# Particle Swarm Optimization(PSO) based Dynamic Load Balancing in Cloud Environment

Dr. Amanpreet Kaur
CEC, Landran (Mohali)
Amanpreet.it@cgc.edu.in

Dr. Parminder Singh
CEC, Landran (Mohali)
singh.parminder06@gmail.com

Harpreet Kaur Toor
CEC, Landran (Mohali)
harpreettoor.appsci@cgc.edu.in

Bhalinder Singh
Ericsson India Pvt.Ltd.
bhalinder.singh@ericsson.com

**Abstract - Cloud Computing is a technique involving multiple resources being requested by versatile cloud users for allocation of shared resources. The tasks of application requests are allotted to virtual machines (VMs). In different situation different machines get different load. So, load balancing becomes necessary among different VMs. A decentralized load balancer is used to identify best number of tasks to be allocated to each VM. During load allocation to VM, for execution of tasks, the most optimal VM is identified for the load which is best capable of handling the task. The cloud user's application task is them mapped onto that VM to reduce the energy consumption and total execution time. In this paper, decentralized load balancing technique is used to distribute the load on each virtual machine which is enhanced using particle swarm optimization (PSO) which is a swarm based heuristic optimization technique. Moreover, the results are analyzed and compared with centralized load balancer for energy efficiency and throughput parameters.**

**Keywords -** Virtual Machines; Centralized; Decentralized; Energy;Particle Swarm Optimization;Load Balancing.

## I. INTRODUCTION

Cloud computing is a budding technology that provide access to computing resources similar to utility services like water and electricity services. Cloud computing provides various computing services with high degree of flexibility to address user's requirements. The services provided by cloud computing paradigm are on pay-as-you-go basis. These services mainly enhance large scale business processes and scientific applications without being apprehensive about investment in infrastructure, licensed software, maintenance and management of hardware/software [1].Cloud computing is based on of virtualization technology which incorporates the creation of virtual machines (VMs) running on a multiple physical machines. Virtualization allows efficient utilization of cloud resources while reducing the overall power consumption, cost, time and other infrastructure.

A cloud environment is complex and heterogeneous due to unpredictable resource requests [2]. It is a challenging task to get accurate information about the state of the system. As the huge set of resources are shared so complex policies are used to manage them. The various factors which affect the management of resources in cloud computing are performance, functionality and cost [3].Due to its versatile and extensive use, Cloud computing has been growing popular among end users and corporate world. However, beyond these advantages, the cloud computing technology suffers many problems which need to be resolved to cultivate best from its profits. Load balancing is one of these issues which plays an important role in growth of this emerging technology and needs to be addressed [4]. Cloud service providers are responsible for providing cloud services using pay-per-use model with cost benefits to its customers. Various large scale popular applications such as social networking sites, E- commerce etc. can provide benefit in terms of minimal costs through using cloud computing which provides as internet based computing services. These services are provided by cloud infrastructure providers that consider the user's requirement and provide them services based on parameters like

Load Balance, Quality of Service (QoS) and other factors which directly affect the consumption of cloud resources by users.

## II. TASK SCHEDULING IN CLOUD

The available resources should be utilized efficiently without affecting the service parameters of cloud. Task scheduling is an important research area in cloud computing which has involved great attention of researchers. Different scheduling algorithms running in cloud environment have been proposed [5]-[7]. However, most task scheduling algorithms that have been proposed are based on an optimization algorithm. Scheduling process in cloud involves 3 steps:

*1. Resource discovering and filtering:* In this phase, the resources present in the cloud infrastructure are discovered by the data center broker who further collects information relating to them.

*2. Resource selection:* In this process, the target resource is selected on the basis of certain parameters of resources and tasks.

*3. Task submission:* During this phase, the task is mapped onto the selected resource.

Various entities involved in task scheduling in cloud computing are:

- *Cloud User:* They refer to the end users who send application requests to cloud. The cloud services are provided to them.
- *Data center Broker:* They intermediate between the data centers and the end users. The broker is responsible to maintain the entire resource information of the cloud datacenter and to manage the resources efficiently. Its purpose is to store the data for remote user.
- *Cloud Information Services:* It is the platform where virtual directory or information regarding the services provided to the end users by cloud service providers is maintained. The data center broker refers to this directory while allocating the resources to the end user. The services list is accessed from these virtual directories (figure 1).

