International Journal in Management and Social Science
Volume 09 Issue 11, November 2021 ISSN: 2321-1784 Impact Factor: 7.088
Journal Homepage: http://ijmr.net.in, Email: irjmss@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal

EXAMINING THE CONCEPT OF LOAD BALANCING AND REBALANCING

ALTAF RAJA
RESEARCH SCHOLAR SUNRISE UNIVERSITY ALWAR,RAJASTHAN

DR. SUMIT BHATTACHARJEE
PROFESSOR, SUNRISE UNIVERSITY ALWAR, RAJASTHAN

ABSTRACT

With the advent of cloud computing, it is now possible to have ready access to computer resources at any time, regardless of whether they are currently being used. The virtual machine is a key component of cloud computing because it provides isolation for business services and transports resources. One of the challenges of cloud computing is determining the best way to physically host virtual machines. The effective distribution of virtual machines (VMs) among available computer servers has emerged as a critical area of study in light of the proliferation of large-scale cloud computing systems. To maximize the effectiveness of VM placement, this study introduces a Virtual Machine Placement and Load Rebalancing (VMP-LR) method based on multi-dimensional resource characteristics in cloud computing systems. The Resource Request Handling Component, the Placement Component, and the Load Monitoring Component are the three primary parts of the proposed VMP-LR. Phase I of VMP-LR is responsible for the work associated with the Resource Request Handling and Placement Components, while Phase II is responsible for the work associated with the Load Monitoring Component. Both sets of suggested algorithms can deal with numerous requests for the same resource. When it comes to virtual machine deployment, load balancing, and rebalancing, VMP-LR takes into account three resources: central processing unit, random access memory, and bandwidth.

Keywords: - Cloud, VM, Computing, Load, Balancing.

I. INTRODUCTION

The period between the end of the 20th century and the beginning of the 21st is sometimes referred to as "the golden age of computing," since this is when the use of computers and digital solutions spread beyond the business and academic worlds into the mainstream. The simultaneous rise of the personal computer and the World Wide Web (WWW) has been a major factor in this growth. In such setups, the actual work is carried out by apps running on remote servers. As more low-cost and high-tech hardware and software became widely available, it became common practice to disperse these servers around many datacentres. But this has led to higher overhead and upkeep expenses. As most enterprise applications have dynamic workloads

with resource requirements that vary according to time, season, and location, research into this topic has found that most applications in these servers use less than 20% of the resources allocated to them. (For instance, a bank's web server is likely to see a spike in demand for its resources around peak banking hours, whereas a media or gaming site's server may see a similar spike after normal business hours. Wastes of resources or inefficient use of available resources are suspected to have a role in this phenomena. For instance, programs that make heavy use of the central processing unit (CPU) tend to be wasteful with the input/output (I/O) resources available to them, whereas apps that are CPU-bound tend to underutilize the CPU's capabilities. Contention for scarce resources happens when applications are given more than they need. This may lead to underutilization of available resources.

In response to the aforementioned problem, cloud computing and virtualization have developed. The flexibility, low cost, and on-demand / pay-as-you-go nature of cloud computing have contributed to its meteoric rise in popularity. The term "cloud" is used to describe a platform that provides a simple graphical interface or API (Applications Programming Interface) to users and applications while hiding the complexity and details of the underlying infrastructure from them.

Virtualization, distributed computing, service-oriented architectures, broadband networks, the browser as a platform, servers, SAN/NAS (Storage Area Network/Network Attached Storage), and open-source software are all components of cloud computing systems. All of these parts work together to improve the services provided to customers. This study focuses on the virtualization process rather than other aspects of it.

II. CLOUD COMPUTING

Computing that takes place over the web or the internet is what is referred to as "cloud computing." Several large corporations now prefer using this kind of computing. Cloud platforms have been developed by major IT companies like Google, Yahoo, Microsoft, Amazon, and IBM, which all make their massive datacentres full of rentable networked computers accessible to consumers. Information on the many facets of cloud computing are provided here.

Definition

Cloud computing is "a model for 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" (Mell et al., 2011), as stated by the National Institute of Standards and Technology. Based on this concept, we may think of cloud computing as networked accesses (or

requests) to potentially shared resources from a pool, supplied on an as-needed basis with minimum intervention from the service provider's administration and low effort from the user. When a cloud provider pools all of its computing resources into one huge pool, they are said to have a resource sharing pool.

III. LOAD BALANCING

Adjusting the workloads of the many components (central processing units, network connections, etc.) that make up the cloud is the primary challenge of cloud computing. To prevent having some nodes excessively loaded while others are idle or performing minimal work, load balancing is a method that distributes the dynamic local workload equally among all the nodes in the whole cloud.

Rebalancing is the act of redistributing work across the nodes of a distributed system in order to eliminate the overloaded and under loaded states that previously existed and maximize resource usage and task response time. Dynamic load balancing algorithms are not concerned with the system's past state or behaviour; rather, they are dependent on the system's current activity. When designing such an algorithm, it's crucial to think about a variety of factors, including but not limited to: load estimate, load comparison, system stability, system performance, node interaction, work transfer type, and node selection. CPU utilization, memory consumption, response time, and network traffic are all possible measures of load.

