JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Cloud Data Center's Virtual Machine Optimization Based on User Needs

Mrs.R P Jadhav,
Ashokrao Mane Group of Institutions,
Vathar, Kolhapur, India

Prof.P.S.Powar
Ashokrao Mane Group of Institutions,
Vathar, Kolhapur, India

Abstract- ***The utilization of virtual machines (VMs) in cloud data centers has become increasingly prevalent for a variety of applications, resulting in the need to allocate these VMs to physical machines (PMs) based on both energy consumption and quality of service (QoS). However, the primary concern of cloud users is their own requirements such as throughput and reliability. This paper proposes an allocation scheme that optimizes user requirements in a cloud data center. The proposed method allocates VMs to PMs based on their usage of hardware resources and the current throughput of PMs in the data center. It also sets CPU utilization thresholds to determine whether migration is necessary, and uses energy consumption before and after allocation to choose which VMs are reallocated. Experimental simulations demonstrate that the proposed method outperforms existing algorithms in terms of total energy consumption, CPU utilization, number of PMs used, and service-level agreement (SLA) violation while ensuring user requirements are met.***

Keywords- ***Data centers of cloud, allocation of VM, consuming Energy, allocating Task, Throughput***

I. INTRODUCTION

Cloud computing is a way of using computer resources from anywhere, anytime. You can use applications, storage, and other things without owning any hardware. This service is provided by data centers that store data and run applications. You only pay for what you use. This is a great choice because you can get resources when you need them, without buying and maintaining your own hardware. It's a popular choice because you only pay for what you use, and you don't have to buy and maintain your own computer equipment.

Cloud computing is very popular and lots of people are using it, which means there's a high demand for computing resources. This demand is causing an increase in energy consumption. To solve this problem, we use virtualization technology. It can create a virtual version of a computer environment, like an operating system, storage device, or network components. It creates many operating systems on one computer. Each of them can run its own operating systems and applications, making it behave like a physical computer. Virtualization technology is used to create many virtual versions of a computer environment on one physical computer. This helps save energy and make computing more efficient.

Virtual machines can be moved from one physical computer to another without any interruption or downtime. This process is called migration. It allows system administrators to work on virtual machines without affecting the system's users. This way they can perform maintenance or upgrades on virtual machines without any disruptions. Overall, virtualization technology and migration are

important tools for managing cloud computing resources. They help maintain system performance and efficiency, while also reducing energy consumption.

II. LITERATURE REVIEW

The paper [1] provides the importance of efficient resource allocation in cloud computing environments to ensure maximum utilization of available resources while minimizing energy consumption and meeting service level agreements. This paper highlights different resource allocation strategies, including priority-based allocation, load balancing, and virtual machine migration techniques.

The authors also discuss the challenges of resource allocation in cloud computing, such as the dynamic nature of workloads, the heterogeneity of resources, and the need to balance performance and energy consumption. The existing resource allocation algorithms are often static and unable to adapt to changing workloads and user requirements. They propose a priority-based dynamic resource allocation approach that considers the priority of tasks and the available resources to allocate resources effectively. The approach involves three phases: priority-based task scheduling, resource allocation, and resource monitoring and optimization.[1]

The paper [2] by Afzal S and Kavitha G provides a comprehensive overview of load balancing techniques in cloud computing. The author propose a hierarchical taxonomical classification of load balancing techniques in cloud computing, which includes global load balancing, DNS-based load balancing, network load balancing, application load balancing, virtual machine load balancing, and container load balancing.

The paper highlighting their respective advantages and disadvantages. For instance, global load balancing ensures low latency and improved user experience by directing user requests to the nearest data center, while DNS-based load balancing distributes traffic across multiple servers through a DNS server. Network load balancing involves using specialized hardware devices such as load balancers or switches, while application load balancing distributes traffic based on specific application requirements such as user location or protocol. Virtual machine load balancing and container load balancing involve distributing workload across multiple virtual machines or containers.

The authors provide insights into the significance of load balancing in cloud computing, noting that it improves resource utilization, reduces downtime, and enhances performance. They also discuss the challenges associated with load balancing in cloud computing, such as scalability, fault tolerance, and security.[2]

The paper [3] propose a mathematical model that aims to optimize the distribution of workload across multiple servers in a cloud environment. The authors then present their proposed mathematical model, which considers the factors of processing time, communication time, and migration time to optimize the load balancing process.

