A Survey on Varies Existing Load Balancing Approach in Cloud Computing Environment

Md. Ahamad Ansari[1], Prof. K. K. Tiwari[2]
ahamad.ansari88@gmail.com[1], krishna19it@gmail.com[2]
Department of Computer Science & Engineering
Surabhi College of Engineering & Technology, Bhopal, India

***ABSTRACT*—Cloud computing is one the fastest growing technology in the IT industries. It emerges as a new paradigm for large scale distributed computing. It provides on-demand services to the user and support for a utility computing. Load balancing is one of the important modules for any computing technology like cluster computing, grid computing, distributed computing, cloud computing etc. But load balancing in cloud is more difficult as compare to other technology because it is so large and user requirement can be change dynamically. It helps in optimizing the resource utilization, hence enhancing the system performance. The prime goal of any load balancing approach is to maximize the resource utilization and reducing the number of active server which will further reduce energy consumption and carbon emission. During the past decades several load balancing approach have been proposed. This paper discussed some existing load balancing approaches in cloud computing and further compares them corresponding advantages, disadvantages and performance metrics are studied in detail.**

Keywords: Cloud computing, Cloud Service Model, Utility computing, Load balancing, Energy efficiency, Green computing.

I. INRODUCTION

Cloud computing is one the fastest growing technology in the IT industry due to it attractive features [1]. It became so popular because of its seductive services like on-demand, easy to use etc. It has moved computing and data away from desktop and portable PCs, into large data centers [2]. A cloud environment consists of multiple data centers. Each data center has several VM and each VM can run multiple applications. When the users request for the resources, cloud provider give the resources to the users according to their requirement. It is a utility model, so user need to pay only for the used resources like electricity bill [3, 4]. It can be deploy in four different way (private, public, community and hybrid) and provide three type of services (software as service, platform as service and infrastructure as service) [4, 5, 6] as shows in figure 1.

Private Cloud: - Private cloud allows users to access cloud services within the network. User can't access cloud services from outside the network. It is suitable for the small organization. User in public cloud can access cloud services from anywhere in the world. It is larger than the private cloud and provides larger number of services. Community cloud is a cloud which is share by the multiple organizations. Hybrid cloud is a combination of one or more cloud like combination of private and public cloud, private and community cloud etc.

Cloud provide services to the users and provide three type of service named software as service (SaaS), platform as service (PaaS) and infrastructure as service (IaaS).

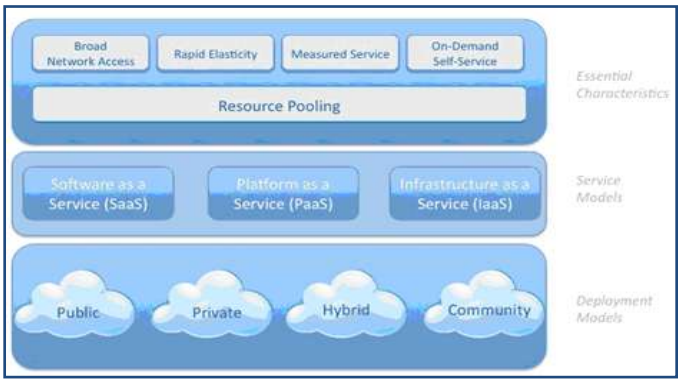

Figure 1: Cloud Computing Architecture

Software as a service mainly delivers the online software application to the client on-demand. User uses these software and application without any installation. Platform as a Service (PaaS) allows developer or gives the capability to create application as a service according to their desire. It allows users to develop their software using programming languages and tools supported by the provider. Infrastructure as a Service (IaaS) provides the capability to have control over complete cloud infrastructure with CPU processing, storage, networks, and other computing resources.

One of the core technologies in cloud is the virtualization which allows the provider to divide the physical resources of the data center. Hypervisor is use to virtualized the PM which create the VM according to the user requirement and assign to the user. Figure 2 shows the basic concept of the virtualization.

