ISSN 2277-3061
Volume 15 Number 9
International Journal of Computer and Technology

A REVIEW ON ENGERY EFFICIENT STRATEGY FOR TASK ALLOCATION IN CLOUD ENVIRONMENT

Sakshi Grover [(1)], Mr. Navtej Singh Ghumman [(2)]
[(1)] Research Scholar, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
sakshi26grover@gmail.com
[(2)] Assistant Professor, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
navtejghumman@yahoo.com

ABSTRACT

Although cloud computing is now becoming more advanced and matured as many companies have released their own computing platforms to provide services to public, but the research on cloud computing is still in its infancy. Apart from many other challenges of cloud computing, efficient management of energy is one of the most challenging research issues. In this paper we review the existing algorithm of dynamic resource provisioning and allocation algorithms and holistically work to boost data center energy efficiency and performance. This particular paper purposes a) heterogeneous workload and its implication on data center's energy efficiency b) solving the problem of VM resource scheduling to cloud applications

Keywords

Cloud Computing, Power Data Center, Green Computing, Load Balacing, Virtual Machine, Energy, Data Center Broker.

INTRODUCTION

Although cloud computing has been widely adopted by the industry, but the research on cloud computing is still at an infancy stage. There are many issues in Cloud computing such as Virtual Machine Migration, Data security, Energy Management, Server Consolidation etc. as discussed in previous section that have not been fully addressed. Energy management is one of the challenging research issues. Cloud Infrastructure is the most important component in a cloud. It may consists tens of thousands of servers, network disks and devices, and typically serve millions of users globally. Such a large-scale data center will consume enormous amount of energy. For example, according to research of Google datacenter used about 2.26 million MW hours of power to operate in 2010, resulting to carbon footprint of 1.46 million metric tons of carbon dioxide. In other words, a single data center can consume power which is equal to a power consumed by small town. In order to reduce power consumption, it is necessary to balance the load among the different nodes.

Green Computing is the practice of implementing procedures and policies that improve the efficiency of computing resources in such a way as to reduce the energy consumption and maintains environmental sustainability. Various existing scheduling techniques are there which manages load among the nodes but are not energy efficient for the Cloud computing platform. Aim of the thesis is to consolidate the load balancing in an efficient way so that the resource utilization can be maximized and the energy consumption of the data center could be minimized that can further result in reducing global warming and hence assist in achieving Green Computing.

CLOUD COMPUTING: AN OVERVIEW

Cloud Computing is a distributed architecture providing computing applications and services via internet. The term cloud refers to as Internet or Network. Cloud Computing is a practical approach to manipulating, accessing, configuring the applications online. It provides online storage of data application and infrastructure without physically acquiring them. Cloud Computing is platform independent, as we need not to install software's on our PC. Cloud Computing is service - oriented as it provide on – demand resources. Cloud Computing is completely internet dependent technology. However, day by day subscribers' needs are increasing for computing resources. But in cloud computing environment, resources are shared so to prevent the wastage of resources it is essential to mange them properly. Another important role of cloud computing is to dynamically balance the load amongst different servers in order to improve resource utilization and avoid hotspots. Therefore, the main problem to be how efficiently manages the resources. So for dynamic resource allocation we are using virtualization technique which can migrate virtual machines to physical machines effectively. Load balancing of the entire system can be handled dynamically by using virtualization technology where it becomes possible to remap VMs and physical resources according to the change in load[5].

Over the past few years, number of online services- like online gaming, search, online video streaming and social networks have exploded. This led to the construction of large scale of data centers which consumes considerable amount of energy. Many existing issues have not been addressed in cloud computing. Energy Management is one of them. According to Amazon's estimation the amount of energy that is consumed by an average data center is equivalent to 25000 household appliances. So higher power consumption may lead to other problems such as wasting energy, reducing the lifetime of devices and emitting carbon dioxide which results in global warming. The other main problem in cloud computing is how efficiently we manage the resources[6]. By doing this some machine goes to idle state and we can turn off these machines to save energy. So it supports the green computing and resources can be allocated properly. The main

objectives of this research are to determine how to achieve effective load balance, how to schedule the resources and how to improve resource utility in a cloud computing.

