ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Skewness: Efficient Dynamic Resource Allocation Using Virtual Machines in Cloud Computing Environment

Raavi Ramesh
M.Tech Student,
Dept of CSE,
TRR College of Engineering, Hyderabad, T.S, India.

Dr. Vaka Murali Mohan
Professor & Head,
Dept of CSE,
TRR College of Engineering, Hyderabad, T.S, India.

Abstract:

Cloud computing is computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid. Cloud computing relies on restricting sharing of resources to achieve coherence and economies of scale, similar to a utility (like the electricity grid) over a network. At the foundation of cloud computing is the broader concept of converged infrastructure and shared services. Cloud computing facilitates Industrial and large business users to increase proportionally and decrease accordingly their data center resource usage depending on requirements. Most of the flaunted benefits in the cloud model originate from resource multiplexing through virtualization technology. In this research paper, we studied and simulated a system that utilizes virtualization technology to distribute data center resources automatically depending on application requirements and maintain green computing by optimizing the number of servers in deployment. We analyzed the concept of "skewness" to determine the unevenness in the multi-dimensional resource utilization of a server. By minimizing skewness, we can merge different types of workloads nicely and develop the overall utilization of server resources. We developed a set of heuristics that avoid overload in the system efficiently at the same time as saving power utilized. Trace driven simulation and experiment results demonstrate that our algorithm achieves good performance.

Keywords:

Cloud computing, Dynamic Resource Allocation, Data center, Virtual machine, Ranking and Load balancing.

INTRODUCTION:

Cloud computing, or in simpler shorthand just "the cloud", also focuses on maximizing the effectiveness of the shared resources.

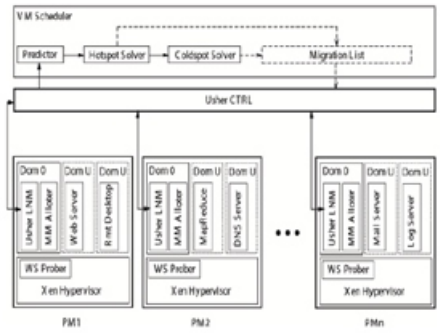
Cloud resources are usually not only shared by multiple users but are also dynamically reallocated per demand. This can work for allocating resources to users. For example, a cloud computer facility that serves European users during European business hours with a specific application (e.g., email) may reallocate the same resources to serve North American users during North America's business hours with a different application (e.g., a web server). This approach should maximize the use of computing power thus reducing environmental damage as well since less power, air conditioning, rack space, etc. are required for a variety of functions. With cloud computing, multiple users can access a single server to retrieve and update their data without purchasing licenses for different applications. The National Institute of Standards and Technology's definition of cloud computing identifies "five essential characteristics": On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Cloud computing is the delivery of computing and storage capacity as a service to a community of end recipients. The name comes from the use of a cloud shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts services with a user's data, software and computation over a network. The remote accessibility enables us to access the cloud services from anywhere at any time. To gain the maximum degree of the above mentioned benefits, the services offered in terms of resources should be allocated optimally to the applications running in the cloud. The elasticity and the lack of upfront capital investment offered by cloud computing is appealing to any businesses. In this paper, we discuss how the cloud service provider can best multiplex the available virtual resources onto the physical hardware. This is important because much of the touted gains in the cloud model come from such multiplexing.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Virtual Machine Monitors (VMMs) like Xen provide a mechanism for mapping Virtual Machines (VMs) to Physical Resources [3]. This mapping is hidden from the cloud users. Users with the Amazon EC2 service [4], for example, do not know where their VM instances run. It is up to the Cloud Service Provider to make sure the underlying Physical Machines (PMs) has sufficient resources to meet their needs VM live migration technology makes it possible to change the mapping between VMs and PMs While applications are running [5], but, a policy issue remains as how to decide the mapping adaptively so that the resource demands of VMs are met while the number of PMs used is minimized.

This is challenging 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. The capacity of PMs can also be heterogeneous because multiple generations of hardware co-exist in a data center. To achieve the overload avoidance that is the capacity of a PM should be sufficient to satisfy the resource needs of all VMs running on it. Otherwise, the PM is overloaded and can lead to degraded performance of its VMs. And also 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 energy.



System Architecture

PROBLEM STATEMENT:

Resource users (cloud users) estimates of resource demands to complete a job before the estimated time may lead to an over provisioning of resources. Resource providers' allocation of resources may lead to an under provisioning of resources. To overcome the above mentioned discrepancies, minuets needed from both cloud providers and users for a RAS. From the cloud user's angle, the application requirement and Service Level Agreement (SLA) are major inputs to RAS.