One VM is assign to one user and single VM can execute multiple VM. Main advantage of the virtualization is migration of the VM. VM migration transfers the VM from one PM to another PM when it is needed. VM is required mainly to handled load balancing and server consolidation.

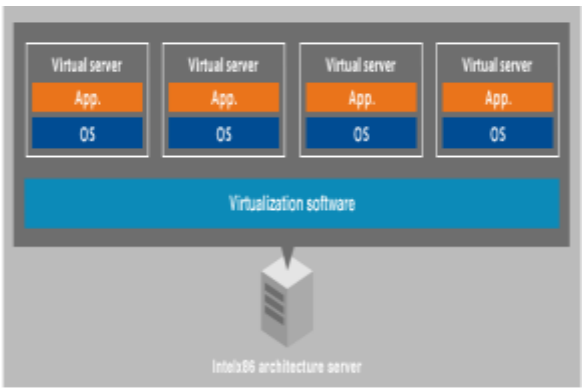

Figure 2: Virtualization

Server consolidation [14] is the technique which consolidates the running services to the minimum number of

active servers. Lower threshold is use for this purpose. when the load on the PM is below the lower threshold all VM running on that PM are migrate to other PM. Server consolidation reduce the energy consumption. Figure 3 shows the basic concept of the server consolidation.

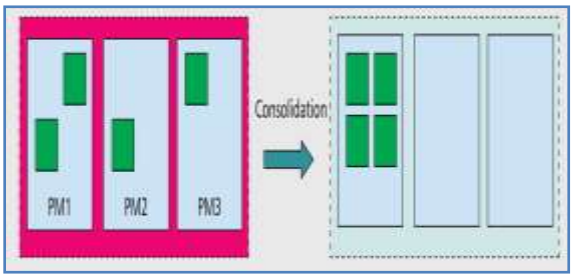

Figure 3: Server consolidation

II. LOAD BALANCING

Load balancing means distributed resources in such a way that avoid the situation where some load are overloaded or some node are underloaded. Load balancing in cloud is one of the main challenging issues due to dynamic behavior of the users in term of resource requirement. Computer consists of several resources like CPU, memory, N/W etc. Load can be required in each resource like CPU, memory, network or delay load. Since resources in cloud are distributed to multiple users to load balancing in cloud is necessary to improve the performance of the system. The goals of load balancing [7] are to-

1. Improve the performance
2. Maintain system stability
3. Build fault tolerance system
4. Accommodate future modification.

Load balancing approaches can be static or dynamic:

A. Static Algorithm

In static algorithm the traffic is divided equally among the servers manually. This algorithm requires a prior knowledge of system resources, so that the decision of shifting of the load does not depend on the current state of system. Static approaches are well suitable for the cloud because resource required by the user change dynamically.

B. Dynamic Algorithm

In dynamic load balancing approaches decision where the VM will be placed is decided at the run time. It avoids the human monitoring. But it is not appropriate for the cloud because it is large and users demand may change dynamically.

C. Need of load balancing in cloud computing

Mainly load balancing approaches is use to avoid situation where single node is overwhelmed and to achieve a high user satisfaction and resource utilization ratio [8]. Main purposes of the load balancing approaches are:

1. Reducing Energy Consumption
2. Reducing Carbon Emission
3. Increasing Resource Utilization
4. Minimize Response Time

III. LITERATURE SURVEY

As the demand for the resources is increase, load balancing is one of the main challenges for the cloud provider because poor load balancing decrease the resource utilization. During the past decades several load balancing approach have been proposed. This paper discussed some existing load balancing approaches in cloud computing and further compares them corresponding advantages, disadvantages and performance metrics are studied in detail.

A. Beloglazov et al. [9], proposed threshold based an energy efficient load balancing approach. According to this paper average power consumed by an idle server is 70% of power consumed by fully utilized server. Hence, appropriate load balancing approach can controlled the power consumed by the data center. This approach used static lower and upper threshold with the difference of 40 between lower and upper threshold. After the experiment they set 30 as a lower threshold and 70 as an upper threshold. This approach reduced the number of migration but main problem with this approach is that they are not working on resource balancing.

