ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 77-82 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Importance of Load Balancing in Cloud Computing Environment: A Review

Yadaiah Balagoni[1], **Dr.R.Rajeswara Rao**[2]

[1]*Assistant Professor, CSE Dept, MGIT Gandipet, Hyderabad.*
yad524.balagoni@gmail.com

[2]*Associate Professor, CSE Dept, JNTU Vizianagaram.*
rajaraob4u@gmail.com

Abstract-**"Cloud computing" is a term, which involves virtualization, distributed computing, networking, software and web services. A cloud consists of several elements such as clients, datacenter and distributed servers. It includes fault tolerance, high availability, scalability, flexibility, reduced overhead for users, reduced cost of ownership, on demand services etc. Central to these issues lies the establishment of an effective load balancing algorithm. Load balancing is the process of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. The load can be CPU load, memory capacity, delay or network load. Load balancing ensures that all the virtual machines (VMs) in the system or every node in the network does approximately the equal amount of work at any instant of time. This technique can be sender initiated, receiver initiated or symmetric type (combination of sender initiated and receiver initiated types).This review provides a systematic study on load balancing, which in turn initiates the researchers for implementation of better load balancing algorithms.**

Keywords- **Cloud Computing, Virtualization, Load Balancing.**

INTRODUCTION

The term *cloud* has been used historically as a metaphor for the Internet. This usage was originally derived from its common depiction in network diagrams as an outline of a cloud, used to represent the transport of data across carrier backbones (which owned the cloud) to an endpoint location on the other side of the cloud. "Cloud computing" is the next natural step in the evolution of on-demand information technology services and products[1]. To an extent, cloud computing is be based on virtualized resources. Cloud computing provides IT capabilities as services-on-demand. This scalable and elastic model provides advantages like faster time-to-market, no capital expenditure and pay-per-use business model. Today Small and Medium scale Business companies are realizing that by simply tapping into the cloud they can gain fast access to best business applications or drastically boost their infrastructure resources, all at minimal cost.

Cloud computing is a style of computing ,which enables a convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Due to the exponential growth of cloud computing, it has been widely adopted by the industry[2]. Formally, a cloud computing is defined as an internet based computing , whereby shared resources ,software and information are provided to computers and other devices on-demand like electricity.

The cloud offers several benefits like fast deployment, pay-for- use, lower costs, scalability, rapid provisioning, rapid elasticity, ubiquitous network access, greater resiliency, hypervisor protection against network attacks, low-cost disaster recovery and data storage solutions, on-demand security controls, real time detection of system tampering and rapid re-constitution of services. Besides this, there are many other existing issues like Load Balancing, Virtual Machine Migration, and Server Consolidation etc.that have not been fully addressed. Virtual Machine Migration enabled by virtualization can help in balancing load, enabling highly responsive provisioning and avoiding hot-spots in datacenters. Server Consolidation helps in improving resource utilization by consolidating various VMs residing on multiple under-utilized servers onto a single server, thereby turning off unused servers, hence reducing energy consumption in a cloud computing environment [8]. Load balancing can help in reducing energy consumption by evenly distributing the load and minimizing the resource consumption. This paper focuses on the importance and prevalent load balancing techniques in cloud computing environment.

CLOUD COMPUTING ARCHITECTURE

As Gartner has posited, the cloud is not architecture, a platform, a tool, or an infrastructure. "It is a style of computing." It is a deployment model, much in the same way SOA is a style of computing. A cloud consists of three major components. Those are clients, datacenters and distributed servers.

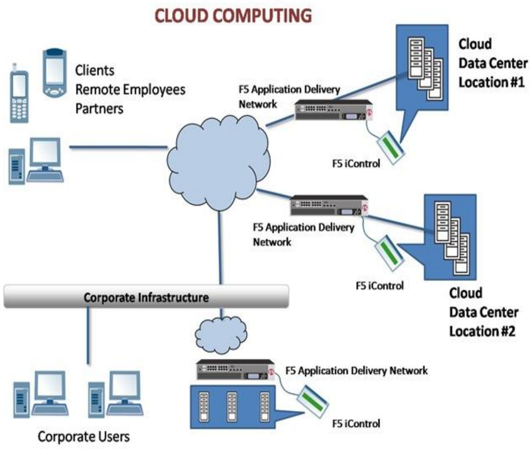
CLOUD COMPUTING
Clients
Remote Employees
Partners
F5 Application Delivery Network
Cloud Data Center Location #1
F5 iControl
F5 Application Delivery Network
F5 iControl
Cloud Data Center Location #2
Corporate Infrastructure
F5 Application Delivery Network
F5 iControl
Corporate Users

