[Volume 1 issue 3 April 2013]
Page No.103-108 ISSN :2320-7167

# Load Balancing Techniques of Cloud Computing

***Karanpreet Kaur, Ashima Narang , Kuldeep Kaur***

SUSCET,Tangori
Punjab( India*)*
GKU,Talwandi Sabo patiala,Punjab INDIA
Punjabi university, Punjab INDI

***Abstract—These days, the organization knows that the power is being consumed by the unutilized Resources that is why the local cloud is becoming very popular. An essential requirement for cloud environments is not only the reduction in power consumption but also focus is also laid on decreasing the operating cost and improve the reliability of the system. The energy-aware computing makes the algorithms run faster and also reduce the computing energy requirements. This paper includes the existing techniques for balancing the load in cloud computing and their comparision on the basis of various parameters like performance, overhead, scalability etc.***

***Keywords – Cloud Computing, Virtual machine, Consolidation, Energy-Aware Scheduling, Load Balancing.***

## I. INTRODUCTION

Cloud computing can be classified as a new paradigm for dynamic provisioning computer services supported by data centers that usually employ virtual machine (VM) technology for consolidation[4]. Cloud computing provides infrastructure, platform and software as services that is available to consumer under the pay as you go model. According to the Service Level Agreement (SLA), customers can access to resources provided by a cloud provider. In distributed data centers, virtualization technology is being used by the clouds to allocate the resources to the customer when required. Clouds are provided to the customers for giving them three models: Software-as- a-Service (SaaS), Platform-as-a-Service (PaaS),and Infrastructure-as-a-Service (IaaS). Load balancing is one of the central issues in cloud computing [5]. It is a technique in which distribution of the dynamic local workload equally across the nodes in cloud in order to avoid the situation where few nodes are overloaded while few are idle. A high user satisfaction and resource utilization ratio is achieved ,therefore it helps in improving the performance and resource utility of the whole system. The popularity of local cloud implementation is increasing due to the fact that commercialized cloud vendor are not much secure according to many organizations. There are many implementations of cloud computing that organizations can use while implementing their own private cloud. Some possible solutions are Open Nebula [16] or Nimbus [18] or cloudbus[17]. This architecture achieves the ease of scalability and availability. The cloud includes several different hardware set ups. a cloud is built by single type of hardware, its nature is to expand by various new hardware throughout its lifetimes. The main part of power consumption in data centers come from computation processing, disk storage, networks and cooling system[15].

This paper is includes Section II need of energy-aware scheduling in clouds. Section III describes energy-aware cloud architectural elements. Section IV describes existing load balancing techniques in cloud computing. Section V describes comparison of existing load balancing technique and Section VI conclusion.

## II. REQUIREMENT OF ENERGY-AWARE SCHEDULING IN CLOUDS

Cloud computing is a client-server architecture composed by large and power-consuming data centers designed to support the elasticity and scalability required by consumers[4]. A huge amount of data is used by the Data center, due to this the maintenance of local cloud is very costly. It consume near about 10 to 100 times more power than a office building[5]. Therefore the appropriate load balancing technique can improve the utilization of the available resources, and hence minimizing consumption of the resource.

• *Reducition in Consumption of Energy* - Load balancing helps in avoiding overheating by balancing the workload across all the nodes of a cloud, hence reducing the amount of energy consumed[6]. For the reduction in power consumption of data centers, consolidation of the computation on servers is important. The basic idea is the reduction in “idle power”, i.e., the power onsumption of the idle servers, so that there is a reduction in the requirements of more working servers that lead to save power power.

## III. ENERGY-AWARE CLOUD ARCHITECTURE ELEMENTS

Figure 1 shows the high-level architecture for supporting energy-efficient service allocation in Cloud computing infrastructure[11]. There are four entities included:

**a) *Consumers/Brokers*:** Cloud consumers or their brokers requests a service from anywhere around the world to Cloud. An important notice is the difference between Cloud consumers and users of deployed cloud services. For example, a company deploying a Web application can be a consumer, that represents different workload as per the different number of "users" using it.