Y. Lua et al. [10] proposed a dynamic load balancing approach named JoinIdle-Queue for dynamically scalable web services. This algorithm provides large scale load balancing with distributed dispatchers by, first load balancing idle processors across dispatchers for the availability of idle processors at each dispatcher and then, assigning jobs to processors to reduce average queue length at each processor. By removing the load balancing work from the critical path of request processing, it effectively reduces the system load, incurs no communication overhead at job arrivals and does not increase actual response time.

Y. Fang et al. [11] discussed a load balancing approach which perform two-level of task scheduling balancing to meet dynamic requirements of users and obtain a 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.

J. Hu et al. [12] proposed a historical data based load balancing approach for the cloud environment. They design a module that store the history data and current state of the PM i.e., resource usage data of the PM. This approach uses genetic algorithm for the VM placement. It helps in resolving the issue of load-imbalance and high cost of migration thus achieving better resource utilization.

R. Addawiyah et al. [13], proposed a load balancing approach for the cloud which is based on the VM migration. Main objective of this approach is to minimize the power consumption for this purpose VM placed only on the basis of CPU utilization. This approach set the value of lower threshold is 10 and the value of upper threshold is 90. That means when the value of CPU utilization is more than 90 then the PM is overloaded. Similarly when the value of CPU utilization is less than 10 then the PM is underloaded.

IV. CONCLUSION

Load balancing means distributed resources in such a way that avoid the situation where some load are overloaded or some node are underloaded. Load balancing in cloud is one of the main challenging issues due to dynamic behavior of the users in term of resource requirement. Several load balancing approach have been proposed during last few years.

Most load balancing approaches use two thresholds named lower and upper threshold. Lower threshold is use to

define when the system is under loaded whereas upper threshold is use to define when the system is overloaded. This paper explained some existing load balancing approach. After studding these algorithms, it can be conclude that system performance is totally depends on the load balancing approach.

REFERENCES

[1] R. W. Luckyet al., "Cloud computing", IEEE Journal of Spectrum, Vol. 46, No. 5,May 2009, pp. 27-45.

[2] 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.

[3] R. K. Gupta et al., "A Complete Theoretical Review on Virtual Machine Migration in Cloud Environment", International Journal of Cloud Computing and Services Science (IJ-CLOSER), Vol.3, No.3, June 2014, pp. 172-178.

[4] R. K. Gupta et al.,," Survey on Virtual Machine Placement Techniques in Cloud Computing Environment", International Journal on Cloud Computing: Services and Architecture (IJCCSA) ,Vol. 4, No. 4, August 2014, pp. 1 -7.

[5] 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.

[6] 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.

[7] R. Mata-Toledo, and P. Gupta, "Green data center: how green can we perform", Journal of Technology Research, Academic and Business Research Institute, Vol. 2, No. 1, May 2010, pages 1-8.

[8] Z. Zhang, and X. Zhang,"A Load Balancing Mechanism Based on Ant Colony and Complex Network Theory in Open Cloud Computing Federation", Proceedings of 2nd International Conference on Industrial Mechatronics and Automation (ICIMA), Wuhan, China, May 2010, pages 240-243.

[9] A. Beloglazov et al. "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", Future Generation Computer Systems (Elsevier), May 2012, pp. 755–768.

[10] Y. Lua, Q. Xiea, G. Kliotb, A. Gellerb, J. R. Larusb, and A. Greenber, "Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services", An international Journal on Performance evaluation, In Press, Accepted Manuscript, Available online 3 August 2011.

[11] 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.

[12] J. Hu, J. Gu, G. Sun, and T. Zhao, "A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment", Third International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), 2010, pages 89-96.

[13] R. Addawiyah et al., "Virtual Machine Migration Implementation in Load Balancing for Cloud Computing", six IEEE international conference, 2014.

[14] Shivani Goel and Tripti Arjariya, " A Review on Different Energy Efficient VM Placement Approaches with their Anomalies in Cloud computing Environment", International Journal of Computer Security & Source Code Analysis (IJCSSCA),Vol.1, Issue, pp. 18-21, 2015.

