ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

Dynamic Resource Allocation by Using Elastic Compute Cloud Service

Kaleeswari [1], Noble Mary Juliet [2]

P.G. Student, Department of Computer Engineering, NPR Engineering College, NPR Nagar, Tamilnadu, India[1]

Professor and Head, Department of Computer Engineering, NPR Engineering College, NPR Nagar, Tamilnadu, India[2]

Abstract: Cloud computing is on demand as it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner to public. In Cloud computing multiple cloud users can request number of cloud services simultaneously. So there must be a provision that all resources are made available to requesting user in efficient manner to satisfy their need. The existing resource allocation mechanism was based on virtualization technique. The problem arises while mapping the virtual machine to physical resources. To address this problem we proposed the new resource allocation scheme based on Elastic Compute Cloud (EC2) service. We are also monitoring the instances happening in EC2 services by using CloudWatch. Hence the proposed system will be more efficient for allocating the resources.

Keywords: Cloud Computing, Dynamic Resource Allocation, Virtual Machine, Virtualization, EC2 Service.

.

I. INTRODUCTION

Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software and information are provided to users over the network. Cloud computing providers deliver application via the Internet, which are accessed from web browser, while the business software and data are stored on servers at a remote location. Cloud computing really is accessing resources and services needed to perform functions with dynamically changing needs. The cloud is a virtualization of resources that maintains and manages itself.

In this system, Elastic Compute Cloud (EC2) service is used for dynamic resource allocation. It is one of the Amazon Web Service (AWS). Amazon Web Services is a collection of remote computing services (also called web services) that together make up a cloud computing platform, offered over the Internet by Amazon.com. The most central and well-known of these services are Amazon EC2 and AmazonS3 (Simple Storage Service). The service is advertised as providing a large computing capacity (potentially many servers) much faster and cheaper than building a physical server farm.

Amazon Elastic Compute Cloud (EC2) is a central part of Amazon.com's cloud computing platform, Amazon Web Services (AWS). EC2 allows users to rent virtual computers on which to run their own computer applications. EC2 allows scalable deployment of applications by providing a web service through which a user can boot an Amazon Machine Image(AMI) to create a virtual machine, which Amazon calls an instance, containing any software desired. A user can create, launch, and terminate server instances as needed, paying by the hour for active servers, hence the term elastic. EC2 provides users with control over the geographical location of instances that allows for latency optimization and high levels of redundancy. Amazon Elastic Compute Cloud is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers. Amazon EC2's simple web service interface allows users to obtain and configure capacity with minimal friction. It provides complete control of computing resources and allows users to run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing users to quickly scale capacity both up and down as users computing requirements change. Amazon EC2 changes the economics of computing by allowing users to pay only for capacity that they actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

Amazon Elastic Compute Cloud provides resizable computing capacity in the Amazon Web Services. Using Amazon EC2 eliminates need of investing the hardware up front, so user can develop and deploy applications faster and can use Amazon EC2 to launch as many or as few virtual servers as their need, configure security and networking and manage storage. Amazon EC2 enables users to scale up or down to handle changes in requirements or spikes in popularity, reducing the need to forecast traffic.

The rest of this paper is organized as follows. In the next section, we discussed the related work. In section III, we presented the proposed system. In section IV, we presented the modules of dynamic resource allocation. Finally, we make a conclusion in section V.

