Hierarchical Load Balancing Algorithms in Cloud: A Survey

B. Priya
Research Scholar, Anna University,
Assistant Professor, Sri Sai Ram Engineering College, Chennai

T. Gnanasekaran, PhD
Professor, Department of IT,
RMK Institutions, Chennai,

ABSTRACT
Cloud Computing is a methodology for distributing and accessing applications across the network. The various parameters considered in Cloud are: fault tolerance, high availability, scalability, and flexibility, reduced overhead for users, reduced cost of ownership, on demand services. The set of rules and policies that control the order by which various jobs are executed in a system form the basis for scheduling. Load balancing algorithms attempts to improve the response time of the user's submitted applications by ensuring maximum utilization of available resources. Load balancing deals with the way by which the various tasks are assigned to the resources thereby improving the system performance. Scheduling of various tasks to the resources in a Cloud environment is an active research area. Resources are dynamic in nature so the load of resources varies with change in configuration of Cloud and hence Load balancing of the tasks in a Cloud environment can significantly influence the Cloud's performance. The hierarchical load balancing concept uses the tree data structure to make decision regarding the placement of tasks on Virtual Machine. In order to utilize the resources efficiently and to satisfy the QoS requirement of the users, several hierarchical load balancing algorithms have been proposed by researchers for various applications. This paper deals with the overview of the load balancing concepts in Cloud with an assessment of the various hierarchical load balancing algorithms.

Keywords
Hierarchical; Virtual Machine(VM); OLB; Load Balancer, LBMM

1. INTRODUCTION
Cloud computing is an internet based computing model that evolved out of the technology of distributed computing. NIST definition of cloud computing "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, server, storage, application, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". [1] Cloud computing minimizes the cost of computation, storage and application hosting.

The term cloud computing describe diverse computing concepts that comprise of several computers connected through a network. It allows the users to utilize the required resources without the knowledge of underlying delivery mechanism. Forrester defines cloud computing as: "A pool of abstracted, highly scalable, and managed compute infrastructure capable of hosting end-customer applications and billed by consumption". Cloud computing is a model for enabling appropriate, on-demand network access to a shared pool of configurable computing for rapid provisioning and release with minimal service provider interaction. In cloud computing tasks are executed with the computing resources for achieving optimal performance, maximum resource utilization and minimum response time.

Cloud computing uses the concept of Virtualization for providing services to the client. Load balancing is an important factor in cloud computing. This concept involves numerous nodes that get the task dynamically distributed between them to attain optimal utilization of resources, thereby increasing the performance without overloading a single node. An efficient load balancing technique provides an ideal environment, which improves the user satisfaction. Load Balancing helps in meeting the QoS requirements as well as maximizing the profit of the Cloud Service Providers with optimal resource usage.

1.1 Cloud And Grid Scheduling
The following are some of the issues in Grid Scheduling:

1. Grid Schedulers do not own resources themselves
   a. They have to negotiate with autonomous local schedulers.
   b. It is associated with authentication/Multi organizational requirements.
2. Grid schedulers have to interface with multiple local schedulers
   a. Some of the local schedulers may have support for reservations, others are queuing-based.
   b. Some may support Checkpointing, migration etc.
3. Application Structure
   a. Adaptation module of the application may need different structures.

Compared to Grid Schedulers, efficient Load Balancing in the cloud involves scheduling the resources and workloads in an effective manner. Different scheduling algorithms are used by load balancers to determine the backend server to which the request has to be directed.

The designated server defines the virtual machine (VM) on the same physical machine by allocating the required resources and scheduling the jobs dynamically. The provider does dynamic reallocation or migration of VM across physical machines for workload consolidation and to avoid over utilization or underutilization of resources.[2]

The rest of the paper is organized as follows: Section II deals with the classification of load balancing concepts in cloud environment with the various metrics for measuring performance. Section III explains the principles and advantages of hierarchical load balancing algorithms. Section IV reviews the various hierarchical load balancing algorithms. Section V compares different hierarchical load balancing algorithms.