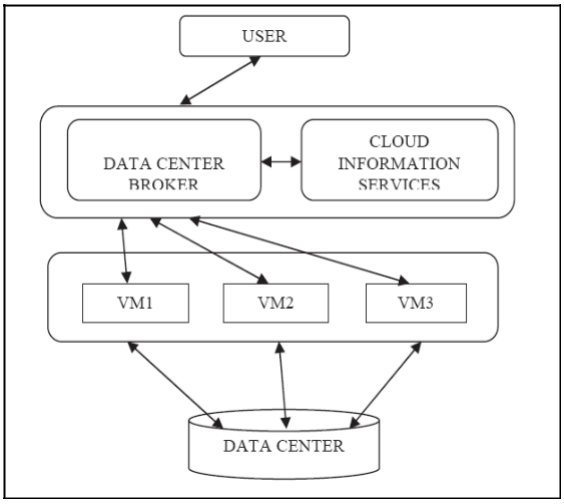


Figure 1. Scheduling Process in Cloud Environment

- *Virtual Machines:* These are the processing machines which processes the request forwarded by the client.
- *Data Center:* All the cloud computing resources-processors, servers, storage, memory, networks are maintained in data centers. These are the data storage and processing place where all the data related activities are being taken place.

## III. LOAD BALANCING

Load balancing techniques are used to make sure that each machine in the cloud datacenter performs approximately the equal number of tasks at any point of time [6]. The dynamic workload of the cloud is distributed evenly among the nodes available for processing. For this, load balancing is done. However, the load can refer to CPU utilization, memory or storage usage or it can network load. The objective is to distribute the load among machines evenly to improve the overall performance task execution over cloud resources (processing power, memory, network and storage). The load is evenly distributed among machines to avoid the situations of overloading and underloading VMs. Techniques are required to optimally manage huge cloud data and streamline the load across machines of data centers so as to achieve high efficiency. Load Balancing techniques are categorized as in figure 2.

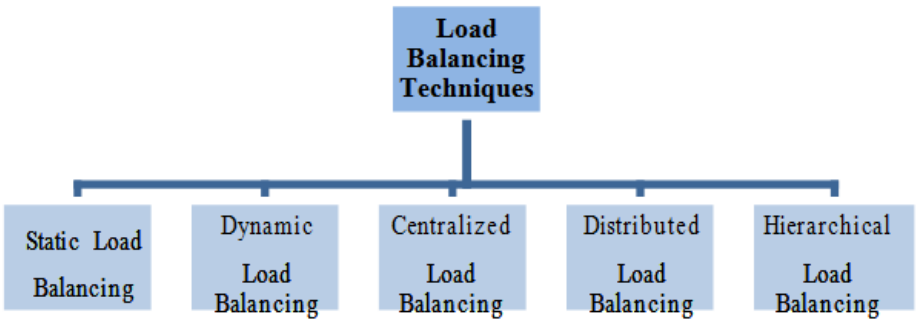


Figure 2. Types of Load Balancing

*A. Static and Dynamic Load Balancing Techniques*

A number of load balancing techniques have been proposed by researchers. Cloud computing environment can be static or dynamic depending on how the cloud is configured by cloud developer as per the demand of cloud service provider. Static environment consist of homogeneous resources without any flexibility with previous knowledge about cloud resources like CPU power, memory, storage available along with the initial details of the user's requirements which remain fixed throughout the execution. While in dynamic cloud environment the resources are heterogeneous, flexible and dynamic. The runtime load information can change the scheduling decisions and demands dynamic load balancing. Similarly, load balancing algorithms also behave differently in static and dynamic environments. Load balancing techniques used in static cloud environment cannot be applied in dynamic cloud environment that demands load changes during execution. Majority of load balancing techniques in cloud computing are dynamic because of the varying user requirements and resource heterogeneity in cloud environment [3-5][8].

