RESEARCH ARTICLE OPEN ACCESS

Energy Aware Load Balancing Technique For Managing Network Workload in Cloud Computing

Abhinav Thorat [1], Prof. Shrinivas Sonkar [2]
ME Student [1], Assistant Professor [2],
Department of Computer Science and Engineering
Amrutvahini COE, Sangamner
Maharashtra - India.

ABSTRACT

In Cloud Computing environment, load balancing is a promising and essential factor for resource utilization. Cloud having the large data centers that contain the multiple VMs, Cloud applications runs on that VMs consumes the amount of energy, so we have to minimize the energy consumption and distribute the network workloads across the VMs presenting on the server by considering the properties like CPU and memory. The number of servers is operating on cloud environment so we can check the properties of each server and distribute the workload among the lightly loaded server. The load balancing algorithm also maintains some important features of server consolidation mechanism. This paper presents a load balancing mechanism in order to provide the secure and reliable utilization of resources into the cloud environment.

Keywords:- Cloud computing, Load balancing, Consolidation, Energy-aware scheduling, Energy proportional systems.

I. INTRODUCTION

Cloud computing is an evaluation of web technology and having lots of virtual resources that can be useful and accessible and also used as resources on demand service basis with or without nominal charges. The energy-aware scheduling operational model used for application scaling As well as load balancing on a cloud application server and also that gives some of the most important features of server Consolidation mechanism. The concept of load balancing comes into existence when the first distributed computing systems were implemented. It should have the means exactly what the name distribute the network workload to a set of cloud data centers to manage the overloaded host, maximize the throughput, minimize the response time, and also increases the system flexibility to faults by avoiding overloading the systems. Load balancing key feature and the central issues in cloud computing. In the technique load balancing, distribute the network workload equally across the different clusters in the cloud in order to avoid the situation where few nodes are overloaded while few are remains idle. The cloud computing is an on-demand computing service, on growing elasticity, large network access, resource pooling, measured services are the essential characteristics in this environment. The dynamic workload results some systems overload and some systems remain unused, therefore it is necessary to distribute this network load efficiency for preventing such odd resource utilization. Therefore, the concept load balancing comes into existence that can distribute network load across different computer virtual machines, network links, a disk driver and some other resources that can improve the throughput and optimal resource utilization avoid overloading and minimize response time. In cloud computing scaling is the technique that allocates additional resources that efficient for cloud application in response to a request consistent with the SLA. There is two scaling methods, i.e. Horizontal and Vertical scaling. Horizontal scaling is the most common and effective method of scaling on a cloud system; it can increase the number of Virtual Machines (VMs) when a load of nodes increases and reduce the number when the load decreases.

II. RELATED WORK

Shunmei Meng [1]defines the Live Migration of Virtual Machines:-Migrating operating system instances across distinct physical hosts is an important and useful tool for administrators or cloud hosts of data centers and clusters: It allows a clean separation between hardware and software, and that can facilitate load balancing, fault management, and low-level system maintenance. While OSes continue to run by carrying out the majority of migration, we can achieve the impressive performance with minimal service downtimes. He can demonstrate the migration of entire OS instances and element on a commodity cluster and recording service downtimes very low as60ms. They can show that performance of the system is sufficient to make live migration for servers running interactive loads.

J. Baliga [2] the availability of the high-speed internet network and IP connections is provided latest network-based services delivery. Therefore, the network-based computing and usage of network resources has become more widespread and rapid and largely expanding the energy consumption of the network. This happens when there is increasing more attention to manages the energy consumption across the information technology and communication sectors. While energy uses by the different servers having received more attention, but there has been given less attention for connecting users to the cloud on to the energy consumption of transmission and switching the networks. In this paper, the author describes the analysis and study of energy consumption in cloud computing that considers the public and private classes in that including the energy consumption in transmission and switching as well as data processing and data storage. He also describes the energy consumption is transport and switching having a significant percentage of total energy consumption in the cloud computing.

Beloglazov [4] describes the most effective and efficient technique to improve the energy efficiency and resources utilization in cloud servers is a dynamic consolidation of virtual machines so it can directly affect the quality of service (QoS) and resource utilization when determining the reallocation of VMs from overloaded services. The QoS is got influenced because of the server get overloaded that causes performance degradation and resource shortage problem of applications. So the heuristic based solutions of this problem are detection overloaded host. In this paper, the author gives a novel approach that can solve the server host overload detection problem that can use the Markov chain model that maximize the mean time of migration also use to handle the workload multi-size sliding window workloads estimation technique is used.

