International Journal of Advanced Engineering and Technology
ISSN: 2456-7655; Impact Factor: RJIF 5.54
Received: 17-04-2019; Accepted: 20-05-2019
www.newengineeringjournal.com
Volume 3; Issue 3; July 2019; Page No. 47-52

Energy efficient algorithms in cloud computing: A green computing approach

Mayur Jain[1], Aayushi Priya[2]
[1] B. Tech Scholar, Department of CSE Sagar Institute of Research and Technology, Bhopal, Madhya Pradesh, India
[2] Professor, Department of CSE Sagar Institute of Research and Technology, Bhopal, Madhya Pradesh, India

Abstract
Cloud Computing has emerged as a popular technology that support computing on demand services by allowing users to follow the pay-per-use-on-demand model. Minimizing energy consumption in cloud systems has many benefits that enable green computing. Energy aware task scheduling in cloud to the users by service cloud providers has non negligible influences on optimal resources utilization and thereby on the cost benefit. The traditional algorithms for task scheduling are not well enough for cloud computing. In such environment, tasks should be efficiently scheduled such a way that the makespan is reduced. Load balancing is a significant area of cloud computing environment which ensures that all connected devices or processors carry out same amount of work in equal time. With an aim to make cloud resources and services accessible to the cloud user easily and conveniently, different algorithms and models for load balancing in cloud computing is being developed. This paper aims to study an energy efficient load balancing algorithms that is intended to minimize performance parameters like make span, latency, total execution time and energy consumption.

Keywords: Cloud computing, load balancing, energy efficient, execution factors, latency, execution time, energy consumption

1. Introduction

Cloud Computing, the internet based computing that won lots of momentum for its flexibility and physical property, provides shared data and process resources as services permitting users to not got to have their own infrastructure and follows the pay-as-you-go model. The cloud system has the potential to transform a large a part of the IT industry since it makes software even a lot of attractive as a service and shaping the means during which IT hardware is designed and purchased [1].

This technique aims at power subsequent generation data centers with each applications delivered services through the web and therefore the hardware in addition as software so as to face with the growing demand of user Quality of Service (QoS) and Quality of experience (QoE). These services are usually computer applications, software development platforms in addition as computing and storage resources. These services have long been referred as software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Nowadays, billions of individuals endlessly use cloud services that embody online gaming, social networking, web hosting, content delivery, streamed contents in addition as scientific applications. Every of those applications present different options that include configuration and deployment requirements. Moreover, cloud suppliers are continuously looking techniques for permitting continuous services providing and in fact for maximising their profit whereas permitting a good QoS/QoE. What is more, the cloud system essentially depends on virtualization, particularly on virtual machines.

Therefore, because of the heterogeneity of computer and storage within the cloud infrastructure scheduling of tasks upon virtual machines may be a crucial task [2, 3]. Because of the user's high demand, the cloud knowledge centers are very difficult. Distributed computing has transformed the information technology industry by some ways and cloud computing is changing into one powerful competitor in distributed system paradigm. In cloud computing user request for the resources as they required from numerous distributed data centers all across the world and pays for it as per the usage of it by using service level agreement (SLA) between user and cloud vendor. Nevertheless Cloud data centers consumes ample quantity of electrical energy occasioning large operating prices and CO2 emission. In 2020, energy consumption by data centers worldwide was predicted to be between 2.5 % and 4 % of the worldwide electricity use and is probably going to grow in upcoming years [5].

An important aspect of big data is that the use of data centers that house computer systems and associated parts, like telecommunications and storage systems. In step with revealed studies the worldwide expenditure on enterprise power offer and cooling of those systems has been estimated to be quite $30 billion. Hence, achieving energy potency in data centers has become an issue of dominant importance [8 [9].

The requirement for optimizing big knowledge computing is presently not an choice however has now become a vital demand for environmentally property computing. During this context, we will use the ideas of green computing that's the study and apply of environmentally sustainable computing. The matter of optimizing computing with reference to environmental considerations may be tackled from different perspectives.