The proposed model uses an optimization algorithm based on the particle swarm optimization (PSO) technique to minimize the makespan, which is defined as the time taken to complete all the tasks in the cloud. The authors present a detailed analysis of their proposed approach and compare it with existing load balancing techniques such as round-robin and least-connection algorithms. The results show that the proposed approach outperforms existing techniques in terms of makespan and resource utilization.

The author highlighting the significance of load balancing in cloud computing and the need for further research to optimize the load balancing process. The proposed approach is a valuable contribution to the field of cloud computing and can be utilized to enhance the performance of cloud-based applications and services.[3]

The paper [4] presents a dynamic load balancing algorithm that balances the workload among virtual machines in cloud computing. The authors, Kumar M and Sharma S, propose a new algorithm that aims to improve resource utilization, reduce response time, and enhance the overall performance of cloud-based applications.

The authors then present their proposed dynamic load balancing algorithm, which uses a threshold-based approach to allocate resources among virtual machines. The proposed algorithm considers factors such as CPU utilization, memory utilization, and network bandwidth to calculate a threshold value for each virtual machine. If the utilization exceeds the threshold value, the algorithm dynamically allocates the resources to other virtual machines with lower utilization rates.

The authors provide a detailed analysis of their proposed algorithm and compare it with existing load balancing techniques such as round-robin, weighted round-robin, and least-connection algorithms. The results show that the proposed algorithm outperforms existing techniques in terms of response time and resource utilization.[4]

The author "Senthamarai N" proposes a new approach that utilizes machine learning techniques to predict the workload and improve resource utilization in the cloud.

The author then presents the proposed migration prediction approach, which uses machine learning algorithms to analyze historical data and predict future workload patterns. The proposed approach utilizes multiple machine learning algorithms such as decision trees, neural networks, and support vector machines to analyze historical data and predict future workload patterns. The approach considers factors such as CPU utilization, memory utilization, and network bandwidth to predict the workload.

The proposed approach compare with existing workload balancing techniques such as round-robin, weighted round-robin, and least-connection algorithms. The results show that the proposed approach outperforms existing techniques in terms of response time and resource utilization.[5]

The paper **[6]** proposes a genetic algorithm-based approach to virtual machine (VM) allocation in cloud systems. The authors argue that current allocation policies often ignore interference, which can lead to suboptimal resource utilization and performance degradation. The authors discuss various approaches such as load balancing, dynamic allocation, and optimization-based approaches that have been proposed in the literature to address interference in VM allocation.

The paper proposes a genetic algorithm-based approach called IAGA, which incorporates interference metrics into the fitness function used by the genetic algorithm.[6]

The authors Jena UK, P.K. the B, and M.R. Kabat a present a load balancing algorithm in a cloud computing environment. The authors argue that load balancing is important to achieve optimal resource utilization and improve system performance in cloud computing environments. The authors highlight limitations of existing approaches such as: B. Requires prior knowledge of the workload and cannot manage dynamic workloads. To overcome these limitations, this paper proposes a hybrid meta-heuristic His algorithm that combines ant colony and particle swarm optimization. The proposed algorithm aims to distribute VM load dynamically and efficiently. The proposed hybrid metaheuristic algorithm offers a promising solution to address the challenges posed by dynamic workloads[7]

III. METHODOLOGY

Significant Features

Our new system will have some important features that make it better than other systems:

1. **Virtual Machines Migration In Cloud**

   Virtual machines can be moved from one computer to another without turning them off. This is really helpful because it keeps everything running smoothly.

2. **To Minimizes the Energy Consumption**

   By migrating the virtual machines lively the system leads to minimization of resource energy. After determining the load on system merging of virtual machines on physical machine which leads to full utilization of it and then turning off the idle physical machine leads to the reduction in consumption of energy.

3. **Hardware Independent System**

   The system doesn't depend on any specific computer hardware.

