E-ISSN: 2997-9382
American Journal of Technology Advancement
https://semantjournals.org/index.php/AJTA

| Research Article

Check for updates

# Virtual Machine Allocation Policy In Cloud Computing

**HARSH**
Assistant Professor, Department of Computer Applications, Panipat Institute of Engineering & Technology, Samalkha, Haryana, India

**Annotation**

Cloud computing is a very powerful concept that can be used to enhance the next generation data center and allow service provider to use data center capability provided by cloud and develop the application based on user requirement. Data center of this cloud computing has huge number of resources and list of applications (with different architecture, configuration and requirement for deployment) want to use those resource. Cloud computing environment uses virtualization concept and provides resources to application by creating and allocating virtual machine to specific application. There for resource allocation policies and load balance policies play very vital role in allocating and managing the resources among various application in cloud computing life cycle.

The next generation of computation service will be provided by the cloud computing services. Cloud computing allows business customers to scale up and down their resource usage based on needs. Many of the touted gains in the cloud model come from resource multiplexing through virtualization technology. Dynamic selection of virtual machines plays an important role in providing services to the consumers. This paper discusses the design and implementation of the dispatcher algorithm for effective utilization of the cloud resources. Also presents a case study which examines the implementation of the dispatcher algorithm, by a server, A proper scheduling and efficient load balancing across the network can lead to improve overall system performance and a lower turnaround time for individual tasks.

**Keywords:** Virtualization, Resource Allocation, Service Level Agreement (SLA), Virtual Machine Allocation.



## INTRODUCTION

Cloud computing provide various services like IAAS (Infrastructure as a Service), PAAS (Platform as a Service), and SAAS software based on a pay-as-you-use model to cloud customers and has potential to transfer a large part of the IT industry, making software even more attractive as a service. Cloud computing is the cutting edge in reckoning. Perhaps individuals can have all that they require on the cloud. From user perspective cloud computing make them able to use and deploy their applications from anywhere on this planet and interest at focused expenses contingent upon clients QOS (Quality of Service) necessities. To provide these services continuously on

demand, internally it uses many of technologies like virtualization, clustering, terminal service, application server and more. Virtualization is a foundational element of cloud computing environment. It can be defined as making of a virtual version of something, such as an operating system or servers or storage devices or network resources.

When the consumers request for the resources cloud service provides must provide the resource available by considering the Service Level Agreement (SLA). So in order to make resource available there is a need of efficient and optimized method for scheduling of resources, developing applications on Virtual Machines (VM). Currently, more work is made on scheduling of consumer applications on cloud. Single SLA such as cost of execution and execution time are considered in these approaches. When the consumer request for job execution on the cloud, it usually divided into several tasks. Following research questions are required to be considered when executing this several tasks.

- ✓ How to measure the workload of several tasks?
- ✓ How to allocate the required resources to execute the several tasks?
- ✓ How to schedule and manage the VM's

Typically, efficient provisioning requires two distinct steps or processes:

(1) Initial static planning step: the initially group the set of VMs, then classify them and deployed onto a set of physical hosts; and

(2) Dynamic resource provisioning: the allocation of additional resources, creation and migration of VMs, dynamically responds to varying workload. Step 2 runs continuously at production time where in contrast Step 1 is usually performed at the initial system set up time and may only be repeated for overall cleanup and maintenance on a monthly or semi-annually schedule.

In Resource allocation (RA) is the process of allocating available resources to the required consumer application for execution over the internet. If resource allocation is not done properly then it will get waste and there will be a failure in providing a service to the consumers.

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. An optimal RAS should avoid the following criteria as follows:

- ✓ Resource contention situation arises when two applications try to access the same resource at the same time.
- ✓ Scarcity of resources arises when there are limited resources.
- ✓ Resource fragmentation situation arises when the resources are isolated. [There will be enough resources but not able to allocate to the needed application.]
- ✓ Over-provisioning of resources arises when the application gets surplus resources than the demanded one.

## RELATED WORK

In [1] author proposed architecture, using feedback control theory, with help of virtualization that do so using virtual machines, In virtual machine architecture all hardware resources are put in one place with memory sharing architecture and requested applications by cloud customer deployed as per SLA (Server lever agreement).In this paper architecture uses controllers: CPU, memory and Input output. Its objective is to control various virtualized assets usage to attain SLA of requisition by using control inputs for every virtual machine resources. Problem under the virtual machine

based architecture is how to provide resources to each application with in response of time management based on workloads. In [5], author proposed two layer architecture uses utility functions in resource allocation in static and also in dynamic manner with help of two agent local agent and global agent. The agent computes current work load and transfer it to global agent which is responsible for proving optimal configuration of resource allocation by analyzing work load given by local agent. For that synchronization mechanism between this two agents is perform important task, if any changes occurred in any one of them then they may have to be changed.

