Resources Allocation in Cloud Computing: A Survey

Karima Saidi[1(✉)], Ouassila Hioual[2,3], and Abderrahim Siam[1]

[1] ICOSI Laboratory, Abbes Laghrour University, Khenchela 40004, Algeria
kariming2008@gmail.com, siamabderrahim@gmail.com
[2] Abbes Laghrour University, Khenchela 40004, Algeria
ouassila.hioual@gmail.com
[3] LIRE Laboratory, Constantine2, Constantine, Algeria

Abstract. Cloud Computing is a model in which resources (Computing, Networking, Storage…) are consumed on any utility such as computing power, storage space, servers and applications. With the growing of the Cloud resources demand and the user's number, the need of the quality of service and the resources allocation become a crucial challenge. This paper presents the challenges of resource allocation and some recent contributions to minimize them, hence the need for paper reviews and surveys to identify the area of our research.

Keywords: Cloud computing · Resource allocation · Quality of Service QoS · Deep learning · MCDA

1 Introduction

Various definitions about the Cloud Computing have been given in the literature. In general, we can say that Cloud Computing is a popular trend that use the technology, it attempts to provide easy and inexpensive access to computing resources. In addition, Cloud Computing is a computing model based on an internet. Cloud virtualization technology plays a very important role, allowing resources to be shared, by allocating virtual machines on demand instead of renting physical machines. The importance of virtualization lies in the fact that it allows almost complete isolation between customers who will actually have the illusion that they have just rented a dedicated physical machine (Yazir et al. 2010). There are four types of virtualization by (Akintoye and Bagula 2017): full virtualization, para-virtualization, native virtualization and operating system virtualization. The main issues related to using the Cloud are resource allocation and task scheduling. The last one can be expressed as the allocation of different types of work using existing resources (Manvi and Shyam 2014), and the allocation of resources as the assignment of tasks to virtual machines and the placement of VMs on physical machines PMs.

The IT resource allocated is based on a Service Level Agreement (SLA) (Alhamad et al. 2010) which is a service level agreement between the cloud customer and the service provider. The latter expresses in detail the quality of service (QoS) such as reliability, response time and throughput. These are performance parameters to be

M. Hatti (Ed.): ICAIRES 2019, LNNS 102, pp. 356–364, 2020.
https://doi.org/10.1007/978-3-030-37207-1_37

respected by the service provider, so resource allocation is the effective allocation and planning of resources to achieve the quality of service performance objectives identified by the SLA (Chana and Singh 2014).

In this work, we try to synthesize a set of contributions that focus on solving the problem of resource allocation in the field of Cloud Computing. The rest of the document is structured as follows: In Sect. 2, some basic concepts and challenges of resource allocation in Cloud Computing are introduced. Section 3 will classify the methods used in resource allocation. Section 4 presents recent contributions in this area with orientation of future research. We're going to conclude the paper with a conclusion.

2 Concepts and Challenges

2.1 Some Basic Concepts in RA

To fully understand the principle of resource allocation in cloud computing, we must first introduce what it really means to "allocate resources efficiently and dynamically". These terms are closely defined by a set of parameters (criteria). As shown in Fig. 1, efficiency and dynamism are keywords that include many useful requirements for resource allocation. We present the following main parameters:

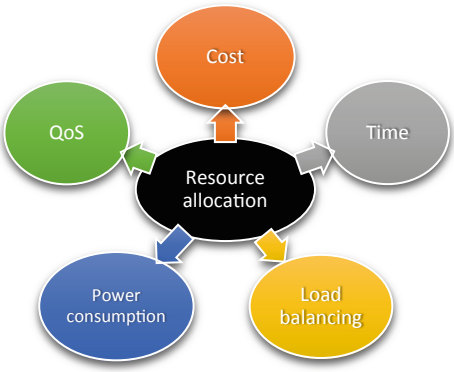

Fig. 1. The parameters of RA

From the perspective of cloud users

- Cost: the allocation of resources must be carried out at a lower cost
- Response time: the allocation of resources must be carried out in a minimum time

For the cloud provider

- Energy consumption
- Load balancing
- The execution times
- QoS quality of service

2.2 The Challenges of Research in RA

Existing problems related to research in the area of Cloud resource allocation that have not been fully resolved according to (Alnajdi et al. 2016) include (see Fig. 2):