II. RELATED WORK

Resource Management is an important issue in cloud environment. The emerging cloud computing paradigm provides administrators and IT organizations with tremendous freedom to dynamically migrate virtualized computing services between physical servers in cloud data centers. Virtualization and VM migration capabilities enable the data center to consolidate their computing services and use minimal number of physical servers. VM migration offers great benefits such as load balancing, server consolidation, online maintenance and proactive fault tolerance. Cloud computing offers utility-oriented IT services to users worldwide. Based on a pay-as-you-go model, it enables hosting of pervasive applications from consumer, scientific, and business domains. However, data centers hosting Cloud applications consume huge amounts of electrical energy, contributing to high operational costs and carbon footprints to the environment. Therefore, to need Green Cloud computing solutions that can not only minimize operational costs but also reduce the environmental impact. So that to define an architectural framework and principles for energy-efficient cloud computing. Based on this architecture, to present the vision, open research challenges and resource provisioning and allocation algorithms for energy-efficient management of cloud computing environments.Virtual machine monitors like Xen provide a mechanism for mapping virtual machines to physical resources. This mapping is largely hidden from the cloud users.VM live migration technology makes it possible to change the mapping between virtual machines and physical machines while applications are running. The capacity of physical machines can also be heterogeneous because multiple generations of hardware coexist in a data center.

III. PROPOSED SYSTEM

This proposed system consists of number of servers, predictor, hotspot and coldspot solvers and migration list. Set of servers used for running different applications. Predictor is used to execute periodically to evaluate the resource allocation status based on the predicted future demands of virtual machines. The Figure 1 shows the overall proposed system design.

Hotspot solver is used to identify whether the system is in the overloaded or not. If the system is overloaded, the hotspot solver migrate the resources into the alternative machine, freeing the original machine for maintenance. Coldspot solver is used to identify whether the system is in idle state or not. If the system is in idle, potential candidate to turn off to save energy. Migration List maintains the set of migrated resources.

The System that uses virtualization technology to allocate data center resources dynamically based on application demands and support green computing by optimizing the number of servers in use. So that, to introduce the concept of skewness to measure the unevenness in the multidimensional resource utilization of a server. By minimizing skewness, users can combine different types of workloads nicely and improve the overall utilization of server resources. Virtual machine monitors (VMMs) like Xen provide a mechanism for mapping virtual machines (VMs) to physical resources. This mapping is largely hidden from the cloud users. It is up to the cloud provider to make sure the underlying physical machines (PMs) have sufficient resources to meet their needs. VM live migration technology makes it possible to change the mapping between virtual machines and physical machines while applications are running.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

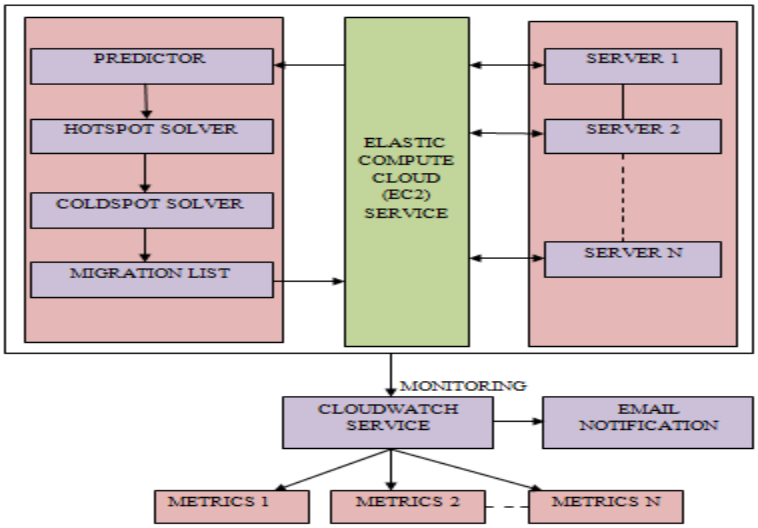

Fig- 1 Architectural design for dynamic resource allocation

To present the design and implementation of an automated resource management system that achieves a good balance between the two goals. The two goals are overload avoidance and green computing. Overload avoidance is the capacity of a physical machine should be sufficient to satisfy the resource needs of all virtual machines running on it. Otherwise, the physical machine is overloaded and can lead to degraded performance of its virtual machines. Green computing is the number of physical machines used should be minimized as long as they can still satisfy the needs of all virtual machines. Idle physical machines can be turned off to save energy.

