A Survey on Resource Allocation Algorithms in Cloud Environment

Ayesha Rani [1], Mahendra Singh sagar [2]

[1]*College of Computing Sciences & Information Technology, TMU, Moradabad*
[2]*College of Computing Sciences & Information Technology, TMU, Moradabad*
[1]ayeshapasha95@gmail.com
[2]mahendra.singh12jan@gmail.com

Abstract— **Cloud Computing is a novel technology for storing and accessing data and programs over the internet. The major issues in cloud computing environment are security because data is stored at different places which can even be the entire globe. User is very much concerned about the data security in cloud technology. This review paper provides brief study of types of cloud and resources allocation techniques.**

Keywords—Cloud computing, Resources allocation, Cloud Services, Infrastructure.

I. INTRODUCTION

Cloud Computing is an on demand services provided to user, business and as well to government agent. Cloud Computing provides scalable and on demand services to users irrespective of their physical area. Users can access cloud computing services anywhere and at any time. Low costs, increased operational efficiencies, scalability, flexibility and many more are the advantages of cloud computing. To provide better services to the users, the services should be delivered in efficient manner. The explanation of "cloud computing" from the National Institute of Standards and Technology (NIST) is that cloud computing enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]. Cloud is an environment that has data centers that provides services over the Internet to satisfy user's requirements [2]. There are three well known models in cloud computing regarding different services. These are Software as a service (SaaS), Platform as a service (PaaS), and Infrastructure as a service (IaaS). SaaS can be described as a process by which Application Service Provider (ASP) provide different software applications over the Internet. This helps the customer to install and operate the application on own computer and use them. In PaaS many set of software programs are present. Customer gets the access to use those programs and help them to organize their own software and application in the cloud [3]. Infrastructure as a service (IaaS) refers to the sharing of hardware resources for executing services using Virtualization technology. Its main objective is to make resources such as servers, network and storage more readily accessible by applications and operating systems. Thus, it offers basic infrastructure on-demand services and using Application Programming Interface (API) for interactions with hosts, switches, and routers, and the capability of adding new equipment in a simple and transparent manner [3].Cloud computing will give services to be consumed easily on demand. Cloud computing is very important for the IT applicants, however, there are still some problems to be solved for users and business enterprises to store data and access applications in the cloud computing environment.

Characteristic of cloud computing are given below:

A. **On-demand self-service:** These are those services which are provided on the biases of user demand when user send required for the cloud resources then cloud computing allow them to access it. This user can accesses cloud services through an online control panel. The cloud services are provided on demand of the users and users can access what they require.

B. **Resource pooling:** Resource pooling in cloud computing environments is used to describe a situation in which providers serve multiple clients, customer or "tenants" with provisional and scalable services. These services can adjust to suit each client's needs without any changes being apparent to the clients and end user. The location of services provider is independence, user do not have any control or knowledge over the location of services provider but can specify location at higher level of abstraction (eg: country, state, or datacentre) [3]. Examples of resources include storage, processing, memory, and network bandwidth.

C. **Broad network access:** Cloud Computing services can be accessed from anywhere and at any time through the internet. Users can get services using laptops, PCs or even mobile phones. The basic requirement to access Cloud Services is the Internet.

D. **Measured service:** These are pay as you go services. We have to pay for those services which we demand. Resources usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

II. TYPES OF CLOUD DEPLOYMENT MODEL

A. **Private cloud**: The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises. The data that is presence in cloud is shared by only the enterprise users or which control the cloud. It is more secure then public cloud.

B. **Public cloud**: Public cloud services are pay you go services on the biases of the services provided. These services are offer over the internet. These services are on the demand of user and it is largely scalability. It is managed by the company that provides services to us. It have virtually unlimited resources. But these services are not secure. These services can be used by any one through the internet and he have to pay for these services.

C. **Hybrid cloud**: Hybrid cloud is associated of both private cloud and the public cloud. It uses the properties of both the cloud. It takes the security features of private cloud and takes the scalability feature from public cloud. So the user who wants security and the scalability then can used hybrid cloud. Hybrid cloud graphical representation is given below. Its consists of public cloud and private. It combined both the cloud[4]. Private cloud is uses by the customer of any organisation only unauthorized person cannot use the private cloud. Public cloud is used by any one. User can used it as pay you go services. Hybrid model is more popular than both the public cloud and private cloud because it reduces cost and provides better security. And it have very large amount of data. So it is used by many people.

D. **Community cloud**: The community cloud infrastructure is own or used by specific community of user. In this community two or more organization comes and shared one community cloud. The community cloud

consists of mission, security, requirement, policy and compliance, consideration. It is owned, managed, operated by one or more organization in the community, a third party or any other organization cannot operate the community cloud.

