# An Approach for VM Allocation in Cloud Environment

Abhijitsinh T. Parmar
Research Scholar, CSE Department
Parul Institute of Engineering & Technology
Vadodara, Gujarat, India

Rutvik Mehta
Assistant Professor, IT Department
Parul Institute of Engineering & Technology
Vadodara, Gujarat, India

## ABSTRACT

In recent years, the Internet can be represented as a cloud and the term "Cloud Computing" in computing research and industry today has the potential to make the new idea of 'computing as a utility' in the near future. The goal of Cloud computing is to provision of computing and storage capacity as a service to a heterogeneous community of end-users. Resource management is key process in cloud computing for cloud service provider. Resource allocation is part of resource management process and main objective of it is to balance the load across Virtual Machine (VM). This paper proposes a novel VM allocation load balancing algorithm. Allocation is made on the basis of the Assignment Problem's solution method concept, which is formulated for cloud. Further, this paper also provides the anticipated results with the implementation of the proposed algorithm.

## General Terms

Cloud Computing

## Keywords

Cloud Computing, Resource Allocation, Assignment Problem, Virtual Machine (VM).

## 1. INTRODUCTION

Cloud Computing is the latest term encapsulating the delivery of computing resources as a service. It is the current iteration of utility computing and returns to the model of 'renting' resources. Cloud computing has appeared as an accepted computing model for processing very large volume of data. The terms Leveraging cloud computing is today, the de facto means of deploying Internet scale systems and much of the Internet is tethered to a small number of cloud providers. The advancement of cloud computing is therefore intrinsic to the development of the next generation of Internet. Cloud computing emerges as a base of all computing directly or indirectly. Due to its attractive advantages and popular services it becomes quite popular nowadays.

There is no standard definition of cloud computing but based on the observation of the essence what Clouds are promising to be, Rajkumar Buyya defines cloud computing as following:

''A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers.''[7] In terms of nature everything we can do on local systems that all things we can do remotely on cloud as it provides application delivered as a services over Internet and hardware and system software delivered through data center.

Cloud Computing has a numerous advantage like Pay as per usage – users have to pay only whatever they consume, zero upfront investment – No need of infrastructure establishment, no worry of Maintenance, highly automated, Flexibility and scalability – Whenever demand is high cloud service is scalable, Mobility – can be used from anywhere by using Internet.

## 2. MOTIVATION

In Cloud, there are many tasks require to be executed by the available resources dynamically to achieve the optimal usage of servers, reduce migration of machines, Effective utilization of resources etc., that's why resource management is most important for both cloud provider and clients. Because of these different intentions, there is need to design, develop, propose a resource allocation algorithm, that is used to outperform appropriate allocation map of tasks on resources. In the resource allocation cloud user may request different resources based on their needs, by using VM scheduler the resource can have allocated. By using predictor, the work load can allocate to physical machine which consist of no Virtual machines in it. Cloud computing can solve complex set of tasks in shorter time by proper resource utilization. To make the cloud to work efficiently, best resource allocation strategies have to be employed. Execution of tasks on resources is one of the most important thing in cloud computing environment where the user's jobs are scheduled to different machines. So, here Assignment problem's alternate solution method of mathematics is formulated for cloud, with the aim of allocation VM resources to task to achieve optimal solution which helps to minimize execution time of task and improve resource utilization.

## 3. RESOURCE ALLOCATION

Resource management is one of the hot topic of cloud computing research nowadays. It includes following issues Resource provisioning, Resource allocation, Resource adaption, resource mapping, resource modelling, Resource estimation out of all this Resource allocation is most affecting issue. Basically resource allocation means distribution of resources economically among competing groups of people or programs. Resource allocation has a significant impact in cloud computing, especially in pay-per-use deployments where the number of resources are charged to application providers. The issue here is to allocate proper resources to perform the computation with minimal time and infrastructure cost. Proper resources are to be selected for specific applications in IaaS. In Cloud Computing VM allocation also referred as problem of resource management which is part of load balancing.