Gandhi [6] introduces an Autoscaling method that can reduce the number of servers that are needed in cloud servers for dynamic capacity management. Auto scaling techniques can scale the capacity of servers, and have vertical and horizontal scaling for adding or removing server when needed. Auto Scale can also maintains the capacity of the cloud data centers to handle the burst and overloading in the request rate and make server efficient and maintain request size. In this paper authors also demonstrate that auto-scale technique can satisfy the SLAs and robustness that makes a rapid improvement over the existing dynamic capacity management policies.

Paya[13] introduces the energy consumption of the virtual machine that can have workload scalability problem. The lightly loaded server requires the more power so the author can propose the concept of load balancing to optimize the energy consumption for a large-scale system that can distribute the network workload among a different set of data centers that can analyze and observe the response time and operate on optimal energy level.

B.Urgaonkar[14]can studied and proves novel dynamic capacity technique for multi-tier internet application that employs the flexible queuing model is to be used to determining that how much resources are allocated to the each tier and predictive and reactive methods combination that used to determine when to provision these resources the experiments demonstrate the techniques for having dynamic workload. This technique doubles the application capacity in five minutes that maintains the response time.

H. N.Van [15] define that the main aim for data centers in cloud computing is to improve the profit and minimizing the power consumption and maintains SLAs. In this paper, the author can describe a framework for resource management that combines a dynamic virtual machine placement manager and dynamic VM provisioning manager.

It can take several experiments that how the system can be controlled to make trade-offs between energy consumption and application performance.

III. PROPOSED SYSTEM

There are three main contributions of this paper that follow:-

1) A new model of cloud servers that is based on different operating regimes with various degrees of energy efficiency (i.e. processing power versus energy consumption).

2) A novel algorithm that can perform load balancing and application scaling for maximizing the number of servers operating in the energy-optimal regime and comparison of techniques analysis for load balancing and application scaling using three different sizes of clusters and two different average load profiles.

3) The objective of the algorithms is to ensure that the largest possible number of active servers operate with their respective optimal operating regime. The actions implementing this policy are (a) migration of VMs from a server operating in the low regime and then switch the server to a sleep state system (b) switching an idle or lightly loaded server to a sleep state system and reactivate servers in a sleep state when increasing the cluster load. (c) Migration of the VMs from an overloaded server, a server operating in the high regime with applications predicted to increase their demand for service for computing.

IV. PROBLEM DEFINITION

The problem is to determine as: - Server gets overloaded due to an excess of a request from different host then it goes into the sleep mode. So by using Server consolidation properties transfer the request from sleep mode to running mode on another server. And Use Auto-scaling and peak energy level.

A. System Architecture

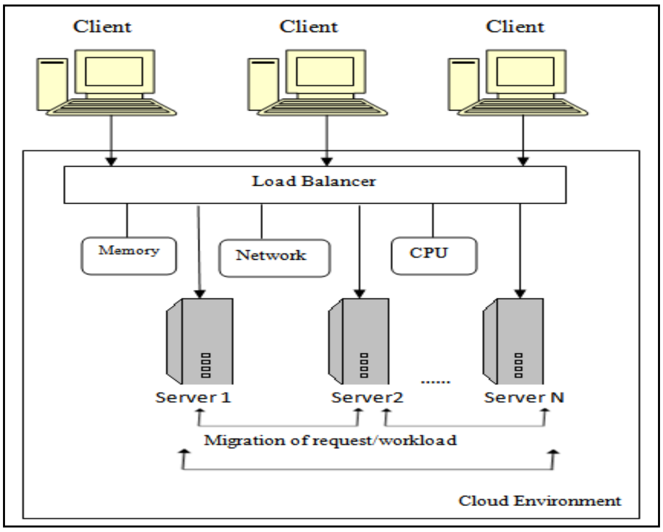

Fig 1: Load Balancing Architecture.

B. System Modules

1) Load Balancing in Cloud Computing

Cloud computing involves web services, visualization, software, networking and distributing computing. Cloud can have several elements such as client data center and multiple distributed servers. It includes on-demand service, high availability, fault tolerance, flexibility, scalability, reduced cost for owners, reduced overhead for clients or users etc. The effective establishment of load balancing algorithm can solve such a problem; the load is basically a CPU utilization, memory capacity, request delay or load in the network. Load balancing is the technique that can distribute the load among the various nodes of different distributed system for improving the job response time and resource utilization that can avoid the situation where some nodes are heavily loaded and others are idle or in a sleep state. All processor every node or system in the network does the same amount of work at any specific time. The main aim is to use effective load balancing algorithm that can minimize the latency and maximize the throughput on the cloud environment.