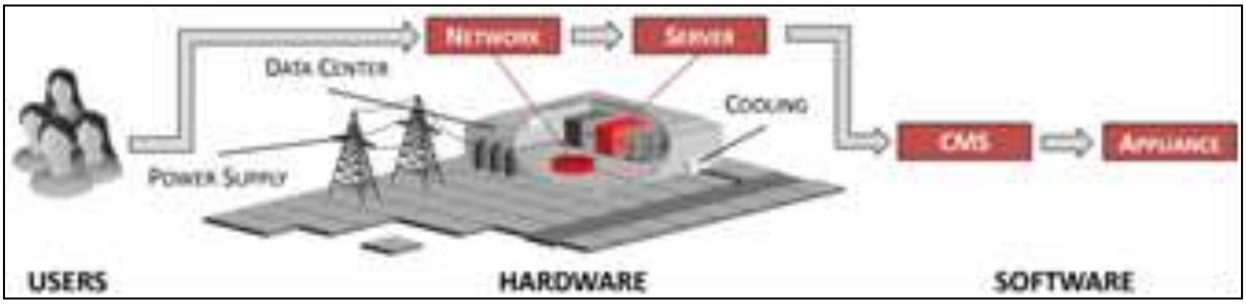

Fig 1: Cloud computing data center domains

Big data systems will be characterized as a parallel and distributed system consisting of the many nodes with software and data components. In parallel computing, the parallel algorithms are outlined that may be executed at the same time on multiple nodes to benefit from the parallel computing power. As such, increasing the process nodes can increase the performance of the parallel programs. An important challenge during this context is that the mapping of parallel algorithms on a computing platform that consists of multiple multiprocessing nodes. A parallel rule will be mapped in numerous alternative routes to the process nodes. Every mapping alternative can perform otherwise for power consumption that's necessary in green computing. During this perspective, choosing a possible mapping of parallel algorithm to computing platforms has become intractable for the human parallel computing engineer. Existing studies on deployment of parallel computing algorithms have chiefly centered on addressing general computing metrics. For analyzing the alternatives with regard to green computing, these metrics don't seem to be adequate and that we got to determine the metrics which will be used to measure the impact of a selected deployment alternative on the environmental parameters (such as power consumption).

2. Related Work

Yadav *et al*. [6], proposed a particle swarm based algorithm that can balance the load in cloud computing so that resources are easily available for users. This paper aims to develop an efficient load balancing algorithm using particle swarm based to minimize performance parameters like make span, latency, total execution time.

Shubham Sidana *et al*. [7] proposed an algorithm to balance load on cloud based on arrangement of resources, according to processing speed for virtual machines and then allocating cloudlets to the resources according to their processing requirement. This algorithm allocates the resources in such a manner that job requiring less processing are not allocated to the machines with high processing power and vice versa.

Gu *et al*. [8] for proposing a scheduling strategy for cloud based environment. Based on the GA, authors proposed a virtual machine load balancing scheme that allows an optimal resource using the cloud data centers.

In the same order of ideas, Ge *et al*. [9] addressed the task scheduling problem with GA by evaluating all jobs in the job queue. The results of their experiments outputted better load balancing.

A multi-objective GA have been used by Liu *et al* [10]. for improving the overall performance of cloud computing. Authors designed a task scheduling model for reducing the system power consumption while improving the profit of service providers by providing a dynamic selection mechanism according to the real time requirements.

Zhu *et al* [11]. introduced a new resource scheduling strategy based on the ACO algorithm for promoting a new business calculation model in cloud computing environments. In their proposed design, classification of users tasks based on priority in order to ensure requirements of different QoS metrics, is achieved for allocating resources and scheduling tasks. Their proposal showed better makespan compared a random distribution algorithm. However, the proposal does not take into account the variation of task size that is an important target of the scheduling algorithms in cloud environments.

Wei *et al* [12]. extended the task scheduling problem to the mobile cloud computing environments by extending the Cloudlet architecture. Authors have taken each tasks profit into consideration in order to maximize the profit of the system, which is an import target of the task scheduling algorithm in the commercial mobile cloud environment. They designed their proposal based on the hybrid ACO algorithm, which has been validated by experiments. Moreover, heterogeneous cloud are usually considered by the cloud providers in order to well manage the utilization of their computing resources.

Dai *et al*. [13] analyzed the structure of heterogeneous cloud, and proposed a framework of multiobjective constrained resource management that extended the computing power and the availability of cloud resources.