In cloud platforms, resource allocation takes place at two levels. First, when an application is uploaded to the cloud, the load balancer assigns the requested instances to physical computers, attempting to balance the computational load of multiple applications across physical computers. Second, when an application receives multiple incoming requests, these requests should be each assigned to a specific application instance to balance the computational load across a set of instances of the same application. For example, Amazon EC2 uses elastic load balancing (ELB) to control how incoming requests are handled. Application designers can direct requests to instances in specific availability zones, to specific instances, or to instances demonstrating the shortest response times.

VM allocation and task scheduling for cloud is a three folded problem which requires: (1) to decide when a VM should be allocated (2) allocating an appropriate physical machine (PM) for it - a problem related to bin packing and (3) scheduling tasks on the VM depending on various client and application given objectives. Usually the provider is controlling the second stage while the first and third are left to the client's decision. This paper focus on first and third stage. [9]

## 4. RELATED WORK

**Jiayin Li et al. [1]** proposed scheme for an adaptive resource allocation algorithm for cloud environment. Algorithm considers preempt able tasks which adjust the resource allocation adaptively based on updated real time task executions. In which static task scheduling is done offline by static resource allocation with the help of Adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) algorithms. The remaining static resource allocation is done by online adaptive procedure with predefined frequency. Here in every re-evaluations process algorithm scheduler re-calculate the finish time of their respective submitted tasks, not the tasks that are assign to that cloud. Finally, results are checked for tight and loose situation in which AMMS has shorter average execution time than ALS.

**Lin, Wang et al. [2]** introduced a dynamic Virtual Machine-Varying Based resource allocation using a threshold. Using this threshold their algorithm decides that the current counts of virtual machines which are assigned to an application are sufficient or not, it is the same for over provisioning. They have defined two other parameters in threshold formulation; one is a rate called normal rate which demonstrates the average amount of workload that one individual virtual instance can tolerate without any over utilization and the other is a parameter that would be defined by system admin based on the work load; those two made the approach very parametric which seems to be a weakness.

**Ray, Sarkar [3]** proposed a load balancing scheme through the concept of resource allocation strategy and then describe the importance of resource allocation in distributed cloud environment. Here author presents the process of allocating the resources for particular job in this dynamic environment. Allocation is made on the basis of the requirement submitted by the consumers or clients. Provider stores the requirement in the repository in xml format. Then final selection of the resource is done based on the resource occupancy matrix, duration of the job and service charge and finally a service level agreement is made between cloud service provider and cloud consumer.

**Pawar et al. [4]** proposed a priority based scheduling algorithm (PBSA) resource provisioning technique. In proposed approach it counts multiple SLA objectives like memory, network bandwidth, required CPU time and resource allocation by preemption mechanism. This work considers high priority task execution for improvement of the resource utilization in Cloud. Here highest priority task and with advance reservation get first chance for execution. This approach capable of improve resource utilization in situation where multiple tasks request high resource to same machine.

**Li, Ge et al. [5]** proposed a comprehensive QDA modeling & scheduling algorithm for the instance intensive workflow task scheduling in cloud environment, which takes users' experiences into consideration. First, the workflow task was modeled by DAG graph. Task parameters and dependencies were determined, and user preference type value was added. Then, the QoS of cloud service resource was modeled to get QoS utility function with user preferences. Finally, combined with staggered sub-deadlines allocation criteria, cloud service resources were sorted according to the corresponding QoS utility function, and then the task scheduling was quickly completed. According to results QDA has much less execution time, better user satisfaction, and improved load balancing rate.

**Zhao et al. [6]** proposed a layered loading balancing scheduling mode by providing the structure of the dispatching resource scheme. Then a comprehensive resource distribution algorithm base on PSO has been designed and implemented in consideration of respective local resource counts, each join points performance, current load distribution. Together with a layered scheduling model and the structure of load balancing system, and then a resource distribution method base on PSO comprehensively considering the task number and current load performance of various local agents has been given out. Results of the proposed system shows that the Discovery method based on loading and PSO in this work can effectively reduce the time of processing user's requests in the Cloud environment.