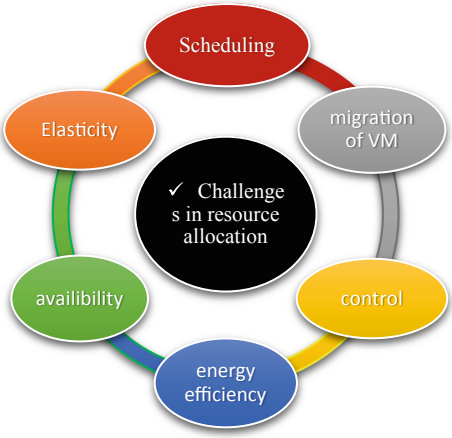

Fig. 2. The different challenges of the resource allocation area

- ***Scheduling***: means the parallel scheduling of tasks in case of increasing demands on resources
- ***VM migration***: it is the user's need to change a provider in order to ensure the high storage of his data
- ***Availability***: in long calculations, it is necessary to propose techniques that automatically deal with the unavailability of resources
- ***Elasticity***: indicates the ability to size and provide resources quickly, i.e. the ability to allocate resources dynamically.
- ***Energy efficiency:*** refers to the amount of carbon emissions released by data centers due to the various and huge IT operations performed in them.
- ***Control***: the need to build a control mechanism that allows resources to be allocated on remote servers.

3 Classification of Methods Used in RA

We can classify the resource allocation methods into two different categories, namely:

(a) The reactive methods that make up the techniques that monitor the effectiveness of resource allocation before deciding on the implementation of an action.
(b) The proactive methods that integrate the prediction of future resource requirements. There are methods based on one model and others without a model.

4 Comparison and Orientation of Future Research

4.1 Comparison

There are many publications in recent years in which the authors discuss some of the challenges of resource allocation. This research presents different algorithms and techniques for resource allocation that have been proposed in a Cloud environment. We present some new contributions between 2016 and 2019 in order to have a clearer vision in our work.

In order to have an analysis of these different research studies based on different parameters, we propose a comparative table (see Table 1).

Firstly, the authors Akintoye and Bagula (2017), have proposed solutions to solve two problems in the Cloud/Fog environment: the first problem is the assignment of tasks to virtual machines. In this case, they use the *HABBP (Hungarian Algorithm Based Binding Policy)* algorithm as a heuristic solution to the problem of linear programming, this allocation solution is one by one where each VM performs only one task and each task (cloudlet) must be assigned to a single VM to minimize the total cost. The 2nd problem is the placement of virtual machines. In this case, the authors used a genetic algorithm called *GABVMP (Genetic Algorithm Based Virtual Machine Placement)* to solve and optimize this problem.

Liu et al. (2017), proposed a hierarchical framework to solve the resource allocation problem at a global level on virtual machines and power management, and at a local level on local servers. This work is based on deep learning (and more precisely Deep Reinforcement Learning DRL). Decision-making is done automatically from the reduction of the state and action spaces for the allocation of global resources. On the other hand, the workload predictor is responsible for providing future workload forecasts to facilitate the local operation of the *LSTM (long short-term memory)* power management algorithm.

The authors Gawali and Shinde (2018), have proposed a heuristic algorithm that efficiently schedules tasks and allocates resources in the cloud. The authors have combined the modified analytical hierarchy process (MAHP), BATS bandwidth aware divisible scheduling, BAR optimization, longest expected processing time preemption (LEPT) and division and conquest methods to perform task and resource planning. The MAHP process allows scientific tasks to be prioritized, and the combination of BATS + BAR optimization methods allows resources to be allocated according to bandwidth constraints and cloud resource load. In addition, LEPT preemption was used to give the status of the virtual machine, and a modified divided methodology to conquer was proposed to aggregate the results after task preemption.

Kumar et al. (2017), have proposed an algorithm based on the combination of two algorithms: the first is the Teaching learning-based optimization algorithm (TLBO) and the second is the Grey Wolves optimization algorithm (GW). The proposed algorithm works more efficiently, compared to others, by balancing time and costs. Theoretically, the proposed algorithm works much better than the other two for task scheduling.