Modern resource-intensive scientific applications and enterprise create growing demand for high performance computing infrastructures. This has led to the construction of large-scale computing data centers consuming enormous amounts of electrical power. Despite of the improvements in energy efficiency of the hardware, overall energy consumption continues to grow due to increasing requirements for computing resources. For example, in 2006 the cost of energy consumption by IT infrastructures in US was estimated as 4.5 billion dollars and it is likely to double by 2011. Moreover, there are other crucial problems that arise from high power consumption. Insufficient or malfunctioning cooling system can lead to overheating of the resources reducing system lifetime or devices reliability. In addition, high power consumption by the infrastructure leads to substantial carbon dioxide (CO2) emissions contributing to the greenhouse effect. A number of practices can be applied to achieve energy efficiency, such as improvement of applications' algorithms, energy efficient hardware, Dynamic Voltage and Frequency Scaling (DVFS), terminal servers and thin clients, and virtualization of computer resources

WORKLOADS

A workload in a cloud environment is a collection of code that can be executed or it is a shared pool of configurable applications, IT resources and data that can be rapidly provisioned [18]. It is an amount of related work that enables end users to complete their specific set of business tasks in a certain period of time.Various types of workloads:

A. Batch or online workload

These workloads are designed to process huge volumes of data. It includes data produced from cell phone bills or from online transactions.

B. Transaction workloads

These workloads were restricted to a single system. Transactional workloads are automation of business processes such as billing and order processing.

C. Analytic Workloads

In these workloads an emphasis is placed on ability to analyze the data embedded across private clouds, public websites and the data warehouse. These workloads require much more real-time computing capability.

D. Database Workloads

These workloads can effect any environment in the data center and the cloud. In some situations data workloads are huge and it requires sophisticated approach however in other data workloads are small and self-contained.

How to Achieve Green Computing

Virtualization : Virtualization is a methodology or framework of dividing the resources of a computer into multiple execution environments, by applying one or more technologies such as software and hardware partitioning, complete or partial machine simulation, emulation, time sharing, quality of service, and many others. By using virtualization it is possible to run several operating system all of their applications at same time. As, day by day the need of resources are increasing. So it is essential to allocate the resources dynamically as static allocations have some boundaries.

Load Balancing : An essential role of cloud computing platform is to dynamically balance the load among the different servers in order to improve resource utilization and to avoid hotspots. Load balancing (LB) is done on both sides i.e. on provider as well as on consumer side. On provider side, load balancing is the problem of allocating virtual machines to servers at runtime. Virtual Machine need to be reassigned so that servers do not get overloaded as demand changes. On consumer side application load can be balanced which provides efficiency to the consumers. On cloud computing platform, load balancing of the entire system can be dynamically handled by using virtualization technology through which it becomes possible to remap virtual machine and physical resources according to the change in load. However, in order to improve performance, the virtual machines have to fully utilize its resources and services by adapting to computing environment dynamically. The load balancing with proper allocation of resources must be guaranteed in order to improve resource utility. Load balancing can be done in such a way that when a particular node is overloaded or goes down with the data, then load is distributed to the other idle nodes to achieve good resource utilization.

There are several LB algorithms for the optimization and improvement of cloud performance parameters such as,

1) Throughput: The total no. of tasks or processes that have completed execution is called throughput. A high throughput is required for better performance of the system.

2) Associated Overhead: The number of overhead that is produced by the execution of the LB algorithm. Minimum number of overhead for successful implementation of the algorithm.

3) Fault tolerant: It is the ability to perform uniformly and correctly even in conditions of failure at any arbitrary node in the system.

