International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

Analysis of Load Balancing Techniques in Cloud Computing

Amandeep Kaur Sidhu[1], Supriya Kinger[2]
[1]Research Fellow, Sri Guru Granth Sahib World University,Fatehgarh Sahib,Punjab.
asaman19@gmail.com
[2]Asst. Professor,Sri Guru Granth Sahib World University,Fatehgarh Sahib,Punjab.
ahujasupriya@gmail.com

ABSTRACT

Cloud Computing is an emerging computing paradigm. It aims to share data, calculations, and service transparently over a scalable network of nodes. Since Cloud computing stores the data and disseminated resources in the open environment. So, the amount of data storage increases quickly. In the cloud storage, load balancing is a key issue. It would consume a lot of cost to maintain load information, since the system is too huge to timely disperse load. Load balancing is one of the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed. It helps in optimal utilization of resources and hence in enhancing the performance of the system. A few existing scheduling algorithms can maintain load balancing and provide better strategies through efficient job scheduling and resource allocation techniques as well. In order to gain maximum profits with optimized load balancing algorithms, it is necessary to utilize resources efficiently. This paper discusses some of the existing load balancing algorithms in cloud computing and also their challenges.

General Terms

Load Balancing in Cloud Computing

Indexing terms

Load Balancing Algorithms, Load Balancing Challenges

Academic Discipline and Sub-Disciplines

Computer Science & Engineering

SUBJECT CLASSIFICATION

Cloud Computing

COVERAGE

This Study covers all the load balancing algorithms and major challenges in the Cloud Computing. Since Cloud computing stores the data and disseminated resources in the open environment. So, the amount of data storage increases quickly. Load balancing is one of the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed.

TYPE (METHOD/APPROACH)

Review or Literature Survey

INTRODUCTION

A Cloud computing is emerging as a new paradigm of large scale distributed computing. It has moved computing and data away from desktop and portable PCs, into large data centres [1]. It provides the scalable IT resources such as applications and services, as well as the infrastructure on which they operate, over the Internet, on pay-per-use basis to adjust the capacity quickly and easily. It helps to accommodate changes in demand and helps any organization in avoiding the capital costs of software and hardware [2] [3]. Thus, Cloud Computing is a framework for enabling a suitable, on-demand network access to a shared pool of computing resources (e.g. networks, servers, storage, applications, and services). These resources can be provisioned and de-provisioned quickly with minimal management effort or service provider interaction. This further helps in promoting availability [4]. Due to the exponential growth of cloud computing, it has been widely adopted by the industry and there is a rapid expansion in data-centres.

ccording to the National Institute of Standards and Technology (NIST) [13], cloud computing exhibits several characteristics:

On-demand Self-service- A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Broad Network Access- Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

Resource Pooling- The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid Elasticity- Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Measured Service- Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

As cloud computing is in its evolving stage, so there are many problems prevalent in cloud computing [11],[16]. Such as:

- Ensuring proper access control (authentication, authorization, and auditing)
- Network level migration, so that it requires minimum cost and time to move a job
- To provide proper security to the data in transit and to the data at rest.
- Data availability issues in cloud
- Legal quagmire and transitive trust issues
- Data lineage, data provenance and inadvertent disclosure of sensitive information is possible

And the most prevalent problem in Cloud computing is the problem of load balancing.

LOAD BALANCING

Load Balancing is a computer networking method to distribute workload across multiple computers or a computer cluster, network links, central processing units, disk drives, or other resources, to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload. Using multiple components with load balancing, instead of a single component, may increase reliability through redundancy. The load balancing service is usually provided by dedicated software or hardware, such as a multilayer switch or a Domain Name System server.

Load balancing is one of the central issues in cloud computing [5]. It is a mechanism that distributes the dynamic local workload evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It helps to achieve a high user satisfaction and resource utilization ratio, hence improving the overall performance and resource utility of the system. It also ensures that every computing resource is distributed efficiently and fairly [6]. It further prevents bottlenecks of the system which may occur due to load imbalance. When one or more components of any service fail, load balancing helps in continuation of the service by implementing fair-over, i.e. in provisioning and de-provisioning of instances of applications without fail.