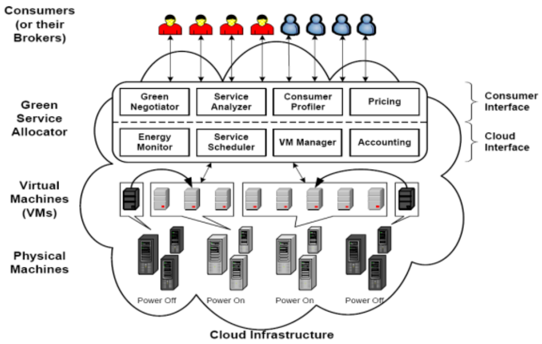

**Figure 1:** High-level system architectural framework

**b) *Green Resource Allocator*:** is an interface between the Cloud infrastructure and consumers. It need interaction of the following components for supportting energy- efficient resource management:

• *Green Negotiator*: it Negotiates with the consumers/brokers to finalize the SLA with specified prices and penalties (for violations of SLA) between the Cloud provider and consumer depending on the consumer's QoS requirements and energy saving schemes. In case of Web applications, for instance, QoS metric can be 95% of requests being served in less than 3
seconds[].

• *Service Analyser*: Interprets and analyses the service requirements of a submitted request before deciding whether to accept or reject it. Hence, it needs the latest load and energy information from VM Manager and Energy Monitor respectively.

• *Consumer Profiler*: Gathers specific characteristics of consumers so that important consumers can be granted special privileges and prioritised over other consumers.

• *Pricing*: it decides how service requests are charged to manage the supply and demand of computing resources and facilitate in prioritising service allocations effectively.

• *Energy Monitor*: it determines that which machine should be power on/off.

• *Service Scheduler*: it Assigns requests to VMs and decides the resources for VMs. It also decides when VMs are to be added or removed to meet demand.

•*VM Manager*: the availability of VMs and their resource entitlements are traced. Also the migrating VMs across physical machines are kept in record by the VM manager

• *Accounting*:. Historical usage information helps in improving the decisions of service allocation.

**c) *VMs*:** Multiple VMs can be dynamically started and stopped on a single physical machine to meet accepted requests, hence providing maximum flexibility to configure various partitions of resources on the same physical machine to different specific requirements of service requests. Multiple VMs can also concurrently run applications based on different operating system environments on a single physical machine. In addition, by dynamically migrating VMs across physical machines, workloads can be consolidated and unused resources can be put on a low-power state, turned off or configured to operate at low-performance levels (e.g., using DVFS) in order to save energy.

**d) *Physical Machines*:** The underlying physical computing servers provide hardware infrastructure for creating virtualized resources to meet service demands.

## IV. EXISTING LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING

Following are the various load balancing techniques that are currently prevalent in cloud computing:

### A. *Dynamic Round-Robin algorithm*

*Dynamic Round-Robin*[1] method is an extension to the Round-Robin method. it uses two rules that help to consolidate virtual machines. The first rule says that if a virtual machine has finished and still other virtual machines that are hosted on the same physical machine, then this physical machine cannot accept any new virtual machine. Such physical machines are called "retiring" state physical machines, that means when the all the other virtual machines finish their execution, then we can shutdown this physical machine.

The second rule says that if there is a "retiring" state physical machine that is used for long period of time, then instead of waiting for that virtual machines to finish, the physical machine is forced to migrate all the other virtual machines to other physical machines, and then shutdown the physical machine after the migration finishes.

The threshold waiting time is represented by the "retirement threshold". A physical machine will be forced to migrate to all the virtual machines and then shut it down as it is in the retiring state but after the retirement threshold, it could not finish all other virtual machines

These two rules are used by the Dynamic Round-Robin strategy so as to consolidate virtual machines implemented by the Round-Robin method. According to the first rule, adding extra virtual machines to a retiring physical machine is avoided. According to the second rule, the consolidation process become fast and it enables Dynamic Round-Robin to shutdown physical machines, such that the number of physical machine used to run all virtual machines is reduced , hence the power can be saved.