## 5. PROPOSED VM ALLOCATION APPROACH

An assignment problem is a particular case of transportation problem where the objective is to assign a number of resources to an equal number of activities so as to minimize total cost or maximize total profit of allocation. The problem of assignment arises because available resources such as men, machines, etc. have varying degrees of efficiency for performing different activities. Therefore, cost, profit or time of performing the different activities is different. Thus, the problem is how the assignments should be made so as to optimize the given objective. [8]

Assignment problem can be solved by following four methods,

1. Enumeration method
2. Simplex method
3. Transportation method
4. Hungarian method

All these four methods are standard and studied but all these four methods are found difficult to formulate for cloud due to its complex calculations. But, new alternate solution method proposed here A study on transportation problem, transhipment problem, assignment problem and supply chain management [8] for assignment problem is found very simple and easy to follow which can derive optimal solution in just few steps, so that concept is formulated in this proposed system for allocation of VM.

One of the main goal of assignment problem is to find optimal solution without spending too much resources. Same way in cloud computing main objective of service provider and client is optimally allocate resources, while meeting user demands and application requirements with maintaining cost minimized associated with it.

In proposed system assignment problem is formulated as a VM allocation problem at the VM level in which VMs assign specific amount of the available processing power to the individual task units in cloud. In this case resources of assignment problem are considered here as a VMs, activities (jobs) are considered as Tasks and cost is considered as an execution time. [9]

## 5.1 Algorithm

```
Input:  Resources (VM) [i=1....n]
        Requests (R) [j=1....n]
Step - 1 For all VM, Calculate Capacity of each VM
            $Vc[i]$ = VM Capacity in MIPS
        End For
Step - 2 For all Request, Calculate Length of each Request
            $R_L[j]$ = Length of Request in MI
        End For
Step - 3 Find Execution Time for each Request to VM
            $Ex\_Time[i][j]=\frac{R_L[j]}{V_C[i]}$ Seconds
Step – 4 Construct Execution Time Matrix Ex[V,R].
//VM Selection Procedure
Step – 5 Find out Minimum Execution Time from each row.
Step – 6 If there is any unique Min. Execution Time in
         Matrix then Select that VM,
            - Assign Request to VM
         else go to step 7.
         - Remove Respected Row and Column from Execution Time Matrix.
         - Update Execution Time Matrix.
Step – 7 Find difference between min. and next min.
         execution time for all that row (VM) which have
         same Requests (R) and select VM for Requests
         which have maximum difference.
         - Assign Request to VM.
         - Remove Respected Row and Column from Execution Time Matrix.
         - Update Execution Time Matrix.
Step – 8 Repeat till all Requests are assigned to VM.
Step – 9 Calculate Total Execution Time.
```

## 5.2 Theoretical Analysis

Theoretical Analysis is done using sample example to understand the working of proposed algorithm and specifications for example are as following,

**Table 1. Request and Resource Specification**

| **VM_Capacity (MIPS)** | **Requests_Length (MI)** |
|---|---|
| 10 | 100 |
| 5 | 45 |
| 20 | 40 |
| 40 | 70 |
| 12 | 144 |

Following is Execution Time Matrix Ex[V,R] of all Requests and VMs.

**Table 2. Execution Time Matrix**

| | Requests | | | | | |
|---|---|---|---|---|---|---|
| Resources (VM) | | **R1** | **R2** | **R3** | **R4** | **R5** |
| | **V1** | 10 | 4.5 | 4 | 7 | 14.4 |
| | **V2** | 20 | 9 | 8 | 14 | 28.8 |
| | **V3** | 5 | 2.25 | 2 | 3.5 | 7.2 |
| | **V4** | 2.5 | 1.125 | 1 | 1.75 | 3.6 |
| | **V5** | 8.33 | 3.75 | 3.33 | 5.83 | 12 |

Here all the rows (VM) are selected and find the minimum execution time for the respective columns.

