# Green Computing On Cloud with Energy Optimal Application Scheduling

**Jeevaraj R[1], Ravindra Prasad S[2]**
[1, 2] Department of CSE
[1, 2] RRIT, Bangalore

***Abstract-*** *Cloud computing is emerging as a new paradigm of large-scale distributed computing. It is a framework for enabling convenient, on-demand network access to a shared pool of computing resources. Load balancing is one of the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed. It helps in optimal utilization of resources and hence in enhancing the performance of the system. The goal of load balancing is to minimize the resource consumption which will further reduce energy consumption and carbon emission rate that is the dire need of cloud computing. This determines the need of new metrics, energy consumption and carbon emission for energy-efficient load balancing in cloud computing. This paper discusses the existing load balancing techniques in cloud computing and further compares them based on various parameters like performance, scalability, associated overhead etc. that are considered in different techniques. It further discusses these techniques from energy consumption and carbon emission perspective.*

***Keywords-*** Cloud computing, Load balancing, Energy efficiency, Green computing.

## I. INTRODUCTION

The cloud computing technology is an attractive field in the area of computer science. All kinds of requirements posed by the user can be accomplished through cloud computing. There are numerous ways to define the term 'cloud computing'. NIST defines cloud computing as “as a model for enabling ubiquitous, convenient, on-demand network access to a shared Pool of configurable computing resources (example: network, servers, storage, applications and services) that can be rapidly that can be rapidly provisioned and released with minimal management effort or service provider interaction. From the definition, it is obvious that, for a cloud computing system to be worthy of the definition, we should first focus on the primary concept that is necessary for the efficiency of the system.

This concept is called load balancing. At present, there is a lot of research underway in the field of load balancing. However, load balancing is not an easy task. There are a number of policies available for load balancing. Some of the policies are static, while others are dynamic. In spite of the many load balancing schemes available, if these load balancing schemes are not effectively used, it becomes impossible to accomplish the definition of cloud computing. This paper discusses how this can be done effectively and efficiently. The tasks provided by the user go through different levels. Only if load balancing is achieved at each and every level, will there be any benefit to the user. Data centres in association with a cloud computing system could be located anywhere in the world. The first priority involves selecting the correct data centre. If this is done effectively, we can say that almost 10% of the work is done and we are one step closer to reaching the definition of a cloud computing system.

Similarly, load balancing schemes have to be correctly used in the remaining levels. The decision regarding which algorithm to use at each level has to be made correctly. As mentioned before, there are lots of load balancing schemes, but their effectiveness varies across levels. Using a particular scheme in one level may be optimal, but using the same algorithm in another level may not be as effective as using an alternative technique. On inspection of each task, it becomes obvious that each node may not be equally capable of handling each task. Understanding the capacity of each node is necessary in assigning tasks to each node. However, if these assignments are not done in a controlled fashion, it may lead to an imbalance situation. Brokers that act between the user and the data centre should take this into consideration. Many companies provide cloud computing services. Satisfying the requirements of the user is a prime factor for every cloud provider. So, the basic objective of each cloud provider is to correctly complete the user’s task within the stipulated time.

Green Computing or Green IT, is the practice of implementing policies and procedures that improve the efficiency of computing resources in such a way as to reduce the energy consumption and environmental impact of their utilization. As High Performance Computing (HPC) is becoming popular in commercial and consumer IT application, it needs the ability to gain rapid and scalable access to high-end computing capabilities. This computing infrastructure is provided by cloud computing by making use of datacentre’s.

It helps the HPC users in an on-demand and payable access to their applications and data, anywhere from a cloud. Cloud computing data-centres have been enabled by high-speed computer networks that allow applications to run more efficiently on these remote, broadband computer networks, compared to local personal computers. These data-centres cost less for application hosting and operation than individual application software licenses running on clusters of on-site computer clusters. However, the explosion of cloud computing networks and the growing demand drastically increases the energy consumption of data-centres, which has become a critical issue and a major concern for both industry and society. This increase in energy consumption not only increases energy cost but also increases carbon-emission. High energy cost results in reducing cloud providers' profit margin and high carbon emission is not good for environment.