B. *Centralized Load Balancing Techniques*

Under this technique, the scheduling and load allocation decisions are taken by a single node (central node). This central node acts as a master which distributes the workload of incoming requests among the slave nodes by considering the information of cloud resources. It can work in static or dynamic cloud environment following the corresponding load balancing technique. The central (master) node executes the load balancing algorithm while other nodes interact with the central node. This technique takes minimum time to analyse the cloud resources, however, the central node can act as a single point of failure with high overhead of managing the entire workload among available nodes. Commonly used static algorithms following the centralized load balancing approach are Round Robin [3], MaxMin [5], MinMin [5], and Central Manager Algorithm [8].

C. *Decentralized load balancing technique*

This technique depends on previous information about the application which is static in nature and the VM workload. Distributed algorithms are best suitable for consistent and stable environments where requests and resources do not change. It does not consider the current state of system but considers factors like processing power, storage capacity, memory usage and recent information about the system performance. Distributed algorithms are based on master – slave concept and before starting the execution, the performance of processor is determined [9-11].

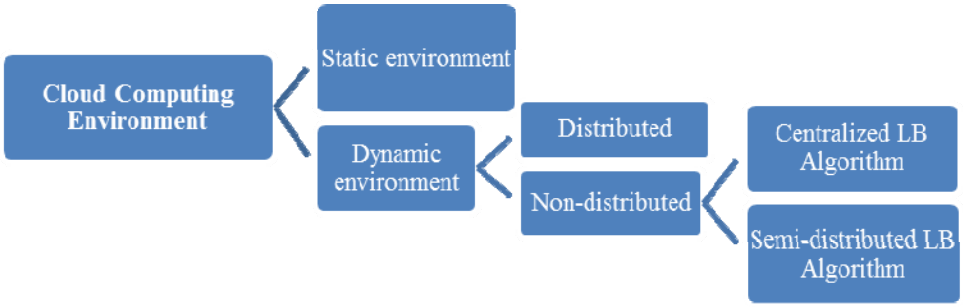


Figure 3. Load Balancing Algorithms in cloud Environment

*D. Virtual Machine (VM) Migration*

Load balancing is used to avoid wastage of resources when machines are underutilized below their capacity. When a large number of requests are targeted to a single Virtual or physical Machine then it gets overloaded with workload above its capacity. This results in increased response time of the applications. This problem can be solved by migrating tasks from one VM (highly loaded) to another which is lightly loaded. Thus the resources are managed automatically in cloud environment. In [12], a distributed data center is used to manage the resources dynamically in cloud environment. A node agent is used to keep track of the resource utilization by each VM of a local data center. Then the migration among VMs of the datacenter is decided after the communication between the local agentand the global agent. However, minimum migrations are to be done since it incurs cost in terms of memory utilization when VMs are migrated from one location to another. So, either minimum or no VM migrations must be involved during load balancing. This problem is solved in [13] which use a genetic algorithm based on both present state and previous data about the system.

In present work, decentralized approach is used for managing the cloud workload (independent tasks)in an energy-efficient manner. To achieve optimized results of load balancing, Particle Swarm Optimization technique is applied in order to minimize the active resources of the data center, thereby increasing and decreasing energy consumption and while avoiding overloading and under loading of VMs.

## IV. RELATED WORK

Hongsheng Su (2007) et al. presented a proper balance of overall load over the available resources in cloud computing paradigm. Authors minimized time and cost involved during execution of tasks on cloud model. Many existing algorithms provide load balancing and better resource utilization [4]. Kaur A. & Kaur B. proposed Hybrid approach based resource provisioning and load balancing framework for workflows execution [19]. The proposed approach optimizes the utilization of VMs by uniformly distributing the load. Pankaj Arora [8]et al. solved cloud workflow based scheduling problem by proposing a Set-Based Particle Swarm Optimization technique in which users define various QoS parameters like reliability, deadline and budget constraints. Srushti Patel et al. proposed a technique for task scheduling and load balancing in multi-processor systems using PSO algorithm which minimize the makespan and improve average utilization of processors in an optimized way[3]. S. Devipriya et al. [13] proposed an algorithm which is based on particle swarm optimization algorithm for reducing the cost and time. They used non-virtualized environments while performing data migration, but these methods are not applicable to cloud environment. S. Thamarai Selvi et al. [3] proposed and implemented a technique of dynamic dispatching system for Cloud Computing Environment based on Particle swarm optimization (PSO). Their experimental results show an improvement in efficiency and utilization of dispatched resource during scheduling in cloud environment [3].

