A Study on Load Balancing in Cloud Computing

* Parveen Kumar,* Er.Mandeep Kaur
Guru kashi University, Talwandi Sabo

Abstract: *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. The resource allocation problem is the major problem for a group of cloud user requests. Another problem is resource optimization within the cloud. The scheduling algorithms are termed as NP completeness problems in which FIFO scheduling is used by the master node to distribute resources to the waiting tasks. The problem like fragmentation of resources, low utilization of the resources such as CPU utilization, network throughput, disk I/O rate. In the future research the GA is implemented to maintain the load.*

Keyword: *GA, clients, SaaS, PaaS, cloud etc.*

I. INTRODUCTION

Cloud computing is an on demand service in which shared resources, information, software and other devices are provided according to the clients requirement at specific time. Its a term which is generally used in case of Internet. The whole Internet can be viewed as a cloud. Capital and operational costs can be cut using cloud computing. In case of Cloud computing services can be used from diverse and widespread resources, rather than remote servers or local machines. There is no standard definition of Cloud computing. Generally it consists of a bunch of distributed servers known as masters, providing demanded services and resources to different clients known as clients in a network with scalability and reliability of data center. The distributed computers provide on-demand services. Services may be of software resources (e.g. Software as a Service, SaaS) or physical resources (e.g. Platform as a Service, PaaS) or hardware/infrastructure (e.g. Hardware as a Service, HaaS or Infrastructure as a Service, IaaS). Amazon EC2 (Amazon Elastic Compute Cloud) is an example of cloud computing services.

II. CLOUD COMPONENTS

A Cloud system consists of 3 major components such as clients, data center, and distributed servers. Each element has a definite purpose and plays a specific role.

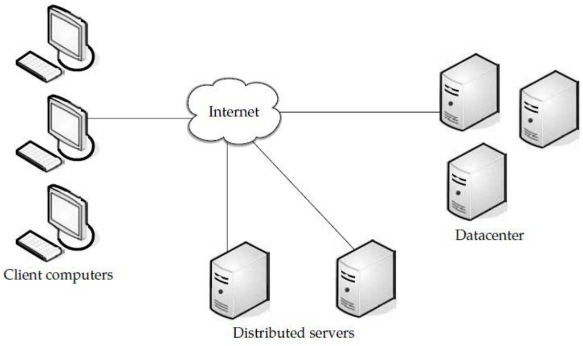


Figure 1: Three components make up a cloud computing solution

Clients

End users interact with the clients to manage information related to the cloud. Clients
generally fall into three categories as given in [1]:

- Mobile: Windows Mobile Smartphone, smart phones, like a Blackberry, or an iPhone.
- Thin: 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: 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.

Data enter

Data center is nothing but a collection of servers hosting different applications. A end user connects to the data center to subscribe different applications. A data center may exist at a large distance from the clients.
Now-a-days a concept called virtualisation is used to install a software that allow multiple instances of virtual server applications.

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.

III. TYPE OF CLOUDS

Based on the domain or environment in which clouds are used, clouds can be divided
into 3 catagories :
_ Public Clouds
_ Private Clouds
_ Hybrid Clouds (combination of bothe private and public clouds)

IV. SERVICES PROVIDED BY CLOUD COMPUTING

Service means different types of applications provided by different servers across the
cloud. It is generally given as ”as a service”. Services in a cloud are of 3 types as given in :
_ Software as a Service (SaaS)
_ Platform as a Service (PaaS)
_ Hardware as a Service (HaaS) or Infrastructure as a Service (IaaS)

Software as a Service (SaaS)

In SaaS, the user uses different software applications from different servers through the Internet. The user uses the software as it is without any change and do not need to make lots of changes or doen’t require integration to other systems. The provider does all the upgrades and patching while keeping the infrastructure running .

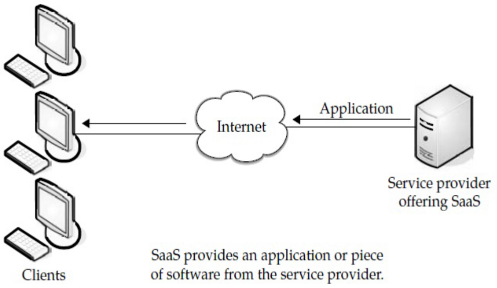

Internet
Application
Service provider
offering SaaS
Clients
SaaS provides an application or piece
of software from the service provider.

Figure 2: Software as a service (Saas)

The client will have to pay for the time he uses the software. The software that does a simple task without any need to interact with other systems makes it an ideal candidate for Software as a Service. Customer who isn’t inclined to perform software development but needs high-powered applications can also be benefitted from SaaS.
Some of these applications include:
_ Customer resource management (CRM)
_ Video conferencing
_ IT service management
_ Accounting
_ Web analytics
_ Web content management

Benefits: The biggest benefit of SaaS is costing less money than buying the whole application. The service provider generally offers cheaper and more reliable applications as compared to the organisation. Some other benefits include: Familiarity with the Internet, Better marketing, Smaller staff, reliability of the Internet, data Security, More bandwidth etc.

