ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

An Approach for Efficient Dynamic Memory Allocation Using Skewness Algorithm in Cloud Environment

Ramesh C[1],Suba Shalini S[2], Suganya S[3], Suvitha N[4]

Assistant Professor, Dept. of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, India[1]

B.E Students, Dept. of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, India[2,3,4.]

ABSTRACT: Cloud computing is an innovative technology which provide service on demand.It offers dynamic resource allocation for reliable service in "pay-as-you-use" or subscription basis to cloud service users. In resource management, dynamic resource allocation becomes a major challenge in satisfying the cloud users. Resources must be allocated efficiently so that it benefits both the cloud users and service providers, hence proper approach is needed for efficient allocation of resource. In the proposed system, the skewness algorithm performs efficient dynamic resource allocation and it has the capability to measure the unevenness in the resource utilization of multiple virtual machines. Based on application demand it allocates data center resources dynamically by using virtualization technology. It also supports green computing by minimizing the number of servers in use. The overall utilization of server resource can be minimized by reducing the skewness. This algorithm avoids server failure due to overload by effective load prediction and load balancing.

KEYWORDS: Cloud computing, Dynamic Resource Allocation, Virtualization, Skewness Algorithm, Hotspot, Cold spot, load balancing

I. Introduction

Cloud computing is an on-demand service and provides services to the customer via internet. One can manipulate and configure the application on online at any time. It does not require installing a specific piece of software to access or manipulating cloud application. The individuals and businesses can use software and hardware that are handled by third parties at remote locations. It allows customers to access the information and computer resources from anywhere where the network connection is available.Resource pooling is achieved in cloud in which it allows multiple tenants to share a pool of resources on pay as you use basis. One can share single hardware, database and basic infrastructure.Rapid elasticity is possible because it is really easy to scale up or down the resources at any time.Resources used by the customers or resources currently assigned to customers are monitored automatically.

Cloud computing allows customers to scale up and down their resources based on needs. It makes the resources as a single point of access to the client and cost is pay per usage. Cloud computing is an emanating technology, where a lake of resources are connected in both private and public networks for provisioning dynamically scalable infrastructure for application.Cloud computing is not application oriented but this is a service oriented. Cloud offers the virtualized resources to the cloud users.It also provides dynamic provisioning and can allocate machines to store data and add or remove the machines according to the workload demands. Cloud computing platforms are provided by Microsoft, Amazon, Google, IBM. Cloud computing is an environment used for sharing resources beyond the knowledge of the infrastructure and can makes it possible to access the applications and its associated data from anywhere at any time[7]. Cloud computing four types of cloud:

- Public cloud
- Private cloud
- Hybrid cloud
- Community cloud

DOI: 10.15680/IJIRCCE.2016. 0402156

A.RESOURCE ALLOCATION

Allocating resources efficiently in cloud has been an attraction for researchers, as it provides an illusion of infinite computing resources to cloud users so that they can increasing or decreasing the resource utilization rate according to the requirement[4]. Resource allocation is the process of assigning available resources to the needed cloud applications. The resource requirement in cloud environment is unpredictable and time variant. Thus, it is important to manage the availability of resources and scheduling all its resources time to time. Various resource allocation policies as explored in are: all or nothing resource allocation approach, partial allocation with blocking, partial allocation with waiting and queue based allocation/scheduling. When the cloud provider has a number of clients, but limited number of resources to serve the requests from those clients, dynamic resource allocation is drastically needed.

B. DYNAMIC RESOURCE ALLOCATION

Cloud computing provides dynamic memory allocation, it issues machines to store data and add or remove the machines depending upon workload. Dynamic resource allocation, which has been extensively used in internet hosting platforms, has examine to be useful in handling multiple time-scale workloads. Unlike other set of resources like VMs are flexibly deployed on physical machines can be automatically generated for various applications. Though traditional physical resources provisioning has been used over the past and over- provisioning or under-provisioning is the problem for most cloud users[10].

In cloud, there are many tasks have been need to be run by the resources to achieve the minimal usage of servers thus reducing the migration of virtual machines, Effective utilization of resources. Due to these different intentions, we need to develop and propose resource allocation algorithms that is used to perform appropriate allocation map of tasks on resources. Though, various resource allocation algorithms have been implemented by researchers most significant algorithms are Skewness, Black-Gray, Vector-dot, etc [11].

In cloud the allocation of resources depends on the infrastructure as a service. In cloud environment, resource allocation happens at two levels namely,

- When an application is transferred to the cloud, the load balancer allocates the needed resources to the PM's to balance the computational load of different applications across physical computers.

