A State-of-Art Load Balancing Algorithms in Cloud Computing

R. Ramya
M. Phil Research Scholar in CS
St. Joseph's College
Trichy, TamilNadu, India

M. Krishsanth
PhD. Research Scholar in CS
St. Joseph's College
Trichy, TamilNadu, India

L. Arockiam
Associate Professor in CS
St. Joseph's College
Trichy, TamilNadu, India

ABSTRACT

Cloud Computing is the ultimate technology in internet. It is a means of acquiring computing possessions, making do and delivering software and services. Cloud Computing allows the customers to apply the application without set up and access their own files on any device with internet. Cloud services can be acquired at any time. The cloud service providers have developed enough to provide services to an ever growing number of users. Sudden peak situation the load balancing mechanism helps to send the requests to the available resources. In this report, we introduce an overview of load balancing and load balancing algorithm.

General Terms

Load balancing Algorithms.

Keywords

Cloud Computing, load balancing and scalability

1. INTRODUCTION

Cloud computing is the fastest developing technology in IT industry. It permits the customers to apply the application without set up and access their own files on any device with internet. It reached commercial success because of pay as you go model. Cloud computing is defined by several sources. Buyya et al. [1] Have defined it as follows: "Cloud is a parallel and distributed computing system consisting of an aggregation of interconnected and virtualized computers that are dynamically provisioned and delivered as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service supplier and clients".

The National Institute of Standards and Technology (NIST) [2] characterizes cloud computing as " cloud computing as a pay-per- use example for enabling available, convenient, on-demand network access to a shared pool of configurable computing resource that can be rapidly provisioned and released with minimal management effort or service provider interaction."

The Gartner group [3] defines cloud computing as " a method of computing in which mostly scalable IT-related capabilities are provided "as a service" using internet technologies to multiple external customers".
Cloud computing has three service models and four deployment models. It has five necessary characteristics are on-demand self-service, extensive network access, resource pooling, rapid elasticity, measured service. The cloud computing framework as [4] Fig. 1 depicts the deployment and service model in cloud computing. The service models as follows.

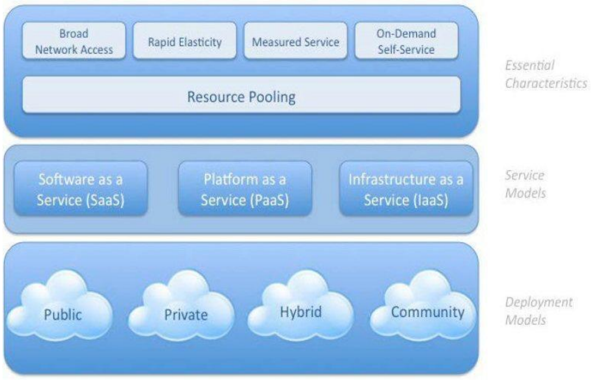

Fig 1: Cloud computing Framework [4]

1.1 Software as a Service (SaaS): SaaS is getting an ever more prevalent delivery model as underlying technologies that support Web services and service-oriented architecture (SOA) mature and new developmental approaches, such as Ajax, become popular. Meanwhile, broadband service has become increasingly available to back up user access from more nations around the globe.

1.2 Platform as a Service (PaaS): It serves the platform for users who is evolving the software. The service delivery model allows the customer to rent virtualized servers and associated services for moving existing applications or producing and trying new ones. PaaS offerings may also include facilities for application plan, application progress, testing, and deployment as well as services such as team collaboration, web service incorporation, and marshalling, database integration, security, scalability, storage, persistence, state administration, application versioning, application instrumentation, and developer community facilitation.

1.3 Infrastructure as a service (IaaS): It serves machines, storage and network resources that developers can manage by installing their own operating system, applications and support resources. The service provider owns the equipment and is responsible for housing, extending and preserving it. The client typically pays on a per usage basis. The definition includes such offerings as practical server space, net connections, bandwidth, IP addresses and load balancers. Iaas can be used by enterprise customers to produce cost efficient and easily scalable IT solutions where the difficulty and expenses of managing the underlying hardware are outsourced to the cloud provider.
The Deployment models are

Private Cloud: it's run within a company's own data center or base for internal and partner employment.

Public Cloud: It's available on a subscription basis (pay as you go) that means whatever the customer use that simply pays the sum.

Hybrid Cloud: It's a combination of secret and public cloud.

Community Cloud: It's a shared by two or more establishments. A community cloud in computing is a mutual effort in which infrastructure is mutual between several systems from a specific society with common concerns. Whether done inside or by a third-party and hosted within or outwardly.

