International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 02, February-2015

Load Prediction of Virtual Machines in a Cloud Environment

S. Saranya,
Assistant Professor,
Computer Science Engineering,
SRM University, Chennai

G. Priyanka
M. Tech,
Computer Science Engineering,
SRM University, Chennai

***Abstract*- Cloud computing has many business customers uses resources usage based on their needs. Through virtualization technology which come from resource multiplexing many touted gains in the cloud model. In this paper, system uses virtualization technology to allocate data centre resources dynamically based on its application demands and it supports green computing by optimizing the number of servers in use. The concept of "skewness" is used to measure the unevenness in the multidimensional resource utilization of a server. By minimizing skewness and combining different types of workloads nicely and improve the overall utilization of server resources. The development of a set of heuristics that prevent overload in the system effectively while saving of energy is used. Trace driven simulation and the experiment results that algorithm achieves good performance.**

Keywords-Cloud Computing, Resource Management, Virtualization, Green Computing,

I. INTRODUCTION

A Cloud Computing is a model for enabling environment, on demand network access to a shared pool of configurable computing resources (e.g., networks, applications, servers, storages and services) that could be rapidly provisioned and released with minimal management effort or service provider interaction. Cloud computing has emerged as a popular solution to provide cheap and easy access to externalized IT (Information Technology) resources. An increasing number of organizations benefit from cloud computing to host their applications[1]. Through virtualization, cloud computing is able to address with the physical infrastructure a large client base with different computational needs. In order to previous paradigms, Cloud computing is not application oriented but service oriented; it offers on demand virtualization resources as measurable and billable utilities.

Cloud computing is receiving a great deal of attention, both in applications and among users. Cloud computing is a subscription based service where can obtain networked storage space and computer resources. Yahoo, Gmail, Hotmail takes care of housing all of the hardware and software necessary to support personal email account is consider to be a part of cloud. To access email open web browser, go to the email client, and log in. Email is not housed on computer, access it through an internet connection and can access it anywhere, if were on a trip, at work, or down the street getting coffee, can check email as long as have access to the internet. Email is different than software installed on computer, such as word processing program[2]. When create a document using word processing software that document stays on the device used to make it unless physically move it. Email client is similar to cloud works. Except instead of accessing just your email, can choose what information have access to within the cloud.

Virtual machine monitors provides Xen for mapping virtual machines. The mapping is largely hidden from cloud users because they don't know where their VM instances run. The VM live Migration itself possible to change the mapping between VMs and PMs. The capacity of PMs also be heterogeneous because of multiple generations of hardware which are exists in data centre

II. RELATED WORK

2.1 Resource allocation and virtual machines

The elasticity and the lack of upfront capital investment offered by the cloud computing is appealing to many business. There is a lot of discussion on the benefits and costs of cloud model and how to move legacy applications onto the cloud platform. This is important because much of the touted gains in the cloud model come from the existing. Due to over provisioning for the peak demand studies have found that servers in many existing data centers are often severely underutilized. In response to load variation the cloud model is expected to make practice of unnecessary by offering automatic scale up and down. Besides reducing the cost of hardware, it also saves on electricity which contributes to a significant portion of the operational expenses in large data centers.

Virtual machine monitors like Xen provide a mechanism for mapping virtual machines to physical resources[3]. Thus the mapping is largely hidden from the cloud users with an instances which they run. It is up to the cloud provider to make sure about the underlying resources have their sufficient physical machine to meet their needs. While applications are running VM live migration technology makes it possible to change the mapping between VMs and PMs. However, a policy issue remains as to decide the mapping adaptively so that the resources demands are met while the number of PMs used is minimized and it is a challenging task when the resource needs of VMs are heterogeneous due to the diverse set of applications they run and vary with time as the workloads grow and shrink.

Capacity of PMs can also be heterogeneous because multiple generations of hardware coexist in a data center. Overload avoidance the capacity of PM should be sufficient to satisfy the resource needs of all VMs running on it as possible[4]. Otherwise, the PM is overloaded and can lead to degrade performance of its VMs.

2.2 Green Computing

The number of PMs used should be minimized as long as they can still satisfy the needs of all VMs. Idle PMs can be turned off to save the energy. There is an inherent tradeoff between the two goals. With in the face of changing resource needs of VMs Should keep the utilization of PMs. Later on low to reduce the possibility of overload in the case of resource needs of VMs increase. For green computing, we have to keep the utilization reasonably high to make efficient use of their energy. We present the design and implementation of an automated resource management system that achieves a good balance between the two goals.