- When an application gets multiple incoming requests, these requests should be allocated to a specific application system to balance the computational load across a set of resources of the same application.

Resource allocation must satisfy the following criteria:

- Resource conflict arises at the time when two applications try to access the same resource at the same time

- Resource fragmentation occurs when the resources are isolated and there would be enough resources but cannot allocate it to the required application due to fragmentation.

- Lack of resources occurs when there are fixed number of resources and also the demand for resources is very high

- The multiple applications requires various types of resources such as CPU and I/O devices should able to satisfy that request.

Over provisioning of resources occurs when the application gets excess resources than the demanded one.

DOI: 10.15680/IJIRCCE.2016. 0402156

II. Related work

Cho-Li Wang et al. [1] designed a fully distributed, multiplexing resource allocation scheme to maintain decentralized resource by integrating volunteer computing into organizing cloud. In the self-organizing each participant can act as consumer and provider, which can operate only light weight process under the randomized policy using content addressable network. SOC,can improve performance of process and fault isolation. The VM-multiplexing resource allocation technique to maintain decentralized resources. This dynamic resource allocation scheme can improve resource utilization and minimize the user usage cost. This method has a complex implementation and the difficulty in re-allocation of physical resources.

K C Gouda et al.[2] the method is priority based resource allocation approaches that it provides flexibility to reorder the request sand dynamicity for allocating resources, even though when there is limited number of resources available. The priority based resource allocation method is used for handling the situation where two or more requests at a same time have same priority. Load and threshold are the two parameters that helped to priorities the requests that already have same priority at a prior level. The problem is that the request in waiting state, with its required load exceeding the threshold, set for the system, needs to have a mechanism to migrate to another system in cloud environment.

Zhen Xiao et al. [3] the allocation of data centers dynamically based on the demand and support green computing by optimizing number of servers in use. These system groups virtual to physical resources can be modified based on the changing demand of cloud users. The use of skewness metric to combining the Virtual Machines with distinct resource characteristics so that the capacities of servers are highly utilized.

Ts`epo Mofolo et al. [6] the algorithm that will keep the migration time minimum thus reducing the number of migrations. This will play a dominant role in ignoring the performance degradation encountered by a migrating VM. Modified best fit decreasing algorithm contributing nearly less performance degradation that occurs due to VM migrations. This algorithm performs host selection and VM reallocation quicker than the previous algorithms. The maintenance of physical servers can be effectively achieved through this algorithm as it tends to efficient consolidation of VMs.

Namita R.Jainet al. [8] the method on related to optimistic resources allocation in cloud environment .The resource allocation are fulfilling customer demands, data center management and application requirements.

III. Existing system

A. Efficient Idle Desktop Consolidation with Partial VM Migration

In cloud computing, allocating the resources efficiently is a demanding job. For each single request an extra message is generated and the demand of each virtual machine changes at run time. It is difficult for the data center to allocate resources dynamically to each spot to enlarge the cloud provider's total revenue. The Idle desktop systems are always power on because applications needs to maintain the network presence.

The technique used ballooning method which does not properly ensures the quick resume and provides the strain to the network. The cloud is naturally dynamic and the usage of resources changes dynamically. Thus any system that tries to implement a SLA need to enfold this dynamic nature. Due to the escalade concerns of privacy and data security, consumers may show hesitation to disclose some details to the cloud providers. Cloud services are focused to load inconsistency and SLA violations that are more likely to happen during the transitions. Due to the nature of these fluctuations are undecidable and hence a constant schedule for evaluating conditions may not be sufficient [5].

B. Dynamic Optimization of Multi Attribute Resource Allocation in Self Organizing Clouds

In the Self Organizing Cloud environment, it provides maximum resource utilization but there are two major issues, One is to locate qualified node to satisfy the user's task resource and the other one is to minimize a job execution time by determining the maximum shares of the different attributes of the resources to allocate the job with variety of QoS constraints such as the expected execution time. Priority algorithm that mainly decides priority among the different types of user request based on many parameters like cost, time required to access the resource, the task type and the number of processors needed to run the job or task. The main disadvantage of this method is that low priority jobs should delay the execution of high priority job.