The offerings, resource status and available resources are the inputs required from the other side to manage and allocate resources to host by RAS. The outcome of any optimal RAS must satisfy the parameters such as throughput, latency and response time. Even though cloud provides reliable resources. It also poses a crucial problem in allocating and managing resources dynamically across the applications.

EXISTING SYSTEM:

Virtual machine monitors (VMMs) like Oxen provide a mechanism for mapping virtual machines (VMs) to physical resources. This mapping is largely hidden from the cloud users. Users with the Amazon EC2 service, for example, do not know where their VM instances run. 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 VMs and PMs While applications are running. The capacity of PMs can also be heterogeneous because multiple generations of hardware coexist in a data center.

Disadvantages:

•A policy issue remains as how to decide the mapping adaptively so that the resource demands of VMs are met while the number of PMs used is minimized.
•This is challenging 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. The two main disadvantages are overload avoidance and green computing.

RELATED WORK:

In [2] author proposed architecture, using feedback control theory, for adaptive management of virtualized resources, which is based on VM. In this VM-based architecture all hardware resources are pooled into common shared space in cloud computing infrastructure so that hosted application can access the required resources as per there need to meet Service Level Objective (SLOs) of application. The adaptive manager use in this architecture is multi-input multi-output (MIMO) resource manager, which includes 3 controllers: CPU controller, memory controller and I/O controller, its goal is regulate multiple virtualized resources utilization to achieve SLOs of application by

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

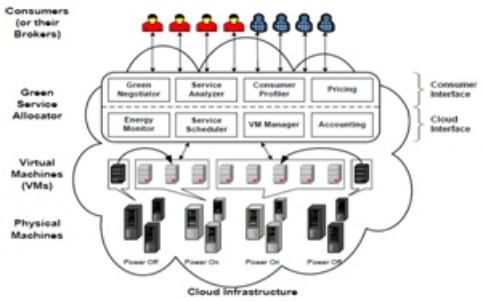
using control inputs per-VM CPU, memory and I/O allocation. The seminal work of Walsh et al. [3], proposed a general two-layer architecture that uses utility functions, adopted in the context of dynamic and autonomous resource allocation, which consists of local agents and global arbiter. The responsibility of local agents is to calculate utilities, for given current or forecasted workload and range of resources, for each AE and results are transfer to global arbiter. Where, global arbiter computes near-optimal configuration of resources based on the results provided by the local agents. In [4], authors propose an adaptive resource allocation algorithm for the cloud system with preempt able tasks in which algorithms adjust the resource allocation adaptively based on the updated of the actual task executions. Adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) algorithms are use for task scheduling which includes static task scheduling, for static resource allocation, is generated offline. The online adaptive procedure is use for re-evaluating the remaining static resource allocation repeatedly with predefined frequency.

The dynamic resource allocation based on distributed multiple criteria decisions in computing cloud explain in [6]. In it author contribution is tow-fold, first distributed architecture is adopted, in which resource management is divided into independent tasks, each of which is performed by Autonomous Node Agents (NA) in ac cycle of three activities: (1) VMPlacement, in it suitable physical machine (PM) is found which is capable of running given VM and then assigned VM to that PM, (2) Monitoring, in its total resources use by hosted VM are monitored by NA, (3) In VM selection, if local accommodation is not possible, a VM need to migrate at another PM and process loops back to into placement. And second, using PROMETHEE method, NA carry out configuration in parallel through multiple criteria decision analysis. This approach is potentially more feasible in large data centers than centralized approaches.

PROPOSED METHODOLOGY:

In this paper, we 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. running on it. Secondly, the PM is overloaded and can lead to degraded performance of its VMs. Overload avoidance: The capacity of a PM should be sufficient to satisfy the resource needs of all VMs

Green computing: The number of PMs should be minimized as long as they can satisfy the needs of all VMs. Idle PMs can be switch off to save energy. We develop a resource allocation system that can avoid overload in the system effectively while minimizing the number of servers used. We introduce the concept of "skewness" to measure the uneven utilization of a servers By minimized skewness we can increase the overall utilization of servers interface of multidimensional resource constraints. We propose a load prediction algorithm that can capture the future resource usages of applications accurately without looking the VMs. The algorithm captures the rising trends of resource usage patterns and helps reduce the placement churn significantly. The cloud computing is a model which enables on demand network access to a shared pool computing resources. Cloud computing environment consists of multiple customers requesting for resources in a dynamic environment with their many possible constraints. In existing system cloud computing allocating the resource efficient is a challenging task. In this paper we proposed allocates resource with less wastage and provides much profit. The developed resource allocation algorithm is based on different parameters as: time, cost, No of processor, request etc. Priority model that mainly decides priority among different user request based on many parameters like cost of resource, time needed to access, task type, number of processors needed to run the job or task. In this model client send the request to the cloud server.