The proposed approach to managing resources for energy minimization differs from previous works as it employs an economic model to tackle the cloud resource management problem. This is achieved by migrating virtual machines while considering potential energy savings. The approach considers a practical cloud environment with various heterogeneous physical machines that offer multiple resource types, as well as users who request diverse VM instances. The proposed mechanism is flexible and energy-conscious, utilizing live virtual machine migration to reduce power usage by powering off physical machines and turning them on as needed

System architecture:

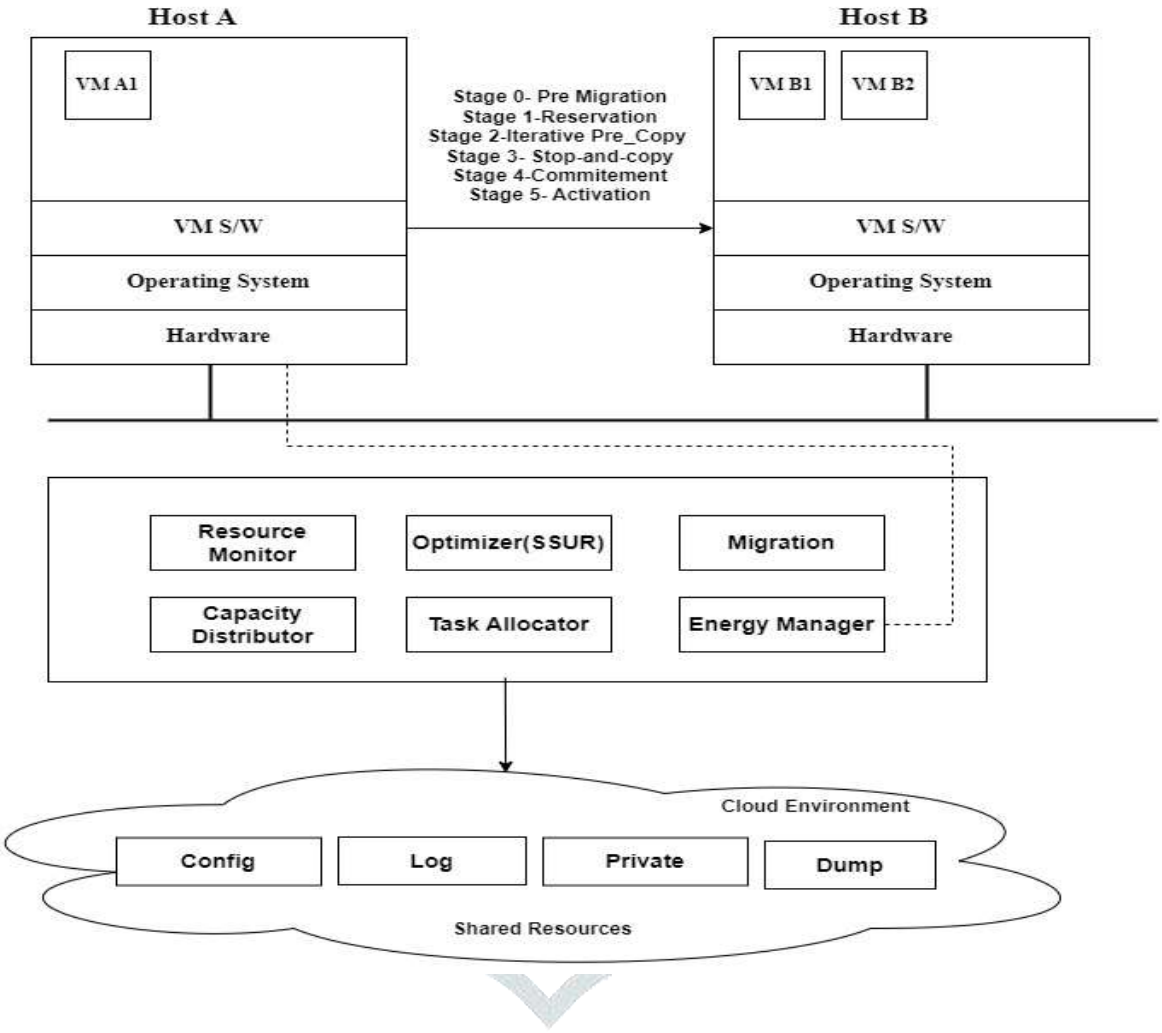

Fig 1. System architecture

The proposed system architecture (as shown in Fig. 1) includes physical hosts and multiple virtual machines. The virtual machines can be migrated from one physical host to another without interrupting their execution by using a pre-copy method. This involves copying pages of memory from the source machine to the destination host iteratively, with a rate-adaptive algorithm controlling the impact of migration traffic on running services. In the final stage, the virtual machine is paused, and any remaining pages are copied to the destination before resuming execution.