III. SIGNIFICANCE OF RESOURCE ALLOCATION

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module. Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal RAS [5]. An optimal RAS should avoid the following criteria as follows:

A. **Resource contention** situation arises when two applications try to access the same resource at the same time.
B. S**carcity of resources** arises when there are limited resources.
C. **Resource fragmentation** situation arises when the resources are isolated. [There will be enough resources but not able to allocate to the needed application.]
D. **Under-provisioning** of resources occurs when the application is assigned with fewer numbers of resources than the demand.
E. **Over-provisioning** of resources arises when the application gets surplus resources than the demanded one.

Resource users' (cloud users) estimates of resource demands to complete a job before the estimated time may lead to an over-provisioning of resources. Resource providers' allocation of resources may lead to an under-provisioning of resources. To overcome the above mentioned discrepancies, inputs needed from both cloud providers and users for a RAS as shown in table I. From the cloud user's angle, the application requirement and Service Level Agreement (SLA)[6] are major inputs to RAS. The offerings, resource status and available resources are the inputs required from the other side to manage and allocate resources to host applications by RAS. The outcome of any optimal RAS must satisfy the parameters such as throughput, latency and response time. Even though cloud provides reliable resources, it also poses a crucial problem in allocating and managing resources dynamically across the applications.

TABLE I. INPUT PARAMETERS

Parameter	Provider	customer
Provider Offerings	√	-
Resource Status	√	-
Available Resources	√	-
Application Requirements	-	√
Agreed Contract Between Customer and provider	√	√

From the perspective of a cloud provider, predicting the dynamic nature of users, user demands, and application demands are impractical [7]. For the cloud users, the job should be completed on time with minimal cost. Hence due to limited resources, resource heterogeneity, locality restrictions, environmental necessities and dynamic nature of

resource demand, we need an efficient resource allocation system that suits cloud environments.

IV. TYPES OF RESOURCES ALLOCATION IN CLOUD COMPUTING

In this paper, we present the methods for efficient resource allocation that will help cloud owner to reduce wastage of resources and to achieve maximum profit. Efficient resource allocation in the cloud is a very challenging task as it needs to satisfy both the user's requirements and server's performance equally. Resource allocation in cloud computing environment is defined as assignment of available resources such as CPU, memory, storage, network bandwidth etc in an economic way. It is the main part of resource management.

A. **Analytic Hierarchy Process:** The AHP provides such a framework that enables us to make effective decisions on complex issues by simplifying and expediting our natural decision-making processes. The AHP organizes feelings, intuition, and logic in a structured approach to decision making[8]. There are two fundamental approaches to solving problems: the deductive approach and the inductive approach. Basically, the deductive approach focuses on the parts whereas the systems approach concentrates on the workings of the whole. The AHP combines these two approaches into one integrated, logic framework. The analytic hierarchy process (AHP) was developed by [9][10]Thomas L. Saaty. Saaty, T.L., The Analytic Hierarchy Process, New York: McGraw-Hill, 1980. The AHP is designed to solve complex problems involving multiple criteria. An advantage of the AHP is that it is designed to handle situations in which the subjective judgments of individuals constitute an important part of the decision process. Basically the AHP is a method of breaking down a complex, unstructured situation into its component parts; arranging these parts, or variables into a hierarchic order; assigning numerical values to subjective judgments on the relative importance of each variable; and synthesizing the judgments to determine which variables have the highest priority and should be acted upon to influence the outcome of the situation. The output of the AHP is a prioritized ranking indicating the overall preference for each of the decision alternatives[11].

Main steps of AHP:

1) To develop a graphical representation of the problem in terms of the overall goal, the criteria, and the decision alternatives. (i.e., the hierarchy of the problem)

2) To specify his/her judgments about the relative importance of each criterion in terms of its contribution to the achievement of the overall goal.

3) To indicate a preference or priority for each decision alternative in terms of how it contributes to each criterion.

4) Given the information on relative importance and preferences, a mathematical process is used to synthesize the information (including consistency checking) and provide a priority ranking of all alternatives in terms of their overall preference.

B. **Priority Based Resources Allocation:** In a cloud computing environment, multiple customers are submitting job request with possible constraints that is multiple users are requesting same resource. For example in a high performance computational environment which mainly deal with scientific simulations such as weather prediction, rainfall simulation, monsoon prediction and cyclone simulation etc

which requires huge amount of computing resources such as processors, servers, storage etc. Many users are requesting these computational resources to run their model which is used for scientific predictions. So at this situation it will be problem for cloud administrator to decide how to allocate the available resources among the requested users[12]. The priority algorithm helps cloud admin to decide priority among the users and allocate resources efficiently according to priority. This resource allocation technique is more efficient than grid and utility computing because in those systems there is no priority among the user request and cloud administrator is randomly taking decision and he is giving priority to those user who have submitted their job first that is based on first come first serve method. But with the advent of cloud computing and by using this implemented priority algorithm, the cloud admin can easily take decision based on different parameters discussed earlier to decide priority among different user request so that admin can efficiently allocate the available resources and with cost-effectiveness as well as satisfaction from users.

