Load Balancing in Cloud System using Max Min and Min Min Algorithm

Rajwinder Kaur
M.Tech Student
SBS state technical campus
Ferozepur, India

Pawan Luthra
Assistant Professor
SBS state technical campus
Ferozepur, India

ABSTRACT
Cloud computing provide unlimited resources and services to the end users. Users can access these resources through internet. They need to pay only for those resources as much they use. In cloud computing load balancing is very important technique. Max-min algorithm and Min-min algorithms are proposed for load balancing in cloud computing. These algorithms are implementing based on study of RASA algorithm. Modified max –min used the concept of basic max -min. Modified Max-min is based on the execution time not on complete time as a selection basis. Load balancing algorithm provides better resource utilization and response time. New Max-min helps to achieving load balancing with lower make span rather than RASA and original Max-min**.**

Keywords
Cloud computing, load balancing algorithms, Max-min, Min-min algorithm

1. INTRODUCTION
Cloud Computing is new computing technology. Cloud providers provide unlimited resources and services to the end users. Cloud computing helps to stores the unlimited data and distribute resources in the cloud environment. The amount of data storage increases quickly in cloud computing environment. The data storage increase very fast so load balancing is a main problem in cloud computing. When numbers of tasks occur same time then load balancing is main issue. Load balancing helps to distribute work between all nodes to ensure that no node is overloaded and no need is free. Load balancing helps in best utilization of resources which also helps in improving the performance the system in cloud system. Many load balancing algorithms helped in better load balancing in cloud system and they provide better way for resource allocation to all tasks. In cloud system to achieve maximum profits ,it is very important that all the resources are assign in proper way. In this paper we discuss about some of the existing load balancing algorithms like max-min min-min algorithm. Cloud computing provide many advantages to the end users.[1][2][3][4]

1.1 Cloud Computing Characteristics
1. Services provide on demand- Cloud providers provide resources to the end users on their demand. They request for their resources as much they want.

2. Resource Pooling- The all resources are pooled to serve end users using cloud model. In cloud system resources are allocated according to consumer requirement.

3. Rapid Elasticity- In cloud computing number of resources is increase quickly and easily.

2. CLOUD DELIVERY MODEL
Cloud computing provide many services to the end users. All the services are provide to the end users through the internet.. Fig .1 represents cloud model[5][6][7][8]

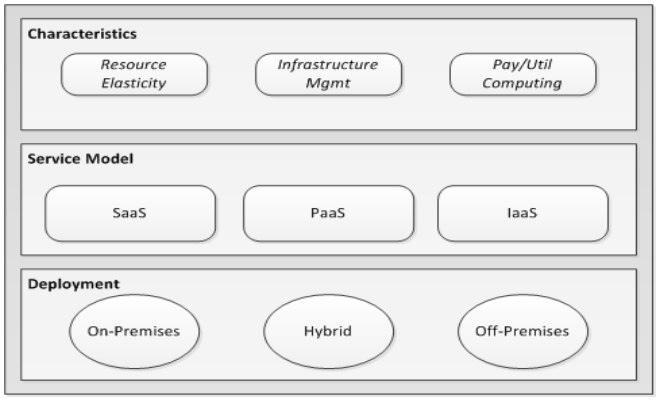

Fig.1 (Cloud Delivery Model)

Cloud computing model consist of three types of services and different types of clouds. There are three cloud services are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) .They all are used to develop public cloud, hybrid cloud, private cloud and community cloud.

2.1 Services of Cloud
Cloud computing provide many services to the end users:

a) Software as a Service: Cloud computing provider provide services to the users but in this user do not need to do installation of hardware and software. In this all the installation controlled by the service providers. There are many services are provide as software as service like Google Docs file, Email cloud.

