



AN ANALYSIS OF CLOUD LOAD BALANCING

Dr A KANNAGI¹, LINU PAULOSE², PREMA MANI³, SHAHINA K K⁴

Professor¹, Assistant Professor^{2,3,4}

Department of Computer Science Engineering, Indira Gandhi Institute Of Engineering And Technology, Nellikuzhi
P.O, Kothamangalam, Ernakulam (Dist) Pincode 686691

Abstract

Cloud computing refers to the use and advancement of computer technology over the Internet. Cloud computing is a computing methodology that involves the provision of dynamically scalable and typically virtualized resources as a service via the Internet. Users are not required to possess knowledge, skill, or control over the technical infrastructure in the cloud that supports them. Efficiently harnessing the possibilities of heterogeneous computing systems relies heavily on effective scheduling. In cloud computing, the load balancing of the whole system may be dynamically managed by using virtualization technology. This allows for the remapping of virtual machines and real resources in response to changes in the workload. In order to enhance performance, the virtual machines must effectively use their resources and services by constantly adapting to the computing environment. Ensuring efficient resource allocation is essential for optimising resource utilisation and achieving load balancing.

Keywords: Cloud computing, Load balancing, Virtual machine, Host, Datacenter, Datacenter Broker

Introduction

Cloud Computing (CC) is a nascent technology that is intricately linked to the Grid Computing (GC) paradigm and other related technologies including utility computing, distributed computing, and cluster computing. Both GC and CC have the objective of attaining resource virtualization. Although GC and CC have a similar objective, they exhibit notable distinctions. The primary focus of GC is to get the highest level of computational performance, while CC aims to optimise the entire computational capacity. CC offers a solution for a variety of organisational requirements by delivering servers and applications that can be scaled dynamically. Prominent cloud computing service providers including Amazon, IBM, Dropbox, Apple's iCloud, Google's apps, Microsoft's Azure, and others have the ability to draw a wide range of consumers worldwide. CC has implemented a novel paradigm that enables users to save or create apps in a flexible manner and access them from any location and at any time simply by connecting to an application over the Internet. CC offers flexible and tailored services to meet the specific needs of customers, allowing them to easily access and interact with cloud apps. CC may be used to provide a framework for creating applications, a system for storing and processing a company's data, and also apps for doing regular chores for the user. When a consumer opts for cloud services, data stored in local repositories will be sent to a distant data centre. Cloud service providers provide services that enable the access and management of data in distant places. It is evident that in order for a user to save or manipulate data in the cloud, they must send the data to a distant server over an internet connection. It is crucial to handle this data processing and storage with the highest level of caution in order to prevent any unauthorised access or disclosure of data.

