Smart Cloud Load Balancing Approaches towards Green Computing

Arbab Kanwal[1], Muhammad Arslan Riaz[2]

[1, 2] Government College University, Faisalabad, Pakistan.
[1]arbab.mit@gmail.com, [2]arslanriaz3939@gmail.com

Abstract— Cloud computing is emerging as a new paradigm for distributed computing. A framework for enabling's convenient network access on demand to a common set of computing resources. Load balancing is one of the major challenges in cloud computing and is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed. It helps optimize the use of resources and thus enhance system performance The goal of load balancing techniques is to reduce load from one node to another by distribution, which is an urgent need for cloud computing. This identifies the need for new standards, energy consumption and carbon emissions to achieve energy-efficient load balancing in cloud computing. This paper discusses current load balancing techniques in cloud computing, comparing them on the basis of different parameters such as performance, scalability, overhead, etc. These approaches are also discussed from the perspective of energy consumption and the perspective of carbon emissions.

Index Terms— Cloud Computing, Green Computing, Load-Balancing Algorithms, Energy-efficient scheduling

u

1 Introduction

On account of the greatest achievement of web in last few years, computing resources is presently more existed and it empowered achievement of a novel registering idea known as Cloud Computing. Cloud Computing atmosphere requires conventional service suppliers to have two various ways These arc infrastructure and service providers Infrastructure providers course °faction of cloud stages and ant assets as indicated by use Service suppliers offer a sources from infrastructure providers to bolster end users Cloud Computing has lured the giant companies, i.e. Amazon, Google and Microsoft considered as a great impact in today's Information Technology companies (Chaczko et al, 2011).

Cloud computing is an on demand facility in which platform infrastructure and software are offered on demand as indicated by the customer's need at specific time Ifs a term regularly used on account of web. One can see whole tab as a cloud Subsequently all the above indicated facilities can be accessed by a resource as a customer to the cloud Presently at the fundamental idea of cloud computing, it has to offer assets / resources i e. VMs as facilities on request Appointing viable VM on request is being directed with the support of the load balancing algorithms in the cloud computing. As these algorithms have a critical impact while picking which VM is to be assigned out on quest of the subscriber

While offering facilities it is conceivable to have number of solicitations at once and in light of that some requestors require to stay in line however they have likelihood to forward their demand to other administration provider. Henceforth with the support of the load balancing algorithm endorser will be fit to choose whether they require staying in the line or getting facilities from the other administration provider. Number of the algorithms for the load balancing in cloud computing are existed for relegating the viable VMs Among such existed algorithms, which one is to be used, is the fundamental choice to be considered.

Some of those calculations / algorithms have been clarified / discussed in this paper. Henceforth, to have utilization of resources and being faithful with each resource, load balancing is being carried out.

2 Cloud COMPUTING

Cloud computing utilize a strategy for the web / internet and central remote servers to oversee and manage information i.e data and applications. Cloud computing licenses organizations and consumers to use applications without installation at any PC with internet access. This procedure licenses for substantially more incapable computing by centralizing storage, processing, memory and bandwidth. Cloud computing is a model of system processing where an application or program works on a connected server or servers rather than on a local computing device i.e. a PC, PDA or tablet Like the traditional client server model or chronic mainframe computing, a customer joins with a server to execution an assignment The distinction with cloud computing is that the registering procedure may work on one or a few connected PCs with the help virtualization. With virtualization, at least one physical server can be designed and partitioned into a few detached "virtual servers" all working freely and appear to the customer to be a single physical device. These virtual servers don't physically exists and can in this way be moved

In every way and flaky up or down on the fly without impacting the end subscriber. The computing resources have moved toward becoming "grainy ", which supply end user and administrator profit including wide access over a few gadgets / devices, on demand service. Resource pooling, fast elasticity and service reviewing capability. Cloud computing is a system of distributed computing that focuses on presenting an expansive scope of subscribers with appropriated access to virtualized software and hardware infrastructure over the internet. It incorporates networking, distributed computing virtualization, web software and web services and web administrations. Idea of cloud computing has focused on enthusiasm of subscribers towards distributed, parallel and virtualization computing systems today. It has appeared as well-known answer to offer simple and easy access to externalized IT resources. Through virtualization, cloud computing is competent to approach with the same physical foundation, a tremendous customer base with different computational necessities. The regular advancement in the region of cloud computing likewise increments basic security concerns Absence of security is the main issue selection of cloud computing.

