# A Review: Resource Allocation Problem in Cloud Environment

Prof. Rupali M.Pandharpatte[#1]
[#] Computer Engineering, Savitribai Phule Pune University
KJ's College of Engineering & Management Research, Pune, India
[1] rupali.pandharpatte@gmail.com

*Abstract*— **Cloud computing is the process of providing and managing computing resources such as hardware and software over the Internet as pay-as-you-go. Cloud Service Provider and User are two main players in cloud. On one hand, cloud provider carries number of computing resources in their large datacenters and rent resources out to users on pay-as-you-go basis. On the other hand, there are large numbers of users who have applications with changing loads and request required resources from providers to run their applications slickly. Typically, the goal of providers is to generate as much revenue as possible with minimum investment and proper resource utilization. On the other hand, users want their jobs done at minimal expense. In the Cloud computing resources need to be allocated and scheduled in such a way that providers achieve their objectives and users meet their applications requirements with minimum expenditure. We call this as a cloud resource allocation problem. Resource allocation is traditionally looked upon as an optimization problem hence resource allocation is NP-Hard; limited resources are available and allocate these resources to competitive events/activities such a way that both parties will achieve their goals. This extensive review aims to elaborate and analyze the numerous solved/unresolved issues in cloud resource allocation.**

**Keyword -** Cloud Computing1, Resources2, Datacenter3, Virtual Machine4, Physical Machine5.

## I. Introduction

Cloud Computing is delivery and Management of software and Hardware resources on Pay-as-you-go basis. The rise of Cloud Computing is rapidly changing the landscape of information technology, and ultimately turning the long-held promise of utility computing into a reality [1]. Amazon Elastic Compute Cloud (EC2) and Microsoft Windows Azure are classical Cloud computing products [2]. Clouds are usually referred to as a large pool of computing and storage resources, which can be accessed via standard protocols with an abstract interface. Cloud is four-layer Architecture: Application layer, Platform layer, Unified Resource Layer and Fabric Layer. Cloud is divided into Deployment Model and Service Model.

### A. Deployment Model:

1. *Public cloud*: a service provider makes resources available to the general public over Internet.
2. *Private Cloud*: Private clouds are built exclusively for a single enterprise.
3. *Hybrid Cloud*: Cloud Infrastructure is a combination of private, public and any type of clouds.
4. *Community Cloud*: The cloud infrastructure is shared among a number of organizations with similar interests and requirements.

### B. Service Model :[3]

1. *Software as a Service (SaaS):* Is a software model in which software are hosted by third party for distribution over internet as "software on demand".
2. *Platform as a Service (PaaS):* Is a cloud computing service in which user can write application, compile, run and manage without installation.
3. *Infrastructure as a Service (IaaS):* Is an infrastructure of cloud which provide resources such as operating system, network, server, CPU etc. using virtualization technique.

- *Basic Cloud Components [2]*

1. *Client (End User):* Clients are the devices that the end user interact with to manage their information on the client.
2. *Provider:* Cloud Service provider is an owner of cloud infrastructure who provides services to requester via internet on pay-as-you-basis.
3. *Datacenter:* A datacenter is a set of physical resources such as CPU, memory, storage, and so on.
4. *Virtual Machine:* Simulation of a Physical Machine in the form of Software is known as a virtual machine.
5. *Physical Machine*: Physical machine is a set of hardware like CPU, RAM, I/O which are request for applications.

6. *Virtualization:* Virtualization is basically process of making a virtual image of cloud resources such as server, desktop, operating system, storage devices or network [4].

This paper shapes Resource Allocation problem in cloud environment, problem formulation, objectives, constraints, solution representation and different techniques to be used to solve resource allocation in cloud.

The rest of the paper organized in as follows: Section II describes Resource allocation problem, solution representation and formulation. Section III review of objectives and constraints, and Section IV gives details about techniques used to solve resource allocation problem and Section V outline the conclusion of our studies.

## II. Resource Allocation In Cloud

Resource Allocation (RA) in a cloud computing is the process of assigning available resources to the needed cloud user application over the internet [5]. A large number of Physical Machines (PM) is deployed in a Datacenters. A Cloud Service Provider can have large number of geographically distributed datacenters.