The above diagram shows how a computer system works when it has many different parts that can change and adapt over time. These parts are called virtual machines, and they are part of a larger system called the cloud.

Log: The program log checks how much a computer's CPU and IO (input/output) are being used and writes it down in a special folder that everyone can access.

Dump: When virtual machines move from one computer to another, a copy of their memory is saved in another folder that everyone can access.

Conf: This folder has important files that tell the system how to work.

Private: This folder has files that only the virtual machine can access, like its own hard drive.

To move a virtual machine without interrupting it, a pre-copy method is used. This involves copying small pieces of memory from the original computer to the new one, one by one. Special hardware is used to make sure everything gets copied correctly, and the speed of the copying is adjusted so it doesn't slow down other things happening on the computer. Once most of the memory is copied, the virtual machine is stopped briefly so any leftover bits can be copied over, and then it starts running again. It's not a good idea to use a different approach where the new computer tries to 'pull' everything over because that can create problems and slow things down.

When we move an operating system to a different computer, we do it very carefully to make sure nothing goes wrong. Even though computer problems can be really bad, we make sure that the operating system is always as safe as it was on the original computer. To do this, we treat the moving process like a conversation between the two computers involved. We make sure that if something goes wrong, the move can be stopped and everything goes back to how it was before.

Stage 0 Pre-Migration: First, we have a virtual machine running on a computer called host A. To make the move faster in the future, we can choose a different computer called host B and make sure it has enough space for the virtual machine to move there.

Stage 1 Reservation: Next, we ask to move the virtual machine from host A to host B. Before doing this, we check that host B has enough space for the virtual machine. If there's not enough space, the virtual machine will continue running on host A as usual.

Stage 2 Iterative Pre-Copy: We start copying all the parts of the virtual machine from host A to host B. We do this in small pieces so it doesn't take too long. After the first time we do this, we only copy the parts that have changed since the last time we copied everything.

Stage 3 Stop-and-Copy: When most of the virtual machine has been copied to host B, we pause it on host A and send all its network traffic to host B. Then we copy any parts that haven't been copied yet. At this point, we have a complete copy of the virtual machine on both host A and host B, but the copy on host A is still the "real" one.

Stage 4 Commitment Host B tells host A that it has a complete copy of the virtual machine. Host A says "ok" and stops running the virtual machine on its computer. Now, the virtual machine runs on host B and is the "real" one.

Stage 5 Activation: Finally, we turn on the virtual machine on host B and it starts running as usual.

IV Observation and Results

Sr. No	Author	Objective	Algorithm	Result
1	Chandrashekhar S. Pawar and Rajnikant B. Wagh	The aim was to create a resource allocation method for cloud systems that prioritizes tasks based on importance and reduces user wait time.	1)Load Balancer 2) Forming a task list based on priorities 3) Cloud min-min scheduling (CMMS) 4) Priority Based Scheduling Algorithm (PBSA)	The proposed approach showed a significant reduction in the average waiting time for users from 180 seconds to 60 seconds. The approach also showed an increase in resource utilization from 78% to 92%.
2	Afzal S And Kavitha G	The objective was to categorize load balancing methods in cloud computing into a hierarchical structure and offer a complete overview of these techniques.	1) A hierarchical taxonomical structure	The research categorized 32 load balancing techniques in cloud computing into a structured overview called a hierarchical taxonomy.

