ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

"Survey on Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment"

Amol Selokar[1], Prof.S.D.Zade[2], Prof.C.U.Chavan[3]

Student, M.Tech (Computer Science & Engineering), PIET, Nagpur, India[1]

Assistant Professor, Computer Science & Engineering Department, PIET, Nagpur, India[2,3]

Abstract: The emergence of cloud computing infrastructures brings new ways to build and manage computing system with the flexibility offer by virtualization technologies. In this context, this focuses on two principal objective First leveraging virtualization and cloud computing infrastructures to build distributed large scale computing platforms from multiple cloud providers allowed to run software requiring large amounts of computation power. Secondly developing mechanisms to make these infrastructures more dynamic. This mechanism provides inter cloud live migration offing new ways to exploit the inherent dynamic nature of distributed clouds. Cloud computing allows business customers to scale up and down their resource usage based on needs. Many of the gains in the cloud model come from resource multiplexing through virtualization technology. In this paper we proposed 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. We introduce the concept of "skewness" to measure the unevenness in the multi-dimensional resource utilization of a server. By minimizing skewness, we can add different types of workloads nicely and improve the overall utilization of server resource. We present a set of heuristics that prevent overload in the system effectively while saved energy used. Trace driven simulation and experimental results demonstrate that our algorithm achieves good performance.

I. INTRODUCTION

The elasticity and the lack of upfront capital investment offered by cloud computing is appealing to fast businesses. There is a lot of discussion on the benefits and costs of the cloud model and on how to move legacy applications onto the cloud computing platform. In this paper we study a different problem how a cloud service provider best can multiplex its virtual resources onto the physical hardware. This is important because much of the lack gains in the loud model come from such multiplexing, Studies have found that servers in huge existing data centers are often severely under -utilized due to over pro-visioning for the demand. The cloud model is expected to make such practice unnecessary by offering automatic scale up and down in response to load variation reducing the hardware cost. This method also saves on electricity which contributes to a significant portion of the operational expenses in large data centers. Virtual machine monitors (VMMs) like Xen provide a mechanism for mapping virtual machines (VMs) to resources. This mapping is much hidden from the cloud users. Users with the Amazon EC2 services for example do not know where their VM instant run. It is up to the cloud provider to make sure the underlying physical machines (PMs) have sufficient resources to their needs. VM live migration technologies makes it possible to change the mapping between VMs and PMs while applications are running. It is 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 minimize. This is challenging the resources Needs of VMs are heterogeneous due to the diverse set of applications they run and vary with time as the workloads expand and shrink. The capacity of PMs can also be heterogeneous because multiple generations of hardware co-exist in a data center.

We are trying to achieve two goals in our algorithm. Overload avoidance: The capacity of a PM should be sufficient to satisfy the resource needs of all VMs working on it. Otherwise the PM is overloaded and can lead to degraded performance of its VMs.Cloud Computing become a de facto standard for computing, infrastructure as a services has been emerged as an important paradigm in IT area. By applying this paradigm we can abstract the underlying physical resource such a CPUs, Memories and Storage and offer this Virtual Resource to users in the formal Virtual Machine. Multiple Virtual Machines are able to run on a unique physical machine. Multiple VMs are able to run on a unique Physical Machine (PM). Another important issues in Cloud computing is provisioning method for allocating resources to cloud consumers. Cloud computing environment consists of two provision. The goal is to achieve an optimal solution for provisioning resource which is the most critical part in cloud computing. To make an optimal decision the demand price and waiting-time uncertainties are taken into account to adjust the trade-offs between on-demand and oversubscribed costs.

The Bender's Decomposition is applied to divide the resource optimization problem into many sub problems to decrease the on demand cost and Reservation Cost. Scenario Reduction Technique is applied to reduce problem by reducing number of Scenarios. It will decrease Reservation cost and Expending cost.

In the meantime, the advent of multi-cores has en-abled the sharing of micro-architectural resources such as shared caches and memory controllers. Contention on such micro-architectural resources has emerged as a major reason for performance variance, as an application can be affected by co-running applications even though it receives the same share of CPU, memory, and I/O. For a single system, there have been several prior studies to mitigate the impact of contention on shared caches and memory controllers by carefully scheduling threads. The prior studies rely on the heterogeneity of memory behaviors of applications within a system boundary. The techniques group applications to share a cache to minimize the overall cache misses for a system. However, if a single system runs applications with similar cache behaviors, such intra-system scheduling cannot mitigate contentions.

However, cloud systems with virtualization open a new opportunity to widen the scope of contention-aware scheduling, as virtual machines can cross legacy system boundaries with live migration. In this paper, we use live VM migration to dynamically schedule VMs for minimizing the contention on shared caches and memory controllers. Furthermore, this study considers the effects of non-uniform memory accesses (NUMA) in multi-socket systems commonly used in cloud servers.

We propose contention-aware cloud scheduling techniques for cache sharing and NUMA affinity. The techniques identify the cache behaviors of VMs on-line, and dynamically migrate VMs, if the current placements of VMs are causing excessive shared cache conflicts or wrong NUMA affinity. Since the techniques identify the VM behaviors dynamically and resolve conflicts with live migration it will not required any prior knowledge on the behaviors of VMs. The first technique, cache-aware cloud scheduling minimizes the overall last-level cache (LLC) misses in a cloud system. The second technique, NUMA-aware cloud scheduling extends the first technique by considering NUMA affinity.