**Table 3. VM Selection Procedure**

| | **R1** | **R2** | **R3** | **R4** | **R5** |
|---|---|---|---|---|---|
| **V1** | 10 | 4.5 | $\boxed{4}$ | 7 | 14.4 |
| **V2** | 20 | 9 | 8 | 14 | 28.8 |
| **V3** | 5 | 2.25 | 2 | 3.5 | 7.2 |
| **V4** | 2.5 | 1.125 | 1 | 1.75 | 3.6 |
| **V5** | 8.33 | 3.75 | 3.33 | 5.83 | 12 |

In Table 3 find min. execution time value VM, so here V1, V2, V3, V4, V5 has same Requests, so min. execution time difference for all VM needs to determine. Min. execution difference for all VM is respectively 0.5(4.5-4), 1(9-8), 0.25(2.25-2), 0.12(1.12-1) and 0.42(3.75-3.33). Since 1 is maximum difference so V2 is assigned to R3 and further delete Row V2 and Column R3.

**Table 4. VM Selection Procedure**

| | **R1** | **R2** | **R4** | **R5** |
|---|---|---|---|---|
| **V1** | 10 | $\boxed{4.5}$ | 7 | 14.4 |
| **V3** | 5 | 2.25 | 3.5 | 7.2 |
| **V4** | 2.5 | 1.12 | 1.75 | 3.6 |
| **V5** | 8.33 | 3.75 | 5.83 | 12 |

In Table 4 again select minimum execution time for remaining VM. Here V1, V3, V4 and V5 has same Task R2, so min. execution time difference for V1, V3, V4 and V5 is determined, which is respectively 2.5, 1.25, 0.62 and 1.55. Since 2.5 is maximum difference so V1 is assigned to R2 and further delete Row V1 and Column R2.

**Table 5. VM Selection Procedure**

|  | **R1** | **R4** | **R5** |
|---|---|---|---|
| **V3** | 5 | 3.5 | 7.2 |
| **V4** | 2.5 | 1.75 | 3.6 |
| **V5** | 8.33 | $\boxed{5.83}$ | 12 |

In Table 5 again select minimum execution time for remaining VM. Here V3, V4 and V5 has same Task R4, so min. execution time difference for all VM is determined, which is respectively 1.5, 0.75 and 2.5. Since 2.5 is maximum difference so V5 is assigned to T4 and further delete Row V5 and Column R4.

**Table 6. VM Selection Procedure**

|  | **R1** | **R5** |
|---|---|---|
| **V3** | $\boxed{5}$ | **7.2** |
| **V4** | **2.5** | **3.6** |

In Table 6 again select minimum execution time for remaining VM. Here V3, V4 has same Task T1, so min. execution time difference for all VM is determined, which is respectively 2.2 and 1.1. Since 2.2 is maximum difference so V3 is assigned to R1 and further delete Row V3 and Column T1. Finally V4 and R5 remains hence assign V4 to R5.

Finally different Resources have assigned Requests uniquely, which is shown below.

| VM | Requests | Execution Time |
|---|---|---|
| V1 | R2 | 4.5 |
| V2 | R3 | 8 |
| V3 | R1 | 5 |
| V4 | R5 | 3.6 |
| V5 | R4 | 5.83 |
| **Total** |  | 26.93 Seconds |

## 5.3 Comparison

Here proposed algorithm is compared with FCFS algorithm, In FCFS algorithm first Request is allocated to the first resources. This algorithm is batch mode so requests served in Queue. If we calculate execution time through FCFS allocation policy then for same sample example results are as following,