The cloud service provider runs the task submitted by the client. The cloud admin decides the priority among the different users request. Each request consists of different tasks. It have the different parameters such as Time-computation, time needed to complete the particular task, Processor request refers to number of processors needed to run the task. The more number of processor faster the completion of task importance-refers to how important the user to a cloud administrator (admin) that is whether the user is old customer to cloud or new customer. Price-refers to cost charged by cloud admin to cloud users.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Cloud computing is a model which enables on demand network access to a shared pool computing resources. A cloud environment consists of multiple customers requesting for resources in a dynamic environment with possible constraints. In existing system cloud computing, allocating the resource usually is a challenging job. The cloud does not show the quality of services.

Advantage:

•We develop a resource allocation system that can avoid overload in the system effectively while minimizing the number of servers used.
•We introduce the concept of "skewness" to measure the uneven utilization of a server. By minimizing skewness, we can improve the overall utilization of servers in the face of multi-dimensional resource constraints.

IMPLEMENTATION
Cloud Computing:

Cloud computing refers to applications and services offered over the Internet. These services are offered from data centers all over the world, which collectively are referred to as the "cloud." Cloud computing is a movement away from applications needing to be installed on an individual's computer towards the applications being hosted online. Cloud resources are usually not only shared by multiple users but as well as dynamically re-allocated as per demand. This can work for allocating resources to users in different time zones.

Skewness:

We use the skewness metric to combine VMs with different resource characteristics appropriately so that the capacities of servers are well utilized. Our algorithm achieves both overload avoidance and green computing for systems with multi-resource constraints. We break down the decision time into two parts: hot spot mitigation (marked as 'hot') and green computing (marked as 'cold'). We find that hot spot Mitigation contributes more to the decision time. We also find that the decision time for the synthetic workload is higher than that for the real trace due to the large variation in the synthetic workload. This is also verified by the right figure which computes the average number of migrations per VM in each decision. The figure indicates that each VM experiences a tiny, roughly constant number of migrations during a decision run, independent of the system size. We use the skewness metric to combine VMs with different resource characteristics appropriately so that the capacities of servers are well utilized. Our algorithm achieves both overload avoidance and green computing for systems with multi-resource constraints.

Resource Management:

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 de-pends on the flexibility; scale and efficiency with which it can utilize the underlying hardware resources while pro-viding 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 underlying resources.

Virtualization:

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.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 to minimize SLA violations.

Green Computing:

Many efforts have been made to curtail energy consumption. Hardware based approaches include novel thermal design for lower cooling power, or adopting power-proportional and low-power hardware. Dynamic Voltage and Frequency Scaling (DVFS) to adjust CPU power according to its load in data centers. Our work belongs to the category of pure-software low-cost Solutions. It requires that the desktop is virtualized with shared storage. Green computing ensures end user satisfaction, regulatory compliance, telecommuting, virtualization of server resources.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

RESULTS AND DISCUSSION:

The goal of the skewness algorithm is to mix workloads with different resource requirements together so that the overall utilization of server capacity is improved. In this experiment, we see how our algorithm handles a mix of CPU, memory, and network intensive workloads. Resource allocation status of three servers A, B, C has total memory allocated 500KB and resource used memory for server A 80KB,serverB 170KB and server C 80K. In Fig. 4 each cloud users provide cloud service Resource allocation in green computing. In Fig.5 display Server usage and resource allocation status for user1 using Bar Chart. The cloud computing is a model which enables on demand network access to a shared pool computing resources. Cloud computing environment consists of multiple customers requesting for resources in a dynamic environment with their many possible constraints. The virtualization can be the solution for it. It can be used to reduce power consumption by data centers. The main purpose of the virtualization is that to make the most efficient use of available system resources, including energy.

A data center, installing virtual infrastructure allows several operating systems and applications to run on a lesser number of servers, it can help to reduce the overall energy used for the data center and the energy consumed for its cooling. Once the number of servers is reduced, it also means that data center can reduce the building size as well. Some of the advantages of Virtualization which directly impacts efficiency and contributes to the environment include: Workload balancing across servers, Resource allocation and sharing are better monitored and managed and the Server utilization rates can be increased up to 80% as compared to initial 10-15%.The results are clear and having good contribution: 1) Allocation of resource is done dynamically. 2) Saves the energy using the green computing concept 3) Proper utilization of servers and memory utilization is taken care using skewness. 4) Minimize the total cost of both the cloud computing infrastructure and running application.

