INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O)
ISSN: 2394-9910(P)
VOLUME 5 ISSUE 2
Apr 2017 - Jun 2017
www.irjeas.org

CLUSTER BASED LOAD BALANCING IN CLOUD

Anurag Shrivastava[1], Jyoti Sondhi[2], Pooja Mukhariya

[1]*Asso. Professor &Head, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[2]*Assistant Professor, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[3] *M.Tech Scholar, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*

Abstract- However, the complexity involved in incoming load to the cloud and the limited availability and characteristics of the resources in the cloud needs some well-designed resource management system to efficiently utilize the resources, without any latency or lag. In this paper we are presenting an effective Virtual machines (VMs) controlling and load distribution algorithm for cloud computing. The proposed technique ensures power saving without producing delay. The proposed work is validated by simulating it for different cloud and load configurations using Matlab and the simulation result verifies the efficiency of proposed algorithm for cloud computing.

Keywords: Cloud Computing, Load Balancing, Clustering.

1. Introduction

In computer networking, the cloud is considered as a cluster of resources, such as processor, memory, software and other computing infrastructure. The cloud provides these resources are as a service to the end users (consumers) on the basis of service plans. The concept of cloud computing reduces the financial load for the infrastructure development and maintenance, since it rented from cloud service provider. Since the cloud is used as a service it is the responsibility of cloud manager, to provide the users requested resources quickly as well as to efficiently operate the resources. The problem of the cloud manager could be seen as the complex optimization problem in which (considering the simplest way) the load coming from different users has to be distributed on the available resources such that the latency and the power requirements should be minimized. The task is difficult to optimized on any single level of the architecture hence it is operated at many levels like firstly load scheduling, queue management, resource allocation, power management etc. This paper presents a clustering approach for load balancing to distribute the load on the cloud efficiently. The rest of the paper is arranged as that second section presents an overview of the some recent literature on the related topic. The third section briefly describe the cloud computing system while the proposed algorithm is presented in fourth section followed by the simulation results in fifth section. Finally in sixth section the conclusion is drawn on the basis of simulation results.

2. Literature Review

Because of the emerging demand cloud services every field related to it gaining interest of researchers and the load balancing is not an exception so many literatures have been already published some of them are presented in this section. Shu-Ching Wang et al [1] introduced a two-phase scheduling algorithm under a three-level cloud computing network is advanced. The proposed scheduling algorithm combines OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) scheduling algorithms that can utilize more better executing efficiency and maintain the load balancing of system. Flexible distributed capacity Allocation and load redirect Algorithms for cloud systems is proposed by Danilo Ardagna et al [2]proposed capacity allocation techniques able to coordinate multiple distributed resource controllers working in geographically distributed cloud sites. Furthermore, capacity allocation solutions are integrated with a load redirection mechanism which forwards incoming requests between different domains. The overall goal is to minimize the costs of the allocated virtual machine instances, while guaranteeing QoS constraints expressed as a threshold on the average response time. Dzmitry Kliazovich et al [3] presented the e-STAB (Energy-Efficient Scheduling for Cloud Computing Applications with Traffic Load Balancing) their work emphases the role of communication fabric and presents a scheduling solution, named e-STAB, which takes into account traffic requirements of cloud applications providing energy efficient job allocation and traffic load balancing in data center networks. Effective distribution of network traffic improves quality of service of running cloud applications by reducing the communication-related delays and congestion-related packet losses. Energy-aware

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 5 Issue 2, Apr 2017- Jun 2017, Page 01-04

resource allocation heuristics for efficient management of data centers for Cloud computing is presented by Anton Beloglazov et al [5]in which they conduct a survey of research in energy-efficient computing and propose: (a) architectural principles for energy-efficient management of Clouds; (b) energy-efficient resource allocation policies and scheduling algorithms considering QoS expectations and power usage characteristics of the devices; and (c) a number of open research challenges, addressing which can bring substantial benefits to both resource providers and consumers. Yiming Han et al [6] presented A Hierarchical Distributed Loop Self-Scheduling Scheme for Cloud Systems to achieve good load balancing by applying weighted self-scheduling scheme on a heterogeneous cloud system. This scheme also considers the distribution of the output data, which can help reduce communication overhead. Mousumi Paul et al [7] proposed Dynamic job Scheduling in Cloud Computing based on horizontal load balancing.

3. Cloud Computing

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). 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 remote services with a user's data, software and computation.