The goal of load balancing is improving the performance by balancing the load among these various resources (network links, central processing units, disk drives.) to achieve optimal resource utilization, maximum throughput, maximum response time, and avoiding overload. To distribute load on different systems, different load balancing algorithms are used.

In general, load balancing algorithms follow two major classifications:

- Depending on how the charge is distributed and how processes are allocated to nodes (the system load);
- Depending on the information status of the nodes (System Topology).

In the first case it designed as designed as centralized approach, distributed approach or hybrid approach in the second case as static approach, dynamic or adaptive approach.

a) Classification According to the System Load

- Centralized approach: In this approach, a single node is responsible for managing the distribution within the whole system.
- Distributed approach: In this approach, each node independently builds its own load vector by collecting the load information of other nodes. Decisions are made locally using local load vectors. This approach is more suitable for widely distributed systems such as cloud computing.
- Mixed approach: A combination between the two approaches to take advantage of each approach.

b) Classification According to the System Topology

- Static approach: This approach is generally defined in the design or implementation of the system.
- Dynamic approach: This approach takes into account the current state of the system during load balancing decisions. This approach is more suitable for widely distributed systems such as cloud computing.
- Adaptive approach: This approach adapts the load distribution to system status changes, by changing their parameters dynamically and even their algorithms. This approach is able to offer better performance when the

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

system state changes frequently. This approach is more suitable for widely distributed systems such as cloud computing.

METRICS FOR LOAD BALANCING IN CLOUD-

Various metrics considered in existing load balancing techniques in cloud computing are discussed below-

- **Scalability** is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.
- **Resource Utilization** is used to check the utilization of re-sources. It should be optimized for an efficient load balancing.

- **Performance** is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays.

- **Response Time** is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.

- **Overhead Associated** determines the amount of overhead involved while implementing a load-balancing algorithm. It is composed of overhead due to movement of tasks, inter-processor and interprocess communication. This should be minimized so that a load balancing technique can work efficiently.

The objective and motivation of this survey is to give a systematic review of existing load balancing algorithms in cloud computing and encourage the amateur researcher in this field, so that they can contribute in developing more efficient load balancing algorithm. This will benefit interested researchers to carry out further work in this research area.

LOAD BALANCING ALGORITHMS

In order to balance the requests of the resources it is important to recognize a few major goals of load balancing algorithms:

a) Cost effectiveness: primary aim is to achieve an overall improvement in system performance at a reasonable cost.

b) Scalability and flexibility: the distributed system in which the algorithm is implemented may change in size or topology. So the algorithm must be scalable and flexible enough to allow such changes to be handled easily.

c) Priority: prioritization of the resources or jobs need to be done on before hand through the algorithm itself for better service to the important or high prioritized jobs in spite of equal service provision for all the jobs regardless of their origin.

Following load balancing algorithms are currently prevalent in clouds:-

Round Robin: In this algorithm [7], the processes are divided between all processors. Each process is assigned to the processor in a round robin order. The process allocation order is maintained locally independent of the allocations from remote processors. Though the work load distributions between processors are equal but the job processing time for different processes are not same. So at any point of time some nodes may be heavily loaded and others remain idle. This algorithm is mostly used in web servers where http requests are of similar nature and distributed equally.

Connection Mechanism: Load balancing algorithm [8] can also be based on least connection mechanism which is a part of dynamic scheduling algorithm. It needs to count the number of connections for each server dynamically to estimate the load. The load balancer records the connection number of each server. The number of connection increases when a new connection is dispatched to it, and decreases the number when connection finishes or timeout happens.

Randomized: Randomized algorithm is of type static in nature. In this algorithm [7] a process can be handled by a particular node n with a probability p. The process allocation order is maintained for each processor independent of allocation from remote processor. This algorithm works well in case of processes are of equal loaded. However, problem arises when loads are of different computational complexities. Randomized algorithm does not maintain deterministic approach. It works well when Round Robin algorithm generates overhead for process queue.

Equally Spread Current Execution Algorithm: Equally spread current execution algorithm [9] process handle with priorities. it distribute the load randomly by checking the size and transfer the load to that virtual machine which is lightly loaded or handle that task easy and take less time , and give maximize throughput. It is spread spectrum technique in which the load balancer spread the load of the job in hand into multiple virtual machines.

