ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
(IJARTET) Vol. 5, Special Issue 14, April 2018

A Survey on Resource Allocation Techniques in Cloud Computing

Lijin P , Dr. S., SivaSathya, .K S Guruprakash,
Pondicherry University
Puducherry

ABSTRACT

Cloud computing has become the most popular technology that provides a fast and cost-effective way to configure and allocate resources to the users on demand. The user gets to pay for every service being used from the cloud. It is vey important to consider the user requirements while allocating resources to the user. Virtualization technology is the core of cloud computing. Virtual machines are created from a set of physical machines and these VMs are allocated to the user based on their requirements. The Cloud user has to provide accurate parameters during configuration in order to avail the resources effectively. This paper presents a survey on different resource allocation techniques for cloud.

Keywords—resource allocation,virtulization,optimization

I. INTRODUCTION

CLOUD Computing is the recent technology that provides a distributed resource access to the user based on their requirements. It is capable of providing multiple resources such as servers, storage etc, that can be accessed on demand over internet as medium. Resources can be allocated on payment basis. There are different service models provided by cloud service provider which are discussed below. There are some cloud computing providers, such as Amazon Web Services (AWS), Google Compute Engine (GCE) and IBM who offer cloud computing services with high availability and scalability. Users can run the resources (i.e., cloud servers) as needed, and pay for the service they have used. Cloud providers provide multiple servers with different configurations of CPU capacity, memory, network capacity, disk I/O performance, and disk storage size. Cloud provides different configurations of server types. In each type, server, capacity, memory and performance will differ.

Table 1
A subset of AWS EC2 instance types and pricing (Amazon EC2 Pricing, 2015).

Type	ECU	Memory (GB)	Price
m3.medium	3	3.75	$0.07/h
m3.large	6.5	7.5	$0.14/h
m3.xlarge	13	15	$0.28/h
m3.2xlarge	26	30	$0.56/h
c3.4xlarge	55	30	$0.84/h
c3.8xlarge	108	60	$1.68/h

Table 2
A subset of GCE machine types and pricing (Google Compute Engine, 2015).

Type	Virtual cores	Memory (GB)	Price (dollar/h)
n1-standard-1	1	3.75	0.07
n1-standard-2	2	7.5	0.14
n1-standard-4	4	15	0.28
n1-standard-8	8	30	0.56
n1-standard-16	16	60	1.12

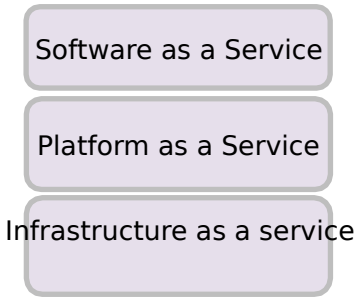
ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
(IJARTET) Vol. 5, Special Issue 14, April 2018

Figure 1: AWS EC2 server types and GCE machine types and pricing is given in table 1

Every user has certain requirements such as cost, storage space, bandwidth, latency, reliability, availability and so on. Cloud providers provide plenty of resources with different configurations as packages. The packages are assigned to the user applications according to requirements provided. There might be a chance of selecting a high configured resource for a small application or low configured resource for large application leading to resource under utilization or over utilization problem. So analysing the user requirements is a major challenge in cloud. To satisfy the users' need, package selection plays an important role in cloud. Another challenge in cloud is to select the best package that match the current user requirement. So this problem is looked upon as an optimization problem. This paper reviews the literature of different resource allocation and optimization techniques in cloud.

A. Cloud Architecture

Cloud architecture is defined by the structure of the system which is composed of middleware, resources, and software components. The architecture of cloud computing could be divided in to three as follows SaaS, PaaS, IaaS.



These three layers describe the services provided by the cloud.

1) Software as a Service (SaaS)

This model distributes and hosts applications over the internet according to the requirements provided by the user. It is the main layer among the three. Cloud provider charges them for the amount and time the software is provided.

2) Platform as a Service (PaaS)

This service model provides an environment where applications can run their task. Developers create and deploy applications on the cloud. Cloud providers provide these environment in which user can work on.