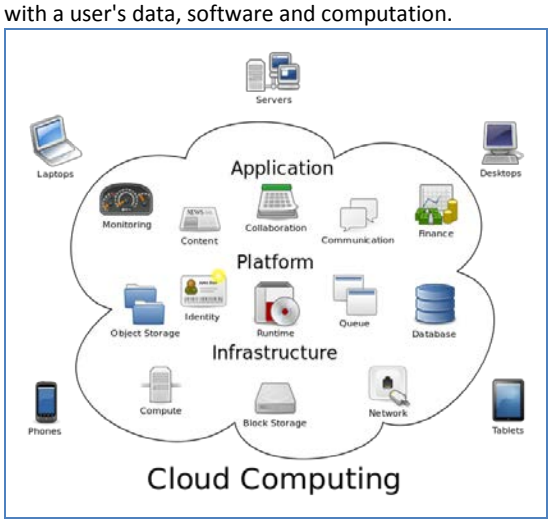

Figure 3.1: Cloud computing logical diagram

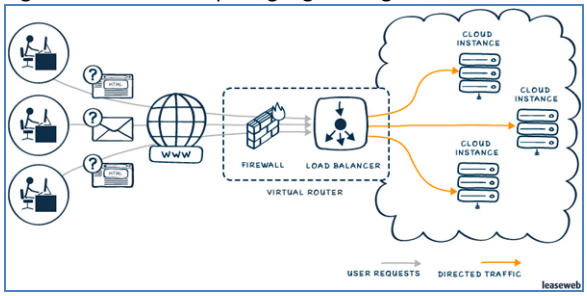

Figure 3.2: Load Balancer in Cloud Computing

3.1 Services on Cloud

There are three types of cloud providers:

1. **Software as a Service:** A SaaS provider gives subscribers access to both resources and applications. SaaS makes it unnecessary for you to have a physical copy of software to install on your devices.

2. **Platform as a Service:** A PaaS system goes a level above the Software as a Service setup. A PaaS provider gives subscribers access to the components that they require to develop and operate applications over the internet [6].

3. **Infrastructure as a Service:** An IaaS agreement, as the name states, deals primarily with computational infrastructure. In an IaaS agreement, the subscriber completely outsources the storage and resources, such as hardware and software that they need [6].

3.2 Standard Load Balancing Algorithms

A. Round Robin Load Balancer

This algorithm works on random selection of the virtual machines. The datacenter controller assigns the requests to a list of VMs on a rotating basis.

B. Throttled Load Balancing Algorithm

Throttled algorithm is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine whichaccess that load easily and perform the operations which is give by the client or user. In this algorithm the client first requeststhe load balancer to find asuitable Virtual Machine to perform the required operation [5].

4. Proposed Algorithm

The proposed algorithm can be explained in the following steps:

1. The load scheduler initially forms the Indexes of Clusters, Indexes of each Level within each cluster, and Indexes of each resource available within each level.

2. Next the load scheduler collects the Indexes of current level in use and indexes of clusters in which current level is present currently.

3. Now it generates a reference table showing the status of different resources,levels and clusters, once a cluster is selected.

4. Suppose, number of clusters = 4, number of levels within selected cluster = 3, number of resources within selected level = 7.

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 5 Issue 2, Apr 2017- Jun 2017, Page 01-04

5. One extra columnis added in reference table which shows the status of that component whether it is free or not and whether it is off or not(in case of cluster and level).

In case of cluster and level,
'0' represents cluster/level is off.
'1' represents cluster/level is in active state.
In case of Components,
'0' represents that Component is free.
'1' represents that Component is busy.

Whenever any cluster is selected for the 1st time its entry is made in the reference table along with its levels and all the resources in those levels.This table is updated immediately after any action is performed on resources/level or cluster either assigning workload (Booting) or turningoff it.
Initially all the clusters are in off state and all the levels except the lowest indexed level.
This is done in order to assure that whenever a new cluster is selected that was in off state, only one wakeup signal will be sufficient to assign workload to resource.

4.1 Procedure for Assigning a Resource to an Incoming Service Request

1. Scheduler will search for the free resources in the currently active levels within the currently active cluster. If free resourcesare present then assign workload to it. Otherwise,

2. Scheduler will look into the table for the free resources from the scratch. If found then, ok. Otherwise,

3. Scheduler will select next lowest indexed cluster and in that the active level and finally workload will be assigned to lowest indexed resources and its entry will be made in the table immediately.
Suppose a cluster entry is present in the table showing that it was previously used but presently it is in power off state , then that cluster will not be selected (unless and until that is the last free available cluster) , keeping in mind the load balancing issues.

4.2 Procedure of Power Controlling at Any Level or Cluster

Suppose any resource is free for time equal to greater than the threshold value, then Scheduler unit will look if all the resources within that level are also free then, that level will be selected to get turned off, but again Scheduler unit will look if all the levels within that cluster are also in turned off state and that level is the last one to turned off then, in spite of turning off that particular level control unit it will turned off the whole cluster, otherwise simply turned off that level.

