ISSN 2277-3061
Volume 15 Number 8
International journal of Computers and Technology

LOAD BALANCING IN CLOUD ENVIRONMENT: A REVIEW

Sheenam Kamboj [1], Mr. Navtej Singh Ghumman [2]
[1] Research Scholar, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
sheenam31.sk@gmail.com
[2] Assistant Professor, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
navtejghumman@yahoo.com

ABSTRACT

An essential role of cloud computing platform is to dynamically balance the load among the different servers in order to improve resource utilization and to avoid hotspots. Load balancing (LB) is done on both sides i.e. on provider as well as on consumer side. On provider side, load balancing is the problem of allocating virtual machines to servers at runtime. Virtual Machine need to be reassigned so that servers do not get overloaded as demand changes. On consumer side application load can be balanced which provides efficiency to the consumers. On cloud computing platform, load balancing of the entire system can be dynamically handled by using virtualization technology through which it becomes possible to remap virtual machine and physical resources according to the change in load. However, in order to improve performance, the virtual machines have to fully utilize its resources and services by adapting to computing environment dynamically. The load balancing with proper allocation of resources must be guaranteed in order to improve resource utility.

Keywords

Cloud Computing, Load Balacing, Virtual Machine, Data Center, Data Center Broker.

INTRODUCTION

Cloud Computing is an emerging technology that can support a broad-spectrum of applications. It is a new computing paradigm, where a large pool of systems are connected in public or private networks, to provide dynamically scalable infrastructure for data, application and file storage. The US Department of Commerce's National Institute of Standards and Technology defines cloud computing as “a model for enabling convenient, on-demand and ubiquitous network access to a shared pool of configurable computing resources(servers, networks, storage, services and applications) that can be rapidly provisioned and released with minimal service provider interaction and minimal management effort from the user side”. Load Balancing[2] is an emerging computer paradigm where data and services placed massively in the cloud and which can be accessed from any connected devices over the internet. It is known as provider of dynamic services using very large scalable and virtualized resources over the internet. Load Balancing is a computer networking method to distribute workload across multiple computer clusters, network links or other resources to achieve optimal resource utilization, maximize throughput, minimize response time and avoid overload. It is a mechanism that distributes the dynamic local work load[3] evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle. Its goal is to improve the overall performance and resource utility of the system.

A. CLOUD COMPUTING: AN OVERVIEW

The term Cloud refers as Network or Internet. In other words, Cloud is something, which is present at remote location. Cloud can provide services over network, i.e., on public networks or on private networks, i.e., WAN,VPN or LAN. Applications such as web conferencing, e-mail and customer relationship management (CRM), all run in cloud. Cloud Computing refers to configuring, accessing and manipulating the applications online. It offers online infrastructure, data storage, and application. It overcomes the platform dependency issues as there is no need install software’s on our PCs. Hence, Cloud computing can be classified as a new paradigm for the dynamic provisioning of computing services supported by state-of-the-art data centers that usually employ Virtual Machine (VM) technologies for consolidation and environment isolation purposes. Many computing service providers including Microsoft, Google, Yahoo, and IBM are rapidly deploying data centers in various locations around the world to deliver Cloud computing services. The cloud computing platform guarantees subscribes that it sticks to the service level agreement by providing resources as service and by needs. The number of online services—such as search, online gaming, social networks and video streaming— has exploded. Therefore, day by day subscribers’ needs are increasing for computing resources and their needs have dynamic heterogeneity and platform irrelevance. Due to data locality issues and the demand for fast response times, resources are shared and if they are not properly distributed then it will result into resource wastage. Computing can be described as any activity of using and/or developing computer hardware and software. It includes everything that sits in the bottom layer, i.e. everything from raw compute power to storage capabilities. Cloud computing [1] ties together all these entities and delivers them as a single integrated entity under its own sophisticated management.

Essential characteristics of Cloud Computing:

• On-demand Self Service: Cloud Computing have capabilities such as network storage, virtual machine can automatically use by a consumer without human interaction with cloud service providers.