Resource allocation system is used to avoid overload in the system effectively while minimizing the number of servers used. Skewness is used to measure the uneven utilization of a server. By minimizing skewness, user can improve the overall utilization of servers in the face of multidimensional resource constraints. To design a load prediction algorithm that can capture the future resource usages of applications accurately without looking inside the VMs. The algorithm can capture the rising trend of resource usage patterns and help reduce the placement churn significantly.

IV. IMPLEMENTATION RESULTS

VIRTUAL MACHINE CREATION

Virtualization in computing is the creation of a virtual (rather than actual) Version of something, such as a hardware platform, operating system, and a storage device or network resources. One or more logical computer system run on the one set of physical hardware.VM live migration is a widely used technique for dynamic resource allocation in a virtualized environment. The process of running two or more logical computer system so on one set of physical hardware. Dynamic placement of virtual servers is used to minimize SLA violations. When user creates a virtual machine, a cloud service is automatically created to contain the machine. User can create multiple virtual machines under the same cloud service to enable the virtual machines to communicate with each other, to load-balance between virtual machines and to maintain high availability of the machines. User can manage the availability of their application that uses multiple virtual machines by adding the machines to an availability set. Availability sets are directly related to fault domains and update domains. A fault domain in Windows Azure is defined by avoiding single points of failure, like the network switch or power unit of a rack of servers. In fact, a fault domain is closely equivalent to a rack of physical servers. When multiple virtual machines are connected together in a cloud service, an availability set can be used to ensure that the machines are located in different fault domains.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

RESOURCE ALLOCATION

Dynamic resource management has become an active area of research in the cloud computing paradigm. Cost of resources varies significantly depending on configuration for using them. Hence efficient management of resources is of prime interest to both cloud providers and cloud users. The success of any cloud management software critically depends on the flexibility, scale and efficiency with which it can utilize the underlying hardware resources while providing necessary performance isolation.

Successful resource management solution for cloud environments needs to provide a rich set of resource controls for better isolation, while doing initial placement and load balancing for efficient utilization of resources. For example, in banking application during peak period (11.00 to 1 o'clock) we can access number of servers from the cloud based upon customer needs. During idle time, we can also shutdown the unutilized server to save energy.

SKEWNESS IMPLEMENTATION

Skewness is used to measure the uneven utilization of a server. By minimizing skewness, user can improve the overall utilization of servers in the face of multidimensional resource constraints. In case of ties, to select the VM whose removal can reduce the skewness of the server the most.For each VM in the list, to see if user can find a destination server to accommodate it. The server must not become a hot spot after accepting this VM. Among all such servers, user select one whose skewness can be reduced the most by accepting this VM. All things being equal, to select a destination server whose skewness can be reduced.Skewness algorithm is to mix workloads with different resource requirements together so that the overall utilization of server capacity is improved.

Skewness algorithm executes periodically to evaluate the resource allocation status based on the predicted future resource demands of VMs. User define a server as a hot spot if the utilization of any of its resources is above a hot threshold. This indicates that the server is overloaded and hence some VMs running on it should be migrated away.

To define a server as a cold spot if the utilizations of all its resources are below a cold threshold. This indicates that the server is mostly idle and a potential candidate to turn off in order to save energy. However, user does so only when the average resource utilization of all actively used servers in the system is below a green computing threshold. A server is actively used if it has at least one VM running. Otherwise, it is inactive. Finally, to define the warm threshold to be a level of resource utilization that is sufficiently high to justify having the server running but not as high as to risk becoming a hot spot in the face of temporary fluctuation of application resource demands.

MONITORING EC2 INSTANCES

Multiple virtual machines can be dynamically started and stopped on a single physical machine according to incoming requests, hence providing the flexibility of configuring various partitions of resources on the same physical machine to different requirements of service requests. By dynamically migrating virtual machines across physical machines, workloads can be consolidated and unused resources can be switched to a low-power mode, turned off or configured to operate at low-performance levels in order to save energy. Energy Monitor is used to observe energy consumption caused by virtual machines and physical machines and provides information about energy consumption to the virtual machine manager to make energy-efficient resource allocation decisions.