Advantages of software as service are: It reduced cost because in this no need to buy the software's . So licensing issues are occur related to software. It also reduced of version problem of software.

b) Platform as a Service: Platform as a service provide a platform to the end users. They can used these platform through the internet. In this user can develop and install their own application. They develop applications according to their needs. In platform as service applications can be implemented quickly without any major complexity.

c) Infrastructure as a Service: It provide Infrastructure of servers and different software to the end customers on their demands. So users not need to buy any particular software and hardware. In this services are provide quickly on the users demand.

2.2 Cloud Deployment Model

There are four cloud deployment models as public, private, hybrid and community clouds. Fig.2 represents different types of cloud.

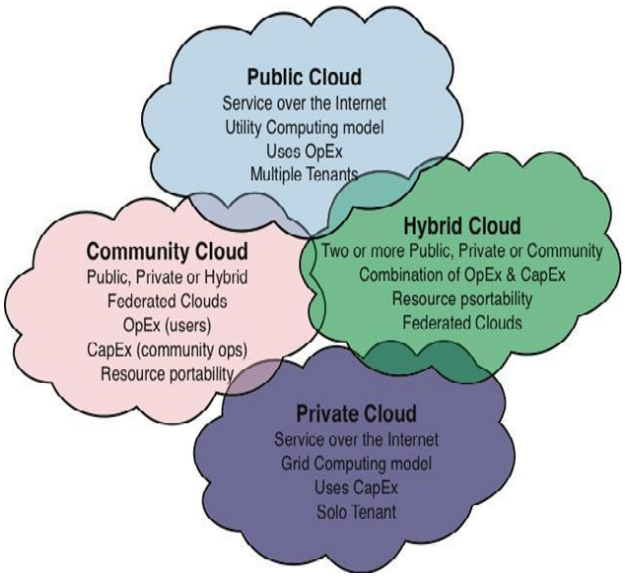

Fig.2 (Cloud Deployment Model)

a) Public Cloud: Public cloud helps to users for access cloud publicly through interfaces. Web browser is a main interface which uses to access the cloud. This cloud computing infrastructure is made available to the general public .[9]

b) Private Cloud: Private cloud is mainly used by a organization. Private cloud provide high level security . In provide more control for deployment of applications. Private cloud provides high level security. Intranet is a type Private cloud. In public cloud all the resources and applications were managed by the provider of services but in private cloud applications and services are combined together and made available for the users. These all services and applications are providing at the organizational level. All the resources and applications are controlled by the organization .Private cloud provide high level security because it is only used by the organization staff, no outsider access private cloud.

c) Hybrid Cloud: Hybrid cloud is a combination of public cloud and private cloud. In private and public work together. Hybrid cloud provides more security. It combines different services and applications and provide these services and resources to the customers through the internet. It allowed the organization to provide its needs in the private cloud and if some common requirements occur it provide the public cloud for computing resources and services.

3. LOAD BALANCING

- In cloud environment load balancing is required Distribute the dynamic local workload evenly between all the nodes. Load balancing helps in better allocation of resource to achieve a high user satisfaction. It helps in implementing fail over and avoiding bottlenecks. [15]
- With the help of Load balancing techniques data can be sent and received without any major delays. In cloud environment many algorithms are available that helps in proper load balancing between all available servers. They can be applied in the cloud system with suitable verifications. Load balancing algorithms are divided into two types in cloud environment: first type of algorithm is Batch mode heuristic scheduling algorithms (BMHA) and second type is online mode heuristic algorithms.
- Many algorithms are provides in BMHA technique. From them First Come First Served algorithm (FCFS), Round Robin scheduling algorithm (RR), Min Min algorithm and Max Min algorithm are main algorithms which provided by BMHA technique.
- In online mode heuristic scheduling algorithm it is very important to estimate proper load , need to do comparison of all load in this technique. CPU load, amount of memory required combine together to calculate the load of machine. In our daily life example of load balancing is websites. Users could experience many major problems without load balancing like delays, timeouts and long system responses.