3 LITERATURE REVIEW

Hussin et al., (2016) they concentrated on green cloud computing through scheduling optimization model. In particular, they explored a connection between performance metrics that picked in scheduling approaches with energy utilization for energy proficiency and efficiency. They trusted that better comprehension on the best way to demonstrate the scheduling performance will prompt green cloud computing. It is astute to viably oversee energy consumption in cloud computing; this would thus be exceptionally valuable cost of computation. In their work, they highlighted three level of system complexities; (1) homogenous or heterogeneous, (2) static or dynamic and (3) distributed. They proposed undertaking task scheduling methodology that goes about as middle person with a specific end goal to screen system performance and processing/handling power in a same time.

Singh and Chana (2014) in this paper, they stressed on the improvements of energy based resources scheduling framework and exhibited a calculation that consider the synergy between different data center infrastructure and performance. In particular, this paper proposed building standards for energy effective administration of clouds and energy proficient resources allocation strategies and planning calculation considering Quality of Services (QoS) viewpoints. The execution of the proposed algorithm has been assessed with the current energy based scheduling algorithms. The trial comes about showed that this approach is powerful in limiting the cost and vitality utilization of clouds applications in this manner moving towards the accomplishment of Green Clouds.

Rajan and Jeyakrishnan (2013) portrayed a review on load balancing plans in cloud environments. There were different load adjusting strategies are utilized as a part of these papers and their comparing favorable circumstances, hindrances and execution measurements are considered in detail. In this paper, they over-viewed different existing load balancing techniques in various environments.

Dhinesh and Venkata (2013) describe that consumption of resources and energy conservation are not generally a prime concatenation of exchange in cloud computing. Not with standing, resources consumption can e kept to a base with appropriate load balancing which helps in decreasing expenses as well as making enterprises greener.

Garg et al. (2011) defined that increased energy consumption increases costs as well as expand carbon-outflow. High energy cost brings about decreasing cloud supplier's net revenue and high carbon outflow is bad for the environment.

4 GREEN COMPUTING IN CLOUDS

Green Computing (Pushpendra et al, 2010), or Green IT, is the complete training of implementing policies and procedures that improve the efficiency of computing resources in such a way as to Mute the energy consumption and environmental impact of their utilization (Kabiraj et al, 2010)

At High Performance Computing (MC) is getting to be noticeably well known in business and consumer IT applications, it needs the capacity to increase fast and adaptable access to top of the line computing capabilities. This figuring infrastructure is given by cloud computing by making utilization of datacenters It helps the FIPC clients in an on-request and payable access to their applications and information, anyplace from a cloud (Garg et al, 2010) Cloud computing data-centers have been empowered by fast PC networks that enable applications to run more proficiently on these remote, broadband PC system contrasted with nearby PCs. These data-centers cost less for application hosting and. operation than individual application software licenses running on clusters computer clusters (Rao et al 2010) However, the explosion of cloud computing and the growing demand drastically increases the energy consumption

which has become a critical issue and a major concern for both industry and society (Nagothu et al. 2010) 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 the environment (Garg et al, 2010) Hence forth, energy-efficient solutions that can address the high energy utilization, both from the point of view of the cloud supplier and the environment arc required This is a desperate need of distributed computing to accomplish Green processing. This entire situation is portrayed in Fig 2 toad balancing can be one such energy-saving arrangement in cloud computing environment

Features of Clouds enabling Green computing

Despite the fact that there is an incredible worry in the community that Cloud computing can bring about higher energy use by the datacenters, the Cloud computing has a green lining. There are a few advances and ideas utilized by Cloud suppliers to accomplish preferable use and proficiency over traditional computing. Along these lines bring down carbon emissions is normal in Cloud processing because of highly energy efficient infrastructure and diminishment in the IT framework itself by multi- tenancy The key driver innovation for energy efficient Clouds is "Virtualization," which permits huge change in energy efficient Cloud suppliers by utilizing the economies of scale related with vast number of organizations having a similar framework. Virtualization is the way toward exhibiting a coherent gathering or subset of processing assets with the goal that they can be gotten to in ways that give benefits over the first arrangement.

By consolidation of underutilized servers as numerous virtual machines sharing same physical server at higher use, organizations can increase high investment funds as space. Administration and energy According to Accenture Report (2010), there are following four: key factors that have enabled the Cloud computing to lower energy usage and carbon emissions from ICT. Because of these Cloud highlights, organizations can diminish carbon outflows by no less than 30% for every client by moving their applications to the Cloud these funds are driven by the high productivity of vast scale Cloud data centers.