• Broad Network Access: Computing capabilities can be accessed over the broadband network using heterogeneous devices like phones, laptops, PDAs.

• Resource pooling: Service providers pool their resources that are shared by multiple user. Resources are assigned and reassigned according to the demand of consumer. The physical may provide virtual machines, storage, processing, network bandwidth.

• Rapid Elasticity: User can quickly acquire resources when it required by scale out and it releases as scale back when it no longer required.

• Measured Services: Resources usage is monitored, reported and controlled by appropriate type of metrics such as monitoring bandwidth usage, CPU hours etc.

B. SERVICE MODELS

Cloud Services generally divided into three categories:

• Software as a Service (SaaS): SaaS can provide different software applications over internet, as a service on demand. It can be describe as an application Service provider (ASP). SaaS costing less money as there is no need to buying software licenses and it eliminates the load of installing, operating and maintaining of software in a computer. SaaS offered by companies are Sales force, Microsoft, Google, Zoho etc.

• Platform as a Service (PaaS): PaaS can provide all resources to build applications and services from Internet. It facilitates run time services for application design, deployment development and testing without the cost and complexity of buying and managing the underlying infrastructure. This platform consists of infrastructure software, and typically includes a middleware, database and development tools. PaaS generally based on HTML or JavaScript. Examples of PaaS are Microsoft Azure, Force.com and Google App Engine.

• Infrastructure as a service (IaaS): Infrastructure as a Service refer as the delivery of hardware (server, network and storage), and associated software (file system, operating systems virtualization technology), as a service. It is using Application Programming Interface (API) to interact with routers, hosts, and switches. Sometimes it also called as Hardware as a Service (HaaS). Examples of IaaS are Amazon S3, Amazon Elastic Cloud Computing (EC2) and GoGrid.

C. DEPLOYMENT MODELS

Four different deployment models of Cloud Computing are shown in figure 1.1:

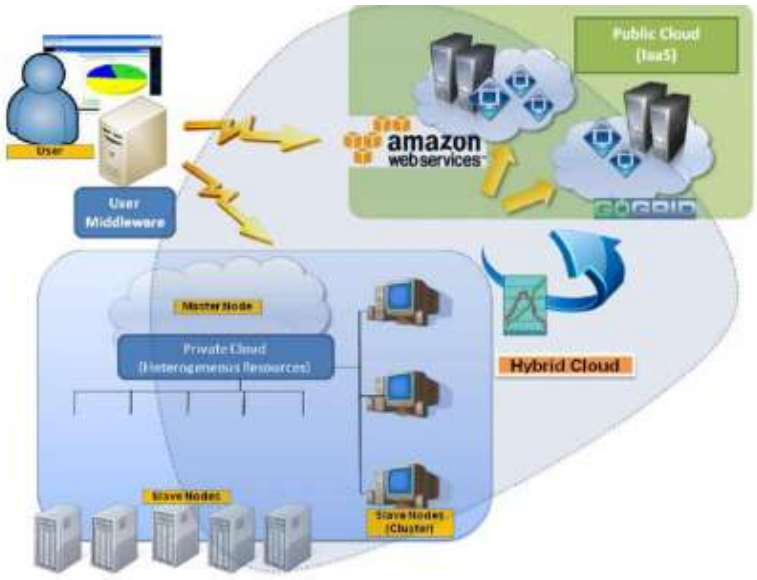

Figure 1.1 Cloud computing deployment model

Public Cloud: Public Cloud is the most common deployment model where all customers share the same infrastructure at the same time with limited security protection, configuration and availability variances. The advantage of the public cloud is that it can easily accessible to the public. In this model, authentication and authorization techniques are not used. Public cloud providers are Amazon, Google and Microsoft Azure. The characteristics of Public Clouds are scalability and multi-tenancy.

Private Cloud: Private Clouds are deployed within an organization to provide IT services to its internal users. In private cloud there are no additional security regulations, band width limitations or legal requirements that can be present in a public cloud environment. It is more secure and expensive as compared to public cloud. This cloud environment provides full guarantee of privacy. Best example of private cloud is Eucalyptus System.