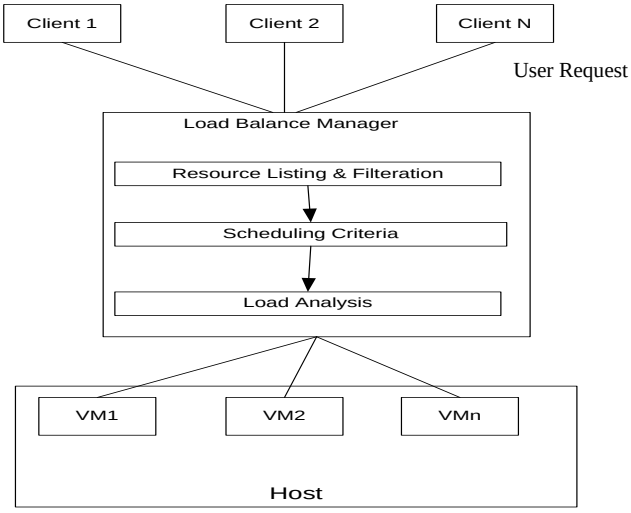

Fig 1: Load Balancing in Cloud

2. CLASSIFICATION OF LOAD BALANCING APPROACHES IN CLOUD ENVIRONMENT

The load balancing approaches in cloud are classified as:

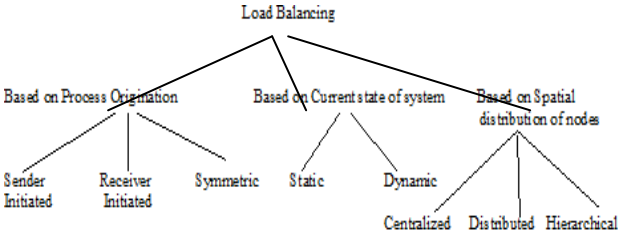

Fig 2: Classification of load balancing approaches

2.1 Process Origination:

a. Sender Initiated:
The request is sent by the client until a receiver is assigned to receive the workload i.e. the sender initiates the process.

b. Receiver Initiated:
Request is sent by the receiver to acknowledge a sender who is ready to share the workload i.e. the receiver initiates the process.

c. Symmetric:
A combination of both sender and receiver initiated type of load balancing algorithm.

2.2 System State [6] :

a. Static Environment:
This is applicable in homogenous environment where a prior knowledge is needed about each node statistics and user requirements. This cannot adapt to runtime changes in the cloud.

b. Dynamic Environment:
This is applicable in heterogeneous environment and the load balancing decisions are initiated at run time. It is complex and time consuming.

2.3 Spatial distribution of nodes[7]:

a. Centralized Load Balancing:
The scheduling decisions are controlled by a single central node. This node has the knowledge of entire cloud network and possible failure. But it is not fault tolerant and can be overloaded.

b. Distributed Load Balancing:
The scheduling decisions do not rely on any particular single node. Multiple nodes with their databases are responsible for load balancing decisions and it incurs communication overhead.

c. Hierarchical Load Balancing:
The nodes at different level denoted by a tree data structure coordinate with the nodes at a level below the hierarchy to make decisions.

2.4 Metrics for Performance Measurement:

The performance of the cloud is evaluated by its characteristics such as resource allocation and efficient scheduling. Essential metrics needed to access the scheduling and load balancing algorithms are [4][5][14]:

1. Communication overhead: The status information that every node has to convey to other nodes.
2. Make span: The total completion time taken to allocate all tasks to a resource. i.e. the measure of the throughput of the system.
3. Average Resource utilization rate: This is related to the average usage of all the resources.
4. Fault Tolerance: It is the ability of the algorithms to perform uniform load balancing in spite of arbitrary node or link failure.
5. Reliability: The ability to schedule the job in predetermined amount of time.
6. Migration Time: Time taken for job or resource to migrate from one node to the other. It should be minimized to enhance the performance of the system.

3. HIERARCHICAL LOAD BALANCING

Following are the principles of hierarchical load balancing algorithms:

1. They are distributed and heterogeneous.
2. Reduced communication delay.
3. Load index based on CPU Utilization, Queue length and Communication delay acts as a decision factor for scheduling the tasks.
4. The various criteria associated are total execution time, number of tasks and the number of nodes.

Advantages of hierarchical load balancing algorithms are:

1. The Hierarchical structure provides a better performance when compared to other load balancing. The information flow is eased through tree and message traffic is well defined.
2. Easy management of Cluster and Node supports heterogeneity and scalability of clouds.

In hierarchical approach, the scheduling and load balancing is performed at various levels. Each level uses a scheduling algorithm to assign work to the next lower level. Every node in the tree is balanced under the supervision of its parent node.

The load balancing executed at lower levels of hierarchy, limits the amount of information passed to the upper level, which decreases the communication delay and the response time. It also reduces the idle time of the processing entities as the load balancing is done faster.