In [2], author proposed an adaptive resource allocation algorithm in cloud computing environment. This paper used adaptive min-min scheduling and list scheduling in but it is for used in static manner. In [3] Author describes major challenges in the resource allocation in cloud computing environment. In this paper described challenges were related to the resource management policy but the allocation method is not described any more. In [4] author proposed cloud computing resource planning in multiple dimension like withrespect to space or with respect to time or with respect to response time. Authors used Amazon EC2 for processing environment and describe be how this would ne benefited for cloud client and supplier.

In [5] author described the major concept in resource management in cloud computing environment with uncertainty. This uncertainty refers to parameter and policy. In [6] author proposed the modeling in CloudSim a modeling and simulation tool kit and describes where as modeling strategy used in this simulator and benefits of this simulator. The broker strategy is described in this paper and described about the various parameters involve in it. In [7] author describes various strategies to overcome the challenges related to configuration as per user demand in, Cloud computing environment. In this paper varies policy described for managing the virtual machine and author proposed one software framework for solving the issue discussed as major challenges. Author used CloudSim for modeling and simulation and describes varied modeling strategies for managing the virtual machine and data center resource in cloud computing environment and describe the simulation result of the CloudSim. In [11] author describe the term grid for providing high performance services for complex computing and data intensive scientific application.[8],[9],[10] described about ‘GridSim’ , ‘SimGrid’ , ‘GangSim’ simulator respectively for simulating the grid computing. SimGrid provides framework for simulating distributed application, GangSim used for modeling and simulator using virtualization concept and GridSim is event based simulator for heterogeneous Grid resource. Scientists have additionally examined virtual machine arrangement crosswise over various crosswise over various cloud suppliers from numerous cloud suppliers under future requests with high accessibility prerequisites.

In [11]-author proposed optimal virtual machine placement algorithm for minimizing the cost that cloud customer have to pay cloud provider, when they need virtual machine from cloud computing environment access as part of cloud service. In [12] author described the multi objective mechanism for scheduling applications that take various cost constrained and availability of resourced in account.

In [13-15] author more focus on resource allocation strategy in selecting cloud provider, but in static manner for selecting a data center from distributed environment where global data center is available, with taking care of timing parameter. In [16] author described briefly about the CloudSim toolkit for modeling and simulation environment. Author described usefulness of the CloudSim by various case studies, virtual machine management in CloudSim and also described about federated Cloud computing model.

## RESOURCE ALLOCATION MODEL

The overall architecture is as shown in figure 1, According to the characteristics of applications, we propose an algorithm which dispatches the request by referencing CPU computing power. The

main effort of this dispatching algorithm is to decide which VM to use and creating a new VM's based on resource availability. It is the place where dispatching decisions are made.