Obstacles:
_ SaaS isn’t of any help when the organisation has a very specific computational need that doesn’t match to the SaaS services
_ While making the contract with a new vendor, there may be a problem. Because the old vendor may charge the moving fee. Thus it will increase the unnecessary costs.
_ SaaS faces challenges from the availability of cheaper hardwares and open source applications.

Platform as a Service (PaaS)

PaaS provides all the resources that are required for building applications and services completely from the Internet, without downloading or installing a software.
PaaS services are software design, development, testing, deployment, and hosting.Other services can be team collaboration, database integration, web service integration, data security, storage and versioning etc.

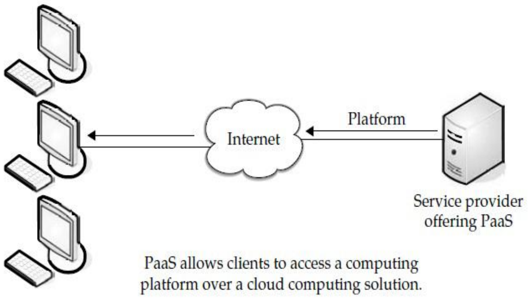

Internet
Platform
Service provider
offering PaaS
PaaS allows clients to access a computing
platform over a cloud computing solution.

Figure 3: Platform as a service (PaaS)

Hardware as a Service (HaaS)

It is also known as Infrastructure as a Service (IaaS). It offers the hardware as a service to a organisation so that it can put anything into the hardware according to its will [1].
HaaS allows the user to “rent” resources (taken from [1]) as

_ Server space
_ Network equipment
_ Memory
_ CPU cycles
_ Storage space

Cloud computing provides a Service Oriented Architecture (SOA) and Internet of Services (IoS) type applications, including fault tolerance, high scalability, availability, flexibility, reduced information technology overhead for the user, reduced cost of ownership, on demand services etc. Central to these issues lies the establishment of an effective load balancing algorithm.

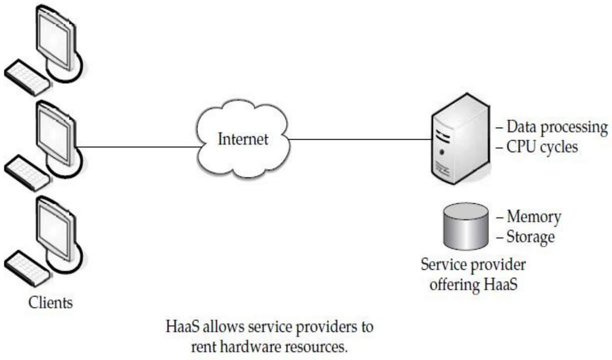


Figure 4: Hardware as a service (HaaS)

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

V. LITERATURE SURVEY

In this paper I have studied the different papers to review my research topic. I have studied different authors papers .each have followed the different techniques and methods.

Florin Pop, Valentin Cristea [2013] in this work Evolutionary computing offers different methods to solve NP-hard problems, finding a near-optimal solution. Task scheduling is a complex problem for large environments like Clouds. Genetic algorithms are a good method to find a solution for this problem considering multi-criteria constrains. This is also a method used for optimization. In these type of environments service providers want to increase the profit and the customers (end-users) want to minimize the costs. So, its all about money and we have minimum two optimization constrains. On the other hand, a good technique to ensure the QoS is to use the reputation of resources offered. This aspect is very important for service providers because represents a ranking method for them. We present in this paper a reputation guided genetic scheduling algorithm for independent tasks in inter-Clouds environments. The reputation is considered in the selection phase of genetic algorithm as evolutionary criteria for the algorithm. We evaluate the proposed solution considering load-balancing as a way to measure the optimization impact for providers and max span as a metric for user performance.[1]

Lucio Agostinho [2011] In cloud computing the allocation and scheduling of multiple virtual resources, such as virtual machines (VMs), are still a challenge. The optimization of these processes bring the advantage of improving the energy savings and load balancing in large data centers. Resource allocation and scheduling also impact in federated clouds where resources can be leased from partner domains. This paper proposes a bio-inspired VM allocation method based on Genetic Algorithms to optimize the VM distribution across federated cloud domains. The main contribution of this work is an inter-domain allocation algorithm that takes into account the capacity of the links connecting the domains in order to avoid quality of service degradation for VMs allocated on partner domains. [2]

VI. PROBLEM DEFINITION

➢The resource allocation problem is the major problem for a group of cloud user requests.

➢Another problem is resource optimization within the cloud.