The three phases of the resource request in the cloud are

Phase 1: This phase involves formation of VMs ready for the scheduler to schedule the job
Phase 2: VMs start the processing of the jobs after allocation.
Phase 3: VMs are deleted**.**

4. RELATED WORK

Geetha C. Megharaj et al[4] proposed a Two-level Hierarchical scheduling model. The two-level hierarchical scheduling model is proposed with the Global Centralized Scheduling Center (GCSC) at higher level and the Local Centralized Scheduling Center (LCSC) at next level. Data center queries global centralized scheduler for allocation of virtual machine when a service request is received. Every local centralized scheduler provides load information to the global centralized scheduler. The load will be transferred to the appropriate local centralized schedulers. The global centralized scheduler also handles the tasks sent by the user and returns the results to the user. The local centralized scheduler gathers the load information from the computing nodes and the load is balanced in that local area. Different computing nodes get assigned different tasks when the local centralized schedulers receives request from the global centralized scheduler. The local centralized schedulers also gather the results from every computing node so to provide the global centralized scheduler.

Shu-Ching Wang et al., [8] proposed a two phase scheduling algorithm OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) under three level cloud computing network. Request manager assigns task to a suitable service manager as part of the first level. As part of the second level the service manager splits the task into subtasks. In the third level the service node is used to execute a subtask. Under the OLB scheduling algorithm, unexecuted tasks are dispatched by the request manager to currently available service manager at random order without taking into account the current workload of the service manager. The task is divided into subtasks by the service manager. Under the LBMM scheduling algorithm, calculation of execution time of each subtask on each service node is carried out and subtasks are distributed to the service node that consumes minimum execution time. The two phase scheduling algorithms results in better execution efficiency with a good load balancing of a system.

The scheduling model in [9] consists of three phases. In the first phase, the BTO (Best Task Order) the execution order for each task request is scheduled. The second phase involves EOLB (Enhanced Opportunistic Load Balancing) scheduling , an appropriate service manager handles allocation of the service node. The third phase comprises EMM (Enhanced Min-Min) scheduling confirms that a suitable service node will be allotted for task execution with minimum execution time.

Susheel Jain et al[10] by minimizing the latency, an efficient, power saving and uniform utilization of resource is reflected in their hierarchical model without impacting the cloud processing speed. The VM operations are controlled in this approach. Selection of VM within the cluster by the scheduler is in a manner that saves the power even in presence of the load balancing issues.

In [11] Minimization of energy consumption of servers and network devices propose a Hierarchical Scheduling algorithm (HAS). In this algorithm, optimization of application placement on servers connected to a common switch is done by a Dynamic Maximum Node Sorting (DMNS) method. To lessen the number of running servers, a Hierarchical crossing-switch adjustment is applied. This results in the reduction of amount of data transfer and the number of running servers.

A weighted self-scheduling scheme[12] is applied to achieve good load balancing to present a Hierarchical Distributed Loop Self-Scheduling Scheme for Cloud. This scheme also reviews the distribution of the output data, to help reduce communication overhead.

Jixiang Yang et al[13] proposed a hierarchical dynamic load balancing strategy. This strategy is based on generalized neural network (HLBSGNN) that considers time-varying characteristics of communication delays in large distributed systems. This hierarchical load balancing strategy minimizes the overhead of the load balancing in large distributed computing systems using communication-optimized hierarchy. In the new strategy, the computational rate of node and time-varying characteristics of communication delay are taken into account and a delay prediction model based on generalized neural network (GNN) theory is designed. This provides an effective optimization method for load balancing strategies.

5. COMPARISION OF VARIOUS HIERARCHICAL ALGORITHMS

Table 1. Hierarchical Schemes - Comparison

S.No	Scheme	Algorithms	Metrics
1	Two level[4] hierarchical Scheduling Model	Global Centralized Scheduling Center (GCSC) at higher level and the Local Centralized Scheduling Center (LCSC) at next level	Communication Cost – Reduced
2	Two Phase[8] Scheduling Algorithm	Opportunistic Load Balancing (OLB) and Load Balancing Min Min (LBMM)	Execution Time - Minimized
3	Three Phase Scheduling[9]	Best Task Order(BTO), Enhanced OLB(EOLB), Enhanced Min Min(EMM)	Makespan – Increased by considering the execution time for each task request for allocation to VM