Two types load balancing algorithms are classified as:

1.static algorithm

2.dynamic approch

3.1 Dynamic Load Balancing Algorithms

They have two types: distributed and non-distributed algorithms.[10][11][12]

1) In distributed algorithms all nodes execute algorithm together, they divide load balancing work between them properly. All the nodes are interacting with each other in two ways: one is cooperative and other is non-cooperative. In the cooperative all the nodes work together but side-by-side to achieve a common goal. .All nodes work together to improve the response time of the system. In Non-Cooperative, all nodes works independently to achieve a common goal for better response time for all tasks.

2) Non-distributed algorithms: In this load balancing is done by one or many resources. Non-distributed dynamic load balancing algorithms have two types: one is centralized approach and other is semi-distributed approach. In centralized approach, the load balancing algorithm is executed only by a single node. In this only central node is responsible for all operation and controlling. In semi distributed approach, all nodes of the system are divided into clusters. Then load balancing is done in centralized form in the clusters. A central node controls all the operations of load balancing. That central node is select by choice in each cluster during load balancing operations.[13][14][15]

3.2 Metrics for Load Balancing Algorithms in Cloud

There are many matrices are used in load balancing .These all matrices are helps in measure the performance of algorithms:

1. Scalability: It means algorithm is able to perform when numbers of tasked are increased quickly in cloud environment. Any algorithm is good when this metric should be improved as compared to other algorithms.

2. Resource Utilizations: It means all the resources are used properly. Better resource utilization it should be better for an efficient load balancing algorithm.

3. Fault Tolerance: It means recovery from all types of failure. The load balancing should be a good fault-tolerant technique. The main faults which occur are like node failure.

4. Response Time: It **is** the time that is taken by a particular algorithm to respond a task. A good algorithm takes minimum time to respond a task.

5. Overhead Associated: It define total amount of overhead involved when implementing a load balancing algorithm in cloud computing environment. Overhead occurs due to the inter process communication between the tasks. If this minimized that load balancing algorithm work properly.

4. MAX MIN MIN MIN ALGORITHM

In these algorithms first minimum execution time and minimum completion time for all the tasks are calculated.[16][17][18]

MET (Minimum Execution Time): MET assigns each task to that resource which performs tasks in minimum time .But it is not considered that resources are available at that time or not.

MCT (Minimum Completion Time): MCT assigns all tasks to those resources which complete them in minimum completion time. It means this assign some tasks to those resources that do not have minimum execution time.

Min-Min: Min-Min start with the set of all unassigned tasks in the makespan. This algorithm work in two phases. First, the minimum expected completion time for all the tasks is calculated. The completion time for all the tasks is calculated on all the machines. In the second phase, the task with the minimum expected completion time from makespan is selected and that tasks assigned to the corresponding resource. Then the task which is completed that is removed from the makespan and this process is repeated until all tasks are completed.

Fig.3 shows basic Max-min algorithm, r_j represents the ready time of resource R_j to execute a task, Cij and E_{ij} represent the expected completion time and Execution time of task. Then task T_k with maximum expected completion time is selected .That task assigned to resource R_j that complete task in minimum execution time.

1.for all submitted tasks in meta-task; T_i

2.for all resources; R_j

3.Cij = Eij + rj

4.While meta-task is not empty

5.find task T_k consumes *maximum completion time*.

6.assignT_k to the resource R_j which gives *minimum execution time*

7.remove T_k from meta-tasks set

Fig.3 (Basic max min algorithm)

Max-Min: Max-Min start with the set of all unassigned tasks in the makespan. This algorithm also works in two phases. First, the maximum expected completion time for all the tasks is calculated. The completion time for all the tasks is calculated on all the machines. In the second phase, the task with the maximum expected completion time from makespan is selected and that tasks assigned to the corresponding resource. Then the task which is completed that is removed from the makespan and this process is repeated until all tasks are completed.