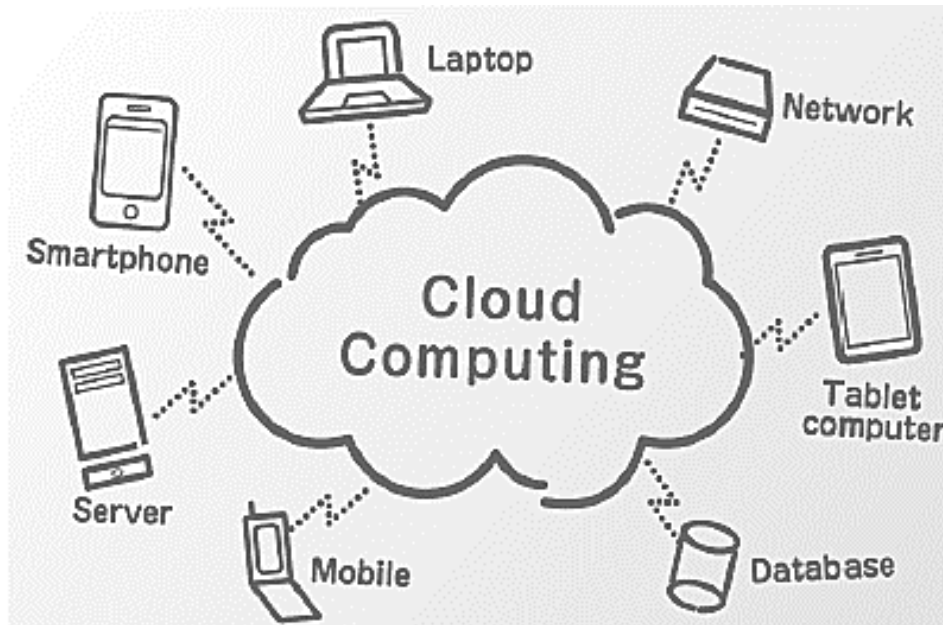


Figure 1. Cloud Computing

This paradigm provides accessible and immediate access to a network, with little administrative requirements, allowing for quick and rapid access to readily available resources. The emerging paradigm has significant economic benefits, including decreased time to market, adaptable computing capabilities, and unlimited computing capacity. The popularity of cloud computing is steadily growing in distributed computing environments. There is an increasing inclination towards using cloud environments for the purpose of storing and processing data. In order to fully harness the capabilities of cloud computing, the transport, processing, retrieval, and storage of data are handled by external cloud providers. Nevertheless, data owners exhibit a high degree of scepticism when it comes to entrusting their data to other entities outside their own sphere of control.

ADVANTAGES OF CLOUD COMPUTING

Several prevalent advantages of cloud computing include:

- **Cost Reduction:** The progressive use of cloud technology helps organisations save on their overall expenses.
- **Enhanced Storage Capacity:** In comparison to personal computer systems, the capacity to store much larger volumes of data is possible.
- **Cloud computing offers a higher level of flexibility** compared to conventional computing technologies by enabling the outsourcing of a full organisational segment or a piece of it.
- **Enhanced mobility:** The ability to access information at any time and from any location, unlike conventional systems where data is stored on personal computers and can only be accessed while in close proximity to them.
- **Shifting IT focus:** Organisations should prioritise innovation by introducing new product strategies, instead of being preoccupied with maintenance tasks like software upgrades or computer concerns. The advantages of cloud computing attract significant attention from the Information and Technology Community (ITC).

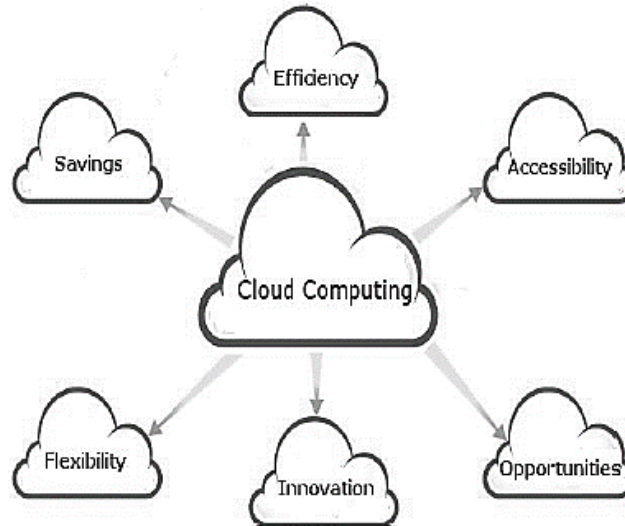


Figure 2. Benefits of Cloud Computing

Cloud computing: Service Models

Cloud computing is accessible via a range of service models. These services are specifically intended to demonstrate certain features and meet the specific needs of the organisation. Based on this, an optimal service may be chosen and tailored to fit an organization's needs. Cloud computing services may be categorised into many types, including Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), Infrastructure-as-a-Service (IaaS), Hardware-as-a-Service (HaaS), and Data storage-as-a-Service (DaaS). The service model provides specific information as outlined below:

Software as a Service (SaaS) refers to a service where the supplier offers the option to utilise one or more applications that are hosted on a cloud architecture. These apps may be accessible using several thin client interfaces, such as web browsers. Users of this service are not required to oversee, administer, or govern the underlying cloud infrastructure, such as the network, operating systems, and storage. Some examples of Software as a Service (SaaS) cloud platforms are Salesforce and NetSuite.