## II. RESEARCH CHALLENGES

The term server consolidation is used to describe:

(1) Switching idle and lightly loaded systems to a sleep state;
(2) Workload migration to prevent overloading of systems; or
(3) Any optimization of cloud performance and energy efficiency by redistributing the workload.

### a. Server consolidation policies.

Several policies have been proposed to decide when to switch a server to a sleep state. The reactive policy responds to the current load; it switches the servers to a sleep state when the load decreases and switches them to the running state when the load increases. Generally, this policy leads to SLA violations and could work only for slow-varying, predictable loads.SLA violations one can envision a reactive with extra capacity policy when one attempts to have a safety margin and keep running a fraction of the total number of servers, e.g., 20% above those needed for the current load. The Auto Scale policy [4] is a very conservative reactive policy in switching servers to sleep state to avoid the power consumption and the delay in switching them back to running state. This can be advantageous for unpredictable spiky loads.

### b. Optimal policy.

We achieve an optimal policy as one which guarantees that running servers operate within their optimal energy regime or for limited time in a suboptimal regime, and, at the same time, the policy does not produce any SLA violations. At this time SLAs are very general and do not support strict QoS guarantees, e.g., hard real-time deadlines. Deferent policies can be ranked by comparing them with an optimal policy. The mechanisms to implement energy-aware application scaling and load balancing policies should satisfy several conditions: (i) Scalability - work well for large farms of servers; (ii) Effectiveness - lead to substantial energy and cost savings; (iii) Practicality - use efficient algorithms and require as input only data that can be measured with low overhead and, at the same time, accurately reacts the state of the system; and last, but not least, (iv) Consistency with the global objectives and contractual obligations specified by Service Level Agreements.

## III. THE SYSTEM MODEL

The model introduced in this section assumes a clustered organization of the cloud infrastructure and target primarily the IaaS cloud delivery model represented by Amazon Web Services (AWS). AWS supports a limited number of instance families, including M3 (general purpose), C3 (compute optimized), R3 (memory optimized), I2 (storage optimized), G2 (GPU) and so on. An instance is a package of system resources; for example, the c3.8xlarge instance provides 32 vCPU, 60 GIB of memory, and 2 320 GB of SSD storage.AWS used to measure the server performance in ECUs (Elastic Compute Units) but has switched recently to, yet to be specified, vCPU units; one ECU is the equivalent CPU capacity of a 1:0 to 1:2 GHz 2007 Opteron or 2007 Xeon processor.

We consider three levels of resource allocation decision making:

(a) The local system which has accurate information about its state; (b) the cluster leader which has less accurate information about the servers in the cluster; and (c) global decisions involving multiple clusters. We are only concerned with in-cluster scheduling coordinated by the leader of the cluster. Inter-cluster scheduling is based on less accurate information as the leader of cluster C exchanges information with other leaders less frequently.

**System and application level resource management.**

The model is based on a two-level decision making process, one at the system and one at the application level. The scheduler of the Virtual Machine Monitor (VMM) of a server interacts with the Server Application Manager (SAM) component of the VMM to ensure that the QoS requirements of the application are satisfied. SAM gathers information from individual application managers of the VMs running on the server.

**Model parameters**.

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

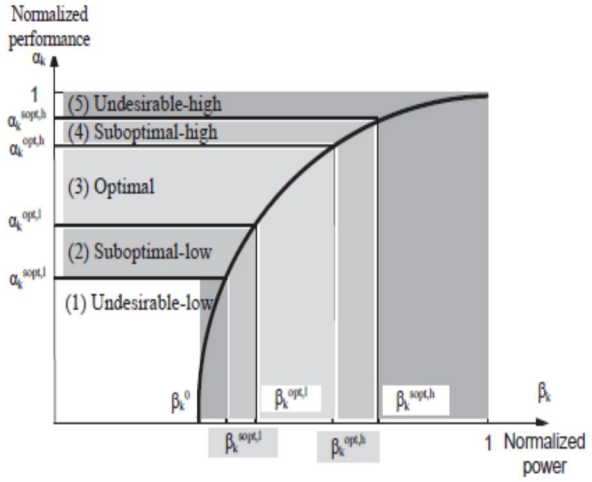


Figure 1: Normalized Performance Versus Normalized Power; boundaries of five operating regimes as shown.