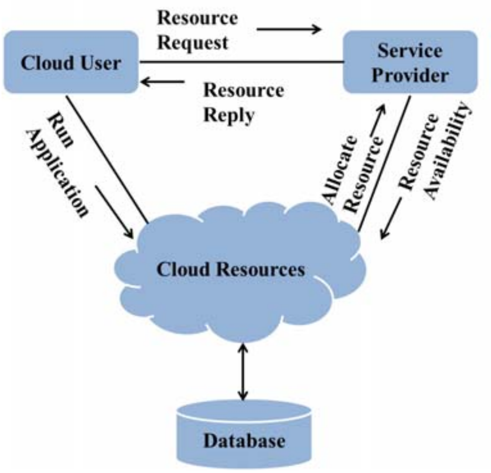


Fig. 1. Resource Allocation System Model

Virtual Machine (VM) requests should be first distributed optimally over datacenters. The requests in each datacenter are then distributed over physical machines if required resources are available [6]. Resource allocation problem is formulated as bin packing Problem. VM allocation in cloud is NP-Hard Problem because at the end we able to find optimal allocation of VMs on PMs.

A. *Types of Resource Allocation[6]*

- *Static Resource Allocation*: Static placement of VMs is done either during system startup or in offline mode. This is the initial placement of VMs in the cloud computing environment.
- *Dynamic Resource Allocation*: If an existing mapping of VMs onto PMs is present, we go for dynamic placement of virtual machines. The main goal of dynamic VM placement is to achieve optimum solutions from the already present mapping of VMs at minimal cost.

B. *Input Parameters of Resource Allocation*

TABLE I. Input Parameters [5]

| Parameter | Cloud User | Provider |
|---|---|---|
| Resource Estimation | √ | |
| Resource Offering | - | √ |
| Resource Status | - | √ |
| Resource Request | √ | - |
| Resource Availability | - | √ |
| Service Level Agreement(SLA) | √ | √ |

When required resources are allocated to user by service provider some criteria should be avoided by service provider. These considerations are as follows: [5]

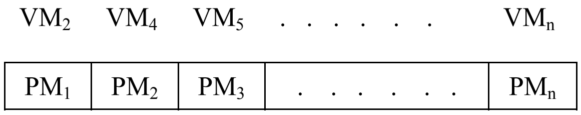
1. *Resource Contention:* Is a conflict over the access to same shared resources by number of users [7].
2. *Scarcity of resources:* Resource availability is limited.
3. *Resource fragmentation*: This situation will arise if resources are available but not accessible or not able to allocate to needed user.

4. *Over-provisioning:* User application gets more resources than requested.
5. *Under-provisioning:* Fewer resources are allocated to user than required.

C. *Problem Description*

- *N* Physical machines (PMs) are present at Cloud Data Center with available capacity $C_n$ (*CPU, RAM, Bandwidth*).
- *M* Virtual Machines (VMs) requests from cloud user to run their application on cloud with resource request $C_m$ (*CPU, RAM, Bandwidth*).
- We have to find an optimal allocation between VMs and PMs that satisfies the VMs resource requirements while minimizing the number of hosts required for VMs. While finding such a allocation, the sum of resource demands of all VMs on that PM does not exceed the total capacity of the PM [8].

D. *Solution Representation*



| $VM_2$ | $VM_4$ | $VM_5$ | . . . . . . | $VM_n$ |
|---|---|---|---|---|
| $PM_1$ | $PM_2$ | $PM_3$ | . . . . . . | $PM_n$ |

Fig. 2. Resource Allocation VM-PM Mapping

A 1D representation of array as shown in Fig. 2 is used to show the number of VM allocation on number of PM. index of array shows number of resource requests (VMs) and value in array shows PM on which requested VMs are placed. Fig. 2 shows the sample solution represents number of virtual machines $VM_1$, $VM_{2,}$...,$VM_N$ and number of physical machines $PM_1$, $PM_2$, ..., $PM_N$. For example, $VM_2$ is assigned to $PM_1$.This representation shows optimal mapping of VMs on PMs.

E. *Problem Formulation*

Resource Allocation problem can be formulated in single objective function as well as multi-objective function.

| | |
|---|---|
| *Set of requested resources* | $VM = (vm_1, vm_2, . . . ,vm_n)$, |
| *Set of Physical machine with available capacity* | $PM = (pm_1, pm_2, . . . pm_n)$ , |
| *Capacity/requirement of VM* | $C_m = (cpu , ram, b/w)$, |
| *Capacity of PM* | $C_n = (cpu , ram, b/w)$ |
| *Objective function [2]* | |
| *Maximize /Minimize* | $y = f(x) = (f1(x), f2(x), . . . , fk(x))$ |
| *Subject to* | $C(x) = ((c1(x), c2(x), . . . , cc(x)))$ |
| *Where* | $x = (x1, x2, . . . , xn)$ Є $X$ |
| | $y = (y1, y2, . . . , yk)$ Є $Y$ |

*x* is decision vector belongs to X(decision space) and *y* is objective vector of objective space Y.