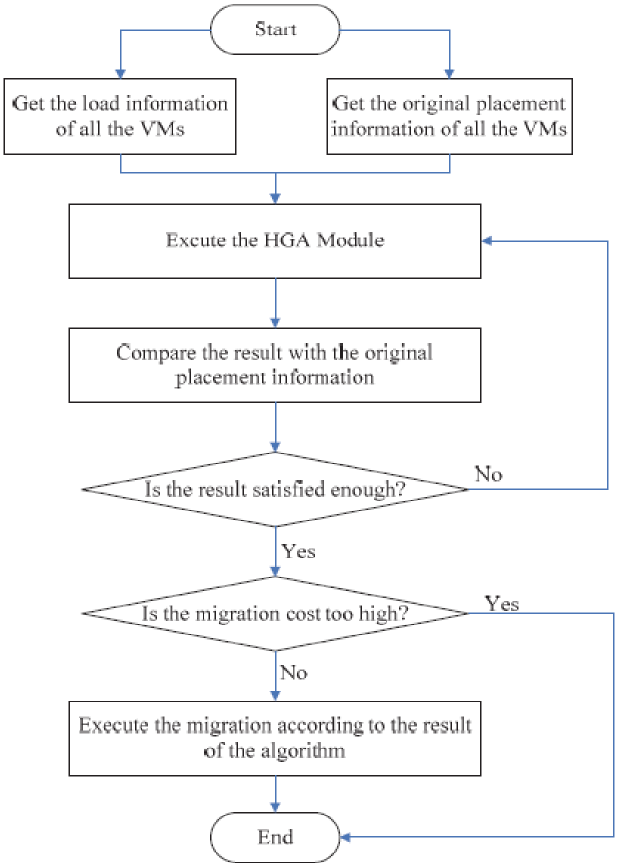
➢The scheduling algorithms are termed as NP completeness problems in which FIFO scheduling is used by the master node to distribute resources to the waiting tasks.

➢The problem like fragmentation of resources, low utilization of the resources such as CPU utilization, network throughput, disk I/O rate.

➢The execution of this algorithm with the virtual machine’s actual migration will save 30-40% of the total physical machine’s occupation as well as smooth the utilization of the loads.

VII.METHODOLOGY

To solve the above problems within the cloud the research work is proposed a cloud computing resource scheduling policy based on genetic algorithm with multiple fitness. To validate the efficiency of the algorithm I proposed a typical cloud computing model which is same as other cloud computing environment. For management of all the physical resources and storage systems there is a cloud manager as the cloud management entry point. On every physical machine a virtualization layer called hypervisor is installed and all the virtual machine is created. The base of this algorithm is that the three load dimensions: CPU load, network throughput and disk I/O load of all the virtual machines carried on one specific physical machine, can be matched and calculated to get the optimal migration advice. So it is necessary that the system must monitor the three dimensions, which fetched from the monitoring-database when the algorithm is executed. To implement this the cloudsim and dot net is used.


Start
Get the load information of all the VMs
Get the original placement information of all the VMs
Excute the HGA Module
Compare the result with the original placement information
Is the result satisfied enough?
No
Yes
Is the migration cost too high?
Yes
No
Execute the migration according to the result of the algorithm
End

VIII. CONCLUSION

In cloud computing the allocation and scheduling of multiple virtual resources, such as virtual machines (VMs), are still a challenge. The optimization of these processes brings the advantage of improving the energy savings and load balancing in large data centers. Genetic algorithms are a good method to find a solution for this problem considering multi-criteria constrains. This is also a method used for optimization. In these types of environments service providers want to increase the profit and the customers (end-users) want to minimize the costs. So, it's all about money and we have minimum two optimization constrains. On the other hand, a good technique to ensure the QoS is to use the reputation of resources offered. In the conclusion load balancing problem is found from the literature survey and in the future GA is implemented to get the better result.

References

[1] Florin Pop, Valentin Cristea "Reputation guided Genetic Scheduling Algorithm for Independent Tasks in Inter-Clouds Environments "27th International Conference on Advanced Information Networking and Applications Workshops, 2013.

[2] Lucio Agostinho, Guilherme Feliciano, Leonardo Olivi, Eleri Cardozo" A Bio-inspired Approach to Provisioning of Virtual Resources in Federated Clouds" IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing,2011.

[3] Jianfeng Zhao, Wenhua Zeng, Min Liu, Guangming Li" Multi-objective Optimization Model of Virtual Resources Scheduling Under Cloud Computing and It's Solution" International Conference on Cloud and Service Computing,2011.

[4] R. Iyer, R. Illikkal, O. Tickoo, L. Zhao, P. Apparao, D. Newell. VM3:Measuring, modeling and managing VM shared resources.. Computer Networks. vol. 53, pp. 2873–2887, Aguest 2009.

[5] A. d. Costanzo, M. D. d. Assunção, R. Buyya. "Harnessing Cloud Technologies for a Virtualized," Distributed Computing Infrastructure, vol. 13, pp. 24-33, Octobor 2009.

[6] G. Tian, D. Meng, J. Zhan. "Reliable Resource Provision Policy for Cloud Computing," Chinese Journal of computer, vol. 33, pp. 1859-1872, Octobor 2010.

[7] D. S. Hochbaum. "Approximation Algorithms for NP-Hard Problems,"Boston, PWS Publishing Company, p. 23, 1997.

[8] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan. "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation. vol. 6, pp. 182-197, April 2002.

[9] C. Shi, Z. Yan, Z. Shi, L. Zhang. "A fast multi-objective evolutionary algorithm based on a tree structure," Applied Soft Computing, vol. 10,pp. 468–480, Feburary 2010.