The remaining part of the paper is coordinated as follows section 2 describes scalability; Section 3 describes the load balancing. Part 4 describes the architecture of load balancing and Section 5 describes the load balancing algorithms.

2. SCALABILITY

Scalability is a cloud's ability to present application process and methods to ever increasing number of users. There are two types of methods as follows [5]

- Horizontal Scalability (Scale out &in)
- Vertical Scalability (Scale up&down)

Horizontal Scalability: It refers to the cloud's ability to connect multiple hardware or software entities. So they make single unit [6].

Vertical Scalability: It refers to the cloud's ability to extend the capability of existing hardware or software by adding the resources [7,8]. The key terms which are linked up to scalability are listed below [9].

- Load balancing
- Workload
- Scheduling
- Resource allocation
- Quality of service
- Service Level Agreement

3. LOAD BALANCING

Load balancing is a proficiency to produce resources, developing parallelism, exploiting throughput invention and to cut down response time through the use of the appropriate distribution of the application [10,11]. The goal of load balancing is minimizing the average response time. The user should not wait in queue. The load balancing process can be defined in three rules [12]: Location rule: it determines which resources domain will be included in the balancing operation. Distribution rule: it establishes the redistribution of the workload among available resources in the area. Selection rule: it decides whether the load balancing operation can be performed preemptively or not. There are various types of load [12,13]

- ◦ CPU load
- ◦ Memory load
- ◦ Network or delay load

The goals of load balancing are [11,14]

- Substantial improvement in performance
- Stability maintenance of the system
- Increase flexibility of the system so as to adapt to the modifications.
- Build a fault tolerant system by creating backups.

Basic Concepts

The following concepts are used in load balancing process [15]

- **Context Switch:** Computing process of storing and restoring state of the CPU and the implementation can be summed up from the same period
- **Throughput:** it is defined as the number of processes completed per unit time.
- **CPU Utilization:** to keep the CPU as busy as possible.
- **Waiting time:** it is the amount of time a process has been waitressing in the ready queue.
- **Reaction time:** it is the time it needs to start answering, not the time it uses up to output the answer.

4. LOAD BALANCING ARCHITECTURE

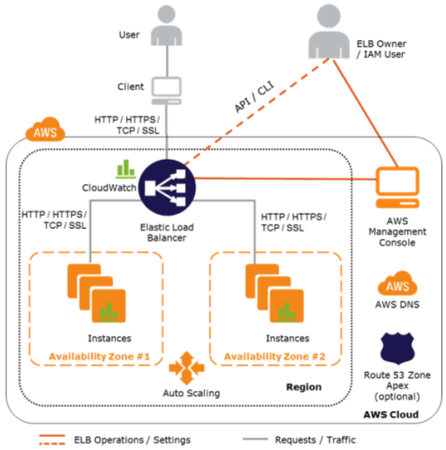

Fig 2: load balancing Architecture[16]

Fig 2 shows the load balancing structural design in amazon.com. There are two reasonable units in the load balancing architecture, namely Load balancers and a controller service.. The load balancers are resources that monitor traffic and handle requests from the net. The controller service supervises the load balancers, put in and convey away the capacity as required and get sure that load balancers perform accurately.

The service is automatically put in or takes out the resources as they needed without any manual participation.

5. TYPES OF LOAD BALANCING ALGORITHM

Load balancing has two types of algorithm [13,11].

- Static Algorithm
- Dynamic Algorithm

5.1 Static Load Balancing Algorithm

The load does not depend on the current state and it requires knowledge about the applications and resources of the system.

The goal of static algorithms is to reduce the overall execution time of synchronous programs with minimum communication delays [11,13].

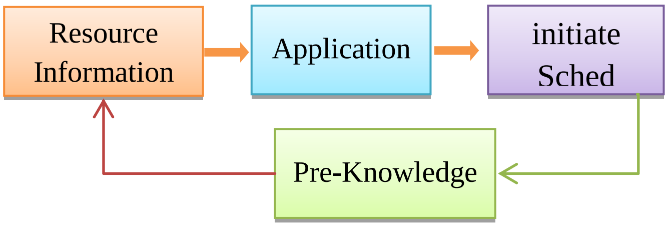

Fig 3: Static Algorithm[17]

Static load balancing algorithms allocate the tasks of a parallel program to workstations based on either the payload at the time nodes are allocated to some task, or based on an ordinary burden of our workstation cluster. The decisions linked to load balance are made at compile time when resource demands are calculated

5.2 Dynamic Load Balancing Algorithm

The load depends on the current state and no demand for prior knowledge. The advantage is if any node fails, it does not end the operation. The goal of dynamic load balancing is to allocate job dynamically [11,13].