Amazon CloudWatch is an easy to use web service that provides visibility into our cloud assets. It is designed to provide comprehensive monitoring for all of the AWS services. Amazon CloudWatch provides monitoring for AWS cloud resources and the applications customers run on AWS. The resources are dynamically allocated by using Amazon Elastic Compute Cloud (EC2) service. The entire resource allocation progress can be tracked running Amazon EC2 Instances.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

By using Amazon CloudWatch we can able to monitor the metrics of dynamic resource allocation and also monitor the EC2 Instances. The user can able to monitor the allocation of resources and applications and also monitor the custom matrices generated by a customer's applications and services. Amazon CloudWatch is used for programmatically retrieve user's monitoring data, view graphs and set alarms to troubleshoot and take automated action based on the state of user's cloud environment. Amazon CloudWatch provides a reliable, scalable and flexible monitoring solution that user can start using within minutes. Finally the user can able to allocate their resources dynamically.

V. CONCLUSION

This paper shows that dynamic resource allocation is growing need of cloud providers for more number of users and with the less response time. Hence the on-demand resource allocation based SLA as per defined task priority helps to satisfy the efficient provisioning of cloud resources to multiple cloud users. The entire resource allocation progress can be tracked running Amazon EC2 Instances. By using Amazon CloudWatch we can able to monitor the metrics of dynamic resource allocation and also monitor the EC2 Instances. The user can able to monitor the allocation of resources and applications and also monitor the custom metrics generated by a customer's applications and services. Amazon CloudWatch is used for programmatically retrieve user's monitoring data, view graphs and set alarms to troubleshoot and take automated action based on the state of user's cloud environment. Amazon CloudWatch provides a reliable, scalable and flexible monitoring solution that user can start using within minutes. The monitored alerts can be used to attain the efficient resource allocation results. In future we can able to analyse the resource allocation process by using this monitoring solution. Monitoring data is retained for two weeks, even if user's AWS resources have been terminated.

REFERENCES

[1] Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen proposed a "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment", IEEE Transactions on Parallel and Distributed System, vol.24, No.6, June 2013.

[2] P.Barham,B.Dragovic,K.Fraser,S.Hand,T.Harris,A.Ho,R.Neugebauer,I.Pratt,and A.Warfield Proposed a "Xen and the Art of Virtualization," Proc. ACM Symp.Operating Systems Principles(SOSP'03),Oct.2003.

[3] C. Clark, K. Fraser, S. Hand, J.G. Hansen, E. Jul,C. Limpach, I.Pratt, and A. Warfield Proposed a "Live Migration of Virtual Machines,"Proc. Symp. Networked Systems Design and Implementation (NSDI '05), May 2005.

[4] C.A. Waldspurger Proposed a "Memory Resource Management in Vmware ESX Server," Proc. Symp. Operating Systems Design and Implementation (OSDI '02), Aug. 2002.

[5] G. Chen, H. Wenbo, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao Proposed a "Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services," Proc. USENIX Symp. Networked Systems Design and Implementation (NSDI '08), Apr. 2008.

[6] Jiayin Li, Meikang Qiu, Jian-Wei Niu, Yu Chen, Zhong Ming, "Adaptive Resource Allocation for Pre-empt able Jobs in Cloud Systems," in 10th International Conference on Intelligent System Design and Application, Jan. 2011, pp. 31-36.

[7] Goudarzi H., Pedram M., "Multi-dimensional SLA based Resource Allocation for Multi-tier Cloud Computing Systems," in IEEE International Conference on Cloud Computing, Sep. 2011, pp. 324-331.

[8] Chieu T.C., Mohindra A., Karve A.A., Segal A.,"Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment," in IEEE International Conference on e-Business Engineering, Dec. 2009, pp. 281-286.

[9] Vincent C. Emeakaroha, Ivona Brandic, Michael Maurer, Ivan Breskovic, "SLA-Aware Application Deployment and Resource Allocation in Clouds", 35th IEEE Annual Computer Software and Application Conference Workshops, 2011, pp. 298-303.