Platform as a Service (PaaS) refers to a service where the provider offers users resources to install on cloud infrastructure. These resources may include apps that are either built or bought by the user. Users of this service have the ability to manage deployed programmes and the application hosting environment. However, they do not have control over the infrastructure, including network, storage, servers, and operating systems. Some examples of Platform as a Service (PaaS) clouds are Google App Engine, Microsoft Azure, and Heroku.

Infrastructure as a Service (IaaS) refers to a cloud computing model where virtualized computing resources, such as servers, storage, and networking, are provided to users via the internet. Consumers are empowered with the ability to manipulate and oversee essential computer resources, such as power, storage, and network, in order to effectively manage various software, including operating systems and apps. By using this particular service, the user has the ability to manage the operating system, storage, installed applications, and maybe exercise some control over certain networking components. Some examples of Infrastructure as a Service (IaaS) clouds include Eucalyptus (an open-source cloud computing system), Amazon EC2, Rackspace, and Nimbus.

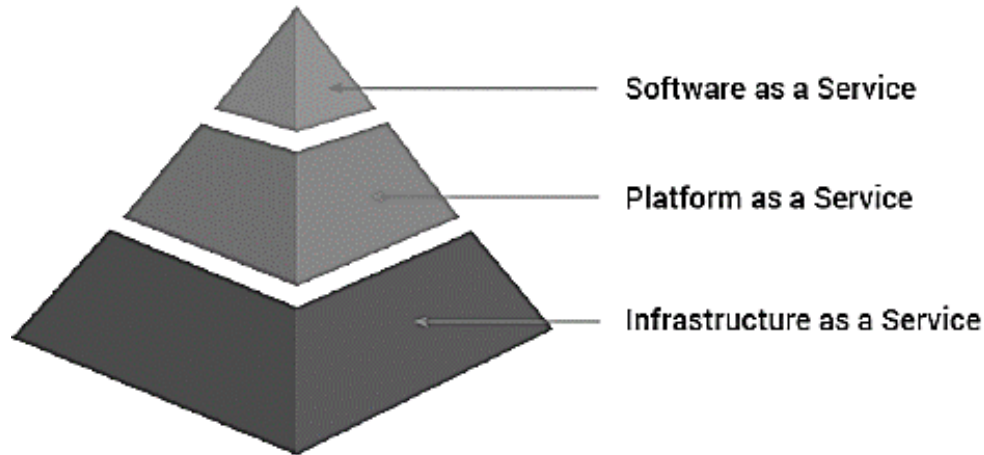


Figure 3. Cloud Computing Service Model

Cloud computing: Deployment Models

Out of the service models mentioned before, SaaS, PaaS, and IaaS are widely favoured by both providers and customers. These services may be implemented on many deployment models, including public cloud, private cloud, community cloud, and hybrid cloud, in order to use the functionalities of cloud computing. Each of these deployment models will be elucidated as follows:

Public cloud: This infrastructure is accessible to huge industrial organisations or the general public. These assets are both managed and controlled by an organisation that sells cloud services.

Private cloud: This kind of cloud deployment is only available to the organisation that creates it. Private clouds may be administered either by a third party or by the organisation itself. In this case, the presence of cloud servers may or may not coincide with the physical location of the organisation.

Hybrid cloud: This deployment approach involves the use of two or more clouds, such as private, public, or community clouds. The constituent clouds, which are combinations of clouds such as 'private and public' or 'public and community', stay distinct but are connected by standardised or prepared technology that allows for the transfer of applications and data.