3) Infrastructure as a Service (IaaS)

Cloud provider manages storage, memory etc. to the user. Resources are allocated to the user and the users are charged for the service they have used so far.

B. Cloud Deployment Model

There are several deployment models which describes how the cloud resources are accessed.

1) *Public cloud:*A public cloud accessible to any entity. The cloud owned by an organization or company is an example for public cloud. A public cloud is less secure.
2) *Private cloud:*It is not accessible to the public. Typically owned by a user of a cloud or a specific entity.Private cloud can't be accessed without authorized access.
3) *Community cloud:*community cloud is the cloud shared by two or more entities such as schools, university, etc.
4) *Hybrid cloud:*Hybrid cloud is the mixture of all the three clouds.

C. Importance of Resource Allocation

In cloud computing, Resource Allocation (RA) means assigning available resources to the cloud applications according to the requirements provided by the user. Resource allocation will not be efficient if the allocation is not managed efficiently. It is very important that resource allocation should

be optimum so as to meet user's requirements. An optimal Resource allocation Technique (RAT) should avoid the following criteria as follows

a) **Resource conflict:** This situation happens when two entities try to access the same resource at the same time.
b) **Scarcity of resources:** It occurs when resources are not available.
c) **Resource fragmentation:** Different resource fragments could be generated during the resource allocation, Due to this, computational resources are not fully utilized.
d) **Over utilization:** Allocating more resources than the required number of resources
e) **Under utilization:**Allocating less number of resources than required
f)

II. SURVEY FINDING

The resource allocation techniques could be divided into several categories based on different perspectives. The classification made in this paper is as follows.

(a) QOS based resource allocation
(b) Cost based resource allocation
(c) SLA based resource allocation
(d) Priority oriented resource allocation
(e) Agent oriented resource allocation
(f) Auction based resource allocation
(g) Energy aware resource allocation
(h) Resource Optimization

a. QOS based resource allocation

Various resource allocation techniques were proposed by different authors to improve the Quality of Service (QoS) goals. A framework for resource allocation and optimum allocation of resources was discussed in [1].This paper presents a method to automatically allocate resources. In this paper, the authors proposed a modelling framework for resource optimization by considering the QOS goals. The framework is composed of the following phases: modelling, model analysis, test automation, code generation, and optimization techniques. ROAR uses a specification language called GROWL, an web application optimization language. GROWL use the configuration space and QOS goal information. The specification is translated into test plan which is the input to the controller module of ROAR.

The ROAR handles the optimization processes such as resource allocation, application deployment, load test execution, performance matrix collection and performance model analysis. The cloud resource utilization matrix and QOS matrix is monitored by the ROAR. The utilization matrix is aligned into QOS matrix and the aligned matrix is inserted into a test state model and model analysis is performed. The process is continued for each resource configuration until a suitable resource configuration is found. Code generation is performed for allocating the resource configuration. Docker contains application configuration or the application itself. Applications are deployed from Docker to servers in the cloud platform. GROWL is a language which is translated in to a specific test plan. The Test plan is an XML file which can be tested by load testing tools like jMeter. Test plan contains certain parameters. Each time performance is analysed with respect to QOS goals. The process is continued until optimized resource configuration is found.

K.S.Guruprakash,S.SivaSathya[2][3] have proposed a method for resource allocation in cloud using web server log file. QOS goal is identified from log file. This method uses web server Log file for finding the requirement. The QOS goals considered here are bandwidth, reliability, availability, computational capacity, usability, correctness, reliability etc. Here the authors focused on automatic user requirement identification based on Log file in order to allocate resource to the consumers which provides better QOS.

b. Cost based resource allocation

There are some cost based resource allocation methods proposed in literature. The author in [4] proposed a method to optimize Resource allocation Cost in Cloud Computing. Cloud provides reservation plan and on-demand plans to the user. Reservation plan price is low when compared to on-demand plan since the users have to pay in advance. With reservation plan, total resource allocation cost can be reduced. User future demands are not considered in this paper and uncertainty is associated with provider resource prices. To solve this problem, an algorithm is proposed and it is applicable to the long-term plan.