Then, in order to highlight a tradeoff among performance, availability, and cost of Big Data application running on Cloud infrastructures, Dai *et al*. [14] have modeled a mechanism of resource management for heterogeneous clouds by a multiobjective optimization algorithm.

Hussain A Makasarwala, *et al* [15]. gives a genetic algorithm (GA) based approach for load balancing in cloud. For population initialization, priority of request is considered based on their time.

After analyzing the existing work it is concluded that in Cloud Computing environment, there are some issues such as memory usage, delay in network due to heavy load or CPU load among cloud resources. Cloud environment is created by producing virtual resources of the actual available resources and sharing them among users or clients. In any situation when the total number of user to the particular virtual machine (VM) exceeds, the load balancing server will schedule the incoming users request on a new virtual machine.

3. VM Scheduling

The assignment of a task by the scheduler is subjected to variety of constraints. Constraints are generally either time constraints or resource constraints. A task could embrace data entry and processing, software access, and storage functions. The datacenter classifies tasks consistent with the service-level agreement and requested services. Every task is then assigned to at least one of the offered servers. In turn, the servers perform the requested task. A response or result is transmitted back to the user [11].

Scheduling could be a balancing situation during which processes or tasks are regular as per the given necessities and used algorithm. The goal of scheduling algorithms in distributed systems is to unfold the load on the processors and to maximize their utilization whereas minimizing total task execution time. Job scheduling, one among the foremost known optimization issues, plays a key role to enhance versatile and reliable systems [12]. The most purpose is to schedule jobs to the flexible resources in accordance with adaptable time that involves sorting out a correct sequence during which jobs are often executed under transaction logic constraints.

In Cloud Computing VM scheduling algorithms are Accustomed schedule the VM requests to the Physical Machines (PM) of the particular data Center (DC) as per the need fulfilled with the requested resources (i.e. RAM, Memory, bandwidth etc). In today's era there are such a large amount of cloud suppliers in market that have different capacity of data Centers and Physical Machines available. Generally planning algorithm works in three levels as given below [9]:

A. For the set of VMs find the appropriate Physical Machine.
B. Determine the proper provisioning scheme for the VMs.
C. Schedule the tasks on the VMs

Figure 2 shows the components of cloud computing scheduling. As shown in the figure, the scheduling model in a cloud datacenter consists of four components, namely, computing entity, job scheduler, job waiting queue, and job arrival process [10].