The Max-min and Min-min algorithms estimate the execution time and overall completion time of all the tasks .Then assign the tasks on better suitable resource. The Max-min algorithm is mostly used in distributed environment. It starts with all unexecuted tasks. Expected execution time and expected completion time of each task on the available resources is calculated. Then select the task with overall maximum expected completion time and assign that task to the resource with minimum overall execution time. Finally recently executed task is removed from the meta-tasks set and then update all calculated times, then repeat until all the tasks are completed. [18][19]

5. IMPROVED MAX-MIN ALGORITHM

Max-min algorithm assign task T_i on the resource R_j where large tasks have highest priority and smaller tasks have lower priority. It means, when we have one long task, then Max-min algorithm could execute many short tasks concurrently while executing large task. The makespan is calculated in this by the execution of long task .It would be similar to the Min-min makespan. We try to minimize waiting time of short jobs through assigning large tasks to be executed by slower resources. On the other hand execute small tasks concurrently on fastest resource to finish large number of tasks during finalizing at least one large task on slower resource. Where meta-tasks contain tasks with different completion and execution time, we proposed a new Max-min algorithm that helps in increasing the efficiency of max min algorithm. Improved max min increases the chances of execution of tasks on resources. Max-min algorithm is followed to implement improved Max-min.

Load balancing algorithms enhances performance in distributed systems. Sometimes these algorithms not help in better makespan. There are many existing load balancing algorithms in cloud computing which used for load balancing. Some new algorithms are also implemented from existing algorithms, this will helps to researchers to carry out further work in this area. Because many new modifications are required in load balancing.[19][20][21]

In this algorithm first calculates the expected completion time of the submitted tasks on each resource. Then the task with the overall maximum expected execution time is assigned to a resource that has the minimum overall completion time. After execution scheduled task is removed from meta-tasks and all calculated times are updated and the processing is repeated until all submitted tasks are executed. The algorithm focuses on minimizing the total makespan .That is the total complete time in cloud distributed environment. In cloud computing environment tasks are executed concurrently on available resources to achieving better load.

The proposed algorithm work similar like RASA Selecting task with maximum execution time leads to choose largest task should be executed. Then Resource are select that consuming minimum completion time means choosing slowest resource in the available resources to execute the tasks. That means allocation of the slowest resource to longest task allows availability of high speed resources for finishing other small tasks at same time, it improve the performance of system. Also, we achieve shortest makespan of submitted tasks on available resources beside concurrently. Max-min is used only when tasks are heterogeneous in their completion time and execution time. Heterogeneous means there are clearly defined large and small tasks.[22][23]

Fig.4 represents the modified max min algorithm. Modified algorithm work in different way as compared to basic max min algorithm. In the basic max min

"Select task with max execution time and assign that all tasks to those resources which take min completion time" and in improved max min it is changed" Select task with max completion time and then assign resource which take min execution time".

1. for all submitted tasks in meta-task; T_i
2. for all resources; R_j
3. Cij = Eij + rj
4. While meta-task is not empty
5. find task T_k costs ***maximum execution time***.
6. Assign T_k to the resource R_j which gives ***minimum completion time.***
7. remove T_k from meta-tasks set
8. update r_j for selected R_j
9. update Cij for all j

Fig .4 (Modified Max-Min Algorithm)

Improved max min algorithm implemented from basic Max-min so it has the same time complexity $O(mn^2)$ .Improved max min produces better makespan with more reliable scheduling schema. Improved Max-min helps in better load balancing of available resources and also helps concurrent execution of submitted tasks with higher probability as compared to original Max-min algorithm.

6. CONCLUSION

We have presented a review of load balancing techniques performed over cloud systems .Load balancing is used to obtain better resource utilization and performance. The two algorithms reviewed are Min Min and Max Min along with the modified max min .This study is concentrated only on tasks and resources.

7. ACKNOWLEDGMENTS

The Success of this research work would have been uncertain without the help and guidance of a dedicated group of people in our institute SBS STATE TECHNICAL CAMPUS, FEROZEPUR INDIA. I would like to express our true and sincere acknowledgements as the appreciation for their contributions, encouragement and support.