Fig1:Cloud Computing Model

Clients

End users interact with the clients to manage information related to the cloud. Clients generally fall into three categories as given below. *Mobile Clients:* Windows Mobile Smartphone, smart phones, like a Blackberry, or an iPhone. *Thin Clients:* They don't do any computation work. They only display the information. Servers do all the works for them. Thin clients don't have any internal memory. *Thick Clients*: These use different browsers like IE or Mozilla Firefox or Google Chrome to connect to the Internet cloud. Now-a-days thin clients are more popular as compared to other clients because of their low price, security, low consumption of power, less noise, easily replaceable and repairable etc.

Datacenter

Datacenter is nothing but a collection of servers hosting different applications. An end user connects to the datacenter to subscribe different applications. A datacenter may exist at a large distance from the clients.

Distributed Servers

Distributed servers are the parts of a cloud which are present throughout the Internet hosting different applications. But while using the application from the cloud, the user will feel that he is using this application from its own machine.

CLOUD SERVICE MODELS

Cloud computing providers offer their services According to several fundamental models: software as a service (*SaaS*), infrastructure as a service (*IaaS*), and platform as a service (*PaaS*). *SaaS* is a model of software deployment whereby a provider licenses an application to customers for use as a service on demand. One example of SaaS is the Salesforce.com CRM application. *IaaS* is the delivery of computer infrastructure (typically a platform virtualization environment) as a service. Rather than purchasing servers, software, datacenter space or network equipment, clients instead buy those resources as a fully outsourced service. One such example of this is the Amazon web services. *PaaS* is one layer above IaaS on the stack and it offers an integrated set of developer environment that a developer can tap to build their applications without having any clue about what is going on underneath the service[5]. It offers developers a service that provides a complete software development lifecycle management, from planning to design to building applications to deployment to testing to maintenance. An example of this would be GoogleApps. The below diagram shows the different layers of cloud Computing architecture

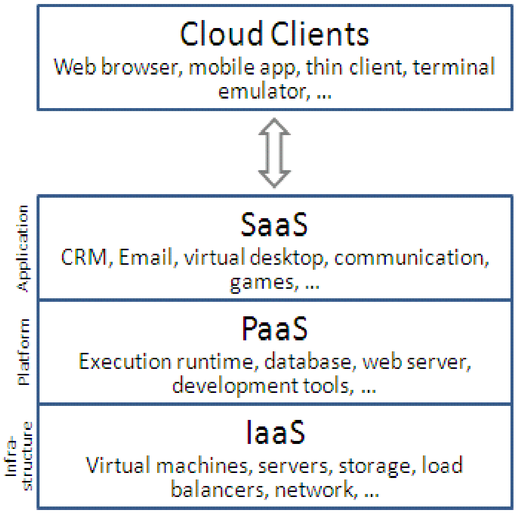
Cloud Clients
Web browser, mobile app, thin client, terminal emulator, ...
SaaS
CRM, Email, virtual desktop, communication, games, ...
PaaS
Execution runtime, database, web server, development tools, ...
IaaS
Virtual machines, servers, storage, load balancers, network, ...
Application
Platform
Infra-structure

Fig 2.Layers and services of Cloud Computing

VIRTUALIZATION

Virtualization means "something which isn't real", but gives all the facilities of a real. Wikipedia says "Virtualization", in computing, is the creation of a virtual (rather than actual) version of something, such as a hardware platform, operating system, a storage device or network resources. It provides ability to run multiple operating systems on a single physical system and share the underlying hardware resources [9]. It is the software implementation of a computer which will execute different programs like a real machine. Virtualisation is related to cloud, because using virtualization an end user can use different services of a cloud.

TYPES OF VIRTUALIZATION

When we say different types of virtualization, we are not talking about vendor specific technologies (such as Hyper-V from Microsoft, VMware or Citrix etc). Each type of virtualization has its advantages. We can classify Virtualization in three main categories as follows:

Hardware Virtualization

Hardware virtualization or platform virtualization refers to the creation of a virtual machine that acts like a real computer with an operating system. Software executed on these virtual machines is separated from the underlying hardware resources. In hardware virtualization, the host machine is the actual machine on which the virtualization takes place, and the guest machine is the virtual machine .The words host and guest are used to distinguish the software that runs on the physical machine from the software that runs on the virtual machine. The software or firmware that creates a virtual machine on the host hardware is called a hypervisor or

Virtual Machine Manager. Different types of hardware virtualization include: *Full virtualization:* In case of full virtualisation a complete installation of one machine is done on another machine. It will result in a virtual machine which will have all the software's that are present in the actual server.