4.1 Dynamic Provisioning.

In conventional setting, datacenters and private framework used to be kept up to satisfy most pessimistic scenario request hence. IT organizations wind up conveying significantly more foundation than required. There are different purposes behind such ova-provisioning

a) It is extremely hard to foresee the demand at once; this is especially valid for Web applications

b) To ensure accessibility of services and to keep up certain level of service quality to aid users

One example of a Web service facing these problems is a Website for the Australian Open Tennis Championship (Srimathi et al., 2012). The Australian Open Website each year receives a significant spike in traffic during the tournament period. The increase in traffic can is to over 100 times its typical volume (22 million visits in a couple of weeks) (Srimathi et al., 2012). To deal with such pinnacle stack amid brief period in a year, running several during The time is not by any stretch of the imagination is not effective. In this manner, the framework provisioned with a preservationist approach results in utilized resources.

Such situations can be promptly overseen by Cloud framework The virtual machines in a Cloud framework can he live moved to another host in the event that client application requires more resources. Cloud suppliers screen and anticipate the demand and in this way allot assets as indicated by demand. Those applications that require less number of assets can be combined on a similar server. In this manner, datacenters dependably keep up the dynamic servers as indicated by current demand, which brings about low energy utilization than the preservationist approach of over-provisioning

4.2 Multi-Tenancy.

Utilizing multi-tenure approach, Cloud computing framework lessens general energy use and related carbon outflows. The SaaS suppliers serve numerous organizations on same foundation and software. This approach is clearly more energy effective than various duplicates of programming software's introduced on various framework Besides, organizations have exceedingly variable demand patterns in general, and henceforth multi-tenancy on a similar server permits the leveling of the general demand request which can limit the requirement for additional framework / infrastructure.

4.3 Server Utilization:

All in all, on commence framework keep running with low usage; once in a while it goes down up to 5 to 10 percent of normal use. Utilizing virtualization advances, various applications can be facilitated and executed on a similar server in isolation, in this manner prompt use levels up to 70%. It drastically diminishes the quantity of active servers. Despite the fact that high use of servers results in more power utilization, server running at higher use can prepare more workload with comparative power use,

4.4 Data center efficiency.

As discussed above, energy efficiency in data centers has a significant impact on energy use of cloud computing liming more energy efficient technologies, CIO providers can greatly improve their data centers the latest data center. The latest data center designs for large Cloud services providers can achieve low PO levels of up to 1.1 to 1.2 about 40% more energy efficient than traditional data centers. The server design in the form of modular containers, water or air cooling, or advanced power management tough the optimal power supply, are all approaches that markedly improved Bo in data centers In addition, cloud computing enables the transfer of services between multiple data centers that work with better Bo values This is achieved using high-speed network, virtualization services and measurement, monitoring and data center accounting.

5 Load Balancing in Cloud Computing

The increase in web traffic and different application in the web world is increasing day by day where millions of data arc created every second. Load balancing has become a very prevalent research field due to need of balancing the load on this heavy traffic Cloud computing use is a concept that use virtual machines instead of physical device to host, store and link the different nodes for their specific purpose. The load balancing is needed on CPU load, memory capacity and network. Load Balancing is done in such a way that the entire load is distributed among various nodes in a distributive system. If there is a failure of any node or host system in the network, it will lead to isolation of web resource in the web work! Load balancing in such situation should be able to provide availability.

The expansion in web traffic and diverse application in the web world is expanding step by step where a huge number of information is made each second. Load balancing has turned into an exceptionally predominant research field because of need of balancing the bad on this substantial movement of heavy traffic. Cloud computing use is an idea that utilizes virtual machine rather than physical gadget to host, store and link the distinctive nodes for their particular reason. The load balancing is required on CPU load, memory capacity and network. It is done such that the whole load is appropriated among different blabs nodes in a distributive framework. If there is a failure of any node or host system in the network, it will prompt isolation of web asset in the web world. Load balancing in such bas to ought to have the capacity to give accessibility, adaptability and scalability.

Load balancing is a process of reassigning the total load to the individual nodes of the computing environment, (Marinov, 2012) this facilitates the network and resources and further improving the system performance. The important parts of this process are estimation and comparison of the stability, load and performance of the system, internodes traffic optimization. To construct load balancing mechanism many techniques and strategies arc used (Kaur et al. 2013). The load need to be distributed over the resources in cloud-based architecture, consequently each resource does almost own equal amount of task at any point of time. The basic goal is to design sonic techniques to balance requests to provide the solutions.

