ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 599-602
doi: 10.18178/wcse.2016.06.103

Resources Management in Cloud Computing

Lotfi Hajjem [1 +] and Salah Ben Abdallah [2]

[1] Institut Supérieur de Gestion Tunis (ISG), Tunisia
[2] Tunis Business School, Tunisia

Abstract. Cloud computing industry has considerably evolved over the last few years. It consists on finding an allocation of shared computing resources to a number of users that maximizes the cloud provider's profit while satisfying all QoS constraints. This problem, known as cloud management was studied within different variants. In this paper, two main variants, resource allocation and resource scheduling in cloud systems will be studied. A literature review of them is given and a comparative study is detailed.

Keywords: cloud computing, resources allocation, resource scheduling.

1. Introduction

Over the last few years, a new business model in the computing world, known as cloud computing, has emerged. According to the official NIST definition, "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". The main advantages of cloud computing are, firstly, the fact that the resources are delivered on-demand and according to the user system's workload, which spares the resources underutilizations. Indeed, a recent survey over six corporate data centers has found that most of the servers were using just 10–30% of their available computing power, while desktop computers have an average capacity utilization of less than 5%. Next, they will not be costly for the organization as they are paid by use. That is, according to Brookings institution, the IT spending saves are in the order of 25-40% when adopting cloud computing. Cloud services are offered to different users to execute their jobs. The different jobs, which can be composed of a set of dependent or independent subtasks, are generally managed by the cloud provider. She/he tries to find an arrangement of the different tasks among the available cloud resources that maximizes its profit while guaranteeing QoS constraints related to response time, services qualities (delay, control overhead), and reliability. This problem was modelled within different forms and many approaches were proposed to solve these variants. In the next section of this paper, we will focus on resource allocation, where an extension of the literature revue presented in [1] will be given, and resources scheduling problems, where a list of the main researches will be presented. And then, in section 3 a comparative study will be given, where the performance metrics used in [1] will be considered. And the last section covers our future works and conclusion.

2. Literature Review

The problem of resource management (RM) in distributed systems and specifically in cloud infrastructures is widely studied in the literature and many solutions were proposed to solve its different variants [1]. In this paper we will bring out a survey on papers treating the problems of resources allocation and tasks scheduling. Mustafa et al. [2] classified RM techniques into energy-aware techniques, SLA-aware

[+] Corresponding author. Tel.: +0021650597050.
E-mail address: lotfi.hajjem@gmail.com.

techniques, Market-oriented techniques, Load-balanced techniques, Network load aware solutions, RM techniques for hybrid /federated cloud and RM techniques for mobile clouds.

2.1. Resource allocation

It consists on allocating a number of physical machines (PM) to the different users' jobs or virtual machines (VM). Generally this problem was modeled as a bin-packing problem in its general form, where jobs represent the items to be packed and physical machines stand for the bins.