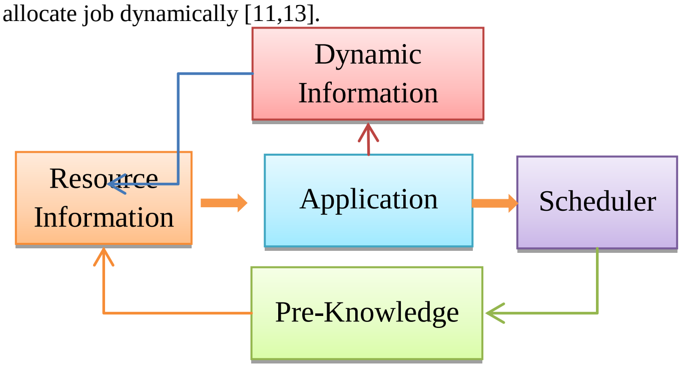

Fig 4: Dynamic Algorithm [17]

Dynamic load balancing algorithms make changes to the distribution of work among workstations at run-time; they use current or recent load information when making distribution decisions**.**

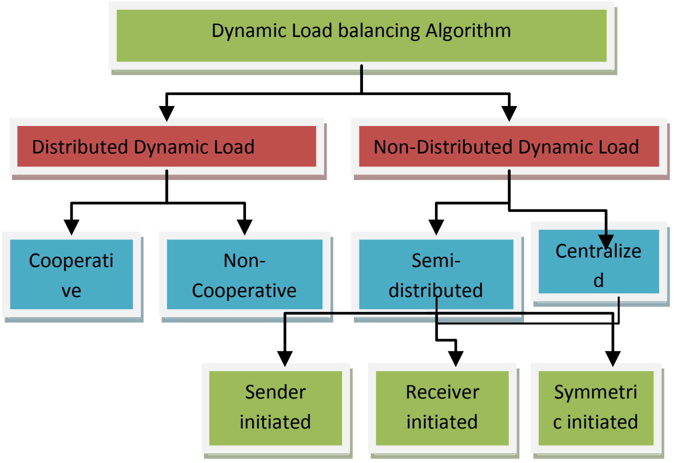

Fig 5: Types of Dynamic Load Balancing Algorithm

The Dynamic algorithm uses the four policies [17]

- Transfer policy
- Selection policy
- Location policy
- Information policy

Transport policy: it is a constituent of the dynamic load balancing algorithm which chooses a job transferring from the local node to a distant client.
Selection policy: it specifies the processors involved in the load exchange.
Location policy: it is constituent of the dynamic load balancing algorithm which chooses a destination node for a transferred task.
Information policy: it is constituent of the dynamic load balancing algorithm responsible for gathering the information about nodes in the arrangement.
Dynamic algorithm has two types [11,18]

- Distributed
- Non-Distributed

5.2.1 Distributed dynamic load balancing

It is executed by all the nodes resent in the system and task of scheduling is shared among them. It has two types

- Cooperative dynamic load balancing
- Non –cooperative dynamic load balancing

Cooperative: the nodes work together to accomplish mutual aims.

Non-cooperative: Each node works independently towards the finish.

5.2.2 Non-Distributed dynamic load balancing

The nodes work individually in order to arrive at a common goal.. It has two types

- Semi-distributed
- Centralized

Semi-distributed: The nodes of the system are divided into clusters.

Centralized: The algorithm is executed only by a solitary node in the whole scheme. It further classified into

- Sender initiated
- Receiver initiated
- Symmetric initiated

In the sender initiated, the client sends the request until the receiver is assigned to him, to receive his work load. In the receiver initiated receiver sends request to acknowledge who is ready to share the workload. In the symmetric, it is a combination of both sender and receiver initiated a type of load balancing algorithm.

5.3Round Robin Algorithm

It works in the round manner where each customer is allotted with a time slice and has to wait for their turn. The time is divided and interval is allotted to each node. The process is waiting for their turn [11, 14, 16].

The advantage is the process is worked sequentially. Each process is allotted for their priority basis. The disadvantage is the process is waiting for their turn. So waiting time is increased.

5.4 Biased Random Sampling Algorithm

It functions based on the virtual graph connectivity between the client. Here each node is assumed as the cloud system. It delivers two types of edges are incoming and outgoing edges. The incoming edge is applied to give the input and output edge is rather the result of the process [13].

5.5 Throttled Load balancing Algorithm

It functions by setting the appropriate virtual machine for setting aside a special project. It's comfortable to allocate job for correct the VM.