Load balancing is the technology of disseminating the load among several resources in any system Hence load needs to be disseminated over the resources in cloud-based architecture, so that every resource does almost the same amount of work at any point of time Basic need is to offer some methods to balance requests to offer the solution of fast reply for request. It is used to achieve a high user satisfaction and resource utilization ratio (Zhang and Zhang, 2010), making sure that no single node is overwhelmed, hence improving the overall performance of the system. Cloud Load Balancers maintain online traffic by disseminating workloads among multiple servers and resources automatically. They increase throughput, decrease response time, and avoid overload. In this paper, an overall survey of the new load balancing methods in the Cloud Computing atmosphere is submitted

Due to the exponential growth of cloud computing, and rapid expansion in data-centers there is a dramatic increase in energy consumption which has straight impact on the environment in terms of carbon footprints. The link between energy consumption and carbon emission has given rise to an energy management issues which should be attain by improving energy-efficiency in cloud computing to achieve Green computing.

6 Existing load balancing techniques in Clouds

Following load balancing approaches are presently current in clouds:

Mehta et al. (2011) proposed new content science load-balancing policy called workload policy and client aware (and Cap). It uses a unique and special property (USB) to determine the unique and special property of application as well as computing nodes. USB helps scheduling to determine the best node suitable for processing requests. This strategy is implemented in a decentralized manner with low public expenditure. Using content information to narrow your search, this technology improves search performance and overall system performance. It also helps in minimize idle time for computing nodes and thus improving their use. Server based load-balancing for Internet distributed services.

Nakai et al. (2011) a new server load-balancing policy has been proposed for web servers distributed worldwide. It helps minimize service response times by using a protocol that limits the redirection of requests to the nearest remote servers without overloading them. It also uses inference to help web servers to carry overload.

6.1 Join-Idle-Queue

Lua et al. (2011) a queue load-balancing algorithm was proposed for dynamically scalable web services. This load-balancing algorithm provides distributions distributed by the first idle processors to load-balancing across senders for the availability of idle processors on each transmitter and then assigns processor functions to reduce the average queue length in each processor. By removing the load-balancing workload from the critical path to handle the request, it effectively reduces system load, insures any overhead connections in arrivals and does not increase the actual response time.

6.2 A Lock-Free Multiprocessing Solutions for LB

Liu et al. (2011) coupling the multiple lock free balancing solution that avoid the use of shared memory in contrast to multiple other load balancing solutions that use shared memory and lock to maintain user session. This is achieved by modifying then Linux Kernel. This solution helps improve the overall performance of the load balancer in multi-core environment by running multiple load balancing operations in a single load balancer.

6.3 Scheduling Strategy on Load-Balancing of Virtual Machine Resources

Hu et al. (2010) the proposal load balancing scheduling strategy was proposed by MIM resources that used historical data and the current state of the system. This strategy achieves better load balancing and reduced dynamic migration using a genetic algorithm. It helps to solve the problem of pregnancy imbalance and the high cost of migration and hence the best use of resources.

6.4 Central Load Balancing Policy for Virtual Machines

Bahadani and Chaudhary (2010) the central load balancing policy (CLPFM) was proposed that balance load equally in a virtual environment distributed on computers/cloud computing. This policy improves the overall performance of the system but does not take into account fault tolerant systems.

6.5 LBVS (Load-Balancing Strategy for Virtual Storage)

Liu et al. (2011) proposed Virtual Load Balancing Strategy (LBVS), which provides a widely used data storage model and storage as a service model based on cloud storage. Virtual storage is achieved using an architecture that is three layers and checks load balancing using two load-balancing units. It helps to improve the efficiency of concurrent access using a symmetrical budget, further reducing response time and enhancing disaster recovery capacity. This strategy also helps to improve the utilization rate of storage resources and the flexibility and durability of the system.

6.6 A Task Scheduling Algorithm Based on Load Balancing

Fang et al. (2010) discuss the task scheduling mechanism of two levels on the basis of load balancing to meet the dynamic requirements of users and access to the use of high resources. It achieves load balancing through the first mapping tasks to virtual machines and then virtual machines to host resources thereby improving mission responsiveness, resource utilization and overall performance of the cloud computing environment.

6.7 Honeybee Foraging Behavior

Randles et al. (2010) looked for decentralized honey-dependent balancing techniques that are an algorithm inspired by a nature for self-regulation. It achieves global load balancing through local server procedures. And it enhances system performance while increasing the diversity of the system, but does not increase productivity with an increase in the size of the system. It is best suited to the circumstances where the diverse populations of the types of the services required,

6.8 Biased Random Sampling

Randles et al. (2010) in a distributed systems and scalable load balancing approaches that uses for random samples from the system domain to get self-regulation and thus achieve a load balance in all nodes of the system. The performance of the system is improved with a large number of similar populations, leading to increase to productivity through the effective utilization of increased system resources. It is deteriorating with increasing population diversity.