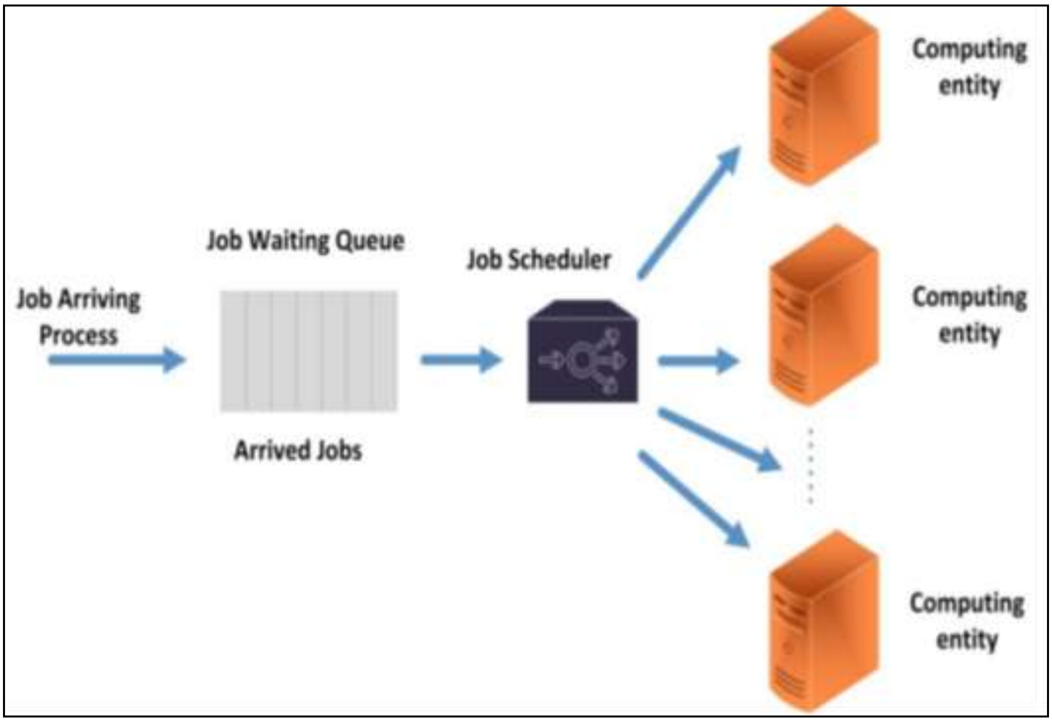

Fig 2: Scheduling Model in Cloud

A. Computing entity is provided through the implementation of a virtualization technique in the cloud computing system. A number of virtual machines that provide computing facilities, such as the operating system and software, are present in the cloud system to process the submitted tasks. A computing entity is characterized by its computing capacity, which indicates the number of instructions it can process in a second [13].
B. Job scheduler is an important component of the scheduling process in a cloud computing.
C. Job waiting queue is the line of jobs for execution waiting to get assigned to a particular machine.
D. Job arrival process is the procedure in which jobs arrive into the scheduling system.

4. Energy efficient load balancing

Load balancing [7] is technique of reallocating the complete load to the distinct nodes of the collective system to form resource utilization effective and to advance the latency of the work latency, all at once eliminating a circumstance throughout that range of the nodes are over loaded whereas some others are under loaded. Consequently Load balancing is an particularly technique that allows networks and assets by suggests that of supplying a most throughput with minimal response time through dividing the traffic between servers. This load taken into thought are C.P.U. load, quantity of memory used, delay or network load.

In a typical cloud system each virtual machine within the cloud data center will the equivalent quantity of work throughout the system and is full with request; thus load balancing is indispensable for increasing the output and lessens the interval to equally distribute workload amidst all offered servers. The load of a machine will be balanced by shifting the workload, dynamically to remote nodes or machines that are under-utilized. Because of this, the user satisfaction is maximized by minimizing latency, escalating resource utilization, lowering the amount of job rejections because of delay and enhancing the performance ratio of the system [3]. Load balancing aid in up the general power efficiency of the information centre. it is used for correctly segregating the traffic between servers, by evading significant overload on the resources of a single targeted server. Data will be sent and received while not high communication delay and thereby minimizes the overall waiting time of the resources. The goal of an energy aware load balancer is to sustain availability to compute nodes for resource requests, whereas reducing the full energy consumed by the cloud infrastructure [4]. Dynamic load balancing is preferable in distributed systems thus on have the system workload to be balanced uniformly among all available nodes in a method that reduces the mean job response time to a larger extent [5]. Load balancing within the cloud computing domain is different from the classical and standard techniques and architectures of load balancing. Load balancing tries to avoid the unevenness in resource distribution within the system and uniformly balances the system load and achieves tremendous enhancements in performance.

The goal of any energy aware load balancer is to ensure availability of requested resources for computation and aid in minimizing the total energy consumed in the cloud data center. Energy efficiency can be defined as a reduction of energy used for a given service or level of activity [4]. However, due to scale and complexity of data center equipment it is extremely difficult to define unique service or activity that could be examined for its energy efficiency. Therefore, we identify four scenarios within a system where energy is not used in efficient way, but instead it is lost or wasted, as shown in Figure 3.

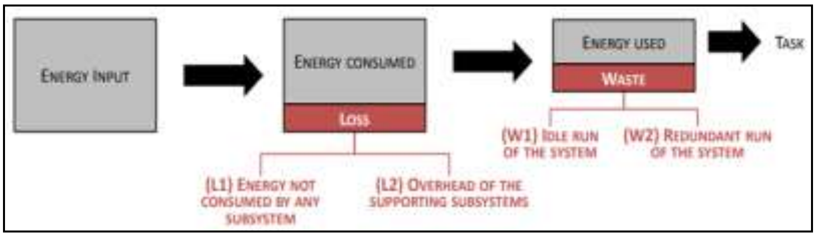

Fig 3: Scenarios where energy is lost or wasted within a system

Energy loss is referred as, (L1) i.e. energy not consumed by any of its subsystems, e.g., energy lost due to transport or conversion. This also includes (L2) energy overhead of the supporting subsystems, such as cooling or lighting within a data center, where provisioning of Cloud services is considered as its main purpose. Energy waste refers to an energy used for its main purpose, however for the (W2) idle run of the system, e.g., processor being turned on but running idle. Additionally, (W2) redundant run of the system is also considered as energy waste, e.g., keeping a cooling system at maximum during night when the temperatures are lower.