The authors in [5] introduced a method for resource allocation in co-operative cloud market. The goal is to reduce the cost while allocating resource to the user. This paper proposes an algorithm to improve the cost. The basic idea of the market-oriented cloud was given by Buyya et al.[21].

c. SLA based resource allocation

There have been some researchers [6][7][8] who have discussed resource allocation based on SLA in the cloud. A client has Service Level Agreement (SLA) in such situations the resource allocation should satisfy SLA requirements. In [6], SLA-Based Resource Allocation framework is presented. They considered multiple cloud providers. This approach uses Nash equilibrium concept of game theory to allocate resources to the users. Resource allocation problem among multiple cloud service provider forms a game. Each CSP has its own servers and each request is sent to appropriate servers.
The authors in [7] discussed reallocation of resources. The authors proposed a architecture of resource reallocation. The architecture is divided into two parts one part includes Customer and datacentre which handles the SLA contract and resource scheduling and in the next part, the control chart is presented to detect SLA violation for host performance. A utility function is used in the evaluation mechanism.

The authors in [8] proposed a method to allocate Resources for SaaS cloud. The authors have proposed algorithms for minimizing the infrastructure cost and SLA violation. There are two actors in this system, SaaS providers and customers. QOS goals are assigned to the resources based on the requirements of the user.

d. Priority based resource allocation

In [9], the author presents a method to execute a high priority job. Also, an idea to reuse VM is presented. New VMs are not created for the running of a newly arrived job. The proposed algorithm is able to run a high priority task or job by suspending a low priority job. The suspended job is resumed in a VM if that VM completely executed one job.

DilipKumar.M et.al[10] has proposed a model for resource allocation. They considered peer to peer cloud.The request for a large amount of CPU is taken as higher priority job. K-means algorithm is used to classify the tasks into high, medium and low priority sets and the task is sorted in the task list based on priority. In the proposed approach, the price is calculated based on the current demand for a resource and its availability. In high contention across the network, resources are discovered from peer clouds.

ShubhakankshiGoutamet.al[11] proposed a method to allocate resources with Fault Tolerance. Pre-emption starts from low priority task to high priority task. An advanced reservation scheme is presented here. The algorithm forms a task list based on priorities and priority based task scheduling is performed. Advanced reservation is given to task having high priorities. The algorithm is also applicable to fault tolerance.

e. Agent oriented resource allocation

Aarti Singh et.al[21] has presented a method for optimizing the resource allocation cost. A framework is presented here which uses the agents to connect each other. An allocation algorithm is proposed which minimizes the cost.In [13], the authors have presented an approach that allocates resources to the user.

The Agent is used to connect the different heterogeneous cloud providers with the cloud users. The user's needs are considered. Cloud users and cloud providers represents the agent. There are a number of actors in the cloud in the above architecture. Each cloud provider has several datacentres, each datacentre is composed of a large number of physical server, and multiple virtual machines are created on the physical servers. Each layer has its own functions. The agent coordinator will find the right selection that satisfies the user request. For achieving this, the method uses multidi mensional comparison of each resource requested by the user and the available resources of the provider.

f. Auction based resource allocation

The author in [14] used auction method to allocate resources in the market-oriented cloud. The Auction is the method for selling and buying the resources by considering the user requirements or choices. The important steps in the method are optimal price det ection and finding the winner of the auction. Modified Paddy Field Algorithm (PFA) is used to detect the best consumer and provider. This paper focused on participant's preference in the auction for selling and buying resources with the satisfied price of chosen resources. The auction method used here is a combinatorial auction (both consumers and providers are participating). The identifiers can find a best consumer and producer from the auction and eliminate dishonest participants from the auction. The society can approach the system when the system participants are trusted. The system is responsible for the provision of high-quality computing resources. An algorithm has been developed for finding the winner of the auction.