A new framework is proposed by Gulshan Soni et al. that provides power aware, scalable, energy efficient Cloud computing architecture using variable resource management, power-aware scheduling techniques, and live migration with minimized VM design [14]. Vedang Shah et al.[15] proposed a hybrid Particle Swarm Optimization (HPSO) whose results show improvement in execution ratio and decrease in average schedule length. M. Sridhar et al. [16] proposed a Distributed Dynamic and Customized Load Balancing (DDCLB) algorithm to handle the arriving user's requests dynamically (as in Amazon EC2 instances). The authors consider CPU utilization of the running instances of EC2 carried out load balancing and also provide elasticity while serving the requests [16]. Ms. Kunjal Garala et al.[17] presented a decentralized technique for energy efficiently managing the VMs along with scalability aspect while provisioning the resources in large enterprise clouds. Each node perform its operation autonomously. A distributed set of load balancing rules are used by each node while managing its workload. Nodes which are underutilized, migrate their workload to other neighbours which can easily handle it within their load constraints. Here, nodes are follow hypercube topology. Then the nodes whose workload has been migrated to other nodes are switched off. Hence, reducing the power consumption and efficiently utilizing the energy resources. Azade Khalili et al. [18] improved makespan time of Particle Swarm Optimization (PSO) algorithm. Authors used PSO with dynamic scheduling with variable inertia weights in cloud environment to minimize makespan. The tasks are mapped and scheduled so that assigned task run on the available resources which helps to maximize resource utilization and minimize the make span. Their results show a linear descending inertia weight with an average of 22.7% reduction in make-span[18].While tasks are migrated from one VM to another, it incurs cost in terms of time, memory and communication cost. Geng Yushui et al. [9] tried to reduce this migration cost in terms of time. Kaur A. & Kaur B. proposed Hybrid approach based resource provisioning and load balancing framework for workflows execution [19]. The proposed approach optimize the utilization of VMs by uniformly distributing the load. [20] addressed the problem of overflow and underflow VM management and identifies which load balancing better improves the performance and quality of service has been answered with a number of experiments.

## V. METHODOLOGY

Cloud provides access to unlimited resources in the form of software, infrastructure and platform for further developing and deploying new software. In this paper cloud infrastructure as a service (IAAS) model has been undertaken from service provider point of view. The request from the cloud clients for resources (especially computing resource) are received by cloud hypervisor (VM monitor) which allocates the available VMs to the requests. The resources are allocated in terms of virtual machines having processing power, memory and storage of physical machine. The present load on each VM is checked and the request (task) is allocated to the machine with minimum threshold load. If the VM is overloaded its load is shifted to other VMs (which are not overutilized). Similarly, the load of underutilized VMs is shifted to other machines and later turned off after migration of their load to other machines to reduce energy consumption and increasing throughput (figure 4).

The decentralized load balancer is responsible for allocating the resources to the tasks. The load balancing is further optimized using PSO technique whose convergence rate is faster than other optimization techniques. The results of power consumed in the entire process from resource allocation to load balancing are compared with the traditional centralized approach for validation of the work.