5. Energy Consumption Analysis

There are ample of parameters for concentrating on power consumption of distributed systems, starting form data structures, topologies of the nodes, replication of data at each nodes, its access rate, memory utilization of nodes at time of reading data as well writing data, transfer of data form hop to hop, whether it is using multicasting or broadcasting all these technical issues and their power usage are important to consider. In cloud infrastructure, a node refers to general multicore server along with its parallel processing units, network topology, power supply unit and storage capacity. The overall energy consumption of a cloud environment can be classified as follows [17]:

$E_{Cloud} = E_{Node} + E_{Switch} + E_{Storage} + E_{Ohters}$

Energy Efficient Resource Scheduling

Several research works have been carried out on energy efficient resource scheduling in virtual machines and grid systems. Many aspects such as: power consumption, reliability, response time etc. need to be considered while designing VM schedulers. A processing unit consumes power or energy as represented below:

$$E_i = \sum_{i=1}^{n} E$$

The number of tasks are considered as i = 1 to n and E_i is the energy consumption at each task i, therefore total energy consumption at all tasks is given by summation of E_i. Further E_i can be derived as [16]:

$E_i = N_w.e_w + N_r.e_r + e_s$

Where,

N_w = Number of memory write during task i

N_r = Number of memory read during task i

e_w = Power consumed during write operation

e_r = Power consumed during read operation

e_s = Normal energy consumption of Memory

Table 1: Comparison of Various Energy-aware Load Balancing Methods [17]

Load Balancing methods	RU	LPC	FRT	SO	HG
Energy-aware Load Balancing and Application Scaling for the Cloud Ecosystem	Y	Y	N	Y	N
Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm	N	Y	Y	N	N
Energy Efficient Load balancing Algorithm for Green Cloud	Y	Y	N	N	N
Predictive Load Balancing in Cloud Computing Environments based on Ensemble Forecasting	Y	N	Y	N	N
Improved GA Using Population Reduction For Load Balancing in Cloud computing	Y	N	Y	N	N

RU=Resource Utilization, LPC = Low Power Consumption, FRT = Fast Response Time, SO = Space Overhead, LHG= Less Heat Generated

6. Load Balancing Parameters

A set of parameters are taken into account when building a VM scheduling algorithm. These parameters play an important role to increase overall cloud performance which is described below:

A. *Makespan* is the total completion time of all tasks in a job queue. A good scheduling algorithm always tries to reduce the makespan. The makespan is defined as the maximum time to complete the ith task on the m_{th} VM [11].

Table I: Makespan Parameters Definitions

Parameters	Definition
t_i	i_{th} task
m_j	m_{th} virtual machine
c_i	time when task t_i arrives
a_j	time when virtual machine m_j is available
e_{ij}	execution time for t_i on m_j
c_{ij}	time when the execution of t_i is finished on m_j $c_{ij}=a_j + e_{ij}$
Makespan	maximum value of c_{ij}

B. ***Average Latency:*** Average latency is the ratio of total waiting time of tasks and number of these tasks.

$$\text{Average Latency} = \frac{\text{Total waiting time of all tasks}}{\text{Number of tasks}}$$

C. ***Execution time span:*** Execution time span is the time duration taken from beginning of first task, start processing and end with last task finished the processing.

Execution Time = Te – Ts

Where, Te is time of ending last task, and Ts is time of start first task.

D. **Average Turnaround Time:** Average turnaround time is the time between submission and completion of all tasks to the total number of submitted tasks.

$$\text{Average TUT} = \frac{\text{Total TUT of all tasks}}{\text{Number of tasks}}$$

E. **Energy consumption:** in cloud data centers is a current issue that should be given more consideration. Many scheduling algorithms were developed to reduce power consumption and improve performance. Thus, cloud services become environment-friendly.