Malekloo et al. (2018), have proposed a multi-objective approach to managing resources in the cloud. It balanced energy consumption with the system's ability to meet QoS quality of service and SLA contract requirements. This strategy is performed

Table 1. Compared recent contributions in RA based on different parameters

Papers	Algorithms Or Approachs	Contributions	Great Challenge	Classify the method	Parameters				
					Cost	Time	Load Balancing	Energy consumption	QoS
Akintoye, S. B. and Bagula, A (2017)	(HABBP) heuristic solution to the linear programing problem Genetic Algorithm Based Virtual Machine Placement (GABVMP)	Used the simulator CloudSim	- Parallel planning of the tasks - Migration of VM to PM	Reactive	✓	✓	✓	×	✓
Liu, N., Li, Z., Xu, J., Xu, Z., Lin, S., Qiu, Q., Tang, G. and Wang, Y (2017)	Deep Reinforcement Learning (DRL),	Used a Global Resource Allocation Framework based on Deep Neural Networks (DNN) and A deep Q-learning framework	- Energy efficiency - Elasticity	Proactive	×	×	✓	✓	✓
Gawali, M. B. and Shinde, S. K (2018)	- MAHP modified analytical hierarchy process - BATS bandwidth aware divisible scheduling - LEPT longest expected processing time preemption	-Cloud Simulator in the experimental phase Uses real Cyber-shake and Epigenomics scientific work-flows as core tasks for the system	Task scheduling	Reactive	✓	✓	×	×	✓
Kumar, P., Yadav, P. S., Bhutani, K., Arora, N., Jain, D. and Dabas, B (2017)	Teaching learning-based optimization algorithm (TLBO) and grey wolves (GW)	Balance the time and the costs	- Tasks scheduling - Elasticity	Reactive	✓	✓	✓	×	✓
Malekloo, M. H., Kara, N. and El Barachi, M (2018)	MACO multi-objective Ant Colony Optimization	Two types of algorithms virtual machine placement and virtual machine consolidation algorithms	- Energy efficiency - VM migration	Reactive	✓	×	×	✓	✓
Gilesh, M. P., Kumar, S. D. and Jacob, L (2018)	Greedy and meta-heuristic algorithm	Virtual datacenter (VDCE) embedding in DCC on the least cost migration using NetworkX library and fnss	- VM migration - Availibility	Reactive	✓	✓	×	×	✓
Wang, W. Jiang, Y. and Wu, W (2017)	Auction mechanism negociation	Allocate virtual machines to PM with a minimum of energy cost	Energy efficiency	Reactive	✓	✓	×	✓	✓
Lin, J. Dai, Y. Chen, X. and Wu, Y (2017)	Machine learning Genetic Algorithm	Software self-adaptation technology e.g. MAPE-K and control loop	Reduce the cost and minimize the resp time	Proactive Reactive	✓	✓	×	×	✓
Jyoti, A. and Shrimali, M (2019)	Multi-agent Deep Learning (MADRL-DRA) Dynamic Optimal Service Load-Aware Service Broker (DOLASB)	Based on load balancing and Service Broker strategy Use the CloudSim simulator	- Elasticity - Tasks scheduling - Energy efficiency	Reactive Proactive	✓	✓	✓	✓	✓
Alsadie, D. Tari, Z. Alzahrani, E.J. and Zomaya, A.Y (2018)	K-means clustring technique	Determine the appropriate VM type based on Google Cloud traces	Energy efficiency	Proactive	×	×	✓	✓	✓

using two types of algorithms: VM placement and VM consolidation algorithms that allow optimal solutions to be found by reducing the total energy consumption of the data center, minimizing the number of active PMs to shut down unused servers, and, in addition, reducing the number of VM migrations.

The authors Gilesh et al. (2018), have proposed a model to address the problem of finding the most cost-effective set of virtual machine migrations in order to minimize the cost of migration and optimize the integration of a virtual data center. The idea was to integrate a set of virtual machine migrations, while respecting cost reduction, into a new virtual data center created in all Cloud data centers. The authors used the Greedy and metaheuristic algorithm to solve this problem.

The authors Wang et al. (2017) have developed an approach to allocating energy-efficient resources by allocating virtual machines to PMs in order to minimize energy costs. This approach is based on two steps: the first is the auction-based allocation of virtual computers and the second was the negotiation-based consolidation of virtual computers using the multiagent system.

The proposed work by Lin et al. (2017), reduces the workload (with minimal response time and cost) associated with system maintenance and configuration. The authors used machine learning to build an effective knowledge model and applied software self-adaptation technology, which is a capability that allows a software system to adjust to respond to frequent changes from external environments. In addition, the authors applied the genetic algorithm to solve the problem of allocating computing resources, which ensures that the solution can be to optimize the configuration of resources. In this work, the virtual machine pool is responsible for managing the number of virtual machines, creating and closing virtual machines at the right time.