4) Migration time: The time taken in transfer or migration of a task from one machine to another machine in the system. For improving the performance of the system this time should be minimum.

5) Response time: It is the minimum time for a distributed system executing a specific load balancing algorithm to take respond.

6) Waiting Time (WT): How much time processes spend in ready queue waiting their turn to get on the CPU.

7) Turnaround Time (TAT): Time required for a particular process to complete, from submission time to completion time.

8) Resource Utilization: It is the degree to which the resources of the system are utilized. A good load balancing algorithm provides maximum resource utilization.

9) Scalability: It determines the ability of the system to accomplish load balancing algorithm with a restricted number of machines or processors.

10) Performance: It represents the effectiveness of the system after performing load balancing. If all the above parameters are satisfied optimally then it will highly improve the performance of the system.

RELATED WORK

Heterogeneous consolidation techniques: Jianfeng Zhan[23] provides phoenixcloud which has two features 1) responsible for the division between service providers or resource provider. 2) It supports coordinated resource provisioning for heterogeneous workloads. Phoenixcloud provides a runtime environment (RE) agreement for a service provider to express RE requirements. RE agreement is a relationship between a resource provider and a service provider. We support three different relationships: same or businessor affiliated. The same relationship means that a single user plays the roles of both service provider and resource provider; the businessrelationship means that a resource provider has the businessrelationship with a service provider; the affiliated relationship means that a user playing the role of resourceprovider is affiliated to a user playing the role ofservice provider. This technique results in increased number of completed jobs.

Lingling Cao [4] proposes Hadoop map reduces which is a popular data processing engine for big data. Hadoop needs the appropriate task scheduling and job scheduling algorithms in order to complete transactions submitted by users efficiently. This technique includes several job scheduling algorithms like first-in-first-come, capacity scheduling and fair scheduling. Lingling Cao, purposes a novel task scheduling algorithm that is adaptive task scheduling strategy based on dynamic workload adjustment(ATSDWA) for heterogeneous Hadoop. WithATSDWA, tasktrackers can make corresponding adjustment of load at runtimeto achieve the optimal state accordance with the computing ability of each node, also preventing the overloading of jobtracker, thus enhancing the overall performance of the heterogeneous clusters.

Dynamic resource allocation: Dr. Shylaja [2] gives method for dynamic scheduling based on the load balancing of virtual machine. VM load balancer can determines which request is next assigned for processing. A load balancing algorithm is dynamic in nature which does not consider the previous state or behavior of the system; it depends on the present behavior of the system. The important things that consider while developing such algorithm are, comparison of load, estimation of load performance of Virtual Machine, selecting of Virtual Machine, nature of work to be transferred. This load considered can be in terms of CPU load, Network load, delay and amount of memory used. Without load balancing users could experience timeouts, delays and long system responses. Most common load balancing algorithms are Active monitoring, Throttled load balancer and Round Robin Algorithm.

Throttled load balancer maintains an index table of each state of virtual machine (ideal or busy). The client or sever first send request to data centre concerning the allocation of virtual machine to perform recommended job. The load balancer scans the index table to found available VM. If VM is found, then throttled load balancer can send the ID of ideal VM to data centre controller and it allocates the ideal VM. If appropriate VM is not found then load balancer will return -1 to data centre.

Active Monitoring load algorithm maintains information about each VMs and the number of requests allocated currently to which VM. When a request arrives to allocate new VM, it first identifies the least loaded VM. Weighted active load balancer modifies AMLB as it assigns weight to each VM in order to achieve better processing time and response time [1].

Round Robin load balancer is the simplest algorithm that uses the concept of time quantum and time slices. Each node is provided with particular time quantum and in this time quantum node will perform their operation. If time quantum is very large than RR algorithm is same as FCFS scheduling algorithm. In this, request is assigns to VMs on a rotating basis. The first request is allocated to randomly picked VM from the group and then the data center controller subsequently assigns request in a circular order. IN RRLB new allocation concept known as Weighted Round Robin Allocation in which weight are assign to each VM. The powerful server gets weight of 2.