2.3 Hot Spot

A server as a hot spot if the utilization of any of its resources is above a hot threshold which indicates that the server is overloaded and hence some VMs running on it should be migrated away. The hot thresholds for CPU and memory resources to be 90 and 80 percent, respectively[5][6]. Thus a server is a hot spot if either its CPU usage is above 90 percent or its memory usage is above 80 percent. The temperature of a hot spot reflects its degree of overload by itself. If a server is not a hot spot, its temperature is zero and it goes to migartion list.

2.4 Cold Spot

A server as cold spot if the utilizations of all its resources are below a cold threshold. Cold threshold mean resource utilization below 25 percentages. This indicates that the server is mostly idle and a potential resource to turn off to save energy. The number of cold spots that can be minimized and sort the list of cold spots in the system based on the ascending order of their memory size. Need to migrate away all its VMs before we can shut down an underutilized server. The memory size of a cold spot as the aggregate memory size of all VMs running on it. Restrict no more than a certain percentage of active servers in a system. This is called consolidation limit.

III. ALGORITHAM AND TECHNIQUES

The concept of Virtualization, in computing is creation of a virtual version of something, such as a operating system, hardware platform, and a storage device or network resources.VM live migration is the technique that always used for dynamic resource allocation in a virtualized environment.

3.1 Resource Management Approach:

Dynamic Resource management has become an active area of research in the cloud computing paradigm. The cost of resources varies significantly depending on configuration for using them[7].Hence efficient management of resources is of prime interest to both cloud users and providers. Successful resource management solution for cloud environments needs to provide a rich set of resource controls for better isolation, while doing initial placement and future load partitioning of underlying resources.

3.2 Skewness Algorithm:

Skewness is a measure of the uneven utilization of a server. By minimizing skewness(by combining after dividing), we can improve the overall utilization of servers in the face of multidimensional resource constraints. Skewness algorithm is to mix workloads with different resource requirements together so that overall utilization of server capacity is improved.

The algorithm has future resource demand of VMs which has a server as hotspot and cold spot. If the utilization of resources is above the level of hot threshold which tells us the server is overloaded hence VMs running are migrated[8] whereas if the utilization of resources is below the level of cold threshold which tells us the server is mostly idle and it turnoff to save energy. If the server is using at least one VMs as running it is actively used as green computing threshold but it becomes high risk in resource demands. In case if we use hotspot have to reduce temperature to low or else there is possibility of reducing skewness.

3.3 Load Prediction Algorithm:

Load prediction algorithm that can capture the future resource usage of applications accurately without looking inside VMs. When load prediction is disabled, the algorithm uses that observed load in its prediction is smaller than that without prediction. The load which is overloaded is reduced by load balancing which balances the heavy load and saves the energy.

IV. SYTEM ARCHITECTURE

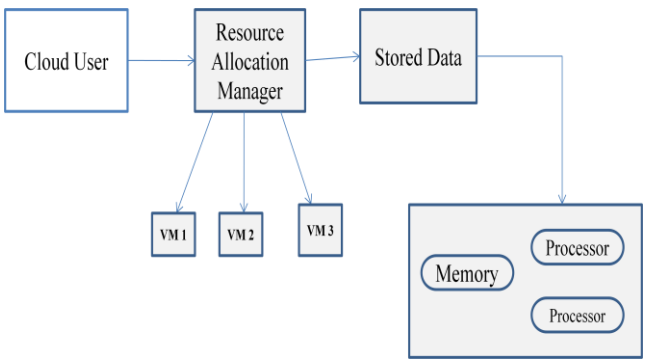

Fig 1. Future Load Prediction

In the Future Load prediction architecture the client will give task to the task scheduler then he sends to the resource manager. Then the resource manager will distribute the tasks to the virtual machines if over load occur in virtual we will migrate that virtual machines.

The scheduler has several components in which the predictor predicts the future resource demands of VMs and the future load of PMs based on the past statistics. The load of a PM by aggregating the resource usage of its own VMs. Thus the details of the load prediction algorithm will also be described in the next paragraph. The LNM at each node first attempts to satisfy the new demands locally, by adjusting the resource allocation of VMs sharing the same VMM. Xen hypervisor can change the CPU allocation among the VMs by adjusting their weights in its CPU scheduler and is responsible for adjusting the memory allocation.