## III. OBJECTIVES AND CONSTRAINTS

To solve resource allocation problem in cloud computing various studies are successfully find solutions for resource assignments and some studies are going to finds various solutions, at the time of studies various objectives are achieved by considering some conditions and satisfying that various constraints/conditions. Table II describes various studies on various objectives to achieve optimal resource allocation in cloud computing.

TABLE II. Objectives

| Sr. No. | Objectives | References |
|---|---|---|
| 1. | Resource Utilization | [2][9][10][11][12][13][14] |
| 2. | Profit Maximization | [15][16][17][18] |
| 3. | Execution time | [19][20][21] |
| 4. | Cost Minimization | [9][15][22][23] |
| 5. | Load Balancing | [10][24] |
| 6. | Energy Consumption | [11] [12] [17][25][26] [27] [28] |
| 7. | Load Migration | [25][27][29] |
| 8. | Number of PMs used | [10][27] |
| 9. | Response Time | [19][24][30] |
| 10. | SLA | [16] [18][31] |
| 11. | Satisfy Customer Demand | [30] |

1. *Resource Utilization*: In cloud computing hardware and software resources are provided by provider to user on pay-as-you-go basis. Effective utilization and management of these resources by minimizing and maximizing some parameters is called as resource utilization.
2. *Profit Maximization*: Profit maximization is objective of service provider by allocating requested resources to needed cloud user and gaining profit.
3. *Execution time*: Execution time is defined as total time required for VM on each PM.
4. *Cost Minimization*: It is minimization of cost of application running on cloud. Cost is in the form of data transfer cost or computational cost.
5. *Load Balancing*: Load balancing is the process of distributing workloads and computing resources such as computers, a computer cluster, network links, central processing units, bandwidth, RAM, or disk drives in a cloud computing environment. Load balancing aims to optimize resource use, maximize throughput, minimize response time, and avoid overload of any single resource [32].
6. *Energy Consumption*: Servers consume a larger fraction of energy in cloud environments and their energy consumption varies with utilization. This consumption also vary with the type of computation going on in the server e.g., data retrieval and data processing[28].
7. *Load Migration*: In Cloud, migration is the process of moving data and applications from on cloud to another cloud because of some constraints, e.g. Load on server. [33].
8. *Number of PMs used*: Allocate VMs on PMs such that numbers of physical machines are used should be minimum by considering load, SLA, etc. it used for optimal resource allocation.
9. *Response Time*: Response time is the time needed to complete the enquiry about resources and start of response [34].
10. *SLA*: A service-level agreement (SLA) is a contract between a service provider and cloud customers that account which type of service provided by provider to user according application need [35].Services are – quality, availability, responsibilities that agreed between the service provider and the service user [36].
11. *Satisfy Customer Demand*: Optimal allocation of requested resources to the user application by considering some aspects like SLA, Quality of services, and profit then only customer will satisfy.

While allocating the resources by considering some objectives there should be some conditions to be satisfied by both parties are listed below:

- *Capacity/Overloading*: The sum of resource demands of all VMs on that PM does not exceed the total capacity of the PM [8].
- *Placement Guarantee Constraints*: Requested VMs should be placed on PM otherwise which results on SLA violation.
- *Number of server used*: If PM/server not in use then switch off that server. It is binary value 1 if used, 0 otherwise.

## IV. Resource Allocation Techniques

To solve resource allocation problem in cloud number of techniques are used evolutionary algorithms, Optimization techniques, Heuristic algorithms, Hybrid algorithms etc. That various techniques are listed below::

TABLE III. Techniques

| Sr. No. | Techniques | References |
|---|---|---|
| 1. | Virtualization | [3][4] |
| 2. | Genetic Algorithm(GA) | [2][17][37][38] |
| 3. | Non-dominated Sorted GA(NSGA) | [2][39] |
| 4. | Game Theory | [6][15] |
| 5. | Heuristic Approach | [19][9][10][25][30][[16][27][40][20] |
| 6. | Biogeography-based optimization (BBO) | [11] |
| 7. | Greedy Particle Swarm Optimization(GPSO) | [23] |
| 8. | Utility Function | [22][15][17] |
| 9. | Optimization Algorithm | [13][21] |
| 10. | Graph Theory | [26] |

## V. Conclusion

This work explored the resource allocation problem in cloud. From this review it is clear that so many researches are going on and completed on cloud computing. We have tried to focus on problem formulation, number of objectives and constraints as well as techniques used to solve such problem. From this analysis it has been observed that there is no one fixed algorithm to solve problem there are so many algorithms try to satisfying the conditions and solve it. Resource allocation in cloud is optimization problem hence it is NP-hard Problem.