Throttled Load Balancing Algorithm: Throttled algorithm [9] is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine which access that load easily and perform the operations which is give by the client or user. In this algorithm the client first requests the load balancer to find a suitable Virtual Machine to perform the required operation.

***A Task Scheduling Algorithm Based on Load Balancing*:** Y. Fang et al. [10] discussed a two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain high resource utilization. It achieves load balancing by first mapping tasks to virtual machines and then virtual machines to host resources thereby improving the task response time, resource utilization and overall performance of the cloud computing environment.

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

Biased Random Sampling: M. Randles et al. [11] investigated a distributed and scalable load balancing approach that uses random sampling of the system domain to achieve self-organization thus balancing the load across all nodes of the system. Here a virtual graph is constructed, with the connectivity of each node (a server is treated as a node) representing the load on the server. Each server is symbolized as a node in the graph, with each in degree directed to the free resources of the server. The load balancing scheme used here is fully decentralized, thus making it apt for large network systems like that in a cloud. The performance is degraded with an increase in population diversity.

Min-Min Algorithm: It begins with a set of all unassigned tasks. First of all, minimum completion time for all tasks is found. Then among these minimum times the minimum value is selected which is the minimum time among all the tasks on any resources. Then according to that minimum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines. Then again the same procedure is followed until all the tasks are assigned on the resources. But this approach has a major drawback that it can lead to starvation [12].

Max-Min Algorithm: Max-Min is almost same as the min-min algorithm except the following: after finding out minimum execution times, the maximum value is selected which is the maximum time among all the tasks on any resources. Then according to that maximum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines[12].

Token Routing: The main objective of the algorithm [14] is to minimize the system cost by moving the tokens around the system. But in a scalable cloud system agents cannot have the enough information of distributing the work load due to communication bottleneck. So the workload distribution among the agents is not fixed. The drawback of the token routing algorithm can be removed with the help of heuristic approach of token based load balancing. This algorithm provides the fast and efficient routing decision. In this algorithm agent does not need to have an idea of the complete knowledge of their global state and neighbour's working load. To make their decision where to pass the token they actually build their own knowledge base. This knowledge base is actually derived from the previously received tokens. So in this approach no communication overhead is generated.

LOAD BALANCING CHALLENGES IN THE CLOUD COMPUTING

Although cloud computing has been widely adopted. Research in cloud computing is still in its early stages, and some scientific challenges remain unsolved by the scientific community, particularly load balancing challenges [15].

- Automated service provisioning: A key feature of cloud computing is elasticity, resources can be allocated or released automatically. How then can we use or release the resources of the cloud, by keeping the same performance as traditional systems and using optimal resources?
- Virtual Machines Migration: With virtualization, an entire machine can be seen as a file or set of files, to unload a physical machine heavily loaded, it is possible to move a virtual machine between physical machines. The main objective is to distribute the load in a datacenter or set of datacenters. How then can we dynamically distribute the load when moving the virtual machine to avoid bottlenecks in Cloud computing systems?
- Energy Management: The benefits that advocate the adoption of the cloud is the economy of scale. Energy saving is a key point that allows a global economy where a set of global resources will be supported by reduced providers rather that each one has its own resources. How then can we use a part of datacenter while keeping acceptable performance?
- Stored data management: In the last decade data stored across the network has an exponential increase even for companies by outsourcing their data storage or for individuals, the management of data storage or for individuals, the management of data storage becomes a major challenge for cloud computing. How can we distribute the data to the cloud for optimum storage of data while maintaining fast access?
- Emergence of small data centers for cloud computing: Small datacenters can be more beneficial, cheaper and less energy consumer than large datacenter. Small providers can deliver cloud computing services leading to geo-diversity computing. Load balancing will become a problem on a global scale to ensure an adequate response time with an optimal distribution of resources.

CONCLUSION

Cloud Computing has widely been adopted by the industry, though there are many existing issues like Load Balancing, Virtual Machine Migration, Server Consolidation, Energy Management, etc. which have not been fully addressed. Central to these issues is the issue of load balancing, that is required to distribute the excess dynamic local workload evenly to all the nodes in the whole Cloud to achieve a high user satisfaction and resource utilization ratio. It also ensures that every computing resource is distributed efficiently and fairly. This paper presents a concept of Cloud Computing along with research challenges in load balancing. Major thrust is given on the study of load balancing algorithm, followed by a comparative survey of these above mentioned algorithms in cloud computing with respect to scalability, resource utilization, performance, response time and overhead associated.