6.9 Active Clustering

Randalls et al. (2010) is self-loading self-balancing technology, the self-assembly algorithm for improving functional tasks by connecting similar services using local rewriting. It enhances system performance with high resources, thereby increasing productivity through these resources effectively. It deteriorates as the diversity of the system increases.

6.10 ACCLB (Load-Balancing Mechanism Based on Ant Colony and Complex Networks Theory)

Zhang and Zhang (2010) the load balancing mechanisms was proposed based on an ant colony and complex rid theory of the Open Cloud Computing Alliance. It uses the micro-world and the scale-free properties of a complex network to achieve better load balancing. This method overcomes homogeneity, is adaptive to dynamic environments, excellent fault tolerance and has good portability and thus helps improve system performance.

6.11 Two-Phase Load-Balancing Algorithms (OLB+LBMM)

Wang et al. (2010) a two phases scheduling algorithms that integrate Opportunistic Load Balancing (OLB) and LBMM(Load Balance Min-Min) algorithms has been proposed to take advantage of better execution efficiency and load balancing of the system. OLB Scheduling Algorithm, keeps each node in working condition to achieve that target load balance and the scheduling algorithm for the LBMM is used to reduce the execution time of each task on the node and the thereby reduce the overall finish time. Thus, this common approach helps the effective use of resources and enhances the efficiency of work.

6.12 Event-Driven

Nae et al. (2010) Load balancing algorithms introduced event-driven real-time MMOG (Massively Multiplayer Online Games). This algorithm after receiving the capacity events as inputs, analyze its components in the resource context and the global state of the same session, thus generation game cycle load balancing procedures. It is able to upgrade and reduce the game session on multiple resources according to variable user load but has occasional quality service breaches.

6.13 CARTON

Stanojevic and Shorten (2009) proposed a mechanism to control the cloud called CARTON unites the use of LB and DRL. A LB (Load Balancing) core is used to distribute functions equally on different servers so that the associated costs can be reduced, DRL (Distributed Rate Limiting) is used to ensure that the resources are distributed in a manner that preserves the allocation of resources fairly. DRL also adapts server capabilities to dynamic workloads so that performance levels across all servers are equal. With very low account and public communications, this algorithm is simple and easy to implement.

6.14 Compare and Balance

Zhao and Huang (2009) The load balancing model is designed and implemented to reduce the migration time of virtual machines through shared storage, to balance load between servers according to the processor or the use of IO, etc., and to keep non-virtual devices in the process. Distributed Load Balancing Algorithm also proposes a compare and balance that is based on sampling and reaches a very fast balance. This algorithm confirms that the migration of VMs is always from high-cost physical hosts to a low-cost host, but assumes that each physical host has sufficient memory, which is a weak assumption.

6.15 VectorDot

Singh et al. (2008) proposed a new load balancing called Fictured. It handles the multidimensional hierarchical complexity of resource loads across servers, network switches, and storage in an agile data center with integrated server and virtualization technologies. Vectordot uses the product point to distinguish the nodes based on the requirements of the item and helps remove overloading on servers, switches and storage nodes.

These existing approaches has been summarized in Table 1

Approaches	Environment	Description	Findings
Decentralized content aware load balancing	Distributed Computing	1. Uses a unique and special property (USP) of requests and computing nodes to help scheduler to decide the best processing the requests. 2. Uses the content information to narrow down the search.	1. Improve the searching performance hence increasing overall performance. 2. Reduce the Idle time of the nodes.
Server-based load balancing for internet distributed services.	Distributed Web Servers	1. Uses a protocol to limit redirection rates to avoid remote servers overloading. 2. Uses a middleware to support this protocol. 3. Uses a heuristic to tolerate abrupt load changes.	1. Reduces service response times by redirecting requests to the closest to servers without overloading them. 2. Mean response time is 29% smaller than RR(round Robin) and 31% smaller than SL (Smallest Latency)
Join-idle-Queue	Cloud Data Centers	1. First assigns the idle processors to dispatchers for the availability of the idle processor at each dispatcher. 2. Then assign jobs to processor to reduce average queue length of jobs at each processor.	1. Effectively reduces the system load. 2. Incurs no communication overhead at job arrivals. 3. Does not increase actual response times.
A Lock-free multiprocessing solution for LB	Multi-Core	1. Runs multiple load-balancing processes in one load balancer.	1. Improves overall performance of load balancer.
Scheduling strategy on load balancing of virtual machine resources.	Cloud Computing	1. Uses generic algorithms, historical data and current state of system to achieve best load balancing and to reduce dynamic migration.	1. Solves the problems of load imbalance and high migration cost.
Central load-balancing policy for virtual machines.	Cloud Computing	1. Uses global state information to make load balancing decisions.	1. Balances the load evenly to improve overall performance. 2. Up to 20% improvement in performance. 3. Does not consider fault tolerance.
LBVS: Load-Balancing Strategy for Virtual Storage	Cloud Storage	1. Uses Fair-Share Replication strategy to achieve replica load balancing module which in turn controls the access load balancing. 2. Uses writing balancing algorithm to control data writing load balancing.	1. Enhance flexibility and robustness. 2. Provides large scale net data storage and storage as a service.
A Task Scheduling Algorithm Based on Load Balancing	Cloud Computing	1. First maps task to virtual machines and then virtual machine to host resources.	1. Improve task response time. 2. Improves resource utilization
Honeybee Foraging Behavior	Large-Scale Cloud Systems	1. Achieves global load balancing through local server actions.	1. Perform well as system diversity increases. 2. Does not increase throughput as system size increases.
Biased Random Sampling	Large-Scale Cloud Systems	1. Achieves load balancing across all systems nodes using random sampling of the system domain.	1. Perform better with high and similar population of resources. 2. Degrades as population diversity increases.
Active Clustering		1. Optimizes job assignment by connecting	1. Performs better with high re-