### *B. A Hybrid algorithm*

For the conservation of energy, Ching-Chi Lin[1] proposed the combination of Dynamic Round-Robin and

First-Fit to form a Hybrid algorithm. The probability distribution
(e.g., a normal distribution). is followed and the number of incoming virtual machines are assumed as a function for time. Hybrid algorithm uses virtual machines's incoming rate for the scheduling of virtual machines. The First- Fit is used by the Hybrid method during rush hours to completely utilize the computing power of physical machines, and then it uses the Dynamic Round-Robin for the consolidation of the virtual machines and thus reduce the consumption of the energy in non-rush hours.

### C. *(PALB)Power Aware Load balancing (PALB)Algorithm*

The PALB algorithm [2] has three elementary sections. The balancing section is completely responsible for the determination of the virtual machines that will be installed. It first gathers the percentage of utilization of each active computer node. If in case, all compute nodes are more than 75% utilization, then PALB initiates a new virtual machine as the computer node having the lowest utilization. When all the computer nodes are about 75% utilization, then all the available computer nodes are working. Otherwise, the booting of the new virtual machine (VM) on the computer node having the highest utilization is required (essentially if it can accommodate the size of the VM). When 25% of the resources are available, then only the threshold value of 75% of utilization was chosen, at least one virtual machine could be accommodated by the use of 12 out of 20 available configurations.

The next upscale section of this algorithm is used for power on additional computer nodes (if there are many available computer nodes). This is done if all currently active computer nodes have utilization over 75%. The downscale section is responsible for powering down idle compute nodes. If the computer nodes are having less than 25% utilization of its resources, PALB gives a shutdown command to that particular node.

### *D. Equally Spread Active Execution .( ESCE ) algorithm*

The estimation of the job size by the cloud manager and then checking for the for the availability of the virtual machine and also the capacity of the virtual machine. Once the available resource (virtual machine) size and the size of the job matches, then immediately the job scheduler allocates identified virtual machine or resource to the job in a queue. The affect of the ESCE algorithm[3] is that an improvement is seen in the response time and the processing time. The equal distribution of jobs is done, now the complete computing system is load balanced and there is no such virtual machines that are underutilized. Due to this merit , there is a reduction in the cost of virtual machineas well as the costof data transfer.

### *E. Task Consolidation Algorithms*

Both ECTC and MaxUtil [4] follow similar steps in algorithm description but the main difference is being their cost functions. for a given task, two heuristics algorithms that every resource identify for the most energy efficient resource. The most energy efficient resource evaluation depends upon the used heuristic i.e. the cost function used by the heuristic. The cost function of ECTC tracks the energy consumption of the current task and subtracts the minimum energy consumption required by the task to runand if some other task is being executing at the very same time. Such that, the energy consumption of the time period when both the tasks are running . Among those tasks, the current task is specifically focused

.

### *F. Load Balancing mechanism based on ant colony and complex network theory( ACCLB) Algorithm*

ACCLB load balancing mechanism[7] based on ant colony and complex network theory from the open cloud computing concepts. the use of the small-world and scale-free characteristics of a complex network is done to achieve efficient load balancing. This technique discourages heterogeneity, is adaptive to dynamic environments, It also encourages fault tolerance and has better scalability that helps in the improvement of the system performance.

### *G. Minimum Cost Maximum f low (MCMF)Algorithm*

This is based on the modified Bin- Packing model[8] that suffers from scalability problems with many examples and increase in the number of PMs as well as the requests. This has motivated for finding an alternate option to the dynamic resource placement problem and hence leads to the Minimum Cost Maximum Flow (MCMF) algorithm.

### *H. Join-Idle-Queue*

for dynamically scalable web services, Y. Lua et al. [12] proposed a Join- Idle-Queue load balancing algorithm. In this algorithm,a large scale load balancing is done with distributed dispatchers. firstly load balancing the idle processors across dispatchers and then, assigns the jobs to the processors for the reduction in the average queue length on each processor. It effectively reduces the system load by removing the load balancing work from the critical path of processing the request,

### *I. OLB+LBMM*

S.-C. Wang et al. [4] proposed a two-phase scheduling algorithm that combines OLB (Opportunistic Load Balancing) and LBMM (Load

Balance Min-Min) scheduling algorithms to utilize better executing efficiency and maintain the load balancing of the system. This approach helps in an efficient utilization of resources and increases the work efficiency. It gives the better results than the existing algorithms.