3	Muhammad Junaid, Adnan Sohail, Rao Naveed Bin Rais, Adeel Ahmed, Osman Khalid, Imran Ali Khan, Syed Sajid Hussain, Naveed Ejaz	The goal was to create a load balancing method for cloud computing that reduces response time, network traffic, and maximizes resource utilization in an optimized manner.	1) Support Vector Machine 2) Swarm Optimisation 3)DFTF Algorithm	The proposed method improved performance metrics compared to existing load balancing methods. It reduced response time by 32%, network traffic by 46%, and increased resource utilization by 27%. The approach also improved load balancing distribution, resulting in more efficient resource usage
4	Kumar M and Sharma S	The goal was to develop a load balancing algorithm for cloud computing that evenly distributes workload among virtual machines, enhances resource utilization, and minimizes response time.	1)A dynamic load balancing algorithm 2) Task scheduling 3) Used a simulation-based approach(CloudSim)	The proposed algorithm outperformed existing load balancing algorithms by improving performance metrics, achieving balanced workload distribution, improving resource utilization, reducing response time, and showing better fault tolerance for a more reliable and stable cloud computing environment.
5	Senthamarai N	To develop a migration prediction approach that can accurately predict overloaded and under loaded workload in cloud environments	1) Resource Checker 2) Resource Estimation Model 3) machine learning algorithms, 4) Proactive Markov Decision Process (MDP)	An accuracy rate of 95% in predicting overloaded and under loaded workload in cloud environments
6	Tarannum Alimahmad Bloch, Sridaran Rajagopal, Prashanth C. Ranga	The aim was to create a VM allocation policy based on genetic algorithms that minimizes interference between virtual machines and improves overall performance.	1) IAGA – Interference Attentive Genetic Algorithm-based VM allocation. 2) A genetic algorithm 3) using CloudSim, an open-source cloud simulation framework,	The simulation results indicate that the IAGA algorithm performs better than existing VM allocation policies in terms of reducing interference and improving resource utilization. It shows up to 26% improvement in resource utilization and up to 43% reduction in interference levels compared to other policies.
7	U.K. Jena a , P.K. Das b, M.R. Kabat a	The aim is to create a hybrid meta-heuristic algorithm that optimizes load balancing in cloud computing environments, improving the overall performance of these systems.	1) Q-learning algorithm 2)QMPSO for load balancing 3)Particle Swarm Optimization (PSO) 4) Genetic Algorithm (GA) techniques.	A new hybrid meta-heuristic algorithm outperforms existing algorithms like RR, TH, and MTH in terms of makespan, execution time, and average response time. The proposed algorithm also enhances the utilization of cloud resources, leading to superior cloud computing system performance.

Table 1: Comparison table

V. Conclusion

The study proposes the SSUR approach for optimizing virtual machine allocation in cloud data centers based on user requirements, considering service response time, energy consumption, and resource utilization. The experimental results show that the proposed approach outperforms existing methods by achieving a better balance between user requirements and data center resource utilization. Additionally, the SSUR approach has the potential to reduce energy consumption and carbon emissions in cloud data centers, promoting efficiency and sustainability. In summary, this study provides a promising solution for improving cloud data center efficiency and sustainability while meeting user requirements.

Bibliography

[1] C. S. Pawar and R. B. Wagh, "Priority based dynamic resource allocation in cloud computing," in *Proceedings - 2012 International Symposium on Cloud and Services Computing, ISCOS 2012*, IEEE Computer Society, 2012, pp. 1–6. doi: 10.1109/ISCOS.2012.14.

[2] S. Afzal and G. Kavitha, "Load balancing in cloud computing – A hierarchical taxonomical classification," *Journal of Cloud Computing*, vol. 8, no. 1. Springer, Dec. 01, 2019. doi: 10.1186/s13677-019-0146-7.

[3] M. Junaid *et al.*, "Modeling an optimized approach for load balancing in cloud," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3024113.

[4] M. Kumar and S. C. Sharma, "Dynamic load balancing algorithm for balancing the workload among virtual machine in cloud computing," in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 322–329. doi: 10.1016/j.procs.2017.09.141.

[5] S. N, "Migration Prediction Approach for Predict the Overloaded and Under Loaded Workload in Cloud Environment," *International Journal of Computer Networks and Applications*, vol. 9, no. 1, p. 51, Feb. 2022, doi: 10.22247/ijcna/2022/211600.

[6] T. A. Bloch, S. Rajagopal, and P. C. Ranga, "IAGA: Interference Aware Genetic Algorithm based VM Allocation Policy for Cloud Systems." [Online]. Available: www.ijacsa.thesai.org

[7] U. K. Jena, P. K. Das, and M. R. Kabat, "Hybridization of meta-heuristic algorithm for load balancing in cloud computing environment," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2332–2342, Jun. 2022, doi: 10.1016/j.jksuci.2020.01.012.