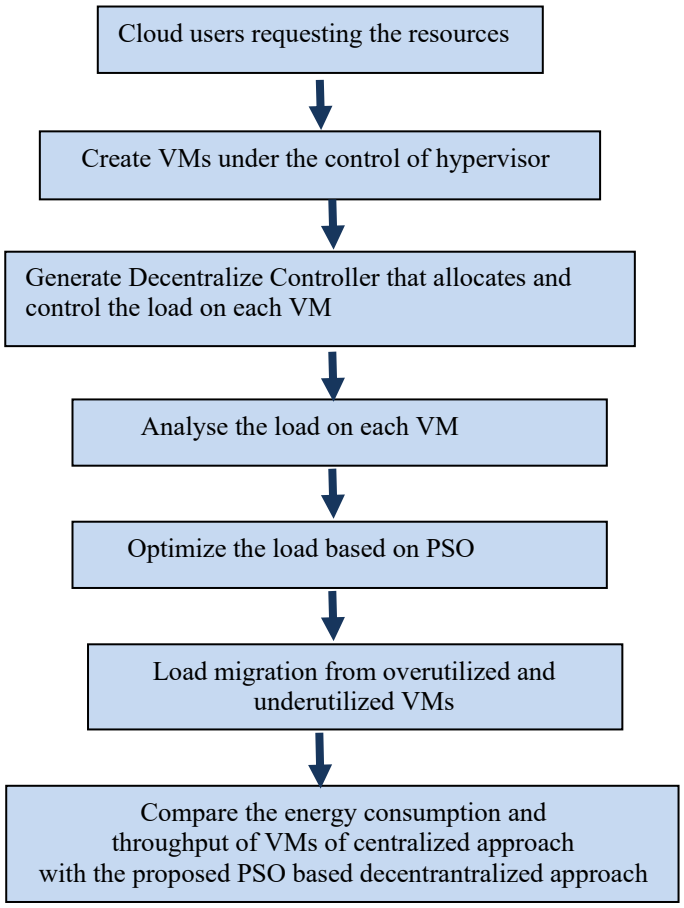


Figure 4. Methodology

## VI. RESULTS AND DISCUSSIONS

To evaluate and analyze various parameters for achieving load balancing using decentralized approach with PSO, CloudSim tool has been used. Energy consumption and throughput results show that decentralized load balancing approach using PSO is better than centralized load balancing approach. Each VM has a maximum capacity to accept the load with defined buffer for temporarily storing the incoming requests. The VM is considered overloaded when this queue is filled and it is no more capable of accepting incoming request. On the other hand, the machine is under-loaded when its load is less than the maximum capacity that it can undertake. The process of migrating the request from one machine to another for avoiding the VM of the system from being overloaded and hence, reducing power consumption by shifting the load of two or more virtual machines to one or more machine(s)lightly loaded machines and setting the machine whose load is shifted to power safe/sleeping mode. Thus improving the throughput and resulting in energy efficiency.The energy utilized by

PSO based decentralized technique shows that less energy is required for load balancing in comparison to the centralized technique. There is an improvement around 18% (Figure 5).

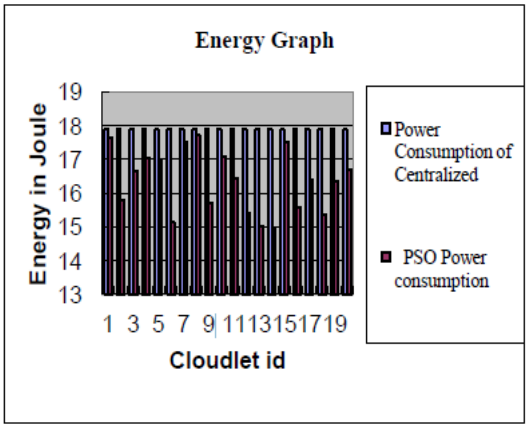


Figure 5. Energy Consumption

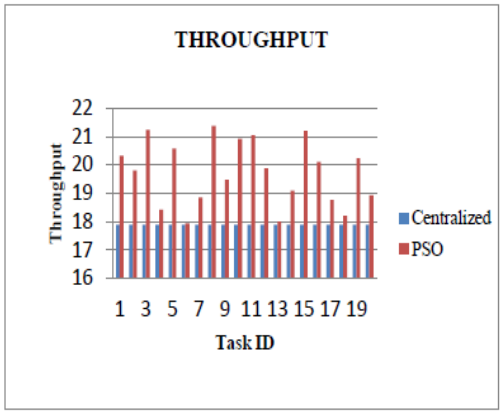


Figure 6. Throughput

In comparison to the centralized, the distributed (PSO) shows better throughput during execution of tasks(figure 6).

## VII. CONCLUSION AND FUTURE SCOPE

The results of energy consumed and throughput achieved of decentralized load balancing technique using Particle Swarm Optimization Algorithm are better than the centralized load balancing technique. The load is distributed and balanced among VMs with minimum VM migration and achieving high resource utilization and throughput.