Christopher Clark[12] Live migration is refer as powerful tool to moving running application or virtual machine between different physical machines without disconnecting from the client or application. In particular terms, this means that we can migrate an on-line streaming media server and on-line game server without requiring clients to reconnect. We achieve this by using a pre-copy approach in which we copied all the memory pages from source machine to destination host without stopping the execution of VM on the source machine. Basically for memory migration we have three phases:

• Push phase: In this phase source VM continuously pushed pages across the network to new destination. The modified pages (Dirty) during this process will re-copy or re-sent.

• Stop-and-copy phase: After the push phase VM will stop, the remaining dirty pages are copied across the destination and VM will be resumed on the destination host.

• Pull phase: The new VM executes and the remaining memory pages of VM to the target this is known as pre-paging. At the target, VM tries to access that pages which yet not been transferred, it generate page fault. These faults are called as network faults. Too many faults can degrade the performance of running applications inside the VM.

Marvin McNett[14], Usher is a cluster management system which designed to reduce the administration burden of managing cluster resources. Usher users can create any number of virtual clusters of any size which improves the ability of users to control, request and customize their resources. Usher balances using a combination of architecture and abstraction. Architecturally, usher designed have few constraints that are No two sites have identical software and hardware configuration, application requirements or service infrastructure. Usher combines a core set of interfaces to implement basic virtual cluster and machine management mechanisms such as migrating, creating and destroying VMs. Usher clients called as Ush provides is an interactive command shell for users to interact with system. Newly created VMs in usher can use a DHCP service to obtain domain names and addresses. Usher enables other powerful policies such as resource guarantees, power management and distribution. Usher has been installed in cluster computing environment at Russian Research Centre in Kurchatov, Russia and UCSD.

Zhen Xiao[15], Skewness algorithm is used to avoid unevenness utilization of resources on the server. It consists of three parts:

1. Hotspot: It is an area in which there is relatively higher temperature than surrounding that means if the utilization of its resources is above a hot threshold. This indicates the overload server. The goal of the proposed algorithm is to eliminate all hot spots if possible or we can keep their temperature as low as possible.

2. Coldspot: It is an area in which there is decrease in temperature that means if the utilization of its resources are below threshold it indicates that most of the servers are idle and can turn off to save energy. We can sort the list of cold spots on the ascending order of their memory size. Since the proposed system needs to migrate away all its VMs before we can shut down an under-utilized server.

3. Green computing: Green cloud computing not only provides a solution to save energy for environment but also reduces operational cost. The challenge is to reduce the number of servers at low load without sacrificing performance. Load skewness algorithm is used when utilization of servers are below green computing threshold

GREEN CLOUD ARCHITECTURE

The aim of green cloud internet data center is to reduce the power consumption on same side it leveraging live virtual machine migration technology and guarantee the performance from users perspective. A major challenge for Green Cloud is to automatically make the scheduling decision for dynamically consolidating/migrating virtual machines among physical servers to meet the workload requirements meanwhile saving energy.