7. Conclusion

The Cloud Computing infrastructure and its flexible nature can be utilized indirectly for energy optimization. Techniques such as powering on/off machines can be used for power regulation. Using such approach, data center's dynamic load can be used for regulating electricity demand and thus production, which is important for keeping optimal frequency of the power grid. Energy efficient load balancing is the most important task in Cloud Computing environment to reach successfully maximum utilization of Resources. Load Balancing is greatest challenge in cloud Computing. The major objective of this algorithm is to minimize the make span and latency. Not only that this approach reduces costs for the data center, it also reduces power losses due to an energy transfer and the load on an energy grid. In this paper a study is performed for some existing energy-efficient load balancing techniques and analyzed different parameters indicating merits and demerits of these algorithms.

8. References

1. M Armbrust *et al*. Above the Clouds: A Berkeley View of Cloud Computing, technical report. Univ. of California, Berkeley, 2009.
2. M Katyal, A Mishra. A comparative study of load balancing algorithms in cloud computing environment." in International Journal of Distributed and Cloud Computing, 2013, 1(2).
3. RG Rajan, V Jeyakrishnan. A survey on load balancing in cloud computing environments, in International Journal of Advanced Research in Computer and Communication Engineering, 2013; 2(12):4726–4728.
4. Teena Mathew, K Chandra Sekaran, John Jose. Study and Analysis of Various Task Scheduling Algorithms in the Cloud Computing Environment, IEEE, 2014, 658-664.
5. Pandaba Pradhan, Prafulla Ku Behera, BNB. Ray. Modified Round Robin Algorithm for Resource Allocation in Cloud Computing", International Conference on Computational Modeling and Security, ELSEVIER, 2016, 878-890.
6. Yadav R, Namdev M. A Study on Particle Swarm based Load Balancing Algorithms in Cloud Computing", IJOSTHE. 2018, 5(1). Retrieved from https://ijosthe.com/index.php/ojssports/article/view/99. DOI: https://doi.org/10.24113/ojssports.v5i1.99. Date accessed: 10 Jan 2019.
7. Shubham Sidana, Neha Tiwari, Anurag Gupta, Inall Singh Kushwaha. NBST Algorithm: A load balancing algorithm in cloud computing, IEEE, 2016, 1178-1181.
8. Gu J, Hu J, Zhao T, Sun G. A new resource scheduling strategy based on genetic algorithm in cloud computing environment," Journal of Computers. 2012; 7(1):42–52.
9. Ge Y. Wei G. GA-based task scheduler for the cloud computing systems," in Web Information Systems and Mining (WISM), International Conference on, vol. 2. IEEE, 2010, 181–186.
10. Liu J, Luo GX, Zhang XM. Zhang F, Li BN. Job scheduling model for cloud computing based on multi-objective genetic algorithm," IJCSI International Journal of Computer Science. 2013; 10(1)134–139.
11. Zhu L, Li Q, He L. Study on cloud computing resource scheduling strategy based on the ant colony optimization algorithm, IJCSI International Journal of Computer Science. 2012; 9(5):545-8.
12. Wei X, Fan J, Lu Z, Ding K. Application scheduling in mobile cloud computing with load balancing, Journal of Applied Mathematics, 2013.
13. Dai W, Chen H, Wang W. Rahec: A mechanism of resource management for heterogeneous clouds, IEEE, 2015, 40-45.
14. Dai W, Qiu L, Wu A, Qiu M. Cloud infrastructure resource allocation for big data applications, IEEE Transactions on Big Data, 2016.
15. Hussain Makasarwala A, Prasun Hazari. Using Genetic Algorithm for Load Balancing in Cloud Computing, IEEE, 2016.

16. Nitin S More, Rajesh B Ingle, Challenges in Green Computing for Energy Saving Techniques, International Conference on Emerging Trends & Innovation in ICT, 2017.
17. Nirmala M, Adi Lakshmi T. Energy-Efficient Load Balancing in Cloud: A Survey on Green Cloud, International Journal of Engineering Trends and Technology (IJETT), 2017.