Review of Literature on Optimization of Load Balancing and Task Scheduling in Cloud Computing Environments Using Artificial Neural Networks

Tija P Thomas, Assistant Professor Department of computer science, Dr.G Shankar Govt Women's First Grade College and P G Study Centre, Ajjarakad Udupi, Karnataka (India)

Abstract

Thousands use websites at some point of time for other. Cloud has limitation in maintaining load obtained from all demands at time any point of time. It results in destroy of the entire network. It is the process in which computing resources and workloads are distributed to more than one server. Workload is divided between two or more servers, hard drives, system interface and other computing resources resulting in good use and system response time. A huge traffic web site requires a high powerful load balancing for smooth performance in business. It helps maintaining network firmness, operation and security against network failures. As more people utilize the cloud, more employment opportunities become available. With constraints such as a limited make-span, a high utilization rate of available resources, minimal execution costs, and a rapid turnaround time for scheduling, this becomes an NP-hard optimization issue. The number of solutions/combinations increases exponentially with the magnitude of the challenge, such as the number of tasks and the number of computing resources, making the task scheduling problem NP-hard. As a result, achieving the optimum scheduling of user tasks is difficult. An intelligent resource allocation system can significantly cut down the costs and waste of resources. For instance, binary particle swarm optimization (BPSO) was created to combat ineffective heuristic approaches. However, the optimal solution will not be produced if these algorithms are not paired with additional heuristic or meta-heuristic algorithms. Due to the high temporal complexity of these algorithms, they are less useful in real-world settings. For the NP problem, the binary variation of PSO is presented for workload scheduling and balancing in cloud computing. Considering the updating and optimization constraints stated in this research, our objective function determines if heterogeneous virtual machines (VMs) Phave the most significant difference in completion time. In conjunction with load balancing, we developed a method for updating the placements of particles. According to the experiment results, the proposed method surpasses existing metaheuristic and heuristic algorithms regarding work scheduling and load balancing. This level of success has been attainable because of the application of Artificial Neural Networks (ANN). ANN has demonstrated promising outcomes in resource distribution. ANN is more accurate and faster than multilayer perceptron networks at predicting targets.

Keywords: Load, Balancing, Cloud, Literature

Introduction:

First, load in load balancing refers not only to website traffic but also of memory capacity, CPU load and network on server. Its main function is to ensure every network of system has the same amount of work. It means neither that the system is under used or overloaded. It makes equal distribution of data based on how busy the server is. Without this the client has to wait long for processing this data this could be a frustrating for them. During this process, data like CPU process and job arrival rate in the processors are modified. Failures in application of this head have severe consequences like data loss. Different companies utilize various load balancers with numerous load balancing techniques. The most commonly used model or techniques is the “Round Robin” load balancing.

Load indicates not only a traffic website but includes network load, memory capacity and CPU load of any server. This method promises that every network in system has similar number of work at a time. Any of them is highly under or over-loaded use. It gives data based on how busy every node or server is.

Using techniques from parallel and distributed computing, cloud computing makes shared computer resources accessible to clients via the Internet. The “pay-as-you-go” business model has nearly democratized cloud computing. Cloud providers, service providers, and end users participate in this phase of software deployment. Cloud service providers offer their customers computational capabilities via virtual computers (VMs). Service providers utilize these virtual machines when it comes to application-level client services. Service providers implement task scheduling algorithms to spread client jobs across virtual machines, reduce response times, ensure a high quality of service (QoS), and maximize resources. Because of this, the job scheduling algorithm is a vital part of any cloud architecture. Cloud computing needs adjustments to the several scheduling techniques utilized in various computer environments. It is possible for a scheduling method optimized for a cluster to perform poorly in the cloud. Before the algorithm can deal with the structure of the cloud environment, the method's parts need to be moved into the problem space. The greater the variety of virtual machines and size of the workloads being managed, the greater the number of available task configurations. Finding the shortest path across all potential permutations is one of the most challenging problems in computer science. Even though metaheuristic algorithms have already been utilized to assist with cloud scheduling, the authors of this work have devised a new load balancing version of the original PSO approach for cloud scheduling. A load balancing method and a metaheuristic algorithm can be used to make both service providers and customers happy (in terms of resource use and user happiness, respectively) (make-span reduction).