Hybrid Cloud: A combination of two or more cloud deployment models, linked in such a way that data transfer takes place between them without affecting each other. With a Hybrid Cloud, service providers can utilize 3rd party Cloud Providers in a partial or full manner thus increasing the flexibility of computing. A well-constructed hybrid cloud can provide secure services like customer payment as well as employee payment processing. Example includes Amazon Web Services.

Community Cloud: Infrastructure is accessible or shared by group of organizations. These clouds are normally based on an agreement between related business organizations such as educational or banking organizations. An example community cloud is Facebook.

LOAD BALANCING

It is the mechanism 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. It also ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. Load Balancing [5] is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server. To make the final determination, the load balancer retrieves information about the candidate server's health and current workload in order to verify its ability to respond to that request. Load balancing solutions can be divided into software-based load balancers and hardware-based load balancers. Hardware-based load balancers are specialized boxes that include Application Specific Integrated Circuits (ASICs) [32] customized for a specific use. They have the ability to handle the high speed network traffic whereas Software-based load balancers run on standard operating systems and standard hardware components.

Some load balancers provide a mechanism for doing something special in the event that all backend servers are unavailable. This might include forwarding to a backup load balancer, or displaying a message regarding the outage. Load balancing gives the IT team a chance to achieve a significantly higher fault tolerance. It can automatically provide the amount of capacity needed to respond to any increase or decrease of application traffic.

It is also important that the load balancer itself does not become a single point of failure. Usually load balancers are implemented in high-availability pairs which may also replicate session persistence data if required by the specific application.

- [Wu et al.,2013b] proposed a task scheduling algorithm based on QoS-driven for cloud computing. Intially, it calculates the priority of the tasks based on various QoS parameters such as completion time, priviledges and then assign the task to a service which has minimum completion time.
- [Domanal and Reddy,2014] proposed a VM-assign load balancing algorithm which maintains an index table for the virtual machines. It finds the least loaded VM and then checks whether it is not the last used VM and assigns the load to the virtual machine. If it is the last used VM, then it nds the other least loaded VM from the index table.
- [Mesbahi et al., 2014] proposed a new multilevel cloud light weight architecture for load balancing in cloud computing. At the lower level lies the VMs. hosts and datacenters. At the middle level lies the VM manager and at the topmost level applies the Head Vm manager and ESB (Enterprise Level Bus). The algorithm proposed considers all the tasks of equal weight in terms of distributing workload and achieves load balancing as well as assures QoS to the users
- [Delavar and Aryan, 2014] proposed a hybrid heuristic algorithm to reduce the completion time. The algorithm computes the priority of tasks based on impact of tasks on each other in work flow graph. Then, it decides the suitable scheduling for these tasks and obtains early response time, load balancing and speedup ratio.
- [Sharma and Peddoju, 2014] proposed a load balancing algorithm based on the response time of the incoming requests. The algorithm only considers the response time of the incoming requests while scheduling the task to a virtual machine. It prevents extra computation and reduces communication cost.
- [Ren et al., 2011] proposed a load balancing algorithm for a cloud computing platform based on weighted least connection algorithm. The proposed algorithm is based on prediction and uses the historical data and uses the smoothing factor to make the recent data have greater impact than long term data. The algorithm is effective in reducing the load on real servers.
- [Wu et al., 2013a] proposed an elastic load balancing algorithm that uses prediction based on historical data. The algorithm is an improvement over traditional algorithm which does not respond to dynamic changes. This algorithm applies for the virtual machines in advance based on heuristic data and revises prediction results to handle the current load effectivel

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.

RESEARCH GAP

- Load balancing is used to achieve optimal resource utilization.
- It helps to increase the throughput and minimise the response time.
- To handle the over utilized and under utilized nodes.