The research work in [15] has proposed a method to ease the Resource Allocation in Cloud. They have presented a new method for resource pricing so that the complexity can be reduced. Yeongho Choi et.al[16] has proposed an Approach for optimizing the Resource Allocation for IOT. An auction-based model is used in this paper. The auction is done on a group of cloud resources. SLA is classified into job-based and class-based technique. In class-based SLA and job based SLA QOS is measured. SLA penalty occurs if service quality affects QOS. Here the research is about job-based SLA since it is most robust. An SLA contains an agreement of various performance matrixes. In this paper SLA is defined in terms of deadline along with the execution time of each job.E.IniyaNehruet.al[17] has proposed Auction Based Dynamic Resource Allocation in Cloud. Combinatorial auction is used to allocate a set of resources. When there is resource contention, the cloud manager will call for the auction of resources among the users. Users then start to send bids for the resources. Users with the highest bid will be the winner. Resources are allocated on the basis of winners sorted list.

g. Energy aware resource allocation

Energy consumption is a major concern in the mobile cloud. There have been many research works [18] [19][20] towards this problem. The research in [18], proposes a framework to reduce the energy loss in cloud datacentres. The proposed framework does the following

i) Predicts the number of virtual machine requests, and the amount of resources.
ii)The number of physical machines (PMs) is estimated so that the user requests are satisfied.
iii) Improves energy consumption of cloud datacentres.

The proposed framework consists of data classification, workload prediction module, and power management module. In[19], the author presented an approach to allocate resources by reducing the energy consumption. A method for Optimizing the Virtual Machine Migration has been presented here. The authors have used time series based forecasting method for prediction of CPU utilization and virtual machine migration. In [20], the research is based on IaaS cloud. They have considered IaaS cloud. They have proposed two models, server power model to optimize the power and resource wastage model to reduce resource wastage of server in IaaS cloud. In resource wastage model, VM request from different

users of different resource specification has been taken and in server power model, energy consumption is considered.

h. Resource optimization

The optimization of cloud resources has been widely researched recently. There have been some research papers that describe cloud resource optimization. The author in the paper [23] has proposed a model for resource optimization and to satisfy the user needs. Particle swarm optimization algorithm uses Map reduce for the optimization of cloud resources. Particle Swarm Optimization (PSO) is an optimization technique, which is about the knowledge of a group of birds moving towards a final goal through communication as well as independent searching. Each particle learns and communicates with each other and they update their new position. Each of the particles has personal best and apart from that, there is one global best. The particle moves with some velocity. The cloud resource allocation optimization model has been subdivided into 3 types which fall into the following categories: time, revenue and user satisfaction. Time model has further classification. Queuing theory is used to measure the average response time in response time model. The proposed algorithm is used to identify the tasks and to allocate resources in the cloud.

ZhengqiuYang1,[24] has used ant colony algorithm to allocate resources to the user. The algorithm is based on ant's foraging behaviour. The algorithm simulates the food searching behaviour of natural ants and adopts the methods to solve the NP problem. The algorithm is a new heuristic optimization algorithm. The algorithm is used to reduce the length of paths. The task agent collects task from the user and ACO algorithm is performed on that. The weight of pheromones is found which is used for communication. Nodes are selected by the Ant colony algorithm. This strategy is used for resource allocation in cloud.

Xiao-long Zheng et.al[25] has proposed Pareto based fruit fly optimization algorithm (PFOA) to allocate resources and task scheduling in cloud. A heuristic is proposed to initialize the population. Non-dominated genetic algorithm also called MOO(multi- objective optimization) is used to update the population. Search based on smell and vision is performed using the algorithm. Searching is continued until improved smell value is not available.The authors in [12] proposed a method to optimize the resource allocation by reducing the Cost. In this paper they have presented an optimization algorithm to reduce the deployment cost while satisfying (QOS) requirements. The algorithm is designed for web application deployment in cloud data center. They have taken input parameters from the cloud provider and web application. The algorithm is designed for choosing the most suitable combination of cloud resources.

Fan-Hsun Tseng et.al[26] has used Genetic algorithm(GA) to predict the resources in the cloud datacentre. Future resources are predicted based on the historical data in past time slots. The prediction is done using GA algorithm. Virtual machines are allocated by using the prediction results.