Our Contribution

1. Improve load balancing, so requests are distributed more fairly based on the machine's processing capacity. Improved VM load balancing resulted in much more significant time reductions than previous research.
2. Examine the complete vector of resources (storage, RAM, and bandwidth) rather than just the CPU when determining whether user requests are suitable with VMs. Consequently, our model is more suited for the cloud.
3. To meet the needs of service providers and customers, there needs to be a fitness function that cuts down on time while using resources better.
4. Previous approaches to work schedules would be simplified if a single-goal strategy that considered the interests of both service providers and customers was implemented.
5. As a result, the PSO and load balancing algorithms can be effectively coupled.

The task scheduling algorithms can lengthen the time needed to complete work while simultaneously reducing the throughput of the overall cloud system. In this regard, the goal of cloud computing is to increase overall performance and make better use of the available computing resources in an environment that contains various types of devices. Several different work scheduling techniques, such as the ant colony optimization algorithm (ACO), the particle swarm optimization algorithm (PSO), and the genetic algorithm (GA), are utilized in a cloud computing environment. In this study, to schedule the activities in a load balanced manner, we have linked the ANN technique with the BPSO strategy to create a hybrid method. Our solution outperforms the traditional BPSO task scheduling algorithm by increasing resource utilization by 22% and decreasing mean time by 33%. For this contribution, we have developed a complete literature analysis and a novel load balancing-enabled job scheduling system, as follows: in **Section 2**, you will find a discussion connected to the BPSO and ANN, in addition to specifics addressing various work scheduling methods for cloud computing. At the end of **Section 2**, there is also a comparison of the different ways to schedule tasks in the form of a table. **Section 3** elaborates on the issue formulation and the ANN-BPSO approach, starting with definitions, an explanation of the BPSO Technique, problem definition, and a discussion of the prospered framework, which consists of BPSO and ANN. The ANN-BPSO model, System model, Inertia weight approach, and suggested Task scheduler are all covered in this part. The experimental setup in the cloud computing environment for task scheduling is discussed in **Section 4**, along with the results, experimental configuration setup, dataset information, findings, and discussions. **Section 5** presents the research's recommended conclusion with the future perspective of task scheduling and load balancing in cloud computing.

Literature Reviewed

Literature, shows the presence of many load balancing procedures, resource designation methods, advancement systems and calculations for building throughput and productivity and improve the reaction time in cloud condition. Every methodology has its own advantages.

Resource Allocation and Scheduling

Resource scheduling and designation assume an indispensable job in CC generally to create performance implementation and use of resources, vitality sparing, clients QoS necessities fulfillment as well as expanding the benefit of cloud suppliers. Moreover, its calculation as well as strategy legitimately impact cloud cost and execution.

Qingxian et al. (2019), have investigated a case wherein the basic data sources (or shared resources) of all stages are known. Building a game that sees each phase as a player, helps incorporation of a game hypothesis with network data envelopment analysis (DEA) for the investigation of the result distribution issue in a three-organize framework. Network DEA models are made for computing the ideal benefits of the framework during resource sharing (i.e, pre-and post- joint effort ideal benefits), and the Shapley esteem technique is applied for the allotment of the expanded benefits of the framework to its stages.Results show the game among stages in a three-organize framework as an added substance. A numerical model is given to show our the technique.

XingChen et al. (2019) have proposed a self-adaptive resource allocation strategy that is really a structure made out of criticism circles, every one of which experiences a planned iterative QoS forecast model and PSO-based runtime decision algorithm. As opposed to past Qo forecast models which foresee a QoS esteem for the last time, thisstrategy improves the anticipated QoS esteem towards the best one. In the forecast, the remaining task at hand is first utilized along with, the designated resource, the genuine QoS esteem and an activity of resource allocation to deliver a QoS esteem. At that point PSO-based runtime decision algorithm is utilized together with the anticipated QoS incentive to decide on future resource allocation activities. The circles are repeated until the PSO-based algorithm proposes no further improvement over the present resource allocation. The methodology is assessed on RUBiS benchmark. Representing this and based on theequivalent verifiable data, the strategy can accomplish a superior QoS expectation exactness that is 15% higher than the present cutting edge. Besides, a gained 5-6% improvement of the viability of cloud application resource allocation has been demonstrated.