## IV. PROPOSED APPROACH

In the proposed approach, different load balancing algorithms are used. Each and every task will pass through different stages. And ultimately, each task will reach its intended destination, referred to as the host. The overall system architecture is shown in Figure2. The three stages are given below:

1. Task to Data centre Controller
2. Data centre controller to each partition
3. Finally the task is assigned to each host

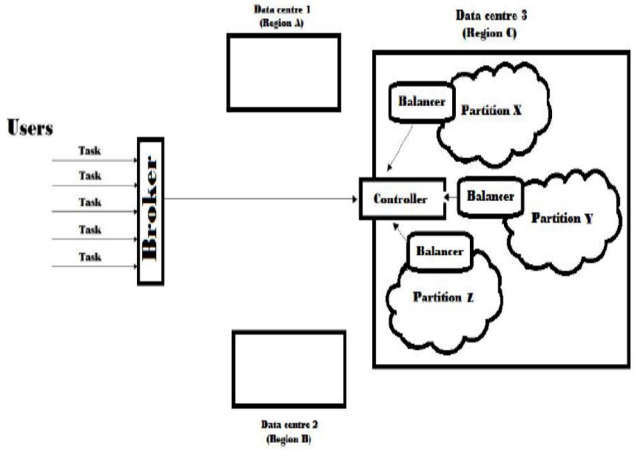


Figure 2: Overall Architecture.

### First Stage: Task to Data centre Controller

In the first stage, each task from the user has to reach a suitable data centre. The decision regarding this allocation is done by considering the locality of the data centre. The nearest data centre is chosen. This is done with the intention of reducing the execution time of the task. Figure3.shows the model of this stage. The responsibility of this task is carried out by the cloud broker. Thus it is necessary for the broker to have a mechanism to locate the nearest data centre

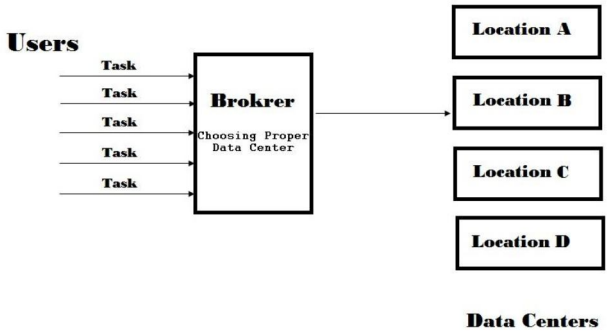


Figure 3:Stage1-Task to Data center Controller

### Second Stage: Data centre controller to each partition

In the next level, note that the task has been allocated to a data centre. Each data centre has its own controller. Each data centre has a number of partitions and each partition has its own nodes (hosts). In the second stage, the controller chooses a suitable partition by considering the load status. For this, each partition has a balancer. This balancer maintains a status of loads on different hosts under that partition. This status has to be periodically updated and sent to the main controller. The partition can be of 3 types:

**1. Idle:** The partition is idle when the percentage of idle nodes exceeds α

**2. Normal:** The partition is idle when the percentage of normal nodes exceeds β

**3. Overloaded:** The partition is overloaded when the percentage of overloaded nodes exceeds γ

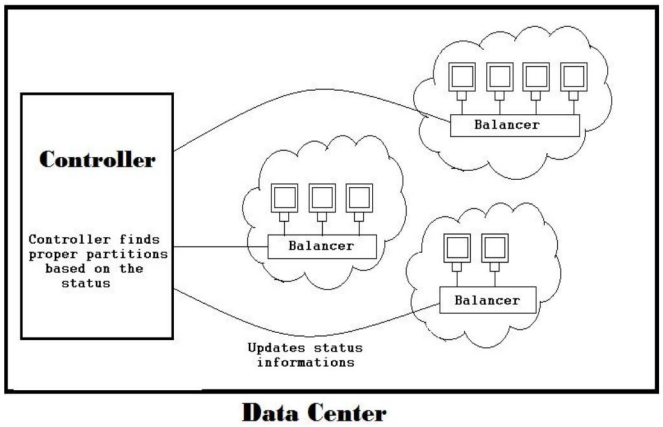


Figure 4: Second Stage- Data center controller to each partition

**Third Stage: Finally task assigned to each host**

In the third stage, the tasks are assigned to corresponding hosts. This can be done in 2 different ways. The load degree has to be computed for each node in the partition. The load degree is computed from some static and dynamic parameters. The static parameters are number of CPUs, memory size etc. Dynamic parameters are memory utilisation ratio, CPU utilisation ratio, network bandwidth