Community cloud: The community cloud is a kind of cloud infrastructure that is used by several organisations to service a particular community with common problems. This may be overseen by either an organisation or a third party and can be implemented either outside or inside the organization's premises.

The use of deployment patterns and services supplied by cloud computing alters the manner in which systems are interconnected and work is performed inside an organisation. It provides the ability for applications, platforms, infrastructure, or any other resource that is organised and used in cloud computing to be dynamically extendable.

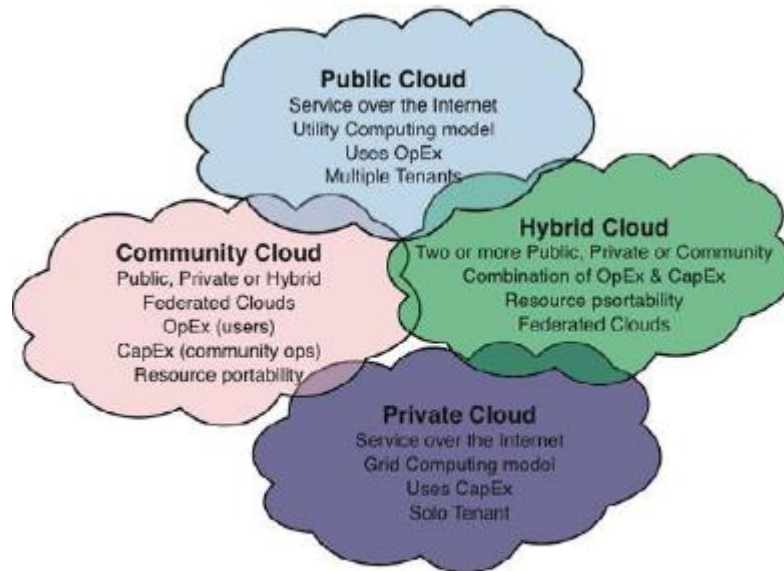


Figure 4. Types of Cloud

Load Balancing

One of the most often used uses of load balancing is to provide high-quality service delivery from several servers, sometimes referred to as a server data centre. Load-balanced systems are often operational inside widely-used internet sites, large chat networks, high-bandwidth file transfer protocol sites, and domain name System (DNS) servers. Furthermore, it serves the purpose of preventing clients from establishing direct communication with back-end servers, hence providing security benefits via the concealment of the underlying network architecture. Some load balancers include a means for enhancing a specific parameter, particularly inside the back-end server. Load balancing provides the IT staff with the chance to achieve a much greater level of fault tolerance. It will provide the necessary capacity to manage any fluctuations in application traffic automatically. Furthermore, it is essential that the load balancer itself does not contribute to any failures. Load balancers on high-availability servers may also duplicate the user's session required by the application. Load balancing involves distributing the workload across a group of computers to achieve optimal response time. This ensures that all nodes are evenly loaded and, as a result, users are serviced more quickly. Load balancing may be implemented using hardware, software, or a combination of both. Load balancing is often the primary cause of an unequal response time from servers. Load balancing strategies aim to optimise resource use, maximise overall success rate, minimise waiting time, and prevent resource overload. Using various algorithms and processes with load balancing, instead than relying on only one algorithm, may enhance reliability and efficiency. Load balancing in cloud computing deviates from traditional load balancing approaches in terms of using data centre servers to process requests based on a first-come, first-served basis. The previous load balancing technique distributes the requests based on the client's incoming requests.

LITERATURE REVIEW

The authors, Nguyen Khac Chien et al. (2016), have introduced a load balancing algorithm that aims to improve the performance of the cloud environment by using an estimation technique for predicting the completion time of services. They have successfully improved the service time and response time for the customer.

Ankit Kumar et al (2016) examines a load balancing technique that efficiently distributes incoming workloads



International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)

A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.