## References

[1] F. Xhafa and N. Bessis, "Cloud Computing: Paradigms and Technologies", Inter-cooperative Collective Intelligence: Techniques and Applications, Studies in Computational Intelligence, Springer- erlag Berlin Heidelberg 2014.

[2] A. C. Adamuthe, R. M. Pandharpatte, and G. T.Thampi, "Multiobjective virtual machine placement in cloud environment," in Proceedings of the International Conference on Cloud and Ubiquitous Computing and Emerging Technologies (CUBE '13), pp. 8–13, IEEE, Pune, India, November 2013.

[3] M. Rouse, "SPI model (SaaS, PaaS, IaaS)", February 2012, [Online]. Available: http://searchcloudcomputing.techtarget.com/definition/SPI-model.

[4] L. Malhotra, D. Agarwal et.al., "Virtualization in Cloud Computing", Information Technology & Software Engineering, JITSE, Vol. 4 , Issue 2,Open Access.

[5] V.Vinothina, Dr.R.Sridaran, Dr.PadmavathiGanapathi, "A Survey on Resource Allocation Strategies in Cloud Computing", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, Issue 6 ,pp.97-104, 2012.

[6] A.K.Paul, "Dynamic Virtual Machine Placement in Cloud Computing", M.Tech. dissertation, Dept. CS. Engg., National Institute of Technology, Rourkela, India, 2005.

[7] Cavan Group, "What Is Cloud Contention and Why Does It Matter?", [Online]. Available:http://www.cavangroup.com/what-is-cloud-contention-and-why-does-it-matter.

[8] A. Shankar, "Virtual Machine Placement in Computing Clouds", dissertation, Dept. CS. Engg., Indian Institute of Technology, Bombay, India, 2010.

[9] W. Lina, J. Wangb, C. Liangc et.al., "A Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing", PEEA 2011, pp.695-703, Elsevier,2011.

[10] S. Patil, K. Bhavani, "Dynamic Resource Allocation using Virtual Machines for Cloud Computing Environment", International Journal of Engineering and Advanced Technology (IJEAT), Vol.3 Issue 6, pp.218-221, August 2014.

[11] Q. Zhenga, R. Li, X. Li, et.al., "Virtual machine consolidated placement based on multi-objective biogeography-based optimization", Future Generation Computer Systems, Elsevier,pp.95-121,2016.

[12] A. Gupta, D. Milojicic, S. Balle, "HPC-Aware VM Placement in Infrastructure Clouds", International Conference on Cloud Engineering IC2E '13, pp. 11-20, March 25 - 27, 2013.

[13] S. Di and C. Wang, "Dynamic Optimization of Multi-Attribute Resource Allocation in Self-Organizing Clouds", IEEE Transactions on Parallel and Distributed Systems, Vol. 24, Issue 3, pp. 464 – 478, 15 May 2012

[14] S. Kapur and K. Dinesh, "Resource Utilization in Cloud Computing using Hybrid Algorithm" , Indian Journal of Science and Technology, Vol. 9, Issue. 43, pp.1-10, November 2016

[15] P. Pillai and S. Rao, "Resource Allocation in Cloud Computing Using the Uncertainty Principle of Game Theory", IEEE Systems Journal, Vol.10, Issue 2, pp. 637 – 648, 09 May 2014.

[16] H. Goudarzi and M. Pedram, "Maximizing Profit in Cloud Computing System via Resource Allocation", International Conference on Distributed Computing Systems Workshops ICDCSW '11,IEEE,pp. 1-6, June 20 - 24, 2011.

[17] A. Mosa and N. Paton, "Optimizing virtual machine placement for energy and SLA in clouds using utility functions", Journal of Cloud Computing: Advances, Systems and Applications, Open Access, 2016.

[18] L. Wu, S. Garg and R. Buyya, "SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments", International Symposium on Cluster, Cloud and Grid Computing,IEEE, pp.195-204,2011

[19] A. Saraswathi , R. Kalaashri, Dr.S.Padmavathi, " Dynamic Resource Allocation Scheme in Cloud Computing", ICGHIA2014,pp.31-37,2015.

[20] L. Kangkang, Zh. Huanyang, W. Jie, D. Xiaojiang, “Virtual machine placement in cloud systems through migration process”, International Journal of Parallel, Emergent and Distributed Systems, 03 September 2015, Vol.30(5), p.393-410.