Partial virtualization: Some but not the entire target environment is simulated. Some guest programs, therefore, may need modifications to run in this virtual environment. *Para virtualization:* A hardware environment is not simulated; however, the guest programs are executed in their own isolated domains, as if they are running on a separate system. Guest programs need to be specifically modified to run in this environment.

Desktop Virtualization

Desktop virtualization is the new trend in corporations these days as organizations move away from the pain of maintaining expensive hardware which is often underutilized. In Desktop Virtualization, typically a Workstation is virtualized with all its applications, user customizations and preferences all in a virtual machine. This virtual machine can be accessed from anywhere in the organization using any workstation hence cutting down on licensing costs for installing software on individual machines. Virtual Desktops also reduce the carbon foot print and increase Total Cost of Ownership when compared to maintaining physical machines.

Storage Virtualization

Virtualizing storage space means you also are virtualizing your Virtual Machine disk files as well. This becomes extremely useful for your Business Continuity Planning and Disaster Recovery, because with a virtualized infrastructure, you can very easily replicate your storage across over to another location or even a different geographical regions. Better yet, Corporations are now looking into implementing fault tolerant clusters for their storage. Newer technologies such as Snap Mirror from NetApp make this possible, the initial replication is bandwidth intensive but the subsequent transfers are just differentials maintained on both sides of the wired cluster.

LOAD BALANCING

Load balancing is one of the major issues in cloud computing environment. It is a mechanism that distributes the workload evenly across all the nodes in the entire cloud to avoid a situation where some nodes are loaded heavily while others are idle or doing little work. Load balancing helps to achieve a high user satisfaction and resource utilization ratio, hence improving the overall performance and resource utilization of the cloud computing system. It also ensures that every computing resource is distributed efficiently and fairly[10]. 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. It also ensures that every computing resource is distributed efficiently and fairly.

NEED OF LOAD BALANCING IN CLOUD COMPUTING

Load balancing in clouds is a process of re-assigning the total load to the individual nodes of the collective system to make effective resource utilization and to improve the response time of the job, simultaneously removing a condition in which some of the nodes are over loaded while some others are under loaded. It is used to achieve a high user satisfaction and resource utilization ratio, hence improving the overall performance of the system. Proper load balancing can help in utilizing the available resources optimally, thereby minimizing the resource consumption. It also helps in implementing fail-over, enabling scalability, avoiding bottlenecks and over-provisioning, reducing response time etc.The primary goals of load balancing are mentioned below:

- To improve the performance of the cloud substantially.
- To have a backup plan in case the system fails even partially.
- To maintain the system stability.
- To accommodate future modification/changes in the system.
- To provide optimal resource utilization.
- To maximum throughput, minimize response time, and avoiding overload.
- To treat all jobs in the system equally regardless of their origin.

TYPES OF LOAD BALANCING ALGORITHMS

Cloud is made up of massive resources. Management of these resources requires efficient planning and proper layout. While designing an algorithm for resource provisioning on cloud the developer must take into consideration different cloud scenarios and must be aware of the issues that are to be resolved by the proposed algorithm. Therefore, resource provisioning algorithm can be categorized into different classes based upon the environment, purpose and technique of proposed solution.

Load balancing based on initiation

Depending on who initiated the process, load balancing algorithms can be of three Categories as given below:
Sender Initiated: If the load balancing algorithm is initialised by the sender
Receiver Initiated: If the load balancing algorithm is initiated by the receiver
Symmetric: It is the combination of both sender initiated and receiver initiated

Load Balancing based on Cloud Environment

Depending on the current state of the system, load balancing algorithms can be divided into two categories as given below:

Static load balancing:

- In this scenario, the cloud requires prior knowledge of nodes capacity, processing power, memory, performance and statistics of user requirements.
- These user requirements are not subjected to any change at run-time.
- Algorithms proposed to achieve load balancing in static environment cannot adapt to the run time changes in load.
- It is easy to simulate but is not well suited for heterogeneous cloud environment.
- It doesn't depend on the current state of the system. It requires the Prior knowledge of the system.