Manasrah, A.M et al. (2018) Cloud computing condition permits resource sharing as well as on-request benefits for customers. Business forms are kept up by work process advancements that feature the difficulties seen in utilization of resources in a productive manner due to conditions between errands. Hybrid GA_PSO's technique reducesexpense and parity heap on the required undertakings. This technique is utilized for the productive use of resources. Exhibition of this technique is contrasted with different techniques like PSO, GA, MTCT,WSGA and HSGA. What's more, it improves the heap adjusting of the work process application over the accessible resources.

Also, the multifaceted nature of the work process scheduling issue, makes building up an enhanced work process scheduling algorithm for work process errands dissemination to the accessible resources inside a sensible overhead, that is, CPU time an exceptionally testing process.

Wei et al. (2018) have indicated the dependence of resource allocation of cloud method on not perfect data Stackelberg game (CSAM-IISG) utilizing HMM in CC condition. CSAM-IISG has appeared to expand the benefit of both resource provider and candidate.

Initially, HMM was utilized to anticipate specialist organization's present offer utilizing the chronicled resources dependant request. Progressive foreseeing of offer helped building up the not perfect data Stackelberg game (IISG). IISG spurs specialistorganizations the pick the ideal offering procedure as per the general utility, accomplishing the most extreme benefits. Dependence on unit costs of various sorts of resources, resource allocation method is suggested to ensure ideal increases for foundation provider. Suggested resource allocation method bolster synchronous allocation for both multiservice suppliers as well as different resources.

Moon, Y.et al. (2017) Computing resources from CC is charged depending on climate. Similarly scheduling resource technique has a complex effect on clients. A novel cloud scheduling technique is utilized. It is like for deciding on this ACO technique that makes a profitable assignment of assets to VM. Techniques of diversification as well as fortification are taken up utilizing slave ants. These ants handle the issue of enhancement. New ACO technique known as SACO with slave ants staying away from overheads provide good execution. This technique deals with the NP-difficult problem even more beneficially. Heterogeneous clusters used right presently are checked for costreduction.

Thanasias et al. (2016) CC has picked up significance in numerous fields conveying different changes to data advancement. IaaS gives flexible providing and denial of computing assets. It is a compelling method for workloads and a temporary, test or vary startlingly. The prerequisite for assets shifts from time to time and oughtto be kept inside the given budget while keeping up efficiency.

The need for resource variation after some resource provisioning such of obtained budget is productively utilized for maintain obtaining job done efficiency is an important challenge. This resource exhibits the task scheduling problem and provisioning resourcefor numerous tasks in IAAS cloud. It gives novel scheduling and provisioning techniques used for task implementation inside the budget for limit log jam due to budget constraint.

Choudhary et al. (2018) had proposed the application of workflow scheduling in CC to provide solution for NP complete issue. Ithas numerous problems, vitality issues, time span and cost. Numerous meta and heuristic techniques are suggested. The technique satisfies the method to some degree showing integration of meta and heuristic technique such as GSA and HEFT. Significant characters are MCR and SLR. Different trial causes are represented in this technique outputs like HGA, HEFT and GSA. Outcomes are produced by ANOVA test. Future findings are in numerous cloud condition.

Madni et al. (2016) had dealt with the allocation of resources using meta-heuristic techniques for IaaS CC condition. Different problems have been noticed in resource allocation meta heuristics techniques, comparative specification as well as experimental instruments used for various method validation. Classification and survey are the bases for additional researches in IaaS CC.

Ma et al. (2014) have presented 5 significant problems in scheduling resource and CC allotment including locality aware,reliability aware and vitality aware resource of scheduling and allocation. SAAS includes scheduling and allocation as well as scheduled work process. They have made thorough analysis and conversion of different current resource allocation as well as the sequence of scheduling and techniques of current issues as away as various specifications.