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing", EECS Department, University of California, Berkeley, Technical Report No., UCB/EECS-2009-28, pages 1-23, February 2009.

[2] R. W. Lucky, "Cloud computing", IEEE Journal of Spectrum, Vol. 46, No. 5, May 2009, pages 27-45.

[3] M. D. Dikaiakos, G. Pallis, D. Katsa, P. Mehra, and A. Vakali, "Cloud Computing: Distributed Internet Computing for IT and Scientific Research", IEEE Journal of Internet Computing, Vol. 13, No. 5, September/October 2009, pages 10-13.

[4] G. Pallis, "Cloud Computing: The New Frontier of Internet Computing", IEEE Journal of Internet Computing, Vol. 14, No. 5, September/October 2010, pages 70-73.

[5] B. P. Rima, E. Choi, and I. Lumb, "A Taxonomy and Survey of Cloud Computing Systems", Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC, Seoul, Korea, August 2009, pages 44-51.

[6] A. M. Alakeel, "A Guide to dynamic Load balancing in Distributed Computer Systems", International Journal of Computer Science and Network Security (IJCSNS), Vol. 10, No. 6, June 2010, pages 153-160.

[7] Zhong Xu, Rong Huang,(2009)"Performance Study of Load Balanacing Algorithms in Distributed Web Server Systems", CS213 Parallel and Distributed Processing Project Report.

[8] P.Warstein, H.Situ and Z.Huang(2010), "Load balancing in a cluster computer" In proceeding of the seventh International Conference on Parallel and Distributed Computing, Applications and Technologies, IEEE.

[9] Ms.NITIKA, Ms.SHAVETA, Mr. GAURAV RAJ; "Comparative Analysis of Load Balancing Algorithms in Cloud Computing", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May 2012.

[10] Y. Fang, F. Wang, and J. Ge, "A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing", Web Information Systems and Mining, Lecture Notes in Computer Science, Vol. 6318, 2010, pages 271-277.

[11] T.R.V. Anandharajan, Dr. M.A. Bhagyaveni" Co-operative Scheduled Energy Aware Load-Balancing technique for an Efficient Computational Cloud" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[12] T. Kokilavani J.J. College of Engineering & Technology and Research Scholar, Bharathiar University, Tamil Nadu, India" Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing" International Journal of Computer Applications (0975 – 8887) Volume 20– No.2, April 2011.

[13] Peter Mell, Timothy Grance, "The NIST Definition of Cloud Computing", NIST Special Publication 800-145, September 2011.

[14] Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh, Christopher Mcdermid (2011)"Availabity and Load Balancing in Cloud Computing" International Conference on Computer and Software Modeling IPCSIT vol.14 IACSIT Press,Singapore 2011.

[15] A. Khiyaita, M. Zbakh, H. El Bakkali and Dafir El Kettani, "Load Balancing Cloud Computing: State of Art" , 9778-1-4673-1053-6/12/$31.00, 2012 IEEE.

[16] Wayne Jansen Timothy Grance" Guidelines on Security and Privacy in Public Cloud Computing" NIST Draft Special Publication 800-144.

Author's Biography

Amandeep kaur sidhu is Student of M.Tech in the Department of Computer Science at SGGSWU, Fatehgarh Sahib, Punjab, India.She have done PGDCA under Punjab University Chandigarh and MSc-IT under the Punjabi University Patiala.

Supriya is currently working with SGGS World University, Fatehgarh, as Assistant Professor. She received her Master's in Technology degree in field of Computer Science from YMCA, India and Bachelor's in Technology degree in Computer Science from Kurukshetra University, India. Her areas of interest include Parallel and Distributed Processing, Cloud Computing and Software Engineering. Currently she is doing research in field of 'Energy Efficient Clouds'. She has more than 8 years of experience in UG & PG teaching and research. She has attended and organized a number of workshops and conferences and has presented papers in field of Green Clouds, Component Based Software Engineering, Information Security and Networks. She is a life member of ISTE.