PROBLEM DESCRIPTION

After reading the exisiting research papers, we have found the below listed problems :

- There is no sorting mechanism defined on the clusters created at the Virtual Machine level which leads to extra overhead of scanning the entire list of VMs in that particular cluster.
- There can be a scenario where the broker has received the multiple cloudlets of same type. So in the existing work, each cloudlet in the list is assigned one by one to each VM, thereby consuming lot of time, computational power and cost.
- There is no priority mechanism defined on the cloudlets.

CLOUD SIM

The CloudSim simulation layer provides support for modeling and simulation of virtualized Cloud-based data center environments including dedicated management interfaces for VMs, memory, storage, and bandwidth. The fundamental issues, such as provisioning of hosts to VMs, managing application execution, and monitoring dynamic system state, are handled by this layer. A Cloud provider, who wants to study the efficiency of different policies in allocating its hosts to VMs (VM provisioning), would need to implement his strategies at this layer. Such implementation can be done by programmatically extending the core VM provisioning functionality. There is a clear distinction at this layer related to provisioning of hosts to VMs. A Cloud host can be concurrently allocated to a set of VMs that execute applications based on SaaS provider's defined QoS levels. This layer also exposes the functionalities that a Cloud application developer can extend to perform complex workload profiling and application performance study. The top-most layer in the CloudSim stack is the User Code that exposes basic entities for hosts (number of machines, their specification, and so on), applications (number of tasks and their requirements), VMs, number of users and their application types, and broker scheduling policies. By extending the basic entities given at this layer, a Cloud application developer can perform the following activities: (i) generate a mix of workload request distributions, application configurations; (ii) model Cloud availability scenarios and perform robust tests based on the custom configurations; and (iii) implement custom application provisioning techniques for clouds and their federation.

CONCLUSION

This paper is based on cloud computing technology which has a very vast potential and is still unexplored. The capabilities of cloud computing are endless. Cloud computing provides everything to the user as a service which includes platform as a service, application as a service, infrastructure as a service.
One of the major issues of cloud computing is load balancing because overloading of a system may lead to poor performance which can make the technology unsuccessful. So there is always a requirement of efficient load balancing algorithm for efficient utilization of resources. Our paper focuses on the various load balancing algorithms and their applicability in cloud computing environment.

REFERENCES

[1] Wu, H.-S., Wang, C.-J., and Xie, J.-Y. (2013a). Terascaler elb-an algorithm of predictionbased elastic load balancing resource management in cloud computing. In Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on, pages 649-654. IEEE.

[2] Wu, X., Deng, M., Zhang, R., Zeng, B., and Zhou, S. (2013b). A task scheduling algorithm based on qos-driven in cloud computing. Procedia Computer Science, 17:1162-1169.

[3] Sharma, A. and Peddoju, S. K. (2014). Response time based load balancing in cloud computing. In Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on, pages 1287-1293. IEEE.

[4] Ren, H., Lan, Y., and Yin, C. (2012). The load balancing algorithm in cloud computing environment. In Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on, pages 925-928. IEEE.

[5] Raju, R., Amudhavel, J., Kannan, N., and Monisha, M. (2014). A bio inspired energy-aware multi objective chiropteran algorithm (eamoca) for hybrid cloud computing environment. In Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on, pages 1-5. IEEE.

[6] Mesbahi, M., Rahmani, A. M., and Chronopoulos, A. T. (2014). Cloud light weight: A new solution for load balancing in cloud computing. In Data Science & Engineering (ICDSE), 2014 International Conference on, pages 44-50. IEEE.

[7] Domanal, S. G. and Reddy, G. R. M. (2013). Load balancing in cloud computingusing modified throttled algorithm In Cloud Computing in Emerging Markets (CCEM), 2013 IEEE International Conference on, pages 1-5. IEEE.

[8] Domanal, S. G. and Reddy, G. R. M. (2014). Optimal load balancing in cloud computing by efficient utilization of virtual machines. In Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, pages 1-4. IEEE.

[9] Delavar, A. G. and Aryan, Y. (2014). Hsga: a hybrid heuristic algorithm for work flow scheduling in cloud systems. Cluster computing, 17(1):129-137.