement (VMP). This work proposes a complex Infrastructure as a Service (IaaS) environment for VMP problems, considering both service elasticity [2] and overbooking of physical resources [3].

To the best of the authors' knowledge, there is no published work simultaneously taking into account elasticity and overbooking, directly related to the most relevant dynamic parameters in the literature on uncertain VMP problem considering multi-objective optimization. In order to model this complex IaaS environment for VMP problems, cloud services (i.e., inter-related VMs) are considered instead of isolated VMs [4].

It is worth remembering that VMP is a NP-Hard combinatorial optimization problem [5]. From an IaaS provider perspective, the VMP problem is mostly formulated as an online problem and must be solved with short time constraints [6].

Online decisions made along the operation of a dynamic cloud computing infrastructure negatively affects the quality of obtained solutions in VMP problems when comparing to offline decisions [7]. In this context, offline algorithms present a substantial advantage over online alternatives. Unfortunately, offline formulations are not appropriate for highly dynamic environments for real-world IaaS providers, where cloud services are requested dynamically according to current demand.

This work presents a two-phase optimization scheme, decomposing the VMP problem into two different sub-problems, combining advantages of online and offline VMP formulations considering a complex IaaS environment. The presented optimization scheme for the VMP problem introduces novel methods to decide when to trigger placement reconfigurations with migration of VMs between PMs (defined as *VMPr Triggering*) and what to do with cloud services requested during placement recalculation times (defined as *VMPr Recovering*).

For IaaS customers, cloud computing resources often appear to be unlimited and can be provisioned in any quantity at any required time [8]. Consequently, this work considers a basic federated-cloud deployment architecture for the VMP problem.

It is important to consider that more than 60 different objective functions have been proposed for VMP problems [6]. In this context, the number of considered objective functions may rapidly increase once a complete understanding of the VMP problem is accomplished for practical problems, where several different parameters should be ideally taken into account. Consequently, a renewed formulation of the VMP problem is presented, considering the optimization of the following four objective functions: (i) power consumption, (ii) economical revenue, (iii) resource utilization and (iv) reconfiguration time.

Due to the randomness of customer requests, VMP problems should be formulated under uncertainty [9]. This work presents a scenario-based uncertainty approach for modeling uncertain parameters, considering a two-phase optimization scheme for VMP problems in the proposed complex IaaS environments.

An experimental evaluation against state-of-the-art alternative approaches for VMP problems was performed considering 80 different workloads in 5 different CPU load scenarios, totalizing 400 experimental scenarios. Experimental results indicate that the proposed VMPr Triggering and Recovering methods of the presented two-phase optimization scheme outperform other evaluated alternatives, improving the quality of solutions.

In summary, the main contributions of this paper are:

- A first proposal for a complex IaaS environment for VMP problems considering service elasticity, including both vertical and horizontal scaling of cloud services, as well as overbooking of physical resources, including server (CPU and RAM) as well as networking resources [4].
- A two-phase optimization scheme for VMP problems, combining advantages of both online and offline VMP formulations in the proposed IaaS environment, introducing a prediction-based VMPr Triggering method to decide when to trigger a placement reconfiguration (*Research Question 1*) as well as an update-based VMPr Recovering method to decide what to do with VMs requested during placement recalculation times (*Research Question 2*).
- A first scenario-based uncertainty approach for modeling the following relevant uncertain parameters of the proposed complex IaaS environment: (i) virtual resources capacities (vertical elasticity), (ii) number of VMs that compose cloud services (horizontal elasticity), (iii) utilization of CPU and RAM memory virtual resources (relevant for overbooking) and (iv) utilization of networking virtual resources (also relevant for overbooking).
- A first formulation of a VMP problem considering the above mentioned contributions, for the optimization of the following four objective functions: (i) power consumption, (ii) economical revenue, (iii) resource utilization, as well as (iv) placement reconfiguration time.
- An experimental evaluation of the presented two-phase optimization scheme against state-of-the-art alternatives for VMP problems, considering 400 different scenarios.

The remainder of this paper is structured in the following way: preliminary concepts and research challenges addressed in this work are introduced in Section 2, while related works and motivation of this work are summarized in Section 3. Section 4 presents the proposed uncertain VMP problem formulation considering four objectives, while Section 5 presents details on the design and implementation of evaluated alternatives to solve the proposed renewed formulation of the VMP problem. Experimental results are summarized in Section 6. Finally, conclusions and future work are left to Section 7.

2. Preliminary concepts and research challenges

The following sub-sections introduce relevant concepts related to the considered IaaS environments for VMP problems, a brief motivation for decomposing the VMP problem into two different sub-problems in a two-phase optimization scheme as well as uncertainty issues related to resource allocation in cloud computing. Additionally, the main challenges and research questions addressed in this work are also briefly introduced.

2.1. IaaS environments for VMP problems

In real-world environments, IaaS providers dynamically receive requests for the placement of cloud services with different characteristics according to different dynamic parameters. In this context, preliminary results of the authors identified that the most relevant dynamic parameters in the VMP literature are [4]: (i) resource capacities of VMs (associated to vertical elasticity) [10], (ii) number of VMs of a cloud service (associated to horizontal elasticity) [11] and (iii) utilization of resources of VMs (relevant for overbooking) [12]. Considering the above mentioned dynamic parameters, environments for IaaS formulations of provider-oriented VMP problems could be classified by one or more of the following classification criteria: (i) service elasticity and (ii) overbooking of physical resources [4]. A cloud service may represent virtual infrastructures for basic services such as Domain Name Service (DNS), web applications or even elastic applications such as MapReduce programs [4].