2) Energy Efficiency of a System

We can measure the energy efficiency of the different system as a ratio of performance per watt of power. From few decades performance of computing system is increased much more rapidly than its energy efficiency. The idle or lightly loaded system consumes less energy or should be nearer to zero and it increased linearly with the system workload. But practically the idle system can consume more than half energy of full load system. We have an optimal energy consumption regime that is far from the typical operating regime of data center servers. When energy proportional system is in idle state is consumes no energy or very little energy and increases gradually as load increases.

3) Resource management policies for large-scale data centers

These policies can be loosely grouped into five classes (i) Admission control (ii) Capacity allocation (iii) Load balancing (iv) Energy optimization and (v) Quality of service (QoS) guarantees. To prevent the system from accepting the workload in violation of high level system policies is explicit goal of an admission control policy; system not accepting the additional workload that can prevent it from completing the work that is already done or in progress We have a knowledge of the global state of the system for limiting the workload. In a dynamic system, this knowledge is when available, is at best case. Allocating the resources for individual instances is known as Capacity allocation.

4) Server Consolidation

We can measure the energy efficiency of the different system as a ratio of performance per watt of power. From few decades performance of computing system is increased much more rapidly than its energy efficiency. The idle or lightly loaded system consumes less energy or should be nearer to zero and it increased linearly with the system workload. But practically the idle system can consume more than half energy of full load system. We have an optimal energy consumption regime that is far from the typical operating regime of data center servers. When energy proportional system is in idle state is consumes no energy or very little energy and increases gradually as load increases.

5) Energy-aware Scaling Algorithms

The main aim of this algorithm is to ensure that the numbers of active server out of all are operating with the optimal operating regime. The actions implementing this policy are (a) migration of VMs from a server operating in the low regime and then it switch the server to a sleep state system. (b) switching idle servers to a sleep state system and reactivate servers in a sleep state when the data center load increases; (c) migration of VMs from an overloaded server, a server operating in the high regime with applications predicted to increases their demands for computing.

C. Proposed Algorithm

Input:

Workload (W) ->W1, W2, W3.....
Resource (R) -> R1, R2, R3...

Output:

Migration List (M) -> M1, M2, M3...

Energy Efficiency Algorithm

```
1) Start
2) Extract Total workload list
   W (Z) -> W1, W2, W3....Wn
3) Access total Resource list
   R (Z) -> R1, R2, R3.....Rn
4) User uploads file
   U (Z): F (Z) -> Un, Fn
5) Check first cloud server workload
6) Limitation of server depends on energy level
7) If server is going to energy threshold
8) File migrates to another server->HOT SPOT Process.
9) Check remaining server workload.
10) Find Min (workload Resources) ->optimization
11) Manage workload of every server ->Green Computing.
12) Check energy level
13) End.
```

V. EXPERIMENTAL RESULT

A. Model parameters:

The cluster leader maintains static and dynamic information about all servers in cluster C.

Static information includes:

S_k - The serverId;

k - A constant quantifying the processing power of server S_k;

The processing power is expressed in vCPUs.

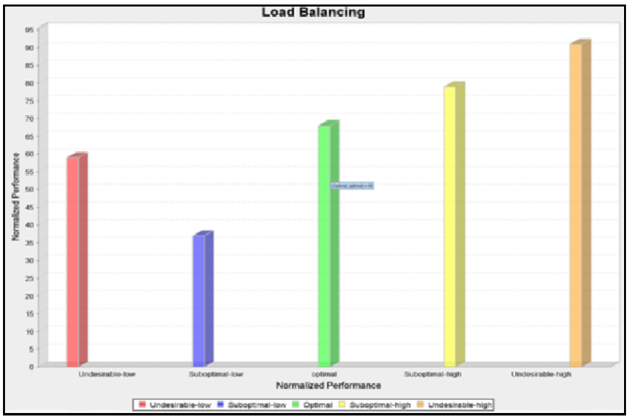

Fig2. Normalized Performance Rate

The model illustrated in Fig.2 uses several parameters:

- X_sopt;l
- X_k _opt;l
- X_k , _opt;h
- X_k , and _sopt;h
- X_k , the normalized performance
- Boundaries of different operating regimes.

X_k - the reallocation interval. Every k units of time the system determines if and how to reallocate resources.The application record of application.

A_i;k includes the application Id and several other parameters:

1) a_i;k(t) - Current demand of application A_i for processing power on server S_k at time t in CPU units. e.g., 0:4.
2) l_i;k - highest rate of increase in demand for processing
power of application Ai on server Sk.
3) p_i;k(t) - migration cost.
4) q_i;k(t) - horizontal scaling cost.