Available online at: www.ijrrset.com

ISSN (Print) : 2347-6729

ISSN (Online) : 2348-3105

JIR IF : 2.54

SJIF IF : 4.334

Cosmos: 5.395

Volume 6, Issue 12 - December 2018 - Pages 127-135

across virtual machines (VMs) in cloud data centres. The method presented in this study has been built using the Cloud Analyst simulator. The performance of the proposed algorithm has been compared with three preexisting algorithms based on reaction time. Cloud data centres have a significant problem in properly managing the demands from millions of users and providing them with efficient service, due to the worldwide distribution of both the data centres and the cloud computing users.

In this study, Bura D, Singh M, Nandal P (2018) address the significant issue of energy usage in cloud computing infrastructures. The researchers introduced a new and innovative power-conscious load balancing technique called ICAMMT to effectively control power use in data centres for cloud computing. We have used the Imperialism Competitive Algorithm (ICA) to identify hosts that are being excessively used. Subsequently, we transfer one or more virtual machines from these servers to other hosts in order to reduce their workload. Ultimately, we see other hosts as hosts that are not being fully used. If feasible, we transfer all of their virtual machines to the other hosts and put them into sleep mode.

The objective of Surbhi Kapoor et al. (2015) is to enhance user satisfaction by reducing task response time and optimising resource utilisation via equitable distribution of cloud resources. The conventional Throttled load balancing algorithm is an effective method for achieving load balancing in cloud computing, since it equally distributes incoming workloads across the virtual machines (VMs). However, the main limitation of this approach is its effectiveness only in settings with uniform virtual machine systems. It does not take into account the individual resource requirements of the jobs and incurs extra cost by scanning the complete VM list whenever a task is received. A method called Cluster-based Load Balancing has been proposed to solve the concerns. This technique is effective in environments with different types of nodes, takes into account the individual resource requirements of jobs, and decreases the amount of scanning needed by separating the computers into clusters.

The objective of Kumar P, Kumar R. (2019) is to allocate workload across various cloud systems or nodes in order to achieve improved resource utilisation. It is the primary method for achieving effective distribution and utilisation of resources. Load balancing has emerged as a significant concern in contemporary cloud computing systems. In order to accommodate the user's extensive range of requests, a dispersed solution is necessary. This is because it is not always feasible or cost-effective to manage one or more inactive services. It is not possible to allocate servers to specific customers on an individual basis. Cloud Computing is a distributed system that consists of a vast network and several components spread over a broad geographical region. Therefore, it is necessary to provide load balancing among its many servers or virtual machines. The researchers have introduced an algorithm that prioritises load balancing in order to mitigate instances of virtual machines being overloaded or underutilised, resulting in a significant improvement in cloud performance.

According to Reena Panwar et al. (2015), cloud computing has become a crucial term in the field of Information Technology and represents the next phase in the development of the Internet. The load balancing problem in cloud computing is a crucial issue that significantly impacts the efficient functioning of cloud computing systems and may impede the quick progress of cloud computing. A multitude of customers from throughout the globe are now seeking a wide range of services at an accelerated pace. Several effective load balancing methods have been developed to allocate requests by selecting appropriate virtual machines. A technique for dynamic load management has been devised to efficiently distribute incoming requests across virtual machines.

Mohamed Belkhouraf et al. (2015) seeks to provide various services to customers, including infrastructure, platform, or software, at a progressively reduced cost for clients. In order to accomplish these objectives, it is necessary to solve some issues, particularly by using the existing resources efficiently to enhance overall performance, while also considering the security and availability aspects of the cloud. Therefore, load balancing in cloud computing, particularly for large distributed cloud systems that handle many clients and substantial volumes of data and requests, is a highly researched topic. The suggested strategy primarily guarantees enhanced overall performance via effective load balancing, uninterrupted availability, and a focus on security.