Table 2: Optimization algorithm

Optimization Algorithm	Used To Solve
Ant Colony Optimization Algorithm Based on Particle Swarm Optimization[24]	Resource Allocation
Fruit Fly Optimization Algorithm(PFOA)[25]	Task Scheduling ,Resource Allocation, Minimize Cost
PSO-ACO Algorithm Based On MapReduce[23]	Resource optimization
Multi Objective Genetic Algorithm(GA)[26]	Resource Prediction, multi objective Resource Allocation

III. TABLE OF COMPARISON

Ref:No	Methodology Used	Advantages	Disadvantages

QOS Based Resource Allocation				
[1]	Developed resource optimization allocation and recommendation system, performed code generation, optimization, test creation	Process continues until QOS goals are satisfied	Manual process	
[2]	SMI parameters extracted from log file, work load data is taken from user data analysis is performed on log data	Automated resource allocation	Applicable for only single cloud service provider, user have to manually give workload data	
[3]	A frame work is created for cloud service recommendation and prediction QOS parameters are taken from log file .A perfect mapping is done on parameters	Automated resource allocation	Considering all parameters is a problem	
[12]	Multi-objective optimization algorithm	Users (QOS) parameters are satisfied, optimization of many objectives like cost etc.	Applicable for only single service provider	
Cost Based Resource allocation				
[4]	OCRP algorithm with stochastic programing model, applied decomposition algorithm	Applicable for Multiple cloud service providers	Uncertainty in stochastic programing model	
[5]	Most Cost Effective Providers' Resources First(MCEPRF) algorithm	Multiple cloud service providers are taken	QOS goals are less satisfied	
SLA Based Resource Allocation				
[6]	Uses game theory Nash equilibrium concept	Resource allocation among multiple cloud service providers	Applicable only for single cloud	
[7]	Proposed an architecture based on SLA requirement	Prevents SLA violation	Performance is reduced while migrating VM, SLA requirements are not satisfied.	

[18]	Resource provisioning framework, clustering, workload prediction, and power management	Energy savings and high utilization	Not a optimal solution
[19]	Time series based forecasting method, virtual machine migration	Energy savings	QOS goals are not satisfied
[20]	Server power model, resource wastage model, VRAS strategy	Reduce resource wastage, optimize power	QOS goals are not satisfied
[23]	PSO-ACO algorithm	Optimization, QOS goals are satisfied	Applicable only for single cloud
Optimization			
[24]	Ant Colony algorithm,PSO algorithm	Future paths are predicted	Not an optimal solution
[25]	Pareto Based Fruit Fly Optimization Algorithm(PFOA)	Task Scheduling, Resource Allocation, Minimize The Cost	Applicable only for single cloud
[26]	Multi Objective Genetic Algorithm(GA)	Resource Prediction, multi objective Resource Allocation	Applicable only for single cloud

IV. CONCLUSION

Several resource allocation techniques for cloud has been discussed in this paper. QOS is a major concern in resource allocation. In Quality of service based technique, requirements are taken from users and resources are allocated accordingly. Service level agreement provides awareness to users involved in processing. Efficient methods are used so that SLA violation can be removed. Priority based resource allocation is used to allocate resources with varying priority. In agent oriented resource allocation, agents are involved to connect with different cloud provider and users' resource selection. In auction based resource allocation an offer price or bid is put on different resources. Resources with the highest bid will be the winner and the resources are allocated accordingly. Several resource optimization techniques are also discussed.

VI. REFERENCES

[1] Yu sun a,jules white b,'' ROAR: A QoS-oriented modeling framework for automated cloud resource allocation and optimization",june 2016,vol 116

[2] K. S. Guruprakash1 and S. Siva Sathya2," Indian Journal of Science and Technology, Vol 9(30), DOI: 10. 17485/ijst/2016/v9i30/99011, August 2016.