In [3] the authors proposed an approach for resources allocation based on the negotiation of SLA regarding three parameters price, time slot and response time. They proposed a utility function for each one of them and an agent based approach is adopted. The agents are allowed to concurrently make multiple proposals (combination of price, time slot and response time) with the same aggregated utility. And the broker's agent selects the one that corresponds more to his preferences. The proposed approach was compared to four other approaches (Greedy, random, round robin and manual zone selection) and has registered better results. The authors in [4] studied the problem of multi-cloud resources allocation where a task can be allocated to multiple clouds at the time. Although, this technique might minimize the costumer costs, additional constraints should be considered and some problems regarding the communication time and data homogeneity and transfer have to be faced. The authors in [5] treated the problem of allocating a number of PMs to the users' VMs. The main objective of the allocation process is to satisfy the Service Level Agreement (SLA) contract regarding CPU performance, memory capacity and response time. To assist the allocation process, the authors adopt Support Vector Regression, which is machine learning method for recognizing patterns and analyzing data, to estimate the number of resources utilization. In addition a Genetic algorithm (GA) was developed where a roulette wheel technique for the selection step, a one-point crossover method that is always applied (crossover rate = 100%) and the mutation operator changes a VM from a PM to another. In [6], a set of providers buy their resources to a number of tasks. The authors consider two cases; there are sufficient and insufficient resources for all tasks. In the former, each provider proposes its bid, which are known by the other ones and the provider with the cheapest bid will be selected. While in the latter case, a double auction approach is proposed, and the consumers with highest bids are selected to execute their tasks. The problem studied in [7] considers a set of cloud users aiming to run their VMs within SaaS providers, which rent resources from IaaS providers. The authors model the problem as a game where each player (cloud users, SaaS providers and IaaS providers) adjusts its strategy according to the ones of the other players. They defined the strategy of each player that optimize the utility of the cloud, i.e. maximize the profits of the SaaS and IaaS providers and minimize the cost of the cloud users. In [8], the authors proposed an approach to minimize the number of active PMs for energy aware reasons. The proposed technique consists on finding a list of VMs to be migrated. This list is composed of some VMs selected from overloaded PMs and all VMs of under loaded PMs. Then, a Power-aware Best Fit decreasing algorithm will be applied to assign the VMs list to other PMs in such a way that no PM will be overloaded and a maximum number of under loaded ones are switched to the sleep mode. A taxonomy of energy-aware resources allocation techniques for cloud computing are presented in [9]. The problem studied in [10] considers a cloud provider having several VM instances of different characteristics available for allocation and run. In addition, there is a number of cloud consumers, where each one requests a bundle of VMs and proposes his bid for it and the winners will be served. This auction process will be repeated and the users may revise their bids according to the previous auction, until getting served (they win the auction) or rejected (there is no enough time for executing their tasks). Three mechanisms were proposed to solve this problem, the fixed price approach and two combinatorial auction-based mechanisms, and a comparative study is given.

2.2. Task scheduling

This problem consists on allocating cloud resources to a set of tasks that can be dependent or not. Some papers suppose that tasks are non-preemptive while others consider the fact that tasks should be preempted in order to be migrated for another host or to liberate resource for a task with higher priority.

The problem studied in [11] considers a number of resources (PMs) and a set of non-preemptive tasks, where each task is composed of subtasks. The main objective is to find an allocation of the PMs to the subtasks that minimizes the cost and the total makespan. To solve this bi-objective problem, an improved differential evolution algorithm (IDEA), which is a combination of the Tagushi algorithm and the DEA approach, is proposed to generate the Pareto set of non-dominated solutions. In [12], the authors consider a system with *n* heterogeneous PMs and a number of VMs of different characteristics. In addition, there is a set of a large number of real-time, aperiodic and independent tasks that should be assigned to the VMs. The main objectives of this problem were to maximize the number of tasks finishing before their deadline and to minimize the total energy consumption. To solve this problem a Proactive Reactive Scheduling approach is proposed. It dynamically exploits proactive and reactive scheduling methods. This approach was compared to four other techniques, Nom-Migration-PRS, Earliest-Deadline-First, Minimum Completion Time and Complete Rescheduling, and experimental results show that it performs better than them. In [13], the tasks scheduling process is performed within a federated cloud, where each cloud provider offers different number of resources at different prices. And the objective is to find in each scheduling interval an arrangement of the different tasks that optimizes their performance and the incurred cost. To solve this problem, simulated annealing and thermodynamic simulated annealing were proposed. The task scheduling problem in federated clouds was also treated in [14, 15]. In [14] two online dynamic scheduling algorithms were proposed. These algorithms take into consideration the variation of tasks workload and adjust dynamically resources allocation accordingly. And in [15] the allocation process is decomposed into time slots. At each time period, an agent based approach is executed to find, if possible, an affectation of the different jobs. This problem was modelled as a game and Nash equilibrium was proven. In [16], the authors proposed a solution to the problem of allocating a number of VMs to a set of dependent tasks. It is supposed that the tasks are non-preemptive and are executed in the same VM. The authors proposed two scheduling strategies minimizing as well the incurred monetary cost as the tasks execution makespan. The first one aims at mapping the tasks to the most cost-efficient VMs based on the concept of Pareto dominance. And the second one extends the first approach to reduce the monetary costs of non-critical tasks, i.e. tasks whose execution time can be extended. The problem studied in [17] consists on affecting a set of users' jobs to a number of cloud nodes. The main objective was to minimize the job's makespan. To solve this problem, two adaptations of the Biogeography Optimization approach are developed. Their generated results were more competitive when compared to a GA, a simulated annealing (SA) and particle swarm optimization techniques. And in [18], M PMs with I resources are allocated to serve J types of pre-emptive jobs. The allocation process is decomposed into time slots. In each one, a routing algorithm, Join-the-Shortest-Queue, is applied to assign new arrival jobs to the servers with the smallest queue length of job type j. And a scheduling approach, the Max Weight Scheduling, is performed to find the configuration of jobs, among the servers' queues, to be served at the current time slot.