Lu Kang et al. (2015) enhance the weighted least connections scheduling method and develop the Adaptive



Scheduling method Based on Minimum Traffic (ASAMT). ASAMT performs real-time minimum load scheduling for node service requests and preconfigures the available idle resources to guarantee the service quality of service (QoS) criteria. OPNET is used for simulating the traffic scheduling method in the context of cloud computing architecture.

Hiren H. Bhatt et al. (2015) develop a Flexible load sharing algorithm (FLS) that incorporates a third function. The third function partitions the system into domains. This function facilitates the selection of additional nodes that exist inside the same domain. Implementing flexible load sharing in certain domains within a distributed system may enhance performance when any node becomes overwhelmed.

Research Gap

Cloud computing utilises dispersed technology to meet a wide range of applications and user requirements. The primary objectives of cloud computing are to facilitate the sharing of resources, software, and information over the internet, while also reducing capital and operational costs. Additionally, cloud computing aims to improve performance in terms of response time and data processing time, maintain system stability, and allow for future system modifications. There are several technical challenges that must be resolved in distributed systems, such as virtual machine migration, server consolidation, fault tolerance, high availability, and scalability. However, the main concern is load balancing. Load balancing is the process of distributing the workload evenly across multiple nodes in a distributed system. This improves resource utilisation and job response time, while preventing situations where some nodes are overloaded while others are underutilised. It also guarantees that each processor in the system or every node in the network does a roughly equal amount of work at any given moment. In order to make the final decision, the load balancer gathers information on the candidate server's health and current workload to confirm its capacity to respond to the request. Load balancing systems may be categorised into software-based load balancers and hardware-based load balancers. Hardware-based load balancers are dedicated devices that use Application particular Integrated Circuits (ASICs) tailored for a particular purpose. Hardware-based load balancers provide the capability to manage high-speed network traffic, whereas Software-based load balancers operate on regular operating systems and normal hardware components.

Problem Formulation

The text does not provide a detailed explanation of any clustering process. The servers have been categorised into several clusters based on their processing capacity, including clusters for high processing power servers, medium processing power servers, and low processing power servers. The text does not provide an explanation for the technique that determines the cluster to which a virtual machine (VM) will be assigned if it has a large amount of RAM but a lower number of MIPS (Million Instructions Per Second). Clustering is only accessible on the cloud provider's end. There is no clustering configured at the cloudlets on the client side.

Cloud Sim

Cloud service companies charge consumers based on the amount of storage or services they get. When doing research and development (R&D) studies, it is sometimes impractical to have access to the physical cloud infrastructure. It is impractical for research scholars, academics, or scientists to repeatedly engage cloud services in order to run their algorithms or implementations. Open source libraries are accessible for research, development, and testing purposes, providing a cloud-like experience. Currently, cloud simulators are extensively used by researchers and professionals in the research sector, eliminating the need to make any financial payments to a cloud service provider.



Cloud simulators may be used to execute several activities, including:

- Modelling and simulating extensive cloud computing data centres.
- Developing models and simulations to analyse and optimise the allocation of resources to virtualized server hosts, allowing for the customisation of policies.
- Modelling and simulating computational resources that are conscious of energy use.
- Analysing and simulating the network topologies and message-passing applications of data centres [18].
- Modelling and simulating federated clouds.
- The ability to introduce simulation components in real-time, as well as pause and resume the simulation.
- Customised rules for assigning hosts to virtual machines (VMs), and policies for distributing host resources among VMs.

The scope and distinctive features of cloud simulations include:

The activities involved in cloud computing include load balancing, data centres, cloudlet creation and execution, resource provisioning, task scheduling, and consideration of storage and cost issues.

CONCLUSION

This study focuses on the untapped potential of cloud computing technologies. The potential of cloud computing is limitless. Cloud computing offers a range of services to users, including platform as a service, application as a service, and infrastructure as a service. Load balancing is a significant concern in cloud computing since an overloaded system may result in subpar performance, perhaps rendering the technology ineffective. An effective load balancing method is always necessary to ensure optimal utilisation of resources. The primary objective of our research is to examine the different load balancing methods and assess their suitability in a cloud computing context.