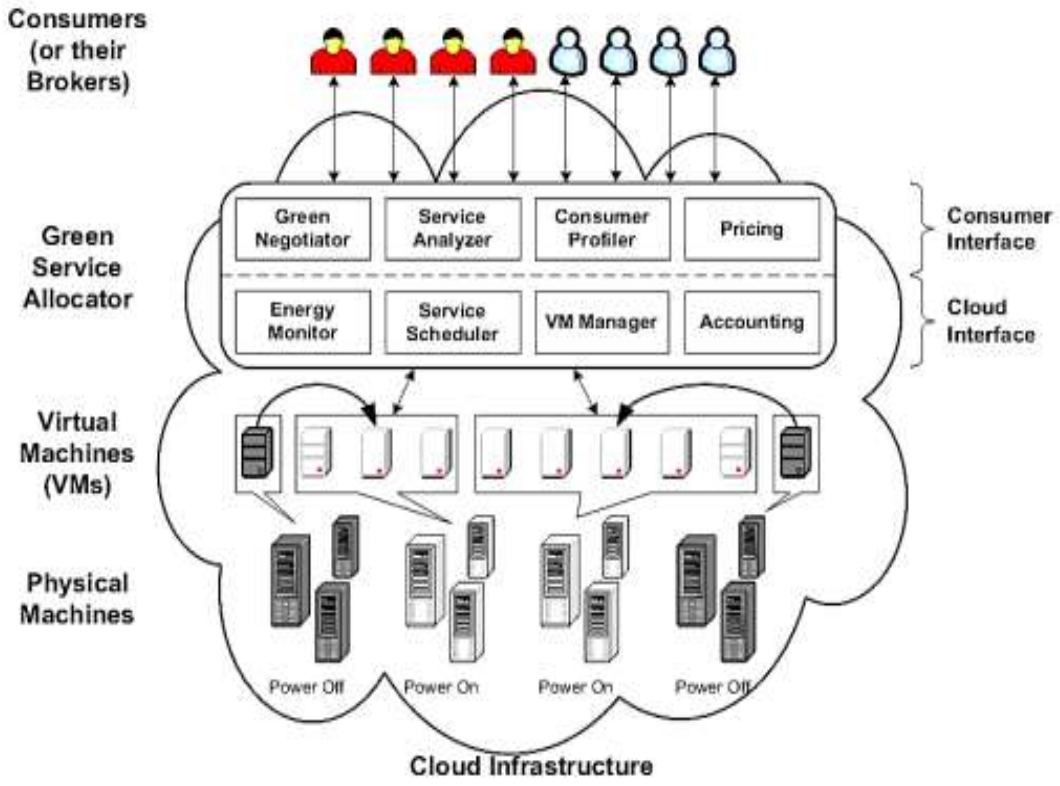

Figure 1. High Level Green Cloud Architecture

PROBLEM DESCRIPTION

The growing demands of consumers for computing services are encouraging the service providers to deploy large number of data centers, all over the world that consume very large amount of energy. Increasing amount of energy consumption by the datacenters is one of the reasons of increase in the level of carbon dioxide in our ecosystem. Research gives an idea that one google search generates as much CO2 as car produces by driving 3 inches and could power a 100 watt light bulb for 11 secs.

All monthly google search generate 2,60,000 kg CO2 which requires 39,00,000 KWh energy. According to gartner the information and communication industry produces 2 % of global carbon dioxide emission .

- In the existing work, no proper load balancing algorithm has been discussed.
- Without load balancing there will be no proper division of load among several VM's.
- Tasks are assigned randomly on round robin basis to the VM's where VM's are arranged in ascending order of their carbon foortprints.

CONCLUSION

This paper is based on cloud computing technology which has a very vast potential and is still unexplored. The capabilities of cloud computing are endless. Cloud computing provides everything to the user as a service which includes platform as a service, application as a service, infrastructure as a service. One of the major issues of cloud computing is load balancing because overloading of a system may lead to poor performance which can make the technology unsuccessful. So there is always a requirement of efficient load balancing algorithm for efficient utilization of resources. Our paper focuses on the various load balancing algorithms and their applicability in cloud computing environment.

REFERENCES

[1] Ms. R. Krishnan, Ms. S.Varghese "Survey Paper for Dynamic Resource Allocation using Migration in Cloud," International Journal of Engineering and Computer Science,2014

[2] Dr. B.S. Shylaja "Dynamic allocation method for efficient Load balancing in virtual machines for cloud computing Environment,"Advanced Computing: An International Journal, Vol.3, No.5,

[3] Jianzhe Tai Juemin Zhang Jun Li Waleed Meleis Ningfang Mi "ARA: Adaptive Resource Allocation for Cloud Computing Environment under Bursty workload".