In order to balance the demands on the available resources, it is essential to understand the primary motivations behind load balancing algorithms.

- To dramatically enhance performance.
- Always be prepared with a fall back strategy, in case the system fails in any way.
- Keeping the system stable and adaptable so that changes in the cloud, such as topological ones, may be readily managed.
- So that it may grow with the system as it evolves.
- The goal is to increase system performance while keeping costs low.
- Giving more priority to tasks that are urgent in order to better meet their needs.
- There are typically three types of load balancing algorithms (Joshi et al., 2014), and they vary depending on who started the process.

- Sender Initiated: The sender initiates the load-balancing procedure.
- The receiver may trigger the load-balancing algorithm in the "Receiver Initiated" mode.
- Both the transmitter and the receiver may trigger the load-balancing process, making it symmetrical.

There are two types of load balancing algorithms that may be used depending on the present configuration of the system.

- **Static:**It is independent of the system's present condition. You need to be familiar with the system beforehand.
- **Dynamic:**The present status of the system is taken into account while making load balancing decisions. It's preferable than a purely static method since it requires no previous information.

There is a large dispersion of nodes in the cloud. Therefore, the kind of algorithm employed is determined by the node that makes the provisioning decision. In a cloud computing setup, load balancing may be handled by one of three possible techniques.

IV. LOAD REBALANCING

The most well-known cloud service providers often provide a variety of VM types and accompanying resource requirements. Some of these virtual machine (VM) instances are bigger than others, while some have a better capacity for a particular resource type compared to the others. Workload prediction and estimate may be achieved by capturing the dynamic resource needs of running virtual machines (VMs) in the cloud.

Complementary resource needs across different resource dimensions are widespread in cloud data centres due to the aforementioned characteristics. Clouds also provide a pay-as-you-go pricing approach, so customers may create and delete as many virtual machines as they need. When current services are terminated, either by cloud customers or because of a host power cycle, an imbalance in the load is created. As a result, the server's ability to handle new requests is degraded, leading to fragmentation.

In a dynamic cloud environment, where VM workloads fluctuate regularly, it is impossible to maintain load balance with even the most optimal placement choices. Load rebalancing methods may be used to handle this situation.

V. MOTIVATION

By allowing for the elastic on-demand supply of computer resources, cloud computing, together with social, mobile, and analytic technologies, has changed the IT sector. In 2015, 9.76% more businesses were using the cloud than in 2014, as stated by the Right Scale Survey.

However, forecasts that cloud use will have increased by 41.3% by the year's end. More than 82 percent of businesses expect significant savings by moving to the cloud, according to a recent survey. The same analysis predicts that in the following year, worldwide DC traffic will expand by a factor of three, while global traffic will increase by a factor of three and a half.

Based on these findings and projections, it is clear that cloud computing is an indispensable technique with high requirements for service quality in terms of security, convenience, low cost, and efficient use of available resources. Because of this need, cloud computing has become an important area of study.

VI. CONCLUSION

In recent years, cloud computing has become an abundant supply of processing power. In order to efficiently run several applications, a cloud computing system may make its resources elastic, scalable, and available on demand. Virtualization technology is an integral part of cloud computing systems. Through the use of virtualization, computers and their associated software may be shared across a number of users and programs. The primary emphasis of this study is on the virtual machine placement step of virtualization.

Placement of virtual machines (VMs) is the act of assigning these requests for resources to physical machines (PMs) in a way that maximizes resource usage and minimizes the time needed to provide the service. Traditionally, manual mapping of VMs to suitable PMs is achievable when the number of VMs and PMS is modest. With the present circumstances, however, the number of VMs and PMs has increased dramatically, making automation of the placement operation essential. In order to generate near-optimal placement plans, existing automated methods must analyze a large number of alternative mappings for a given set of virtual machines (VMs) and physical machines (PMs).

REFERENCES:-

1. Agrawal, K. and Tripathi, P. (2015) Power Aware Artificial Bee Colony Virtual Machine Allocation for Private Cloud Systems, International Conference on Computational Intelligence and Communication Networks Pp.947-950.

2. Amalarethinam, G.D. and Beena, T.L.A. (2016) Workflow scheduling for publc cloud using genetic algorithm (WSGA), IOSR Journal of Computer Engineering, Volume 18, Issue 3, Ver. V, Pp. 23-27.

3. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I. and Zaharia, M. (2009) Above the Clouds: A Berkeley View of Cloud Computing, Technical Report No. UCB/EECS-2009-28, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/ EECS-2009-28.html, Pp. 1-25.

4. Banerjee, S., Mukherjee, I. and Mahanti, P.K. (2009) Cloud Computing Initiative using Modified Ant Colony Framework, World Academy of Science and Technology, Vol. 56, Pp. 221-224.

5. Beloglazov, A. and Buyya, R. (2010a) Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers, Proceedings of the 8th ACM International Workshop on Middleware for Grids, Clouds and e-Science, Vol. 4, Pp. 1-8. -review

6. Beloglazov, A. and Buyya, R. (2010b) Energy efficient allocation of virtual machines in cloud data centers, Proceedings of 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, Pp. 577-578.