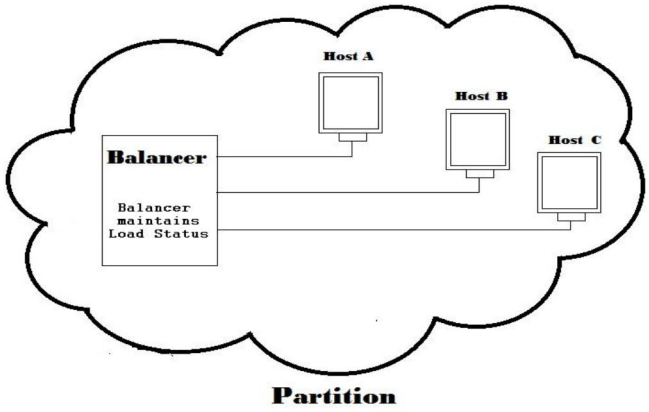


Figure 5: Third Stage-Task assigned to each host.

## V. GREEN COMPUTING IN CLOUDS

Green Computing [6], or Green IT, is the practice of implementing policies and procedures that improve the efficiency of computing resources in such a way as to reduce the energy consumption and environmental impact of their utilization.

As High Performance Computing (HPC) is becoming popular in commercial and consumer IT applications, it needs the ability to gain rapid and scalable access to high-end computing capabilities. This computing infrastructure is provided by cloud computing by making use of data centres.

It helps the HPC users in an on-demand and payable access to their applications and data, anywhere from a cloud [7].

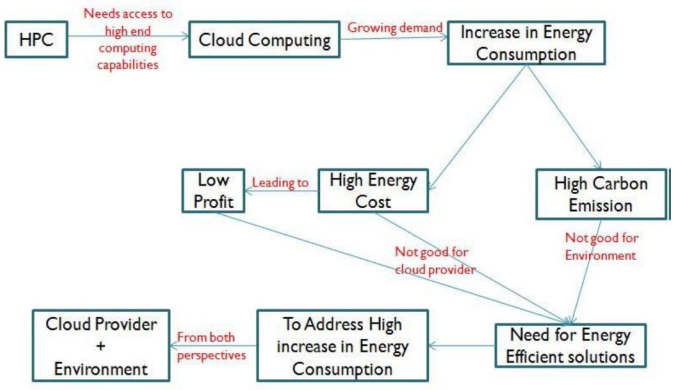


Figure 6: Green Computing in Clouds

Here is our proposed algorithm for Job submission and scheduling:

- Client submits the job to the job scheduler.
- Job scheduler checks for the availability of resources through VM monitor.
- VM monitor checks for scheduling table containing threshold value & net congestion on the slaves.
- If the threshold value < the maximum & no congestion then accept the job & send it to the job queue, else reject.
- If the job is accepted, for every job, go to VM assigner.
- If there are under loaded CPU of VM, assign the job to that VM and update the database of the scheduler.
- Else CPU of VM is overloaded and returns the job to the queue.

Here we can categorize the Load to LOW, MEDIUM and HIGH classes. The node is LOW when it can accept load from other nodes. A node is HIGH loaded when it is detected to be higher than the upper threshold as defined in the previous protocol. The main difference is in the MEDIUM load definition. A node is MEDIUM if it has a load above the maximum limit of LOW load and can accept load from other nodes to reach the upper threshold, which is called MEDIUM MAX. Thus, MEDIUM loaded nodes do not need to give loads to LOW loaded nodes since they are not overloaded. Beside the categorization of nodes, clusters can be classified as LOW, MEDIUM and HIGH as regards to the sum of all local node loads in order to get a global point of view. Two main target of this scheduling for load balancing protocol is to distribute loads of HIGH load nodes and group of nodes which are called clusters to LOW & MEDIUM nodes and clusters to reach global stable load distribution.