Dynamic load balancing

- In dynamic environment the cloud provider installs heterogeneous resources.
- Decisions on load balancing are based on current state of the system. No prior knowledge is needed.
- In this scenario cloud cannot rely on the prior knowledge of the system.
- The requirements of the users may change at run-time.
- Dynamic environment is difficult to be simulated but is highly adaptable with cloud computing environment.

Load balancing based on Spatial Distribution of Nodes

Nodes in the cloud are highly distributed. Hence the node that makes the decision also governs the Category of algorithm to be used. Depending upon which node is responsible for balancing of load in cloud computing environment. Load balancing algorithms are classified into the following category.

Centralized Load Balancing

In centralized load balancing technique all the allocation and scheduling decision are made by a single node. This node is responsible for storing knowledge base of entire cloud network and can apply static or dynamic approach for load balancing.

Distributed Load Balancing

In distributed load balancing technique, no single node is responsible for making resource provisioning or task Scheduling decision. Multiple domains monitor the network to make accurate load balancing decision. Every node in the network maintains local knowledge base to ensure efficient distribution of tasks in static environment and re-distribution in dynamic environment.

Hierarchical Load Balancing

In Hierarchical load balancing technique, different levels of the cloud involved in load balancing decision. Such load balancing techniques mostly operate in master slave mode. These can be modelled using tree data structure wherein every node in the tree is balanced under the supervision of its parent node. Master / manager can use light weight agent process to get statistics of slave/child nodes. Based upon the information gathered by the parent node scheduling decision is made.

METRICS USED FOR EVALUATING LOAD BALANCING ALGORITHMS IN CLOUDS

The following are the some of the metrics which are considered while evaluating or comparing the existing load balancing techniques in cloud computing environment[3]. Some of them are listed below.

- *Throughput:* is used to calculate the no. of tasks completed per a unit time. It should be high to improve the performance of the system.
- *Overhead Associated:* determines the amount of complexity/ overhead involved while implementing a load-balancing algorithm. It is consists of overhead due to movement of tasks, inter-processor and inter-process communication etc. This should be minimized so that a load balancing technique can work efficiently.
- *Fault Tolerance:* is the ability of an algorithm to perform uni-form load balancing in spite of arbitrary node/ link failure. The load balancing should be a good fault-tolerant technique.
- *Migration time:* is the time to migrate the jobs or resources from one node to other. It should be minimized in order to enhance the performance of the system.
- *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.
- *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
- *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.

CONCLUSION

In this paper we have surveyed on load balancing. In cloud computing load balancing is the main issue. Because when client is requesting for service it should be available to the client. When node is overloaded with job at that time load balancer has to set that load on another free node. So, to balance the load is necessary in cloud computing. So in our paper we have discussed about the importance and the different classifications of load balancing algorithems in cloud environment.And we have also discussed the virtualization and cloud computing basics.

REFERENCES

[1]. Marios D. Dikaiakos, George Pall is, Dimitrios Katsaros, Pankaj Mehra, Athena Vakali, 2009 "Cloud computing : Distributed Internet Computing for IT and Scientific Research" IEEE Internet Computing, Published by the IEEE Computer Society.

[2] Peter Mell, Timothy Grance, The NIST Definition of "Cloud Computing" National Institute of Standards and Technology - Computer Security Resource Center-www.csrc.nist.gov.

[3] Luo, S., Lin, Z. & Chenm, X. (2011). *Virtualization security for cloud computing service.* 2011International Conference on Cloud and Service Computing 978-1-4577-1637-9/11/$26.00 ©2011IEEE. Shenzhen, China: ZTE Corporation.

[4] Zhang, Q., Cheng, L. & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1),7-18. DOI 10.1007/s13174-010-0007-6.

[5]. Radojevic, B. & Zagar, M. (2011). *Analysis of issueswith load balancing algorithms in hosted (cloud) environments*. In proceedings of 34th InternationalConvention on MIPRO, IEEE.

[6]. Basic concept and terminology of cloud computing- http://whatiscloud.com

[7]. L. Wang, J. Tao, M. Kunze,"Scientific Cloud Computing: Early Definition and Experience", the 10th IEEE International Conference Computing and Communications 2008.

[8].Load Balancing in Cloud computing, http://community.citrix.com/display/cdn/Load+Balancing

[9]. Sotomayor, B., RS. Montero, IM. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," in IEEE Internet Computing, Vol. 13, No. 5, pp: 14-22, 2009.

[10]. 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, pp: 108-113, July 2010.