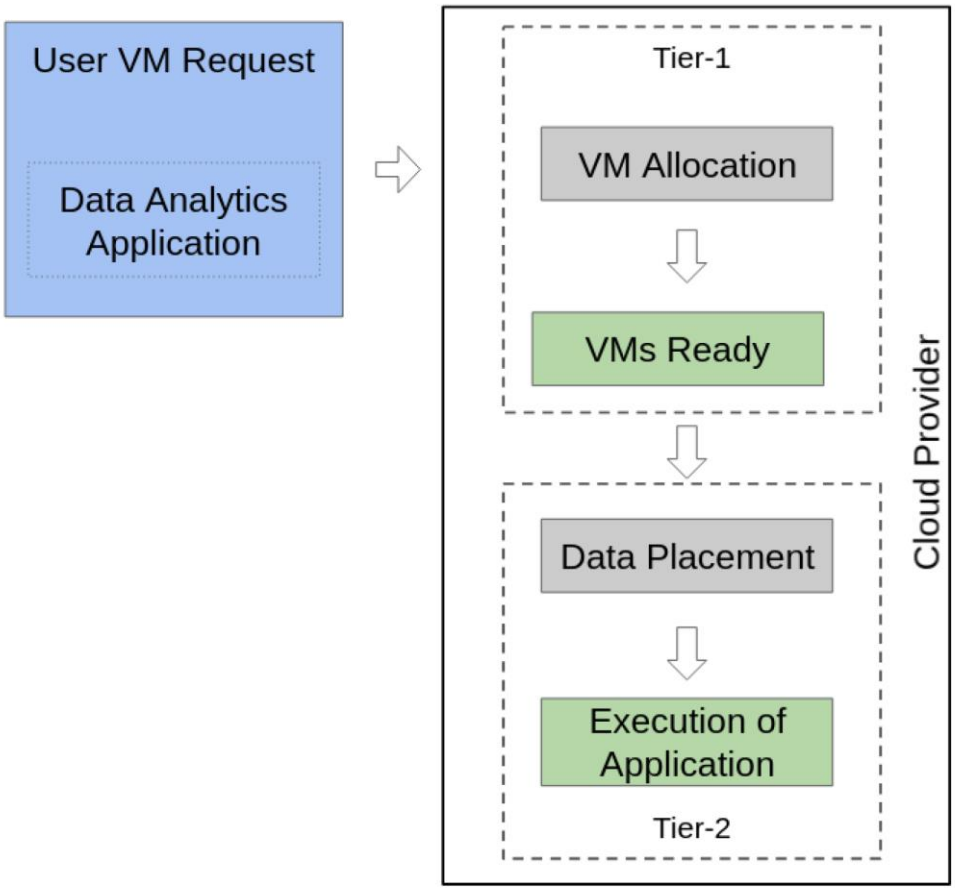


**Fig 1: Resource Management work flow in Cloud Computing**

Once a VM is chosen and the connection is constructed, all remote invocations go through this link are served by this VM. Here we can have the channel objects periodically discard connections in purpose for the reconstruction of connections to less load servers. The network processor records the IP and port information of the client and the selected VM in the VM connection table called VMCT for each constructed connection.

The remote request for resources with the same source IP, the same source port, and the same destination port will be directed to the same destination IP according to VMCT. The response packets from the servers are also directed to the correct VMs by this VM table. The destination port mentioned VMCT is used to identify remoting services Different services distributed and then go to the different ports in our customized consumer channel objects. Dispatching algorithm is to find the least load server for dispatching. Different scheduling methods can be plugged in for this step. In the following, we propose a method to schedule tasks to the server minimizing the estimated task time.

## DYNAMIC DISPACHER ALGORITHM

Load balancing is a technique to enhance resources, utilizing parallelism, exploiting throughput improvisation, and to cut response time through an appropriate distribution of the applications. To minimize the decision time is one of the objectives for load balancing which has yet not been achieved. Proper task scheduling is the only efficient way to guarantee that submitted task are completed reliably and efficiently in case of process failure, processor failure, node crash, network failure, system performance degradation, communication delay, addition of new

machines dynamically even though a resource failure occurs which changes the distributed environment. Generally, load balancing mechanisms can be broadly categorized as centralized or decentralized, dynamic or static, and periodic or non-periodic.

**Algorithm:** Dynamic Dispatcher

**Input:** Get the list of work load and several jobs to execute

Measure the work load

List of available VM's

**Initialize Used VM's**

**If** Resource required <= Used VM

**then** After executing clear the Used VM's

**Else**

**For** VM in List of available VM's

**Do**

Dynamically calculates the required resources and

Dispatches the resources by

Deploying the VM's

Done

**Output:** Mapping of VMs to Physical Machines

## LOAD BALANCING POLICIES

Load balancing algorithms can be based on many policies; some important policies are defined below.

**Information Policy:** This policy specifies what workload information should be collected, when it is to be collected and from where.

**Triggering Policy:** This policy determines the appropriate period to start a load balancing operation.

**Resource Type Policy:** This policy classifies a resource as server or receiver of tasks according to its availability status.

**Location Policy:** This policy uses the results of the resource type policy to find a suitable partner for a server or receiver.

**Selection Policy:** This policy defines the tasks that should be migrated from overloaded resources (source) to most idle resources (receiver).

The main objective of load balancing methods is to speed up the execution of applications on resources whose workload varies at run time in unpredictable way. Hence it is significant to define metrics to measure the resource workload. Every dynamic load balancing method must estimate the timely workload information of each resource.

## EXPERIMENTAL RESULTS

The experiment is conducted to perform 20 tasks distributed in three different scenarios, first is consumer request to image blurring, the server reads the image and waits for the checking the resource availability. As soon as the server receives text message to blur, the server looks for a VMs utilization and time required will be calculated. Then server dispatches the task on run time

to volunteer one or two depends on the RAM resource utilization. Second scenario is consumer request to watermark image on image, the server reads the target image and waits for the image.

Third scenario is consumer request to find the number of occurrences of word in a set of files. The experimental result is as shown in Table 1.

| VM | Time taken (ms) to perform Image blurring | Time taken (ms) to perform Watermark image with image | Time taken (ms) to perform find No. occurrences of Word in a set of files |
|---|---|---|---|
| **1** | 0.17145 | 0.15318 | 0.0434275 |
| **2** | 0.02139 | 0.05847 | 0.125624 |
| **3** | 0.01146 | 0.04638 | 0.215413 |
| **4** | 0.03215 | 0.02353 | 0.135512 |
| **5** | 0.02113 | 0.01264 | 0.163517 |
| **6** | 0.01124 | 0.03125 | 0.112412 |