	Large-Scale Cloud Systems	similar services by local rewiring	sources. 2. Utilizes the increased system resources to increase throughput 3. Degrades as system diversity increases.
ACCLB (Load Balancing mechanism based on ant colony and complex network theory)	Open Cloud Computing Federation	1. Uses small-world and scale-free characteristics of complex network to achieve better load balancing.	1.Overcomes heterogeneity 2. Adaptive to dynamic environments. 3. Excellent in fault tolerance. 4. Good scalability.
Two-phase load balancing algorithm (OLB+LBMM)	Three-level Cloud Computing Networks	1. Uses OLB (Opportunistic Load Balancing) to keep each node busy and uses LBMM (Load-Balance Min-Min) to achieve the minimum execution time of each task.	1. Efficient utilization of resources. 2. Enhances work efficiency
Event-driven	Massively Multiplayer Online Games	1. Uses complete capacity event as input, analyzes its components and generates the game session load balancing actions.	1. Capable of scaling up and down a game session on multiple resources according to the variable user load. 2. Occasional QoS breaches as low as 0.66%
CARTON	Unifying framework for cloud control	1. Uses load balancing to minimize the associated cost and uses distributed rate, limiting for fair allocation of resources.	1.Simple 2. Easy to implement 3.Very low computation and communication overhead
Compare and Balance	Intra-Cloud	1.Based on sampling 2. Uses adaptive live migration of virtual machines.	1. Balances load amongst servers 2. Reaches equilibrium fast 3. Assures migration of VMs from high cost physical hosts to low-cost host. 4. Assumptions of having enough memory with each physical host.
VectorDot	Datacenters with integrated server and storage virtualization	1. Use dot product to distinguish node based on the item requirements.	1. Handles hierarchical and multi-dimensional resource constraints. 2. Removes overloads on server, switch and storage.

7 METRICS FOR LOAD BALANCING AND ENERGY EFFICIENCY IN CLOUDS

For an energy-efficient load balancing in clouds, various parameters like performance, response time, scalability, throughput, resource utilization, fault tolerance, migration time, associated overhead, energy consumption and carbon emission have also been considered,

7.1 Carbon Emission (CE)

Determines the carbon emission of all the resources in the system. As energy consumption and carbon emission go hand in hand, the more the energy disbursed, higher is the carbon impression. So, for an energy-efficient load balancing solution, it should be concentrated.

7.2 Energy Consumption (EC)

Determines the carbon emission of all the resources in the system Load balancing helps in avoiding overheating by balancing the workload across all the nodes of a cloud, hence reducing energy consumption.

7.3 Fault Tolerance

It's the power of an algorithm to perform uniform load balancing in spite of absolute node or link failure. The load balancing should be a good fault right technique.

7.4 Migration Time

It's the time to migrate the jobs or resources from one node to other. It should be minimalized in order to enhance the performance of the system.

7.5 Overhead Associated

Determines the amount of overhead involved while implementing a load-balancing algorithm. It is composed of overhead due to movement of tasks, inter-process and inter-process communication. This should be minimized so that a load-balancing technique can work efficiently

7.6 Performance

It is used to visualize balance the overall potency of the system. It's to be improved at inexpensive cost like that reduce response time whereas keeping acceptable delays.

7.7 Response Time

It’s the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.