The advantage is to assign the VM for the job. The disadvantage is if it's not finding correct VM it did not specify the VM for job [14].

5.6 Active Clustering Algorithm

It works as based on the clustering concept. The same type of nodes is grouped and working as the group. It increases the throughput of the system [13].

5.7 Equally Spread Current Execution Algorithm

It takes a load balancer which monitors the jobs which are needed for execution. The algorithm performs using cloud analyst simulation [14, 19].

The disadvantage is if the business asks for execution otherwise it did not allocate the process for that business.

5.8 Honey Bee Foraging Algorithm

This algorithm is utilized to allocate job dynamically. This algorithm also balances the priorities of tasks on the machines in such a way that the amount of waiting time in the waiting line is minimal. This load balancing technique works well for heterogeneous cloud computing systems [20].

5.9 Index Name Server Algorithm

The algorithm is applied to determine the data duplication and data redundancy. It integrates the access point selection optimization. The algorithm is utilized to find out whether the connection can handle additional clients or non. There are classified three levels: B (a) means the connection is very busy and cannot maintain additional resources. B (b) means the connector is not busy and additional connector can be appended. B (c) implies that the association is set. It uses Distribution Hash Table [21].

The advantage is three levels are maintained, which is employed to allocate the burden. The disadvantage is only certain parameters are taken such as space and time.

5.10 Dual Direction downloading from FTP servers (DDFTP) Algorithm

The algorithm operates by splitting a file m into m/2 separation. Each server node starts the processing the task allotted to it based on certain patterns. One server will start from block 0 and keep downloading the file incrementally while another server will take up from block m and download file incrementally. Meanwhile, both servers are over with their task [22,23].

The advantage is to bring down the network communication needed between the node and nodes. The disadvantage is the reproduction of data files requires high memory in all clients.

5.11 Load balancing Min-Min (LBMM) Algorithm

LBMM has a three level load balancing framework. The first level of the LBMM architecture is the request manager which is responsible for receiving the task and assigning it to one service manager. When the service manager gets the request an divide task into sub projects to speed up processing that request. The service manager assigns sub tasks into service node which is responsible for execution of chores. The project is allotted based on some attributes such as remaining CPU space, remaining memory [23, 24].

The advantage is reliable task assignment to clients. The disadvantage is slower than other algorithms because the task must pass three layers to process.

6. COMPARISON OF THE LOAD BALANCING ALGORITHMS

Table 1 illustrates the advantage and disadvantage of reviewing algorithms. It shows the positive and negative point of each algorithm.

Table 1: Advantage and disadvantage of algorithms

Algorithm	Advantage	Disadvantage
Round Robin [11,14,16]	Priority basis Sequence processing	Increase waiting time
Equally Spread Current Execution Algorithm [14,9]	Allocate process based on job waiting	If the job asks for execution, then only allocate process otherwise it did not allocate the process for that occupation.
Throttled Load balancing [14]	Allocate job for suitable VM	The finding is difficult
Honey Bee Foraging [20]	Allocate job dynamically Priorities of task well for heterogeneous system	Priority basis
Active Clustering [13]	Same nodes are grouped Easy to process	Heterogeneous nodes are not counted.
Index Name Server [21,23]	Initially prove handle some algorithms	Only certain parameters are considered such as time and distance
Dual Direction downloading from FTP servers (DDFTP) [22,23]	Reduce network overhead Reliable to download files	Requires high storage in all nodes
Load balancing Min-Min (LBMM) [23,24]	Reliable to assign the task	Slower than other algorithms

7. CONCLUSION

Cloud computing is fastest growing technology in IT. It permits the users to access the application without setup and access their own files on any device with internet. Load balancing is one of the foremost issues of cloud computing. So load balancing algorithms help to improve the carrying into action of the load balancing. It shortens the waiting time. It might be to ameliorate the quality of service. Load balancing algorithms are improving utilization of computing resources. In this paper, we presented the overview of load balancing. It gives the comparison of the load balancing algorithms. In future load balancing algorithms are conquer shortcoming and improve efficient use of computing resources.

8. REFERENCES

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg and I. Brandic, Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility, Future generation system, 2009.

[2] P. Mell and T. Grance, "The NIST Definition of Cloud Computing", National Institute of Standards and Technology, Information Technology Laboratory, Technical Report version 15 ,2009.

[3] Definition of cloud computing-http://www.fastcloud.org

[4] National Institute of Standard and technology. csrc.nist.gov/groups/ SNS/cloud-computing/cloud-def-v15.doc, 2009

[5] Rahul Sharma, Mohit Mathur, " Achieving vertical scalability: A hindrance to cloud computing", Proceedings of the 4th National Conference; INDIACom, 2010.