Current research work is based on PSO using which best possible solution is being identified. This PSO based decentralized load balancing technique imparts better results in comparison to centralized load balancing technique. As future work, dynamic Load balancing can be analyzed for further enhancements using Artificial Bee Colony Algorithm (ABC) and BAT optimization techniques. The parameters like scalability, reliability, node functioning can be enhanced by various dynamic load balancing algorithms

## REFERENCES

[1] M. Vanessa, A. Tchernykh, and D. Kliazovich., "Dynamic Communication-Aware Scheduling With Uncertainty of Workflow Applications in Clouds", Communications in Computer and Information Science (2016), pp.169-187, 2016.
[2] G. Yushui and Y. Jiaheng,“Cloud data migration method based on PSO algorithm”, 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science, pp.143-146, 2015.
[3] S. Patel and P. Mishra, “A Survey of Resource Allocation Policies in Cloud Computing", International Journal of Computer Science and Information Technologies, vol. 4(3), pp. 416-419, 2013.
[4] S.Thamarai Selvi and, R. Eberhart,“Resource Allocation Issues and Challenges in Cloud Computing”, in 2014 International Conference on Recent Trends in Information Technology, pp. 1942–1948, 2014.
[5] R. Piplode and M. Naghibzadeh, "An Overview and Study of Security Issues & Challenges in Cloud Computing" The Third IEEE/IFIP International Conference on Internet, Uzbekistan, 2012.
[6] Vignesh, V., Sendhil Kumar, K. S., & Jaisankar, N., “Resource Management and Scheduling in cloud environment”, International journal of scientific and research publications, Vol. 3(6), pp. 1-6, 2013.
[7] Sonia Sindhu, “Task Scheduling in Cloud Computing”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4(2), February 2015.
[8] T.Guerout, S.. Medjiah, D. Costa, & Monteil, T., “Quality of service modeling for green scheduling in Clouds”, Sustainable Computing: Informatics and Systems, vol. 4(4), pp. 225-240, 2014.
[9] Lu, F., Parkin, S., & Morgan, G., “Load balancing for massively multiplayer online games”, . In Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games, Oct 30-31, Singapore, 2006.
[10] H. Su, “The Available Transfer Capability Based on a Chaos Cloud Particle Swarm Algorithm”, IEEE Sixth International Conference on Grid and Cooperative Computing, 2007. GCC 2007, Los Alamitos, CA, pp. 221-227, 2007.
[11] Mohsen & H. Delda, “Balancing Load in a Computational Grid Applying Adaptive, Intelligent Colonies of Ants”. Informatica, Vol. 32 pp. 327–335, 2008.
[12] H. Matsumoto, G. Tzortzakis, and Alex Delis, “Dynamic Resource Management in cloud environment”. Fujitsu Sci. Tech. J, Vol. 47(3), 270-276, 2011.
[13] S.Devipriya and Han Ruilian, “Study on cloud computing task schedule strategy based on MACO algorithm”, Computer Measurement & Control, Vol.19 (5), pp.1203-1211, 2013.
[14] G. Soni and A. Taleb-Bendiab, “A Novel Approach for Load Balancing in Cloud Data Center” in Proceedings of IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Perth, Australia, April 2014.
[15] Shah, V., & Trivedi, H., “A Distributed Dynamic and Customized Load Balancing Algorithm for Virtual Instances”, In 2015 5th Nirma University International Conference on Engineering (NUiCONE) IEEE, 2015.
[16] M.Sridhar, G. Karypis, V. Kumar, "Hybrid Particle Swarm Optimization Scheduling for Cloud Computing", In Proceedings of 2015 IEEE International Advance Computing Conference (IACC), pp. 1196-1200, 2015.

[17] K. Garala and W. Hong Sun, "A Performance Analysis of Load Balancing Algorithms in Cloud Environment", IEEE transactions on systems, man, and cybernetics—part A: systems and humans, Vol. 33(5), 2015.
[18] A. Khalili and S. M. Babamir, "Makespan Improvement of PSO-based Dynamic Scheduling in cloud environment", In 23rd IEEE Iranian Conference on Electrical Engineering (ICEE), vol.15, pp.613-617, 2015.
[19] Kaur, A., & Kaur, B., "Load balancing optimization based on hybrid Heuristic-Metaheuristic techniques in cloud environment", Journal of King Saud University-Computer and Information Sciences,(in press), 2019.
[20] Kaur, A., Kaur, B., & Singh, D., "Meta-Heuristics Based Load Balancing Optimization in Cloud Environment on Underflow and Overflow Conditions", Journal of Information Technology Research (JITR), 11(4), pp. 155-172, 2018.