7.8 Resource Utilization

It’s used to check the utilization of resources. It should be visualized and optimized for an efficient load balancing

7.9 Scalability

It’s capability of an algorithm to perform load balancing for a system with any finite number of nodes. This matric should be improved

7.10 Throughput

It’s used to calculate the no. of tasks whose execution has been completed. It should be high to improve the performance of the system.

Based on the above metrics, the existing techniques of load-balancing have been compared in Table.

Comparison of existing load balancing techniques based on measurement parameters Table 2

Load-Balancing Techniques	CE	EC	Fault Toler-ance	Migration Time	Over-head	Performance	Re-sponse time	Resource Utiliza-tion	Scalabil-ity	Through-put
Join-Idle-Queue	**F**	**F**	**F**	**F**	**F**	**T**	**F**	**T**	**F**	**F**
A Lock-Free Multiprocessing Solutions for LB	**F**	**F**	**F**	**F**	**F**	**T**	**F**	**F**	**F**	**T**
Scheduling Strategy on Load-Balancing of Virtual Machine Resources	**F**	**F**	**F**	**F**	**T**	**F**	**F**	**T**	**F**	**F**
Central Load Balancing Policy for Virtual Machines	**F**	**F**	**F**	**F**	**F**	**T**	**F**	**T**	**F**	**T**
LBVS (Load-Balancing Strategy for Virtual Storage)	**F**	**F**	**T**	**F**	**F**	**T**	**T**	**F**	**T**	**F**

A Task Scheduling Algorithm Based on Load Balancing	**F**	**F**	**F**	**F**	**F**	**T**	**T**	**T**	**F**	**F**
Honeybee Foraging Behavior	**F**	**F**	**F**	**T**	**F**	**T**	**T**	**T**	**F**	**F**
Biased Random Sampling	**F**	**F**	**F**	**F**	**T**	**T**	**F**	**F**	**T**	**T**
Active Clustering	**F**	**F**	**F**	**F**	**F**	**T**	**F**	**F**	**T**	**T**
ACCLB (Load-Balancing Mechanism Based on Ant Colony and Complex Networks Theory)	**F**	**F**	**T**	**F**	**F**	**T**	**F**	**T**	**T**	**F**
Two-Phase Load-Balancing Algorithms (OLB+LBMM)	**F**	**F**	**F**	**F**	**F**	**T**	**F**	**T**	**F**	**F**
Event-Driven	**F**	**F**	**F**	**F**	**F**	**F**	**F**	**T**	**T**	**F**
CARTON	**F**	**F**	**F**	**F**	**T**	**T**	**F**	**T**	**F**	**F**
Compare and Balance	**F**	**F**	**F**	**T**	**T**	**F**	**F**	**T**	**F**	**F**
VectorDot	**F**	**F**	**F**	**F**	**F**	**F**	**F**	**T**	**F**	**F**

CONCLUSION

Therefore, there was a need to develop energy-efficient load balancing technology that can improve cloud computing performance along with maximum resource utilization, which in turn reduces energy consumption as well as carbon emission to a degree that would help achieve green computing. There we study load balancing techniques in cloud like performance, response time, scalability, throughput, resource utilization, fault tolerance, migration time, associated overhead, energy consumption and carbon emission

REFERENCES

[1] Accenture Microsoft Report. (2010) Cloud computing and Sustainability: The Environmental Benefits of moving to the Cloud

[2] Bhadani, A., & Chaudhary, S. (2010, January). Performance evaluation of web servers using central load balancing policy over virtual machines on cloud. In Proceedings of the Third Annual ACM Bangalore Conference (p. 16). ACM.

[3] Fang, Y., Wang, F., & Ge, J. (2010). A task scheduling algorithm based on load balancing in cloud computing. Web Information Systems and Mining, 271-277.

[4] Garg, S. K., Yeo, C. S., Anandasivam, A., & Buyya, R. (2011). Environment-conscious scheduling of HPC applications on distributed cloud-oriented data centers. Journal of Parallel and Distributed Computing, 71(6), 732-749 (Garg et al, 2011)

[5] Garg, S. K., & Buyya, R. (2012). Green cloud computing and environmental sustainability. Harnessing Green IT: Principles and Practices, 315-340.

[6] Verma, P., Shekhar, J., & Asthana, A. (2014). A model for evaluating and maintaining load balancing in cloud computing. IJCSMC, 3(3), 501-509.