**Table 7. Execution Time Matrix for FCFS Algorithm**

|  | **R1** | **R2** | **R3** | **R4** | **R5** |
|---|---|---|---|---|---|
| **V1** | $\boxed{10}$ | 4.5 | 4 | 7 | 14.4 |
| **V2** | 20 | $\boxed{9}$ | 8 | 14 | 28.8 |
| **V3** | 5 | 2.25 | $\boxed{2}$ | 3.5 | 7.2 |
| **V4** | 2.5 | 1.125 | 1 | $\boxed{1.75}$ | 3.6 |
| **V5** | 8.33 | 3.75 | 3.33 | 5.83 | $\boxed{12}$ |

| VM | Requests | Execution Time |
|---|---|---|
| V1 | R1 | 10 |
| V2 | R2 | 9 |
| V3 | R3 | 2 |
| V4 | R4 | 1.75 |
| V5 | R5 | 12 |
| **Total** |  | 34.75 Seconds |

Here if allocation is done on the basis of the assignment problem's solution method then total execution time taken is 26.93 seconds which is **7.82** seconds **less** than FCFS of 34.75 seconds. So here we can see the improvement of total execution time.

# 6. EXPERIMENTAL RESULTS

CloudSim is used in this paper to implement and simulate the proposed approach in cloud environment. Performance of proposed approach is than compare with FCFS algorithm in terms of Total Execution Time. This research work considers Datacenter, VM, host and Cloudlet components from CloudSim for implementation of a proposed algorithm. Datacenter component handles service requests. VM consist of application elements which are connected with these requests, so Datacenters host should allocate VM requested by user.

Evaluation of proposed approach is done in three different scenarios. Initially in Scenario – A, 5 Requests and VMs having small length and capacity are considered. In Scenario – B, 25 Requests and VMs with the respectively wide range of Lengths and Capacity are considered. In Scenario – C, 50 Requests and VMs having small Lengths and Capacity are considered. Following Table 8 shows the result analysis of Total Execution Time of all scenario and Fig. 1 shows the graph of result analysis.

**Table 8. Result Analysis**

|  | **FCFS Algorithm** | **Proposed Algorithm** |
|---|---|---|
| **Scenario – A** | 34.75 Seconds | 26.93 Seconds |
| **Scenario – B** | 396.00 Seconds | 217.81 Seconds |
| **Scenario – C** | 147.57 Seconds | 118.58 Seconds |

After evaluating and testing in different scenarios Total Execution Time for Proposed Approach is less than FCFS algorithm in every scenario. Here as per Scenario – B greater the range of requests and VM then impact of improvement for proposed approach is high. This approach is OneVMperTask type of approach in which each VM has a single Request, so this approach is mostly suitable for sequential requests in which request length is very large and another requests comes only after current one executed. So each requests are executed on one separate VM.

# 7. CONCLUSION AND FUTURE WORK

In Cloud Computing VM allocation can be referred as problem of resource management which is part of load balancing. One of the main objective of VM allocation is to plan sufficient capacity of all resource to maximize the resource utilization while satisfying client's resource demand. Various approaches related to VM allocation are reviewed in this paper to understand the concepts of VM allocation. In this paper new approach is proposed based on the concept of alternate solution method for assignment problem by considering VM utilization and user's resource demand.