[3] K. S. Guruprakash1 and S. Siva Sathya2," *International Journal of Applied Engineering Research* ISSN 0973-4562,Number 16 (2015) pp 37770-37776 © Research India Publications. http://www.ripublication.com

[4] Sivadon Chaisiri, Student Member, IEEE,"Optimization of Resource Provisioning Cost in Cloud Computing", *IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 5, NO. 2, APRIL-JUNE 2012.*

[5] K Hemant Kumar Reddy1, Geetika Mudali1 and Diptendu Sinha Roy2*,"A novel coordinated resource provisioning approach for cooperative cloud market",
Reddy et al. Journal of Cloud Computing: Advances,Systems and Applications (2017) 6:8 DOI 10.1186/s13677-017-0078-z

[6] Yanzhi Wang "A Game Theoretic Framework of SLA-Based Resource Allocation for Competitive Cloud Service Providers"/2014 Sixth Annual IEEE Gree n Technologies Conference

[7] Jen-Hsiang Chen "Resource Reallocation based on SLA Requirement in Cloud Environment"/2015 IEEE 12th International Conference on e-Business Engineering

[8] Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya, "SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments", 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing.

[9] Saraswathi AT a, Kalaashri.Y.RA b, Dr.S.Padmavathi c1," Dynamic Resource Allocation Scheme in Cloud Computing"

[10] Dilip Kumar S. M,Naidila Sadashiv,R. S. Goudar,"Priority Based Resource Allocation and Demand Based Pricing Model in Peer-to-Peer Clouds"/2014 IEEE

[11] Shubhakankshi Goutam,Arun Kumar Yadav,"Preemptable Priority Based Dynamic Resource Allocation in Cloud Computing with Fault Tolerance"/20 15 IEEE

[12] Seyedehmehrnaz Mireslami, "Simultaneous Cost and QoS Optimization for Cloud Resource Allocation"*IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 14, NO. 3, SEPTEMBER 2017*

[13] Mohamed El-kabir Fareh, Okba Kazary, Manel Femmamy, and Samir Bourekkache,"An Agent-Based Approach for Resource Allocation in the Cloud Computing Environment"/2015 IEEE

[14] Neethu B, K.R RemeshBabu,"Dynamic Resource Allocation in Market Oriented Cloud using Auction Method"/2016 IEEE

[15] Lena Mashayekhy,Mahyar Movahed Nejad,"An Online Mechanism for Resource Allocation and Pricing in Clouds"/*IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 4, APRIL 2016*

[16] Yeongho Choi1 and Yujin Lim2, " Approach for Resource Allocation on Cloud Computing for IoT", International Journal of Distributed Sensor Networks Volume 2016, Article ID 3479247, 6 pages http://dx.doi.org/10.1155/2016/3479247

[17] E. Iniya Nehru, RanjithBalakrishnan, "Auction Based Dynamic Resource Allocation in Cloud"/2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[18] Mehiar Dabbagh,Bechir Hamdaoui,Ammar Rayes, "Energy-Efficient Resource Allocation and Provisioning Framework for Cloud Data Centers"/*IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 12, NO. 3, SEPTEMBER 2015*

[19] Anita choudhary, M.C Govil,Girdhari Singh,"Energy Efficient Resource Allocation Approaches with Optimum Virtual Machine Migrations in Cloud Environment"/2016 fourth International conference on Parallel,Distributed and Grid computing.

[20] Yaohui Changl,Chunhua Gui, Fei Luol,"A Novel Energy-Aware and Resource Efficient Virtual Resource Allocation Strategy in IaaS Cloud"/2016 2nd *IEEE International Conference on Computer and Communications*

[21] Aarti Singh,"A novel agent based autonomous and service composition framework"/https://doi.org/10.1016/j.jksuci.2015.09.001

[22] Rajkumar Buyya, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility" Future Generatio n Computer Systems, Volume 25, Number 6, ISSN: 0167-739X,june 2009

[23] Zhi-guang Ao,"Research on Cloud Resource Optimization Model Based on Users Satisfaction"/2016 IEEE

[24] Zhengqiu Yang1,"study on cloud resource allocation strategy based on particle swarm ant colony optimization algorithm"/2012 IEEE

[25] Xiao-long Zheng,"A Pareto based Fruit Fly Optimization Algorithm for Task Scheduling and Resource Allocation in Cloud Computing Environment"/2016 IEEE

[26] Fan-Hsun Tseng,"Dynamic Resource Prediction and Allocation for Cloud Data Center Using the Multiobjective Genetic Algorithm"/2017 IEEE