[7] LD, D. B., & Krishna, P. V. (2013). Honey bee behavior inspired load balancing of tasks in cloud computing environments. Applied Soft Computing, 13(5), 2292-2303

[8] Hussin, M., Mahmood, R. A. R., Husin, N. A., & Norowi, N. M. (2016). A performance optimization model of task scheduling towards green cloud computing. International Journal of New Computer Architectures and their Applications (IJNCAA), 6(1), 1-8

[9] Hu, J., Gu, J., Sun, G., & Zhao, T. (2010, December). A scheduling strategy on load balancing of virtual machine resources in cloud computing environment. In Parallel Architectures, Algorithms and Programming (PAAP), 2010 Third International Symposium on (pp. 89-96). IEEE.

[10] Srimathi, V., Hemalatha, D., & Balachander, R. (2012). Green Cloud Environmental Infrastructure. International Journal of Engineering And Computer Science, 1(3), 168-177

[11] Kabiraj, S., Topkar, V., & Walke, R. C. (2010). Going green: a holistic approach to transform business. arXiv preprint arXiv:1009.0844.

[12] Kaur, K., Narang, A., & Kaur, K. (2013). Load balancing techniques of cloud computing. International Journal of Mathematics and Computer Research.

[13] Liu, X., Pan, L., Wang, C. J., & Xie, J. Y. (2011, May). A lock-free solution for load balancing in multi-core environment. In Intelligent Systems and Applications (ISA), 2011 3rd International Workshop on (pp. 1-4). IEEE.

[14] Kliotb, G., Lua, Y., Xiea, Q., Gellerb, A., & Larusb, J. R. Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. An international Journal on computer Performance and evaluation, In Press, Accepted Manuscript, Available online, 3.

[15] Marinov, M. (2012). Intuitionistic fuzzy load balancing in cloud computing. In 8th Int. Workshop on IFSs, Banská Bystrica(Vol. 18, No. 4, pp. 19-25).

[16] Mehta, H., Kanungo, P., & Chandwani, M. (2011, February). Decentralized content aware load balancing algorithm for distributed computing environments. In Proceedings of the International Conference & Workshop on Emerging Trends in Technology (pp. 370-375). ACM.

[17] Nae, V., Prodan, R., & Fahringer, T. (2010, October). Cost-efficient hosting and load balancing of massively multiplayer online games. In Grid Computing (GRID), 2010 11th IEEE/ACM International Conference on (pp. 9-16). IEEE.

[18] Nagothu, K. M., Kelley, B., Prevost, J., & Jamshidi, M. (2010, September). Ultra low energy cloud computing using adaptive load prediction. In World Automation Congress (WAC), 2010(pp. 1-7). IEEE.

[19] Nakai, A. M., Madeira, E., & Buzato, L. E. (2011, April). Load balancing for internet distributed services using limited redirection rates.

In Dependable Computing (LADC), 2011 5th Latin-American Symposium on (pp. 156-165). IEEE.

[20] Rajan, R. G., & Jeyakrishnan, V. (2013). A survey on load balancing in cloud computing environments. International Journal of Advanced Research in Computer and Communication Engineering, 2(12), 4726-4728.

[21] Rao, K. T., Kiran, P. S., & Reddy, L. S. S. (2010). Energy efficiency in datacenters through virtualization: A case study. Global Journal of Computer Science and Technology.

[22] Randles, M., Lamb, D., & Taleb-Bendiab, A. (2010, April). A comparative study into distributed load balancing algorithms for cloud computing. In Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on (pp. 551-556). IEEE.

[23] Singh, S., & Chana, I. (2014). Energy based efficient resource scheduling: a step towards green computing. Int J Energy Inf Commun, 5(2), 35-52.

[24] Singh, A., Korupolu, M., & Mohapatra, D. (2008, November). Server-storage virtualization: integration and load balancing in data centers. In Proceedings of the 2008 ACM/IEEE conference on Supercomputing (p. 53). IEEE Press.

[25] Stanojevic, R., & Shorten, R. (2009, June). Load balancing vs. distributed rate limiting: an unifying framework for cloud control. In Communications, 2009. ICC'09. IEEE International Conference on (pp. 1-6). IEEE.

[26] Wang, S. C., Yan, K. Q., Liao, W. P., & Wang, S. S. (2010, July). Towards a load balancing in a three-level cloud computing network. In Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on (Vol. 1, pp. 108-113). IEEE.

[27] Zhang, Z., & Zhang, X. (2010, May). A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation. In Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on (Vol. 2, pp. 240-243). IEEE.

[28] Zhao, Y., & Huang, W. (2009, August). Adaptive distributed load balancing algorithm based on live migration of virtual machines in cloud. In INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on (pp. 170-175). IEEE.

[29] Chaczko, Z., Mahadevan, V., Aslanzadeh, S., & Mcdermid, C. (2011, September). Availability and load balancing in cloud computing. In International Conference on Computer and Software Modeling, Singapore (Vol. 14).