### *J. Min-Min Algorithm*

It starts with a set of unassigned tasks. firstly, minimum time for the completion for all the tasks is found.
Then the minimum number of times , the minimum value is selected in which the minimum times among the tasks on the resources. After that according to that minimum time, the scheduling of the task is done on the corresponding machine. After that the execution time of all other tasks is

updated on that machine by adding the execution time of all the assigned task to the number of execution times of other tasks for that machine and all the assigned tasks are removed from the list of the tasks that are to be assigned to the machines. The same pattern is followed again until all the assigned tasks are on the resources. But this approach has a major drawback that it can lead to starvation [7].

## V. COMPARISON OF EXISTING LOAD BALANCING TECHNIQUE

Below table show the comparative study of different load balancing. Difference made on bass of techniques that are used in respective algorithms, advantages and disadvantages.

Table 1: Comparisons of different load balancing algorithms

| Algorithm | Description | Advantages |
|---|---|---|
| Dynamic Round-Robin[1] | The first rule Avoid adding extra virtual machines to a retiring physical machine. The second rule speeds up the consolidation process and enables Dynamic Round-Robin to shutdown physical machines, | 1.power consumption is reduced.<br>2.Save power 3% more than power- sever implementation |
| Hybrid[1] | Combination of Dynamic Round Robin and First-Fit algorithms | 1.Reduce Power consumption.<br>2.Easy to implement.<br>3.Response time is high. |
| PALB[2] | maintains the state of all compute nodes, and based on utilization percentages, decides the number of compute nodes that should be operating. | 1.Simple<br>2.Easy to implement.<br>3.save energy. |
| ESCE[3] | The random selection based distributed problem round robin. Selection depend on least load. | 1.Response time is high.<br>2.Processig time also high.<br>3.Simple and easyto implement. |
| ***OLB+LBMM*** | S.-C. Wang et al. [4] proposed a two-phase scheduling algorithm that combines OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) scheduling algorithms to utilize better executing efficiency and maintain the load balancing of the system. | |
| ***Min-Min Algorithm*** | It starts with a set of unassigned tasks. firstly, minimum time for the completion for all the tasks is found. Then the minimum number of times , the minimum value is selected in which the minimum times among the tasks on the resources. After that according to that minimum time, the scheduling of the task is done on the corresponding machine | 1.The same pattern is followed again until all the assigned tasks are on the resources. |
| ECTC[4] | Two heuristics | 1.Energy consumption is |

| | | |
|---|---|---|
| | check every resource and identify most energy efficient resource for that task. | reduced. |
| MaxUtil[4] | Task consolidation decision based on resource utilization. | 1.Energy consumption is reduced.<br>2.Utilization of small no of resources. |
| ACCLB[7] | Uses small-world and scale-free characteristics of complex network to achieve better load balancing | 1.Overcomes heterogeneity<br>2. Adaptive to dynamic environment<br>3.Excellent in fault tolerance<br>4Good scalability |
| MCMF[8] | It is based on adirected graph representation ofthe dynamic resource allocation problem. | 1. Simple<br>2. Easy to implement.<br>3.Cost Effective in resource utilization. |
| Join-Idle-Queue[12] | 1.First find availability of the idle processors at each dispatcher<br>2. Then assigns jobs toprocessors to reduce average queue length of jobs at each processor | 1. Effectively reduces the system load<br>2.Incurs no communication overhat job arrivals. |

## VI. CONCLUSION

Cloud Computing has widely been adopted by the industry or organization though there are many existing issues like Load Balancing, Virtual Machine Consolidation, Energy Management, etc. which have not been fully implemented. Central to these issues is the issue of load balancing, that is required to distribute the excess dynamic local workload equally to all the nodes in the whole Cloud to achieve a high user satisfaction. It also ensures that every computing resource is distributed efficiently and fairly. Existing Load Balancing techniques that have been studied, mainly focus on reducing overhead, service response time and improving performance etc., and some of the techniques have considered the energy consumption factors. Therefore, there is a need to develop an Energy-aware load balancing technique that can improve the performance of cloud computing along with maximum resource utilization, in turn reducing energy consumption.