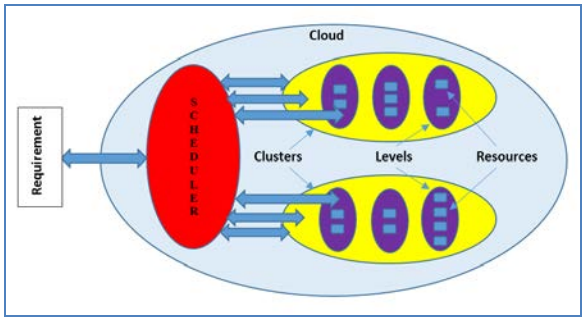

Figure 4.1: Block Diagram of the Proposed System.

5. Simulation Results

The implementation and simulation of the proposed algorithm is performed using MATLAB. The simulation is executed for random load requests and cloud is formed by using 16 VMs. Finally the simulation results are presented in the form of graphs and tables.

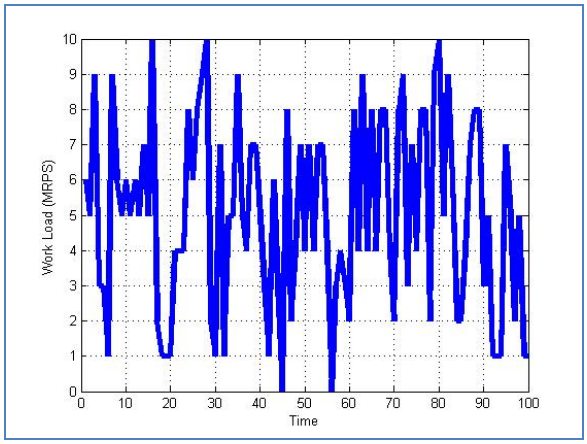

Figure 4.2: the incoming workload (million requests per second)into cloud.

Table 4.1: VMs Utilization in Percentage (Each Element Represents a VM).

20	13	13	0
17	13	13	5
18	13	9	7
30	25	15	11

Table 4.2: Cluster Utilization in Percentage

Cluster Index	Utilization (%)
1	24
2	23
3	21
4	35

Table 4.3: Individual VMs Booting (Number of Times).

3	5	4	0
5	4	4	3

4	5	4	2
6	5	4	3

Table 4.4: Cluster Booting (Number of Times)

Cluster Index	Number of Bootings
1	3
2	3
3	2
4	3

6. Conclusion

This paper presents a new hierarchical approach for the power saving and efficient and uniform utilization of resources whileminimizing latencyand without compromising processing speed in cloud. The simulation results show that the proposed technique uniformly distributes the work loadto all VMs (as shown in table1). It also reduces the number of booting (as shown in table 3 and 4) of individual VMs which reduces the delay in processing. These results validates that the proposed algorithm is can provide a better solution for power and latency optimizedcloud systems.

References

[1] Shu-Ching Wang, Kuo-Qin Yan, Wen-Pin Liao and Shun-Sheng Wang "Towards a Load Balancing in a Three-level Cloud Computing Network", Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on (Volume:1):9-11 July 2010.

[2] Danilo Ardagna, Sara Casolari and Barbara Panicucci "Flexible Distributed Capacity Allocation and Load Redirect Algorithms for Cloud Systems",Cloud Computing (CLOUD), 2011 IEEE International Conference on: 4-9 July 2011.

[3] Dzmitry Kliazovich, Sisay T. Arzo, Fabrizio Granelli, Pascal Bouvry and Samee Ullah Khan "e-STAB: Energy-Efficient Scheduling for Cloud Computing Applications with Traffic Load Balancing", 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing.

[4] Luis M. Vaquero, Luis Rodero-Merino and Rajkumar Buyya "Dynamically Scaling Applications in the Cloud", ACM SIGCOMM Computer Communication Review archive Volume 41 Issue 1, January 2011 Pages 45-52.

[5] Anton Beloglazov, Jemal Abawajy and Rajkumar Buyya"Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", Science Direct: Future Generation Computer Systems Volume 28, Issue 5, May 2012, Pages 755–768.

[6] Yiming Han and Anthony T. Chronopoulos "A Hierarchical Distributed Loop Self-Scheduling Scheme for Cloud Systems", 2013 IEEE 12th International Symposium on Network Computing and Applications.

[7] Mousumi Paul, Debabrata Samanta and Goutam Sanyal "Dynamic job Scheduling in Cloud Computing based on horizontal load balancing", Int. J. Comp. Tech. Appl. ISSN:2229-6093, Vol 2 (5), SEPT-OCT 2011.

[8] Kejiang Ye, Xiaohong Jiang, Dawei Huang, Jianhai Chen and Bei Wang "Live Migration of Multiple Virtual Machines with Resource Reservation in Cloud Computing Environments", 2011 IEEE 4th International Conference on Cloud Computing.