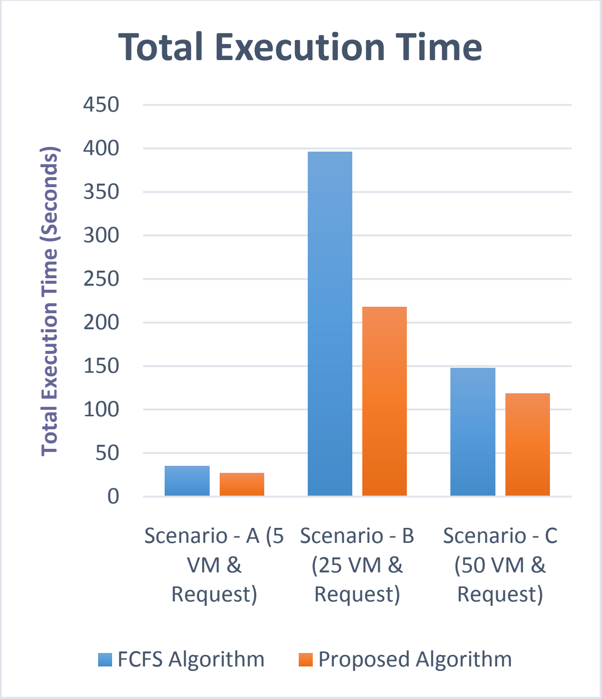

**Fig. 1: Result Analysis Graph**

Theoretical Analysis and Result Analysis of FCFS and proposed approach shows that proposed approach improve the execution time and utilize resources better. In proposed approach it is also found that as the number of requests increase a decrease in execution time is observed. This paper focuses on reducing execution time and balance load by allocating VMs effectively. Many other QoS parameters are not considered like Migration, Cost, etc. and Evaluation of this approach in Real Cloud Environment is a part of future work.

## 8. REFERENCES

[1] Jiayin Li, Meikang Qiu, Yu Chen., "Adaptive Resource Allocation for Preemptable Jobs in Cloud Systems", IEEE 10th International Conference on Intelligent Systems Design and Applications, 2010.

[2] Weiwei Lin, James Z. Wang, Chen Liang, Deyu Qi, "A Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing", Procedia Engineering, Volume 23, Pages 695-703, ISSN 1877-7058, Elsevier, 2011.

[3] S. Ray and A. De Sarkar, "Resource Allocation Scheme in Cloud Infrastructure," 2013 IEEE Int. Conf. Cloud Ubiquitous Comput. Emerg. Technol., pp. 30–35, Nov. 2013.

[4] Pawar, C.S.; Wagh, R.B., "Priority Based Dynamic Resource Allocation in Cloud Computing," 2012 IEEE International Symposium on Cloud and Services Computing (ISCOS), vol., no., pp.1,6, 17-18 Dec. 2012.

[5] Huifang Li; Siyuan Ge; Lu Zhang, "A QoS-based scheduling algorithm for instance-intensive workflows in cloud environment," IEEE The 26th Chinese Control and Decision Conference (2014 CCDC), vol., no., pp.4094,4099, May 31 2014-June 2 2014.

[6] Hongwei Zhao; Wang Chenyu, "A Dynamic Dispatching Method of Resource Based on Particle Swarm Optimization for Cloud Computing Environment," IEEE 10th Web Information System and Application Conference (WISA), vol., no., pp.351,354, 10-15 Nov. 2013.

[7] R. Buyya, C. Shin, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms : Vision, hype, and reality for delivering computing as the 5th utility," Futur. Gener. Comput. Syst., vol. 25, no. 6, pp. 599–616, 2009.

[8] A study on transportation problem, transshipment problem, assignment problem and supply chain management by Gaglani, Mansi Suryakant http://hdl.handle.net/10603/3970.

[9] M. E. Frincu and S. Genaud, "On the efficiency of several VM provisioning strategies for workflows with multi-threaded tasks on clouds," Springer, 2014.

[10] Phyu Thwe, "Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm", International Journal of Scientific & Technology Research (IJSTR), Volume 2, Issue 3, March 2013.

[11] Dilpreet Kaur, A.P. Sukhpreet Kaur, "User Future Request Prediction Using KFCM in Web Usage Mining", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 2, Issue 8, August 2013.

[12] Kaushal Kishor Sharma, Prof. Kiran Agrawal, "A Hybrid Approach for Predicting User's Future Request", Proceedings of Fourth International Conference on Communication System and Network Technologies, IEEE, 2014.

[13] C. Dimopoulos, C. Makris, Y. Panagis, E. Theodoridis, A. Tsakalidis, "A Web page usage prediction scheme using sequence indexing and clustering techniques", Data & Knowledge Engineering 69, 371-382, 2009.

[14] Etzioni, O. "The World-wide web: Quagmire or gold mine," in Communication of the ACM, 1996, 65-68.