The authors Jyoti and Shrimali (2019), considered load balancing and the Service Broker strategy as two main areas to solve the problem of resource allocation. First, a multi-agent deep learning model (MADRL-DRA) was used. In this model, the local user agent (LUA) is used to predict the environmental activities of the user task and allocate the task to the virtual machine (VM) based on priority. Then, a load balancing (LB) is performed in the VM, which increases the flow rate and reduces the response time of the resource allocation task. Secondly, DOLASB (Dynamic Optimal Service Load-Aware Service Broker) is used in the GUA (Global User Agent) to plan the task and provide services to users based on the cloud brokers available to minimize the costs of cloud customers and at the same time generate a profit for Cloud Service Broker CSB. Finally, the authors proposed the BD-MIP algorithm that provides an optimal solution to the problems of optimizing multi-service configuration, virtual machine allocation and CSB.

The important contribution of Alsadie et al. (2018) is to design an approach to find virtual machines of the appropriate size to optimize resource utilization, thereby reducing energy waste in data centers. The authors used the K-means Clustering technique. Energy efficiency was a major challenge in the proposed approach, it allowed fewer VM instances to be used and proved to be able to reduce the number of rejected tasks.

4.2 Orientation of Future Research

The allocation of resources in Cloud Computing is one of the most relevant issues to be addressed. In this research, we try to guide our future work by citing some challenges in this area.

After a thorough analysis of the existing work, several gaps and disadvantages were identified. For (Akintoye and Bagula 2017), the proposed allocation model does not address the case where several tasks are assigned to a single virtual machine, nor does the placement of virtual machines study load balancing in each physical machine. On the other hand, (Jyoti and Shrimali 2019), the proposed model is based on load balancing and service broker to solve the challenges of scheduling tasks in parallel. In addition, addressing the reduction of energy waste has proven to be a major challenge for cloud researchers to optimize the use of resources according to real needs and to balance the workload in a suitable way. For example, the authors (Liu et al. 2017), (Alsadie et al. 2018) proposed models to determine the type of VM appropriate to the IT resource requirements of the task group. Among the parameters that have been taken into consideration in (Lin et al. 2017) the prediction of response times using machine learning. Other parameters were treated to increase the efficiency of the allocation and other areas were introduced, such as (Liu et al. 2017), (Jyoti and Shrimali 2019) introduced deep reinforcement learning (DRL) and obtained more precise results.

On the other hand, among the approaches that transfer VMs to PMs in the same Cloud considering the cost of migration, we can cite the works of (Wang et al. 2017), (Malekloo et al. 2018). There are other research studies by (Jyoti and Shrimali 2019), (Gawali and Shinde 2018) that have focused on solving some challenges such as parallel task scheduling.

(Kumar et al. 2017) and (Pradhan et al. 2016) proposed changes to resource allocation algorithms by satisfying client requests and reducing wait times. (Pradhan et al. 2016) has modified the oldest algorithm, the Round Robin, which is a simple step in obtaining an optimal planning model. In addition, several meta heuristic algorithms have been combined, as well as several mechanisms and techniques are used (see Fig. 3) according to the table above to allocate resources efficiently and meet users' expectations.

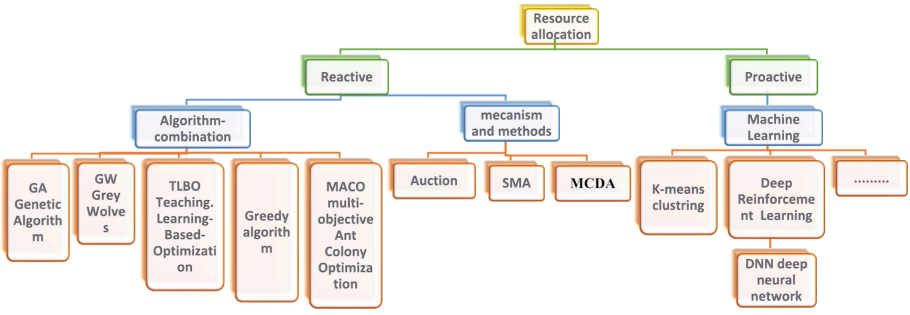

Fig. 3. The taxonomy of resource allocation methods

Our work is based on the combination of MCDA multi-criteria decision support methods as (Gawali and Shinde 2018) but our role is to help a decision-maker to select one of several alternatives based on decision criteria, and in the field of machine learning more specifically preference learning or possibly deep learning.