3. Comparative study

In this section, a comparative study of the papers presented in the last section will be described. This study is based on the metrics presented in [1]. To each one of these metrics is assigned a value between high, medium and low according to some parameters related to the problem complexity and the results analysis. In addition, this comparative study is also based on the classification adopted in [2].

TABLE I: PERFORMANCE METRICS EVALUATION

Schemes	**Metrics**					**RM Technique**
	Reliability	***Ease of deployment***	***QoS***	***Delay***	***Control overhead***	
Resource allocation						
An SLA-based RA schema in distributed data center [3]	High	Medium	Medium	Low	Medium	SLA-awareness Load-balancing
Optimal application allocation on multiple cloud [4]	Medium	Low	High	Medium	Low	Hybrid / federated cloud
An adaptive resource management scheme [5]	Medium	Medium	High	Low	Low	Energy efficiency SLA-awareness
A scalable and automatic mechanism for RA [6]	High	Low	Medium	Medium	High	Price/Revenue handling
Efficient RA for optimizing objectives of cloud users, IaaS provider and SaaS provider [7]	Low	Medium	Medium	Low	Medium	Price/Revenue handling
Energy-aware RA algorithms [8]	High	Medium	High	Medium	Low	Energy efficiency SLA-awareness
Combinatorial auction-based allocation of VM [10]	High	Medium	Medium	Low	Medium	Price/Revenue handling
Declarative Automated Resource Orchestration [11]	Medium	Low	High	Medium	Medium	Load-balancing
Resource scheduling						
Real-time tasks under uncertainty[12]	Medium	Medium	High	Low	Medium	Energy efficiency
Multi-criteria tasks scheduling in heterogeneous cloud using SA [13]	High	Medium	Medium	Low	Medium	Hybrid / federated cloud

Online optimization of task scheduling in IaaS clouds [14]	Medium	Low	Medium	Low	Low	Hybrid / federated cloud Load-balancing
A distributed framework based on selfish agents [15]	High	Medium	High	High	Medium	Hybrid / federated cloud
Cost efficient task scheduling [16]	Medium	Medium	High	High	Low	Price/Revenue handling
Bio-geography optimization for job scheduling [17]	Medium	Low	High	Medium	Low	Load-balancing
Heavy traffic optimal resource allocation algorithms for cloud computing clusters [18]	Medium	Low	Medium	Medium	Medium	Load-balancing

4. Conclusions and Future Works

In this paper, which can be considered as an extension of the survey presented in [1], a comparative study of a number of papers treating the problems of resource allocation and resource scheduling in cloud computing is detailed. In our next researches, we will focus mainly on resource allocation in multi-clouds.