C. Auction Based Resources Allocation: Cloud resource allocation by auction mechanism is based on sealed-bid auction. The cloud service provider collects all the users' bids and determines the price. The resource is distributed to the first kth highest bidders under the price of the (k+1)th highest bid. This system simplifies the cloud service provider decision rule and the clear cut allocation rule by reducing the resource problem into ordering problem. But this mechanism does not ensure profit maximization due to its truth telling property under constraints. The aim of resource allocation strategy is to maximize the profits of both the customer agent and the resource agent in a large datacenter by balancing the demand and supply in the market. It is achieved by using market based resource allocation strategy in which equilibrium theory is introduced (RSA-M) [13]. RSA-M determines the number of fractions used by one VM and can be adjusted dynamically according to the varied resource requirement of the workloads. One type of resource is delegated to publish the resource's price by resource agent and the resource delegated by the customer agent participates in the market system to obtain the maximum benefit for the consumer. Market Economy Mechanism is responsible for balancing the resource supply and demand in the market system.

v. CONCLUSION

Cloud computing technology is increasingly being used in enterprises and business markets. In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper summarizes the techniques of resources allocation and its impacts in cloud system. Some of the strategies discussed above mainly focus on CPU, memory resources but are lacking in some factors. Hence this survey paper will hopefully motivate future researchers to come up with smarter and secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

REFERENCES

[1] Peter Mell and Tim Grance, *"Draft NIST Working Definition of Cloud Computing,"* http://csrc.nist.gov/groups/SNS/cloud-computing/index.html, 2009

[2] R. Buyya, et al., "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, "Future Generation Computer Systems, 2009, 25, p. 599-616.

[3] M. Armbrust, et al., "A view of cloud computing," Communications of the ACM, 2010, 53, p. 50-58

[4] Buyya Rajkumar, Chee Sin Yeo, Venugopal Srikumar, Broberg James, Brandic Ivona, *"Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility".* Future Gener Comput Syst 25(6):599–616, 2009

[5] Patricia Takako Endo et al. :Resource allocation for distributed cloud :Concept and Research challenges(IEEE,2011),pp.42-46 .

[6] Linlin Wu, Saurabh Kumar Garg and Raj kumarBuyya: SLA –based Resource Allocation for SaaS Provides in Cloud Computing Environments (IEEE, 2011), pp.195-204

[7] Patricia Takako Endo, Andre Vitor de Almeida Palhares, Nadilma Nunes Pereira, 2011. Resource Allocation for Distributed Cloud: Concepts and Research Challenges, IEEE, july 2011.

[8] Daji Ergu, Gang Kou, Yi Peng, Yong Shi, Yu Shi. "*The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment"* The Journal of Supercomputing Springer Science+Business Media, pp 835-848, 2011

[9] Saaty, T.L. (1990) How to Make a Decision: The Analytic Hierarchy Process. *European Journal of Operational Research*, **48**, 9-26. http://dx.doi.org/10.1016/0377-2217(90)90057-I

[10] Saaty, T.L. (2003) Decision-Making with the AHP: Why Is the Principal Eigenvector Necessary. *European Journal of Operational Research*, **145**, 85-http://dx.doi.org/10.1016/S0377-2217(02)00227-8

[11] Ergu, D., Kou, G., Peng, Y., Shi, Y. and Shi, Y. (2011) The Analytic Hierarchy Process: Task Scheduling and Resource Allocation in Cloud Computing Environment. *The Journal of Supercomputing*, **213**, 246-259.

[12] Li Yang et al, A new Class of Priority-based Weighted Fair Scheduling Algorithm, Physics Procedia ,33 (2012) 942 – 948

[13] Xindong YOU, Xianghua XU, Jian Wan, Dongjin YU:RAS-M :Resource Allocation Strategy based on Market Mechanism in Cloud Computing(IEEE,2009),pp.256-263.

[14] O. M. Elzeki, M. Z. Rashad, M. A. Elsoud, "Overview of Scheduling Tasks in Distributed Computing Systems", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[15] M.Asad Arfeen, Krzysztof Pawlikowski, Andreas Willig .2011, A Framework for Resource Allocation Strategies in Cloud Computing Environment, 2011 35th IEEE Annual Computer Software and Applications Conference Workshops.