## References

[1].Ching-Chi Lin, Pangfeng Liu, Jan-Jan Wu. “Energy-Efficient Virtual Machine Provision Algorithm for Cloud System”, IEEE 4th International Conference on Cloud Computing,81-88, 09/2011.

[2]. Jeffrey M. Galloway, Karl L. Smith, Susan S. Vrbsky. “Power Aware Load Balancing for Cloud Computing”, Proceedings of the World Congress on Engineering and Computer Science 2011 Vol.

IWCECS 2011, October 19-21, San Francisco, USA,.

[3]. Jaspreet kaur "Comparison of load balancing algorithms in a Cloud" 2012, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 3, pp.1169-1173.

[4].R. Yamini, "Power Management In Cloud Computing Using Green Algorithm", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012), March 30, 31,2012.pp-128-133.

[5]. R. Yamini, "Energy Aware Green Task Assignment Algorithm In Clouds", Ianternational Journal For Research In Science And Advance Technology, Issue-1,Volume-1,pp-23-29.

[6]. Anton Beloglazov, Rajkumar Buyya, "Managing Overload Host For Dynamic Consolidation Of Virtual Machines Cloud Data Centers Under Quality Of Service Constraints", IEEE Transaction On Parallel And Distributed Systems, 2012.

[7]. Zehua Zhang, Xuejie Zhang, "A Load Balancing Mechanism Bassed On Ant Colony And Compel Network Theory In Open Cloud Computing Federation", IEEE-International Conference On Automation, May 2010, pp-240-243.

[8]. Makhlouf Hadji, Djamal Zeghlache, "Minimum Cost Maximum Flow Algorithm For Dynamic Resource Allocation In Cloud", IEEE-Fifth International Conference In Cloud Computing, Aug-2012, pp-876-882

[9].Wenhong Tian, Yong Zhao,Minxian Xu, Chen Jing, "A Dynamic And Integrated Load Balancing Scheduling Algorithm For Cloud Data Center", Proceeding of IEEE CCIS Feb2011,pp-311-10]. Zenon Chaczko , Venkatesh Mahadevan , Shahrzad Aslanzadeh and Christopher Mcdermid "Availability and Load Balancing in Cloud Computing", 2011 International Conference on Computer and Software Modeling IPCSIT vol.14 , IACSIT Press, Singapore.

[11]. Rajkumar Buyya, Anton Beloglazov, and Jemal Abawajy,"Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges", Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2010), Las Vegas, USA, July 12-15, 2010.

[12]. Trieu C. Chieu, Hoi Chan, "Dynamic Resource Allocation Via Distributed Decisions In Cloud Environment", Eight IEEE International Conference on e-Business Engineering, Sept-2011, pp-125-130.

[13]. Sivadon Chaisiri, Bu-Sung Lee, "Optimization Of Resource Provisioning Cost In Cloud Computing", IEEE transaction on services computing, vol. 5, No. 2 June 2012

[14]. Ayman G. Fayoumi, "Performance Evaluation Of A Cloud Based Load Balancer Severing Pareto Traffic", Journal of Theoretical and Applied Information Technology ,15th October 2011. Vol. 32 No.1

[15]. Anton Beloglazov and Rajkumar Buyya, Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers, Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science (MGC 2010, ACM Press, New York, USA), In conjunction with ACM/IFIP/USENIX 11th International Middleware Conference 2010, Bangalore, India, November 29 - December 3, 2010.

[16].OpenNebula http://opennibula.org/ [17]. Cloudbus http://www.cloudbus.org/ [18]. Nimbus http://www.nimbus.com/