**Table 1: Time taken to complete the tasks by each VM's**

## CONCLUSION

This paper describes aspects of cloud computing and introduces numerous concepts which illustrate its grand capabilities. Cloud computing is definitely a promising tendency to solve high demanding applications and its related problems. Main objective of the cloud computing environment is to balance load and achieve high performance. Dynamic nature and complexity of network make load balancing very complex and vulnerable to faults. To maintain entire load of nodes is very hard due to dynamic nature of resources in a network environment. There are a number of factors, which can affect the server performance like load balancing, heterogeneity of resources and resource sharing in the network environment. It focuses on load balancing and presents factors due to which load balancing is initiated, compares existing load balancing algorithms and finally proposes an efficient dispatcher algorithm for network environment.

## REFERENCES


1. Lizhewang, JieTao, Kunze M., Castellanos, A.C,Kramer, D.,Karl,w, "High Performance Computing and Communications", IEEE International Conference HPCC,2008,pp.825-830.
2. ZhixiongChen,JongP.Yoon,"International Conference on P2P, Parallel,Grid,Cloud and Internet Computing",2010 IEEE:pp 250-257.
3. P. T. Endo, "Resourc alocation for distrbuted cloud: Concept and Research challenges", IEE, pp. 42-46.
4. J. Y. Shei, M. Taiefi and A. Khreisheah, "Resource Planing for Paralel Procesing in the Cloud", IEEE 13th International Conference on High Performance and Computing, (2011).
5. S. Majumdar, "Resource Management on cloud: Handling uncertainties in Parameters and Policies", CSI Communication, (2011), pp.16-19.
6. B. Oiza, "A Proposed Serviece Broker Stratagy in CloudAnalyst for Cost-Effactive Data Centre Selection Dhavael Limbeani", International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622www.ijera.com, vol. 2, no. 1, (2012), pp. 793-797 793.
7. R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose and R. Buyya, "CloudSim: A Toolkit for the Modeling and Simulation of Cloud Resource Management and Application Provisioning Techniques".

8. C. E. L. Duemitresecu and I. R. Fostera, “GangSim: a simulator for grid scheduling studies”, Proceedings of the IEEE International Symposium on Cluster Computing and the Grid, (2005).
9. “Scheduling distributed applications: the SimGrid simulation framework”, Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid, (2003).
10. R. Buyya and M. Murshed, “GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing.Concurrency and Computation”, Practice and Experience, Wiley Press, vol. 14, no. 13-15, (2002).
11. I. Foster, C. Kesselman and M. Kaufmann, “The Grid: Blueprint for a New Computing Infrastructure”, (1999).
12. S. Chaisiri, B. S. Lee and D. Niyato, “Optimal virtual machine placement across multiple cloud Providers”, In: Services computing conference, APSCC, IEEE Asia-Pacific, (2009).
13. M. E. Frincu and C. Craciun, “Multi-objective meta-heuristics for scheduling applications with high availability requirements and cost constraints in multi-cloud environments”, Fourth IEEE international conference on utility and cloud computing, (2011).
14. J. Tordsson, R. S. Montero, R. M. Vozmediano and I. M. Llorente, “Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers”, Future Gener Comput Syst., vol. 28, no. 2, (2011), pp. 358–367.
15. R. N. Calheiros, R. Ranjan and A. Belo, “CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms”.
16. R. Buyya, R. Ranjan and R. N. Calheiros, “Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities”.
17. Zhen Kong et.al : Mechanism Design for Stochastic Virtual Resource Allocation in Non-Cooperative Cloud Systems: 2011 IEEE 4th International Conference on Cloud Computing :pp,614-621.
18. W. E. Walsh, G. Tesauro, J. O. Kephart, and R. Das, “Utility Functions in Autonomic Systems,” in ICAC ’04: Proceedings of the First International Conference on Autonomic Computing. IEEE Computer Society, pp. 70–77, 2004.
19. Yazir Y.O., Matthews C., Farahbod R., Neville S., Guitouni A., Ganti S., Coady Y., “Dynamic resource allocation based on distributed multiple criteria decisions in computing cloud,” in 3rd International Conference on Cloud Computing, Aug. 2010, pp.91-98.
20. Goudarzi H., Pedram M., “Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems,” in IEEEInternational Conference on Cloud Computing, Sep. 2011, pp. 324-331.
21. Kai Lu, Riky Subrata and Albert Y. Zomaya, Networks & Systems Lab, School of Information Technologies, University of Sydney “An Efficient Load Balancing Algorithm for Heterogeneous Grid Systems Considering Desirability of Grid Sites”.
22. Chieu T.C., Mohindra A., Karve A.A., Segal A., “Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment,” in IEEE International Conference on e-Business Engineering, Dec. 2009, pp.281-286