We choose the field of artificial intelligence because it has given almost zero chances of errors and failure rates, and it also allows for high accuracy in the allocation of resources in Cloud Computing (Madni et al. 2017). Currently, methods of learning and automatic preference prediction are among the most recent research areas in disciplines such as machine learning. The sub-domains of machine learning are used because they allow the rapid and automatic creation of models capable of analyzing large and complex data and obtaining faster and more accurate results, even on a very large scale (Jyoti and Shrimali 2019). The objective of using the two sub-domains is to build a decision model based on preferences (Jyoti and Shrimali 2019).

We can classify our problem as classification problems that determine them through supervised learning in which the entry and exit spaces are clearly distinguished from each other. In this type of problem, input instances are mapped to preference models (Kotsiantis et al. 2007).

5 Conclusion

The number of articles published in recent years which are based on the cloud has been enormous, especially those that address the issue of resources allocation. And many more are being revised and published due to the increasing number of resource requests in the Cloud environment. With the large number of articles, it is difficult for new researchers in the field to identify potential areas for discussion. This study aimed to address this challenge.

References

Yazir, Y.O., et al.: Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis. In: Proceedings of the 3rd International Conference on Cloud Computing IEEE, pp. 91–98, July 2010

Akintoye, S.B., Bagula, A.: Optimization of virtual resources allocation in cloud computing environment. In: Proceedings of the Africon IEEE, pp. 873–880, September 2017

Manvi, S.S., Shyam, G.K.: Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. J. Network Comput. Appl. IEEE, (41), 424–440 (2014)

Alhamad, M., Dillon, T., Chang, E.: Conceptual SLA framework for cloud computing, In: Proceedings of the 4th International Conference on Digital Ecosystems and Technologies IEEE, pp. 606–610, October 2010

Chana, I., Singh, S.: Quality of service and service level agreements for cloud environments: issues and challenges. In: Mahmood, Z. (ed.) Cloud Computing. Computer Communications and Networks, pp. 51–72. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10530-7_3

Alnajdi, S., Dogan, M., Al-Qahtani, E.: Asurvey on resources allocation in Cloud computing. Int. J. Cloud Comput. Serv. Archit. IJCCSA **5**(6), 1–11 (2016)

Liu, N., et al.: A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning. In: Proceedings of the 37th International Conference on Distributed Computing Systems ICDCS IEEE, pp. 372–382, June 2017

Gawali, M.B., Shinde, S.K.: Task scheduling and resource allocation in cloud computing using a heuristic approach. J. Cloud Comput. **1**(7), 1–16 (2018)

Kumar, P., Yadav, P.S., Bhutani, K., Arora, N., Jain, D., Dabas, B.: Allocating resource dynamically in cloud computing, In: Proceedings of the International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) ICTUS IEEE, pp. 249–254, December 2017

Malekloo, M.H., Kara, N., El Barachi, M.: An energy efficient and SLA compliant approach for resource allocation and consolidation in cloud computing environments. Sust. Comput. Inf. Syst. **17**, 9–24 (2018)

Gilesh, M.P., Kumar, S.D., Jacob, L.: Bounding the cost of virtual machine migrations for resource allocation in cloud data centers. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 201–206, April 2018

Wang, W., Jiang, Y., Wu, W.: Multiagent-based resource allocation for energy minimization in cloud computing systems. Trans. Syst. Man Cybern. Syst. IEEE **2**(47), 205–220 (2017)

Lin, J., Dai, Y. Chen, X, Wu, Y.: Resource allocation of cloud application through machine learning: a case study. In: Proceedings of the International Conference on Green Informatics ICGI IEEE, pp. 263–268, August 2017

Jyoti, A., Shrimali, M.: Dynamic provisioning of resources based on load balancing and service broker policy in cloud computing. Cluster Comput., 1–19 (2019)

Alsadie, D., Tari, Z., Alzahrani, E.J., Zomaya, A.Y.: Dynamic resource allocation for an energy efficient VM architecture for cloud computing. In: Proceedings of the Australasian Computer Science Week Multiconference on ACSW 2018 ACM, Brisband, Queensland, Australia, pp. 1–8, January 2018

Pradhan, P., Behera, P.K., Ray, B.N.B.: Modified Round Robin Algorithm for Resource Allocation in Cloud Computing. In: Proceedings of the International Conference on Computational Modeling and Security: Procedia Computer Science, no. 85, pp. 878–890 (2016)

Madni, S.H.H., Latiff, M.S.A., Coulibaly, Y., Abdulhamid, S.M.: September, Recent advancements in resource allocation techniques for cloud computing environment: a systematic review. Cluster Comput. **3**(20), 2489–2533 (2017)

Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. Emerging Artif. Intell. Appl. Comput. Eng. **160**, 3–24 (2007)