CONCLUSION:

This paper addresses the theoretic study of various dynamic resource allocation techniques in cloud computing environment. Description of the techniques is summarized the advantages with parameters of the various techniques in cloud computing environment.

The cloud computing allows business customers to scale up and down their resource usage based on need. Many of the gains in the cloud model come from resource multiplexing through virtualization technology. In this paper we propose a system that uses virtualization technology to allocate data center resources dynamically based on application needs and support green computing by optimizing the number of servers in use. We proposed the concept of "skewness" to measure the un-evenness in the multidimensional resource utilization of a server. By minimized skewness, we can combining different of workloads and improve the over-all utilization of server resources. We develop a set of heuristics that prevent overload in the system effectively while saving energy used. Trace driven simulations and experimental results demonstrate that ours algorithm achieves good performance.

REFERENCES:

[1] M. Armrest et al., "Above the Clouds: A Berkeley View of Cloud Computing," technical report, Univ. of California, Berkeley, Feb. 2009.

[2] L. Siegel, "Let It Rise: A Special Report on Corporate IT," The Economist, vol. 389, pp. 3-16, Oct. 2008.

[3] P. Braham, B. Draconic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Oxen and the Art of Virtualization," Proc. ACM Sump. Operating Systems Principles (SOSP '03), Oct. 2003.

[4] "Amazon elastic compute cloud (Amazon EC2)," http://aws. amazon.com/ec2/, 2012.

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

[6] M. Nelson, B.-H. Lim, and G. Hutchins, "Fast Transparent Migration for Virtual Machines," Proc. USENIX Ann. Technical Conf., 2005.

[7] M. Monett, D. Gupta, A. Vanda, and G.M. Volker, "Usher: An Extensible Framework for Managing Clusters of Virtual Machines," Proc. Large Installation System Administration Conf. (LISA '07 T. Wood, P. Shiny, A. Venkataramani, and M. Yousif, "Black-Box and Gray-Box Strategies for Virtual Machine Migration,"

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Proc. Sump. Networked Systems Design and Implementation (NSDI ’07), Apr. 2007.

[8] C.A. Waldspurger, “Memory Resource Management in VMware ESX Server,” Proc. Sump. Operating Systems Design and Implementation (OSDI ’02), Aug. 2002.

[9] G. Chen, H. Wino, J. Liu, S. Nat, L. Ragas, L. Xiao, and F. Zhao, “Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services,” Proc. USENIX Sump. Networked Systems Design and Implementation (NSDI ’08), Apr.

[10] P. Padilla, K.-Y. Hour, K.G. Shin, X. Zhu, M. Usual, Z. Wang, S. Signal, and A. Merchant, “Automated Control of Multiple Virtualized Resources,” Proc. ACM European conf. Computer Systems (Neurosis ’09), 2009.

[11] N. Borough, A. Kocher, and K. Beatty, “Dynamic Placement of Virtual Machines for Managing SLA Violations,” Proc. IFIP/IEEEInt’l Sump. Integrated Network Management (IM ’07), 2007.

[12] “TPC-W: Transaction Processing Performance Council,” http:// www.tpc.org/tpcw/, 2012.

[13] J.S. Chase, D.C. Anderson, P.N. Taker, A.M. Vanda, and R.P. Doyle, “Managing Energy and Server Resources in Hosting Centers,” Proc. ACM Sump. Operating System Principles (SOSP ’01), Oct. 2001.

[14] C. Tang, M. Steindler, M. Spreader, and G. Pacifica, “A Scalable Application Placement Controller for Enterprise Data Centers,” Proc. Int’l World Wide Web Conf. (WWW ’07), May 2007.

[15] M. Zaharias, A. Kaminski, A.D. Joseph, R.H. Katz, and I. Stoical, “Improving Map Reduce Performance in Heterogeneous Environments,” Proc. Sump. Operating Systems Design and Implementation (OSDI ’08), 2008.

[16] M. Izard, V. Prabhakaran, J. Curry, U. Wielder, K. Stalwart, and A. Goldberg, “Quincy: Fair Scheduling for Distributed Computing Clusters,” Proc. ACM Sump. Operating System Principles (SOSP ’09), Oct. 2009.

[17] M. Zaharias, D. Borthakur, J. Sen. Sara, K. Elmeleegy, S. Shankar, and I. Stoical, “Delay Scheduling: A Simple Technique for Achieving Locality and Fairness in Cluster Scheduling,” Proc. European Conf. Computer Systems (Neurosis ’10), 2010.

[18] T. Sandworm and K. Lai, “Map reduce Optimization Using Regulated Dynamic Prioritization,” Proc. Int’l Joint Conf. Measurement and Modeling of Computer Systems (SIGMETRICS)