C. Heuristic Based Resource Allocation Using Virtual Machine Migration

The partial migration has two conditions, one is working set of an idle VM is smaller than total VM memory allocation and second is to wait until all state been transferred to the server before going to sleep condition. In this method Virtualization and VM migration allow the data centre resource allocation services and to use minimum number of physical machine. In the above works there is a dilemma in SLA violation which it is not received thorough the inquiry phase and this algorithm will keep the migration in a very less time and reduce the number of migrations. The heuristic based VM migration based on the following: One is finding when a physical server to be overloaded or under loaded and selection of VMs that can be migrated from an overloaded server to under loaded server. These are called as live migration of VMs which conserves free resources to reduce SLA violation and thus reducing the utilization by less number of migration with efficient live migration of VM migration of VMs which conserve free resources to prevent SLA violation thus reducing the utilization by less number of migration with proficient live migration of VM.

IV. PROPOSED SYSTEM

A cloud environment usually contains a huge number of machines that are connected by a high-speed network. The users connect to those sites hosted by the cloud surroundings through the public internet. The site is usually accessed through a Unified Resource Locator that is translated to a network address through a overall directory service such as DNS. A request to the cloud site is routed through the internet to a machine inside the data center that either processes the request or forwards it.

Clouds are a group of shared environments where the group of cloud users utilize the same equipment and can request number of cloud services simultaneously from the cloud service provider. There must be a provision that all resources are made accessible to requesting user in efficient manner to satisfy their need. Cloud networks are shared in a way making it tough for both the cloud users and operators to reason about how network resources are allocated.

In the existing method, for each single request an extra message is generated and it is difficult for the data center to allocate resources dynamically to each spot, it is found to be time consuming. A policy issue remains as how to decide the flexible mapping so that the demands for the cloud resources of VMs are met while the number of PMs used is minimized. This is the most effective way when the need for resources of virtual machines are dissimilar because of the disparate set of applications they run and vary with time as the workloads grow and shrink. In this proposed method, we present the design and implementation of an automated resource management system that manages a good balance in overload avoidance [12].

The main aim is to develop a resource allocation system that can prevents overload in the system by shortening the number of servers used. The skewness method is to measure the unevenness in the usage of a server. By lessening skewness, we can improve the overall utilization of servers in the multifaceted resource restraints. The improvised algorithm can capture the future resource usages of applications absolutely without looking inside the VMs and can capture the rising trend of resource usage patterns thus shortening the placement churn significantly[9].

The Dynamic resource allocation system uses virtual machines to allocate the resources based on the application demands and the system supports green computing and avoids overload by reducing the number of servers used. The “skewness” concept is to determine the unevenness in the multiple resource utilization of a server and to avoid overload by adding different workloads thus optimizing the utilization of server resources. For the effective load distribution we need to use adapt load balancer and the load prediction algorithm predicts the future load and prevents overload in the system. In live VM migration we can save the energy used and achieve good performance. The

Dynamic resource allocation system architecture shown below in which we have single PM with VMs and PM is treated as server which provide the required resources to the client to perform some task.

In the resource management system we implement resource allocation algorithm which dynamically allocates resource for the cloud user in server. For the effective resource allocation, adapt load balancer is used in the algorithm which makes decision on the basis of skew value of all VMs and predicts the future load on the server. To forecast the future resource needs of VMs. As said earlier, the goal is on Internet applications. One way is to look inside a VM for application level statistics, e.g., to parse the logs of pending requests. By doing so it requires modification of the VM which may not always be possible. Instead, one method is to make our prediction based on the previous external characteristic of VMs. One method is to calculate an Exponentially Weighted Moving Average (EWMA) by use of a TCP-like scheme. The EWMA formula is used to predict the CPU load on the DNS server. We measure the load every minute and predict the load in the next minute.

Skewness Algorithm

Skewness is used to resolve the unevenness by combining different types of load and improves the utilization of multiple resources on a server. If a PM runs many memory-intensive VMs with low CPU load and due to which lot of CPU resources will be wasted because we have limited memory for an extra VM. [13]

$$\text{Skewness (p)} = \sqrt{\sum_{i=1}^{n}(\frac{r_i}{\bar{r}} - \overline{1)}^{2}}$$

Where,

n - number of resource

r_i - usage of the i-th resource.

R -the average utilization of all resources for server p.

By reducing the skewness, we can combine the different workloads effectively and improve the overall utilization of servers. Skewness can be measured by Hot spot and Cold spot.

- Hot spot: If consumption of any resources is above the hot threshold, it indicates that the server is overloaded and some VMs running on it should be migrated.