We evaluate our proposed schedulers using selected Speculum 2006 applications in various combinations. The experimental results show that the cache-aware scheduler can significantly improve the performance compared with the worst case. With our preliminary NUMA optimization, the performance is slightly improved for our benchmark applications, compared with that of the cache-aware scheduler. Green computing: The number of PMs used should be minimized as long as they can still satisfy the needs of all VMs and Idle PMs can be switch off to save energy.

There is an inherent tradeoff between the two goals in the face of changing resource needs of VMs, overload avoidance. We should keep the utilization of PMs low to reduce the possibility of overload in case the resource needs of VMs increasing later, For green computing we should keep the utilization of PMs reasonably high to make efficient use of their energy.

II. RELATED WORKS

Cloud Resources Provisioning scheme, which is flexible enough to adopt to the various general MCC reference use cases being described. The main feature of the employed MCC Services Admission Control algorithm lies in the fact that it jointly handles radio and computing resources rather than confronting the problem as two independent resource management sub-problems The queuing model to optimize the resource allocation for multimedia cloud in priority services scheme**.** which Specifically formulate and solve the resource cost minimization problem and the service response time minimization problems. An optimal cloud resource provision (OCRP) algorithm is proposed by formulating a stochastic programming model, The (OCRP) algorithm can provisioning computing resources for being used in multiple provisioning stages as well as a long term plan, for example four stages in a quarter plan and twelve stages in a annual plan. The demand and price uncertainty is considered in OCRP. In this paper we purposed that different approaches to obtain the solution of the OCRP algorithm are considered including deterministic equivalent formulation, average approximation, and Benders decomposition. Numerical studies are extensively performing in which the results clearly show that with the OCRP algorithm cloud consumer can successfully minimized the average cost of resource provisioning in cloud computing environment. The OCRP algorithm shows that we can find an optimal solution for resource provisioning and VM placement. It uses only two un-certainties only the need and price. In this paper RCRP algorithm is used which is an extension of OCRP where four uncertainty factors r considered. Grid providing services which are not of desired quality. One of the major uncertain factors of grid is single point of failure where one unit on the grid de-grades which will cause the entire system to de-grade. Hence suggests cloud which is used for adaption of many services. The benefit of cloud is that it will prevent single point of failure and also will decrease hardware costing.

Resource allocation strategies (RAS) at a glance The input parameters to RAS and the way of resource allocation vary based on the services and infrastructure and the nature of applications which will demand resources. The schematic diagram depicts the classification of Resource Allocation Strategies (RAS) proposed in cloud paradigm. The following section discusses the RAS employed in cloud.

A. Execution Time

Different kinds of resource allocation mechanisms are proposed in cloud. The actual task execution time and pre-emptible scheduling is considered for various resource allocations. It overviews the problem of resource contention and increases resource utilization by using different modes of renting computing capacities. But estimating the execution time for a job is a hard task for a user and errors are made very often. But the VM model considered in is heterogeneous and proposed for IaaS.

B. Policy

Since centralized user and resource management lacks in scalable management of users, resources and organization-level security policy, we proposed a decentralized user and virtualized resource management for IaaS by adding a new layer called domain in between the user and the virtualized resources. Based on role based access control (RBAC), virtualized resources are allocated to users through domain layer.

C. Virtual Machine (VM)

A system which can automatically scale it's infrastructure resources is designed the system composed of a virtual network of virtual machines capable of live migration across multi- domain physical infrastructure Cloud computing services providers deliver their resources based on virtualization to satisfy the need of users. In cloud computing, the amount of resources required can vary preserve request. Therefore the providers have to offer Different amounts of virtualized resources per request. To provide worldwide service, a provider may have data centers that are geographically distributed through-out the world. The user locations vary in geographical locations. Since cloud computing services are delivered over the internet there may be un-desirable response latency between the users and the database. Hence, for the best recent service, the provider needs to find a data center and physical machine that has a light workload and is geographically close to the users. The proposed model finds the best match for the user requests based on two evaluations:

The geo-graphical distances between the user and database center and the workload of data center this the model allows the users to find a data center that is guaranteed to be the closest distance and have the light workload and it finds a light workload physical machine within the data center for a provider.

III. 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.

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

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.

IV. 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

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

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

[2] 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," Proc. ACM European conf. Computer Systems (EuroSys '09), 2009.

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

[4] C.A. Waldspurger, "Memory Resource Management in VMware ESX Server," Proc. Symp. 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," Proc. USENIX Symp. Networked Systems Design and Implementation (NSDI '08), Apr. 2008.

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

[6] M. McNett, D. Gupta, A. Vahdat, and G.M. Voelker, "Usher: An Extensible Framework for Managing Clusters of Virtual Machines," Proc. Large Installation System Administration Conf. (LISA '07), Nov. 2007.

[7] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Black-Box and Gray-Box Strategies for Virtual Machine Migration," Proc. Symp. Networked Systems Design and Implementation (NSDI '07), Apr. 2007.

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

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

[10] 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," Proc. ACM Symp. Operating Systems principles (SOSP '03), Oct. 2003.