[6] Horizontal scalability- http://searchcio.techtarget.com

[7]vertical-scalability: http://www.techopedia.com/definition/15323/

[8] Vertical Scalability: http://searchcio.techtarget.com

[9] Somasundharam, Prabha, Arumugam, "Scalability issues in cloud computing", Advanced Computing (ICoAc), 2012, pp 1-5

[10] Yingchi Mao, Xi Chen, Xiaofang Li, "Max- Min Task Scheduling Algorithm for Load Balance in Cloud Computing", Springer, 2014, pp 457-467

[11] Shiny, "Load Balancing in cloud computing", IOSR-Journal of Computer Engineering", Vol 15, issue 2,2013, pp 22-29

[12] Kavitha, Shandip Kumar, Sahil Verma, "Fault tolerant Approach for Load balancing in Grid Environment", IJERT, Vol 1, issue 9 2012, pp 1-6

[13] Yatendra Sahu, R. K. Pateriya, "Cloud computing overview with load balancing Techniques", International Journal of Computer Applications, 2013, pp 40-44

[14] Jamuna, Anand Kumar, "Optimized cloud partitioning technique to simplify load balancing", International journal of Advanced research in Computer Science and Software Engineering, Vol 3, issue 11,2013, pp 820-822.

[15]loadbalancing architecture-http://awsmedia.s3.amazonaws.com/2012-02-24

[16] Pooja Samal, Pranati Mishra, "Analysis of variants of Round Robin Algorithms for Load balancing in cloud computing", International Journal of computer science and Information Technologies", Vol 4 (3), 2013, pp 416-419

[17] Dinesh, Rajesh "Load balancing in grid", injure, Vol 2, issue 2,2012, pp 445-450

[18] Argha Roy, Diptam Dutta, "Dynamic Load balancing: Improve efficiency in cloud computing", International Journal of Emerging Research in Management and Technology, Vol 2, issue 4,2013, pp 78-82

[19] Jaspreet Karur, "Comparison of load balancing algorithms in a cloud", International Journal of Engineering Research and Applications", Vol 2, issue 3,2012, pp 1169-1173

[20] L. D. Dhinesh Babu, P. Venkat Krishna,"Honey bee behavior inspired load balancing of tasks in cloud computing environments", Elsevier, 2013.

T-Y., W-T. Lee, Y-S. Lin, Y-S. Lin, H-L. Chan and J-S. Huang, "Dynamic load balancing mechanism based on cloud storage" in proc. Computing, Communications and Applications Conference (ComComAp), IEEE, January 2012, pp 102-106.

[22] Al-Jaroodi, J. And N. Mohamed. "DDFTP: Dual-Direction FTP," in proc. 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), IEEE, May 2011, pp 504-503.

[23] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi, Jameela Al-Jaroodi, "A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms", IEEE Second Symposium on Network Cloud Computing and Applications, 2012, pp 137-142

[24] Wang, S-C., K-Q. Yan, W-P. Liao and S-S. Wang, "Towards a load balancing in a three-level cloud computing network," in proc. 3rd International Conference on. Computer Science and Information Technology (ICCSIT), IEEE, Vol. 1, July 2010, pp 108-113.

9. AUTHOR'S PROFILE

R.Ramya is doing MPhil research in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. Her area of research is Cloud Computing. She is presently working on Scalability issues in Cloud Computing. Her area of interest is Computer Networks.

M .Kriushanth is a full time Ph.D research scholar in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has teaching experience of 2 years. He has attended many International and National Conferences, Seminar and Workshops. His area of research is Cloud Computing. He is presently working on Scalability issues in Cloud Computing. His areas of interest Computer Networks and Web Technologies.

Dr. L. Arockiam is working as Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 25 years of experience in teaching and 17 years of experience in research. He has published more than 187 research articles in the International / National Conferences and Journals. He has also presented 2 research articles in the Software Measurement European Forum in Rome. He has chaired many technical sessions and delivered invited talks in National and International Conferences. He has authored a book on "Success through Soft Skills". His research interests are: Big Data, Cloud Computing, Software Measurement, Cognitive Aspects of Programming, Data Mining and Mobile Networks. He has been awarded "Best Research Publications in Science" for 2010, 2011, & 2012, "Best Teacher Award" for 2012-13, 2013 -14 and ASDF Global Awards for "Best Academic Researcher" from ASDF, Pondicherry for the academic year 2012-13.