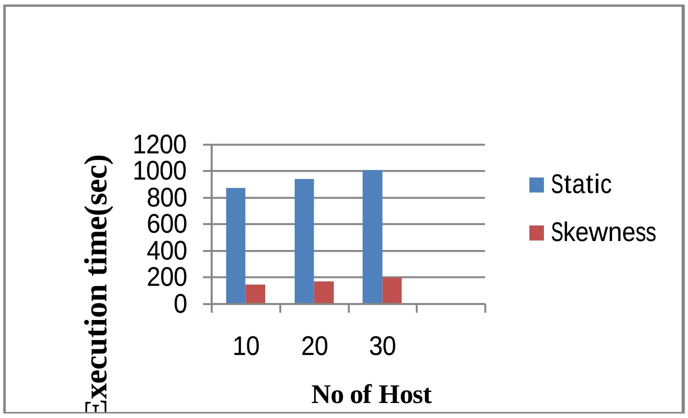
- Cold spot: If consumption of resources is below the cold threshold, it indicates that the server is usually idle and it should be turn off to save energy.

The load prediction algorithm can forecast the future resource required for applications without looking inside the virtual machines. The Load prediction algorithm predicts the CPU load and measure the load every time and predict the load in the next minute using previous history.

V. Simulation Results

An overview of skewness comparison chart is as follows:

Fig.1 : Simulation Time



The above representation (Fig.1), shows thatthe time taken for allocating resource to the host is less unlike the existing static method which gives more allocation time for resources. When the number of hosts increase thereby increasing the need for resources and fast allocation period, the skewness method could be adopted which aims to achieve minimum time for executing the request.

VI. Conclusion

Usage of cloud shows a rapid increase day by day which has led to an essential use of limited resources via cloud optimally. The Proposed method uses Skewness algorithm to predict un even utilization of resources and to improve maximum utilization of the server. This skewness algorithm is one of the best algorithms to allocate resource dynamically in cloud. Cloud computing must involve the deployment of dynamic consolidation methods that aim at minimizing the number of migrations crucially. This algorithm exhibits relatively less performance degradation that occurs due to VM migrations. Further, this proposedwork achieves optimized performance interm so fserver resource utilization with minimum energy consumption and task migration among VMs.

References

[1] Sheng Di and Cho-Li Wang, ”Dynamic Optimization of Multi-Attribute Resource Allocation in Self-Organizing Clouds”, IEEE Transactions on parallel and distributed systems, vol. 24, No. 3, pp. 464-478, 2013.
[2] K C Gouda, Radhika T V, Akshatha M., ”Priority based resource allocation model for Cloud computing”, International Journal of Science, Engineering and Technology Research (IJSETR),vol.2, No.1, 2013.
[3] Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen, ”Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment”, IEEE Transactions on parallel and distributed systems, vol. 24, No.6, pp. 1107-1117, 2013.
[4] Gunho Leey, Byung-Gon Chunz, Randy H Katzy, ”Heterogeneity- Aware Resource Allocation and Scheduling in the Cloud”, International Conference on Distributed Computing Systems, pp. 510-519, 2013.
[5] Nilton Bilay, Eyalde Laray, Kaustubh Joshi, Andres Lagar-Cavilla, Matti Hiltunen_and Mahadev Satyanarayananz, ”Efficient Idle Desktop Consolidation with Partial VM Migration”, Journal of computer application, pp. 210-224, 2012.
[6] Ts`epomofolo, Suchithra R, ”Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective”, International Refereed Journal of Engineering and Science (IRJES), vol.2, No.5, 2013.
[7] Qi Zhang, Eren G¨urses, Raouf Boutaba, JinXiao, ”Dynamic Resource Allocation for Spot Markets in Clouds”, Journal of computer science, pp. 1-6,2012.
[8] Namita R.Jain, Rakesh Rajani, “Virtualization Technology to Allocate Data Centre Resources Dynamically Based on Application Demands in Cloud Computing” International Journal of Software and Hardware Research in Engineering, vol. 2,No. 1, pp. 52-57, 2014.
[9] Rajeswari M, Savuri Raja M, Thamizhselvan I, Suganthy M, “Resource and Power Management in Cloud”, International Journal Of Scientific Research And Education, vol.2, No. 3, pp. 460-469, 2014.
[10] Arm brust et al M, “Above the Clouds: A Berkeley View of Cloud Computing,”technical report, Univ. of California, Berkeley, pp.1-23, 2009.
[11] ”Vibrant Resource Allocation Algorithms using Virtual Machine in Cloud-Survey”, International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, No.1, pp.709-714, 2014.
[12] Mahesh B Nagpure, Prashant Dahiwale, Punam Marbate, “An Efficient Dynamic Resource Allocation Strategy for VM Environment in Cloud”, IEEE Transaction on International Conference on Pervasive Computing (ICPC), vol.1, pp. 4799-6272,2015.

DOI: 10.15680/IJIRCCE.2016. 0402156