[4] Xiaolong Xu, Lingling Cao, and Xinheng Wang, Senior Member, IEEE "Adaptive Task Scheduling Strategy Based on Dynamic Workload Adjustment for Heterogeneous Hadoop Clusters."

[5] L. Dhivya, Ms. K. Padmave "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment"

IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 1,2014.

[6] Bhupendra Panchal, Prof. R. K. Kapoor "Dynamic VM Allocation Algorithm usingClustering in Cloud Computing" International Journal of Advanced Research in Computer Science and SoftwareEngineering 2013.

[7] Joseph L. Hellerstein "HARMONY: Dynamic Heterogeneity−Aware Resource Provisioning in the Cloud".

[8] Malgorzata Steinder, Ian Whalley, David Carrera, Ilona Gaweda and David Chess "Server virtualization in autonomic

Management of heterogeneous workloads'.

[9] Atefeh Khosravi, Saurabh Kumar Garg, and Rajkumar Buyya" Energy and Carbon-Efficient Placement of Virtual Machines in Distributed Cloud Data Centers".

[10] Rajkumar Buyya, Anton Beloglazov1, and Jemal Abawajy "Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges".

[11] Hong Xu, Student Member, IEEE, andBaochunLi, Senior Member, IEEE"Anchor: A Versatile and Efficient Framework for Resource Management in the Cloud".

[12] Christopher Clark, Ying Song, Yuzhong Sun, Member, IEEE, andWeisong Shi, Senior Member, IEEE "A Two-Tiered On- Demand Resource Allocation Mechanism for VM-Based Data Centers" .

[13] R Suchithra "Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective" International Refereed Journal of Engineering and Science

[14] Marvin McNett, Diwaker Gupta, Amin Vahdat, and Geoffrey M. Voelker "Usher: An Extensible Framework For Managing Clusters of Virtual Machines",University of California, San Diego

[15] Zhen Xiao, Senior member, IEEE, weijia song and Qi chen "Dynamic Resource allocation using Virtual Machines For Cloud Computing Environment ," IEEE Transaction on parallel and distributed systems, vol.24, No.6 june 2013.

[16] Alex Delis'Nefeli: Hint-based Execution of Workloads in Clouds" International Conference on Distributed Computing Systems,2010.

[17] Kyle Chard, Member, IEEE, Kris Bubendorfer, Member, IEEE, "Social Cloud Computing: A Vision for Socially Motivated Resource Sharing" IEEE Transactions on Services Computing, VOL. 5, NO. 4, Oct-Dec 2012

[18] Weisong Shi, Senior Member, IEEE, ChuliangWeng, WenyaoZhang, and XiutaoZang"Cost-Aware Cooperative

Resource Provisioning for Heterogeneous Workloads in Data Centers" Vol. 62, NO. 11, Nov 2013.

[19] Michael Cardosa, Aameek Singh, HimabinduPucha Exploiting Spatio-Temporal "Tradeoffs for Energy-Aware

MapReduce in the Cloud"Vol. 61, NO. 12, Dec 2012.

[20] Anthony A. Maciejewski, Fellow, IEEE and Howard Jay Siegel, Fellow, IEEE"Power and Thermal-Aware Workload Allocation in Heterogeneous Data Centers".

[21] Seematai S. Patil, KogantiBhavani"Dynamic Resource Allocation using Virtual Machines for Cloud Computing Environment" International Journal of Engineering and

Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-3 Issue-6, August 2014

[22] Sukhpal Singh and InderveerChana"Energy based Efficient Resource Scheduling: A Step Towards Green Computing" International Journal of Energy, Information and Communications Vol.5, Issue 2 (2014), pp.35-52

[23] Jianfeng Zhan, Lei Wang, Weisong Shi, Shimin Gong "PhoenixCloud: Provisioning Resources for Heterogeneous Cloud Workloads". IEEE transaction service on cloud computing.