B. Explanation:-

- This classification captures the current system load and allows us to distinguish the actions to be taken to return to the optimal regime.
- A system operating in the suboptimal-low regime is lightly loaded; the server is a candidate for switching to a sleep state.
- The undesirable-high regime should be avoided because a scaling request would immediately trigger VM migration and, depending on the system load, would require activating one of the servers in a sleep state.
- This classification also captures the urgency of the actions taken; suboptimal regimes do not require an immediate attention, while the undesirable-low does.
- The time spent operating in each suboptimal regime is also important.

VI. CONCLUSIONS AND FUTURE WORK

By considering the resources like requests on network, memory of requested data and CPU utilization of server and by checking and analysing such a resources, we can balance the network workload by migrating or switching the workload to the efficient and secure virtual machines of different server and get the optimal efficiency for improving the throughput and minimize the response time of the system using the green computing that uses the auto scaling technique by the server.

ACKNOWLEDGMENT

It gives me an immense pleasure to express my sincere and heartiest gratitude towards my guide Prof. S.K.Sonkar for his guidance, encouragement, moral support and affection during the course of my work. I am especially appreciative of his willingness to listen and guide me to find the best solution, regardless of the challenge. This work is also the outcome of the blessing guidance and support of my parents and family members and friends. I am also thankful to all who have contributed indirectly and materially in words and deeds for completion of this work.

REFERENCES

[1]. A. Paya and D. C. Marinescu. “Energy-aware Load Balancing and Application Scaling for the Cloud Ecosystem.” 2168-7161 (c) 2015 IEEE Transactions on Cloud Computing.

[2]. A. Beloglazov, R. Buyya “Energy-efficient resource management in virtualized cloud data centers." Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Comp., 2010.

[3]. A. Beloglazov and R. Buyya. “Managing overloaded hosts for dynamic consolidation on virtual machines in cloud centers under the quality of service constraints." IEEE Trans. on Parallel and Distributed Systems, 24(7):1366- 1379, 2013.

[4]. J. Baliga, R.W.A. Ayre, K. Hinton, and R.S. Tucker.” Green cloud computing: balancing energy in processing, storage, and transport." Proc. IEEE, 99(1):149-167, 2011.

[5]. A. S. Thorat and S.K. Sonkar “A review on energy efficient load balancing techniques for secure and reliable cloud ecosystem” IJARIIE-ISSN(O)-2395-4396, Vol-2 Issue-1 2016

[6]. A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. “AutoScale: dynamic, robust capacity management for multi-tier data centers." ACM Trans. On Computer Systems, 30(4):1{26, 2012.

[7]. A. Paya and D. C. Marinescu. “Energy-aware load balancing policies for the cloud ecosystem." HTTP: //arxiv.org/pdf/1307.3306v1.pdf, December 2013.

[8]. H. N. Van, F. D. Tran, and J.-M. Menaud. “Performance and power management for cloud infrastructures." Proc. IEEE 3rd Int. Conf. on Cloud Comp., pp. 329{336, 201

[9]. D. Gmach, J. Rolia, L. Cherkasova, G. Belrose, T. Tucricchi, and A. Kemper. “An integrated approach to resource pool management: policies, efficiency, and quality metrics." Proc. Int. Conf. on Dependable Systems and Networks, pp. 326{335, 2008.

[10]. A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. “Are sleep states effective in data centers?" Proc. Int. Conf. on Green Comp., pp. 1{10, 2012.

[11]. Jeffrey S. Chase, Darrell C. Anderson, Prachi N. Thakar, Amin M. Vahdat, “Managing Energy and Server Resources in Hosting Centers”SOSP '01 Proceedings of the eighteenth ACM symposium on Operating systems principles, the year 2001

[12]. Gong Chen, Wenbo He, Jie Liu, Suman Nath, Leonidas Rigas, Lin Xiao, Feng Zhao, “Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services” NSDI'08 Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, 2008

[13]. A.D. Gawali and S.K.Sonkar “Dynamic Resource Allocation in Cloud Computing using Virtualization Technology” International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, pp. 2013

[14]. S.K.Sonkar and Dr.M.U.Kharat “A Survey on Resource Management in Cloud Computing Environment” IJATCSE, ISSN (ONLINE): 2278 – 3091, Volume 4. No.4 (2015).

[15]. M. Elhawary and Z. J. Haas. ”Energy efficient protocol for cooperative networks.” IEEE ACM Trans. on Net- working, 19(2):561574, 2011.

[16]. B. Urgaonkar and C. Chandra. ”Dynamic provisioning of multi-tier Internet applications.” Proc. 2nd Int. Conf, on Automatic Comp., pp. 217228, 2005.