4	Hierarchical Scheduling Algorithm (HAS) [11]	Dynamic Maximum Node Sorting (DMNS)	Energy Consumption - Minimized
5	Hierarchical Resource Switching and Load Assignment Algorithm[10]	Levels within Clusters	Latency – Reduced Resource Utilization – Efficient Power Consumption – Reduced
6	Hierarchical Load Balancing Strategy based on Generalized Neural Network (HLBSGNN)[13]	Delay Prediction Model – GNN	Communication Overhead - Reduced

6. CONCLUSION AND FUTURE WORK

In this paper the state of the art of the Hierarchical load balancing algorithms in cloud is reviewed. The various hierarchical schemes are compared with consideration of the various parameters. It is intended to propose a hierarchical load balancing model in cloud to increase the make span and reduce the communication overhead.

7. REFERENCES

[1] Peter Mell,Timothy Grance, "Cloud Computing" by National Institute of Standards and Technology – Computer Security Resource Center-www.csrc.nist.gov.

[2] Vijayalakshmi A. Lepakshi, Dr. Prashanth C S R, "A Study on Task Scheduling Algorithms in Cloud Computing", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 11, May 2013

[3] Han Xiangchun, Chen Duanjun, Chen Jing, "one centralized scheduling pattern for dynamic load balance in Grid", 2009 international forum on Information Technology and Applications.

[4] "Two Level Hierarchical Model of Load Balancing in Cloud", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459,Volume 3 Issue 10, October2013.

[5] Foster, I., Y. Zhao, I. Raicu and S. Lu, "Cloud computing and Grid Computing 360-degree compared," in proc. Grid Computing Environments Workshop, pp: 99-106, 2008.

[6] Nitin Kumar Mishra, Nishchol Mishra, " Load Balancing Techniques: Need, Objectives and Major Challenges in Cloud Computing- A Systematic Review", International Journal of Computer Applications (0975 – 8887) Volume 131 – No.18, December 2015.

[7] Mayanka Katyal, Atul Mishra "A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment", International Journal of Distributed and Cloud Computing Volume 1 Issue 2 December 2013.

[8] Shu-Ching Wang, Kuo-Qin Yan, Wen-Pin Liao and Shun-Sheng Wang "Towards a Load Balancing in a Three-level Cloud Computing Network", Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on (Volume:1): 9-11 July 2010.

[9] Shu-Ching, Wang Kuo-Qin, Yan*(Corresponding author) Shun-Sheng, Wang Ching-Wei, Chen, "A Three-Phases Scheduling in a Hierarchical Cloud Computing Network", 978-0-7695-4357-4/11 $26.00 © 2011 IEEE

[10] Ambika Mishra, Prof. Susheel Jain, Prof. Anurag Jain, " A Hierarchical Resource Switching and Load Assignment Algorithm for Load Balancing in Cloud System", International Journal of Scientific & Engineering Research, Volume 5, Issue 3, March-2014 1179 ISSN 2229-5518

[11] Gaojin Wen, Jue Hong, Chengzhong Xu, Pavan Balaji, Shengzhong Feng, Pingchuang Jiang, "Energy-aware hierarchical scheduling of applications in large scale data centers 2011 IEEE International conference on Cloud on Service Computing.

[12] Yiming Han and Anthony T. Chronopoulos "A Hierarchical Distributed Loop Self-Scheduling Scheme for Cloud Systems", 2013 IEEE 12th International Symposium on Network Computing and Applications.

[13] Jixiang Yang, Ling Ling and Haibin Liu, "A Hierarchical Load Balancing Strategy considering Communication Delay Overhead for Large Distributed Computing Systems", Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2016, Article ID 5641831, 9 pages http://dx.doi.org/10.1155/2016/5641831

[14] V. P. Narkhede, Prof. S. T. Khandare, "Fair Scheduling Algorithm with Dynamic Load Balancing Using In Grid Computing" International Journal Of Engineering And Science Vol.2, Issue 10 (April **2013)**, Pp 53-57 Issn(e): 2278-4721, Issn(p):2319-6483

[15] Rajesh George Rajan, V.Jeyakrishnan, "A Survey on Load Balancing in Cloud Computing Environments", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013

[16] Sajjan R.S, Biradar Rekha Yashwantrao, "Load Balancing and its Algorithms in Cloud Computing:A Survey", International Journal of Computer Sciences and Engineering, Volume 5, Issue 1, 2017.