5. References

[1] S. S. Manvi and G. K. Shyam," Resource management for Infrastructure as a Service(IaaS) in cloud computing: A survey", Journal of Network and Computer Applications, 2013

[2] S. Mustafa, B. Nazir, A. Hayat, A.R. Khan and S. A. Madani, "Resource management in cloud computing: Taxonomy, prospects, and challenges", Computers and Electrical Engineering, Vol. 47, pp. 186-203, 2015.

[3] S. Son, G. Jung and S. C. Jun, "An SLA-based cloud computing that facilitates resource allocation in the distributed data centers of a cloud provider", Journal of Supercomputing , 2013, Volume 64, Issue 2, pp 606-637

[4] S. S. Woo and J. Mirkovic, "Optimal application allocation on multiple public clouds", Computer Networks, 2014.

[5] C. J. Huang, C. TaiGuan, H. M. Chen, Y. W. Wanga, S. C. Chang , C. Yu Li and C. H. Weng, "An adaptive resource management scheme in cloud computing", Engineering Applications of Artificial Intelligence, 2013.

[6] X. Wu, M. Liu, W. C. Dou, L. Gao and S. Yu, "A scalable and automatic mechanism for resource allocation in self-organizing cloud", Peer-to-Peer Netw. Appl., 2014.

[7] C. Li and L. Li, "Efficient resource allocation for optimizing objectives of cloud users, IaaS provider and SaaS provider in cloud environment", Journal of Supercomputing, 2013.

[8] A. Horri, M. S. Mozafari and G. Dastghaibyfard, "Novel resource allocation algorithms to performance and energy efficiency in cloud computing", Journal of Supercomputing, 2014

[9] A. Hameed, A. Khoshkbarforoushha, R. Ranjan, P. P. Jayaraman, J. Kolodziej, P. Balaji, S. Zeadally, Q. M. Malluhi, N. Tziritas, A. Vishnu, S. U. Khan and A. Zomaya, "A Survey and Taxonomy on energy efficient resource allocation techniques for cloud computing systems", Journal of Computing, Springer, 2014.

[10] S. Zaman and D. Grosu, "Combinatorial auction-based allocation of virtual machine instances in clouds", J. Parallel Distrib. Comput, 2013, pp. 495–508

[11] J. T. Tsai, J. Fang and J. Chou, "Optimized task scheduling and resource allocation on cloud computing environment using improved differential evolution algorithm," Computers & Operations Research, vol. 40, 2013.

[12] H. Chen, X. Zhu, H. Guo, J. Zhu, X. Qin and J. Wu, "Towards energy-efficient scheduling for real-time tasks under uncertain cloud computing environment," The Journal of Systems & Software, vol. 99, pp 20–35, 2014.

[13] I. A. Moschakis and H. D. Karatza, "Multi-criteria scheduling of Bag-of-Tasks applications on heterogeneous interlinked clouds with simulated annealing," The Journal of Systems & Software, vol. 101, pp 1–14, 2015.

[14] J. Li, M. Qiu, Z. Ming, G. Quan, X. Qin and Z. Gu, "Online optimization for scheduling preemptable tasks on IaaS cloud systems," Journal Of Parallel and Distributed Computing , vol. 72, pp 666–677, 2012.

[15] F. Palmieri, L. Buonano, S. Vinticinque, R. Aversa and B. Di Martino, "A distributed scheduling framework based on selfish autonomous agents for federated cloud environments," Future Generation Computer Systems, 2013.

[16] S. Su, J. Li, Q. Huang, X. Huang, K. Shuang and J. Wang, "Cost-efficient task scheduling for executing large programs in the cloud," Parallel Computing, vol. 39, pp 177–188, 2013.

[17] S. Kim, J. Byeon, H. Yu and H. Liu, "Biogeography-based optimization for optimal job scheduling in cloud computing," Applied Mathematics and Computation, vol. 247, pp 266–280, 2014.

[18] S. T. Maguluri, R. Srikant and L. Ying, "Heavy traffic optimal resource allocation algorithms for cloud computing clusters," Performance Evaluation, vol. 81, pp 20–39, 2014.