Load balancing for handling the hierarchical and multi dimensional resource constraints in systems but storage is not possible using single dimension as resource constraints are too slow while balancing the load. Two dimension constraint is finding possibility using the virtual machine monitor in a data center network

An energy-aware[10] server provisioning strategy which dynamically turns on/off servers in order to active servers to dynamic user load while ondemand process does not exist in distributed computing environment. Ondemand process is not possible in utility computing which is large scale as pay per use model to be a frame work in process
To determine where a reduce task should run. Given a reduce task the needs to fetch map output from a set of nodes by shuffling time and computation time as map reduce is fetching less in heterogeneous domain network[11][12]. Using map reduce technology a framework to be done on distributed computing so as homogeneous domain also set to be possible.

The client would get implemented by optimizing dynamic resource allocation for scheduling applications in public cloud through large data centres mapping is adaptively done so that PMs get minimized as compared to VMs [13]. Resource management will combine all the loads when partitioned in cloud environment

VM reduces the amount of physical capacity required to support a specified rate of SLA (Service Level Agreement) violations for a given workload by as much as 50% as compared to static consolidation approach[14].SLA region provides layers between physical machines easily. Load in dynamic memory could be done by implementing skewness

V. CONCLUSION

In this paper, we presented that Cloud Computing has several disadvantages which reduce the performance of Overload avoidance and green computing. Cloud computing has emerged as a popular solution to provide cheap and easy access to externalized IT (Information Technology) resources. An increasing number of organizations benefit from cloud computing to host their applications. Through virtualization, cloud computing is able to address with the physical infrastructure a large client base with different computational needs. In order to previous paradigms, Cloud computing is not application oriented but service oriented; it offers on demand virtualization resources as measurable and billable utilities. The system which is implemented by optimizing dynamic resource allocation for scheduling applications in public cloud through large data centres.

ACKNOWLEDGMENT

First and foremost, I would like to thank the God Almighty for showering his blessing throughout our life. I take this chance to express our deep sense of gratitude to our Management, and express our profound thanks to our beloved Professor and Head of the department Dr.E.Poovammal for her able administration and keen interest, which motivated me along the course and Mrs.S.Saranya Assistant Professor for her kind Guidance.

REFERENCES

1. M. Armbrust et al., “Above the clouds: A berkeley view of cloud computing,” University of California, Berkeley, Tech. Rep., Feb 2009.
2. L. Siegele, “Let it rise: A special report on corporate IT,” in The Economist, Oct. 2008.
3. P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho,R. Neugebauer, I. Pratt, and A. Warfield, “Xen and the art of virtualization,” in Proc. of the ACM Symposium on Operating Systems Principles (SOSP’03), Oct. 2003.
4. “Amazon elastic compute cloud (Amazon EC2),http://aws.amazon.com/ec2/.”
5. C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach,I. Pratt, and A. Warfield, “Live migration of virtual machines,” in Proc.of the Symposium on Networked Systems Design and Implementation (NSDI’05), May 2005.
6. M. Nelson, B.-H. Lim, and G. Hutchins, “Fast transparent migration for virtual machines,” in Proc. of the USENIX Annual Technical Conference, 2005.
7. M. McNett, D. Gupta, A. Vahdat, and G. M. Voelker, “Usher: An extensible framework for managing clusters of virtual machines,” in Proc. of the Large Installation System Administration Conference(LISA’07), Nov. 2007.
8. T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, “Black-box and gray-box strategies for virtual machine migration,” in Proc. Of the Symposium on Networked Systems Design and Implementation (NSDI’07), Apr. 2007.
9. C. A. Waldspurger, “Memory resource management in VMware ESX server,” in Proc. of the symposium on Operating systems design and implementation (OSDI’02), Aug. 2002.

10. G. Chen, H. Wenbo, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services," in Proc. of the USENIX Symposium on Networked Systems Design and Implementation (NSDI'08), Apr. 2008.
11. P. Padala, K.-Y. Hou, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, and A. Merchant, "Automated control of multiple virtualized resources," in Proc. of the ACM European conference on Computer systems (EuroSys'09), 2009.
12. N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," in Proc. of the IFIP/IEEE International Symposium on Integrated Network Management (IM'07),2007.
13. "TPC-W: Transaction processing performance council,http://www.tpc.org/tpcw/."
14. J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle,"Managing energy and server resources in hosting centers," in Proc. Of the ACM Symposium on Operating System Principles (SOSP'01), Oct.2001.