## V. CONCLUSION

We have proposed a load balancing approach for cloud computing environment. It has been noted that traditional load balancing algorithms are usually not flexible and cannot match the dynamic changes to the attributes during the execution time. Dynamic algorithms are more flexible and take into consideration different types of attributes in the system both prior to and during run-time.

Load Balancing techniques that have been studied, mainly focus on reducing overhead, service response time and improving performance etc., but none of the techniques has considered the energy consumption and carbon emission factors. Therefore, there is a need to develop an Energy-efficient load balancing technique that can improve the performance of cloud computing along with maximum resource utilization, in turn reducing energy consumption as

well as carbon emission to an extent that will help achieve Green Computing.

## ACKNOWLEDGMENT

We would like to thank Management of RRIT for providing such a healthy environment for the successful completion of this work and express my gratitude to Mr. Ravindra Prasad S (HOD & Assoc. Professor, RRIT, Bangalore) for providing continuous support and encouragement. Last but not the least I thank all my friends who has continuous support in all my works.

## REFERENCES

[1] Saurabh Kumar Garg and Rajkumar Buyya, "Green Computing and Enviromental Sustainability"

[2] Gregor Von Laszewski, Lizhe Wang, Andrew J. Younge and Xi He, "Power-Aware Scheduling of Virtual Machines inDVFS-enabled Clusters".

[3] FeiFei Chen, Jean-Guy Schneider, Yun Yang, John Grundy, and Qiang He "An Energy Consumption Model and Analysis Tool for Cloud Computing Environments", GREENS 2012, Zuricg, Switzerland, pp. 45-50.

[4] A. Suphalakshmi and Sreejith M, "An intelligent, energy conserving load balancing algorithm for the cloud environment using ant's stigmergic behavior", International Journal of Communications and Engineering Volume 04– No.4, Issue: 03 March2012.

[5] Er. Navdeep Kochhar and Er. Arun Garg," ECO-FRIENDLYCOMPUTING GREENCOMPUTING", International Journal of Computing and Business Research ISSN (Online): 2229-6166.Volume 2 Issue 2 May 2011.

[6] Liang Liu, Hao Wang, Xue Liu, Xing Jin, WenBo He, QingBo Wang and Ying Chen, "Green Cloud: A New Architecture for Green Data Center", pp. 29-38.

[7] B. Snyder. "Server virtualization has stalled, despite the hype."http://www.infoworld.com/print/146901 (Accessed on December 6, 2013).

[8] B. Urgaonkar and C. Chandra. "Dynamic provisioning of multi-tier Internet applications." Proc. 2nd Int. Conf, on Automatic Computing, pp. 217-228, 2005.

[9] H. N. Van, F. D. Tran, and J.-M. Menaud. "Performance and power management for cloud infrastructures." Proc. IEEE 3rd Int. Conf. on Cloud Computing, pp. 329–336, 2010.

[10] A. Revar, M. Andhariya, D. Sutariya, M. Bhavsar, Load Balancing In Grid Environment Using Machine Learning-Innovative Approach, International Journal Of Computer Applications 8 (10 (Oct)) (2010) 975–8887.

[11] Shridhar G.Damanal and G. Ram Mahana Reddy, Optimal Load Balancing In Cloud Computing By Efficient Utilization of Virtual Machines – Ieee 2014.

[12] Shridhar G. Domanal And G. Ram Mohana Reddy," Load Balancing In Cloud Computing Using Modified Throttled Algorithm" Ieee, International Conference. Ccem 2013. In Press.

[13] Brototi Mondal, Kousik Dasgupta And Paramartha Dutta, "Load Balancing In Cloud Computing Using Stochastic Hill Climbing-A Soft Computing Approach" In Procedia Te Chnology 4 ( 2012 ) 783 - 789,Elsevier C3it-2012.

[14] Q. Cao, B. Wei and W. M. Gong, "An Optimized Algorithm for Task Scheduling Based On Activity Based Costing In Cloud Computing," In International Conference on Esciences 2009, Pp. 1-3.

[15] Monika Choudhary, Sateesh Kumar, A Dynamic Optimization Algorithm for Task Scheduling In Cloud Environment, Peddoju International Journal Of Engineering Research and Applications (Ijera), May-Jun 2012, Pp.2564-2568.