8. REFERENCES

[1] Sahu, Yatendra and Pateriya, RK, Cloud Computing Overview with Load Balancing Techniques., International Journal of Computer Applications, 2013, 65

[2] Mell, Peter and Grance, Tim, "The NIST definition of cloud computing", National Institute of Standards and Technology, 2009,vol53, pages50, Mell2009

[3] http://www.ancoris.com/solutions/cloudcomputing.html, "Cloud Computing |Google Cloud -Ancoris."'

[4] http://www.personal.kent.edu/~rmuhamma/OpSystems/Myos/prioritySchedule.htm,"PrioritySchedulingOperating Systems Notes."

[5] S. S. MOHARANA, R. D. RAMESH, D. POWAR, "Analysis of load balancers in cloud computing".

[6] http://blog.nexright.com/?cat=6"Cloud Computing « Nexright Blog".

[7] A. Sidhu, S. Kinger, "Analysis of load balancing techniques in cloud computing", INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY 4 (2) (2013) pages737–741.

[8] Parsa, Saeed and Entezari-Maleki, Reza," RASA: A new task scheduling algorithm in grid environment" ,World Applied sciences journal, 2009,vol 7, 152--160

[9] http://www.qualitytesting.info/group/cloudcomputing/forum/topics/software-as-a-s, "Software as a Service (SAAS) - Quality Testing"

[10] http://letslearncloud.wordpress.com/"Cloud Computing | Learn Cloud and its tips."

[11] Chaudhari, Anand and Kapadia, Anushka, " Load Balancing Algorithm for Azure Virtualization with Specialized VM", algorithms,vol 1,pages 2, Chaudhari

[12] Nayandeep Sran,Navdeep Kaur , "Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing ",vol 2,jan 2013

[13] Bala, Anju and Chana, Inderveer, "A survey of various workflow scheduling algorithms in cloud environment", 2nd National Conference on Information and Communication Technology (NCICT), 2011

[14] Chaczko, Zenon and Mahadevan, Venkatesh and Aslanzadeh, Shahrzad and Mcdermid, Christopher, "Availability and load balancing in cloud computing", International Conference on Computer and Software Modeling, Singapore, chaczko2011availability.

[15] www.google.com

[16] http://www.writeaprogram.info/c/os-programs/priority scheduling/"Priority Scheduling Algorithm Example - Write a Program.''

[17] R. Alonso-Calvo, J. Crespo, M. Garcia-Remesal, A. Anguita, V. Maojo, "On distributing load in cloud computing: A real application for very-large image datasets", Procedia Computer Science 1 (1) (2010) pages 2669–2677.

[18] S.-C.Wang, K.-Q. Yan, S.-S.Wang, C.-W. Chen, "A three-phases scheduling in a hierarchical cloud computing network", in: Communications and Mobile Computing (CMC), 2011 Third International Conference on,IEEE, 2011, pp. 114–117.

[19] Y. Hu, R. Blake, D. Emerson, "An optimal migration algorithm for dynamic load balancing", Concurrency: Practice and Experience 10 (6)(1998) pages 467–483.

[20] Etminani, Kobra and Naghibzadeh, M, A min-min max-min selective algorihtm for grid task scheduling, Internet, 2007. ICI 2007. 3rd IEEE/IFIP International Conference in Central Asia on, 2007.

[21] O. Elzeki, M. Reshad, M. Elsoud, "Improved max-min algorithm in cloud computing, International Journal of Computer Applications"vol 50 (12) (2012)pages 22–27..

[22] U. Bhoi, P. N. Ramanuj, " Enhanced max-min task scheduling algorithm in cloud computing".

[23] H. Chen, F. Wang, N. Helian, G. Akanmu, "User-priority guided min-min scheduling algorithm for load balancing in cloud computing".