References

1. Yakhchi, S., Ghafari, S., Yakhchi, M., Fazeli, M., & Patooghy, A. (2015). ICA-MMT: A Load Balancing Method in Cloud Computing Environment. IEEE.
2. Kapoor, S., & Dabas, D. C. (2015). Cluster Based Load Balancing in Cloud Computing. IEEE.
3. Garg, S., Kumar, R., & Chauhan, H. (2015). Efficient Utilization of Virtual Machines in Cloud Computing using Synchronized Throttled Load Balancing. Proceedings of the 1st International Conference on Next Generation Computing Technologies (NGCT-2015), 77-80.
4. Panwar, R., & Mallick, D. B. (2015). Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm. IEEE, 773-778.
5. Belkhouraf, M., Kartit, A., Ouahmane, H., Idrissi, H. K., Kartit, Z., & Marraki, M. E. (2015). A secured load balancing architecture for cloud computing based on multiple clusters. IEEE.
6. Kang, L., & Ting, X. (2015). Application of Adaptive Load Balancing Algorithm Based on Minimum Traffic in Cloud Computing Architecture. IEEE.
7. Chien, N. K., Son, N. H., & Loc, H. D. (2016). Load Balancing Algorithm Based on Estimating Finish Time of Services in Cloud Computing. ICACT, 228-233.
8. Bhatt, H. H., & Bheda, H. A. (2015). Enhance Load Balancing using Flexible Load Sharing in Cloud Computing. IEEE, 72-76.
9. Moharana, S. S., Ramesh, R. D., & Powar, D. (2013). Analysis of Load Balancers in Cloud Computing. International Journal of Computer Science and Engineering (IJCSE), 102-107.
10. Patel, M. P. V., Patel, H. D., & Patel, P. J. (2012). A Survey On Load Balancing In Cloud Computing.



International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)

A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.

Available online at: www.ijsrset.com

ISSN (Print) : 2347-6729

ISSN (Online) : 2348-3105

JIR IF : 2.54

SJIF IF : 4.334

Cosmos: 5.395

Volume 6, Issue 12 - December 2018 - Pages 127-135

International Journal of Engineering Research & Technology (IJERT), 1-5.

11. Kaur, R., & Luthra, P. (2013). Load Balancing in Cloud Computing. International Journal of Network Security, 1-11.
12. Nishant, K., Sharma, P., Krishna, V., Nitin, & Rastogi, R. (2012). Load Balancing of Nodes in Cloud Using Ant Colony Optimization. IEEE, 3-9.
13. Xu, Y., Wu, L., Guo, L., Chen, Z., Yang, L., & Shi, Z. (2011). An Intelligent Load Balancing Algorithm Towards Efficient Cloud Computing. AI for Data Center Management and Cloud Computing: Papers from the 2011 AAAI Workshop (WS-11-08), 27-32.
14. Sidhu, A. K., & Kinger, S. (2013). Analysis of Load Balancing Techniques in Cloud Computing. International Journal of Computers & Technology, 4(2), 737-741.
15. Elzeki, O. M., Reshad, M. Z., & Elsoud, M. A. (2012). Improved Max-Min Algorithm in Cloud Computing. International Journal of Computer Applications, 22-27.
16. Bura, D., Singh, M., & Nandal, P. (2018). Analysis and development of load balancing algorithms in cloud computing. International Journal of Information Technology and Web Engineering (IJITWE), 13(3), 35-53.
17. Kumar, P., & Kumar, R. (2019). Issues and challenges of load balancing techniques in cloud computing: A survey. ACM Computing Surveys (CSUR), 51(6), 1-35.
18. Bezawada, A., Marella, S. T., & Gunasekhar, T. (2018). A systematic analysis of load balancing in cloud computing. International Journal of Simulation--Systems, Science & Technology, 19(6).