[21] T. Nguyen, N. Quang-Hung, N. H. Tuong, V. H. Tran, and N. Thoai,“Virtual machine allocation in cloud computing for minimizing total execution time on each machine,” in Proceedings of International Conference on Computing, Management and Telecommunications (ComManTel), 2013, Jan 2013, pp. 241–245.

[22] N. Kumara, S. Saxenab, “A Preference-based Resource Allocation In Cloud Computing Systems”, 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), pp.104-111, Elsevier,2015.

[23] F. A. Omara, S. M. Khattab and R. Sahal, “Optimum Resource Allocation of Database in Cloud Computing” , Egyptian Informatics Journal, vol. 15, (2014) , pp. 1–12 .

[24] R. Bhaskar , S. Deepu and Dr. B. Shylaja , “Dynamic Alloocation Method for Efficient Load Balancing in Virtual Machines for Cloud Computing Environment”, Advanced Computing: An International Journal ( ACIJ ), Vol.3 No.(5), pp. 53-61,September 2012.

[25] A .Kumar, C. Sathasivam ,P. Periyasamy , “ Virtual machine placement in cloud computing”, Indian Journal of Science and Technology; Vol.9 No.29, pp. 1–5, Aug. 2016.

[26] A. Zhou; S. Wang; B. Cheng; Z. Zheng; F. Yang; R. Chang; M. Lyu; R. Buyya, "Cloud Service Reliability Enhancement via Virtual Machine Placement Optimization," in IEEE Transactions on Services Computing , vol.PP, no.99, pp.1-1,20 January 2016.

[27] R. Camati, A. Calsavara, L. Lima, “Solving the Virtual Machine Placement Problem as a Multiple Multidimensional Knapsack Problem”, ICN 2014 : The Thirteenth International Conference on Networks, pp.253-260, 2014.

[28] A . Uchechukwu, K. Li, and Y. Shen, “Energy consumption in cloud computing data centers,” International Journal of Cloud Computing and services science , vol. 3, no. 3,pp.31-48, 2014.

[29] Z. Xiao, W. Song and Q. Chen, "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment," in IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 6, pp. 1107-1117, June 2013.

[30] P. Pradhan, P. Behera, and B. Ray, “Modified Round Robin Algorithm for Resource Allocation in Cloud Computing”in Procedia Computer Science, 85, pp.878-890,2016.

[31] Y. Choi and Y. Lim, “Optimization Approach for Resource Allocation on Cloud Computing for IoT”, International Journal of Distributed Sensor Networks, pp.1-6,2016.

[32] M. Rouse, “cloud load balancing”, 2014, [Online]. Available:http:// http://searchcloudcomputing.techtarget.com/definition/cloud-load-balancing.

[33] M. Rouse, “cloud migration”,2014, [Online]. Available:http://searchcloudapplications.techtarget.com/definition/cloud-service-migration.

[34] M. Rouse, “response time”, 2014, [Online]. Available: http://searchnetworking.techtarget.com/definition/response-time.

[35] M. Rouse, “service-level agreement (SLA)”, 2014, [Online]. Available: http://searchitchannel.techtarget.com/definition/service-level-agreement.

[36] SLA. In Wikipedia. Retrieved 9 March 2017, from https://en.wikipedia.org/wiki/Service-level_agreement.

[37] M. chaabane, K. Okba, and S. Bourekkache, “A Genetic Algorithm for Resource Allocation with Energy Constraint in Cloud Computing”,in Proceedings of 2016 International Conference on Image Processing, Production and Computer Science (ICIPCS'2016), pp.62-69, March 26-27, 2016.

[38] S. Ravichandran and Dr. E.R. Naganathan, “Dynamic Scheduling of Data Using Genetic Algorithm in Cloud Computing”, International Journal of Computing Algorithm, Vol 2, Issue 01,pp.127-133, June 2013.

[39] J. Zhao, W. Zeng, M. Liu and G. Li, “Multi-objective optimization model of virtual resources scheduling under cloud computing and it's solution,” 2011 International Conference on Cloud and Service Computing, Hong Kong, 2011, pp. 185-190.

[40] P. Prince, D. Ruphavathani and S. P. Lovesum, “ A Security Aware Resource Allocation Model for Cloud Based Healthcare Workflows”, Indian Journal of Science and Technology, Vol 9, issue 45, pp.2-6,December 2016.

## AUTHOR PROFILE

Prof. Rupali M. Pandharpatte working as Assistant Professor at KJCOEMR, Pune. Area of Specialization is Cloud Computing and Algorithm.