Zhang and Su (2014) have done research on basic cloud information target resource scheduling network with its problems. They have seen cloud data focus thought structure and scheduling resource. They have portrayed a method for scheduling resources for cloud data target with a condition to powerful scheduling of cloud resources and less power task scheduling.

This research work shows problems in current field of data target compared to different scheduling resources. Use and profit of resources are low for cloud suppliers and vitality use in data community is high. There is need to improve resource scheduling for data centers for future work.

The basic technology of CC is resource scheduling in resource management. It deals with techniques to enhance efficiency with dynamic scheduling relying on threshold, improved genetic technique with double available and increased ACO for scheduling as suggested by Huang et al. (2013). Areas taken up have been committed with Map decline scheduling research include graph models, dynamic requirement, temporary weight modify, adaptive scheduling, utility based optimization, customization, forecast, equality numerous clients, audit of map reduce entomb reliance and enhance lessening stage. A major task was to increase overall efficiency, reaction time and increase output producing fairness and locality. Open area of work for newapplications had increased the makespan and improved fairness in various clients (Elghoneimy et al., 2012).

Wu et al. (2013) have suggested the QoS scheduling task with the aim of implementing huge required tasks on resources which has numerous least time. Requirements are concluded to satisy special Qos parameters. The technique is in contrast to Min-Min technique andBerger method as well as the makespan of suggested method had been found to be superior to the other two.

Amit Nathani et al. (2011) have suggested a technique in scheduler named Haizea for resource allocation such as best exertion, deadline touchy, adverted reservation and immediate. Haizea is resource lease manager which uses resource leases as abstracts of resource allocation and actualizes leases by VMs allot. The important aim of the authors is to limit the resource dismissal rate and the shuffle cost to give above said resource allocation sequence for the IAAS cloud. It uses 2thoughts, namely, backfilling and swapping for deadline touchy resource allocation criteria. The main idea is leasing 4 specifications for trails, namely, duration, start time, number of hubs and deadline.

Kejiang Ye et al. (2011) have suggested resource reservation dependant on live shift structure of different VMs. Focused machine in structure has 4 VMs, namely, migration decision maker and controller, Resource monitor and reservation control.

the authors have targeted migration performance improvement by live shifting of VMs and suggested 3 optimization methods, namely, source machine optimization, numerous machine parallel migration and workload aware shifting criteria. The authors have used specifications like total migration time, workload efficiency and downtime for enhancing shifting performance. He claims resource reservation method is need at source machine as well as focus machine.

Congfeng Jiang et al. (2011) have presented a compelling resource allocation issue that depends on the real time data on workload as well as efficiency request for processing services. They have suggested the stochastic method of resources in virtual condition and scheduling heuristics techniques and resource allocation with service level constraints. Targeted machine efficiency has been taken as efficiency feedback to source for enhancing the viability of things to approach a dynamic workload. This improves the resource allocation method suggested by authors.

Linlin Wuet al. (2011) have suggested the resource allocation technique for SAAS suppliers which limits framework cost as well as SLA violation for SAAS purchasers to ensure service satisfaction. The authors have considered buyers QoS specifications, as for instance, framework specifications and reaction time in server start time. They have presented 3 cost driven techniques from 2 shoppers as well as SAAS suppliers view. The first technique was one which increases profit by adjusting the number of SLA violations. Consecutive techniques increased profit by cost reuse of VMs limit, with a huge space. The third technique increased profit by cost reuse VMs limit withless space. The second and the third suggested by authors were simulated on cloud sim condition.

Different types of resource allocation techniques have been suggested in cloud. Gunho Lee et al. (2011) have suggested a structure for optimized resource allocation in IAAS based cloud structure. CurrentIAAS structure is unaware of facilitated apps significant. This pathallotted resources free of requirements with a major effect in the efficiency for disseminated data serious apps. Structure that has "what if" method to manage allocate decision taken by IAAS has been suggested for locating this resource allocation problem. The structure used an expectation motor with lightweight simulator for calculation of the efficiency of the given resource allocation and GA to find optimized sequence in large search space.

The hybrid strategy was utilized by Ahmad M. and his co-workers [1] to balance the load in a cloud environment with diversely accessible resources. This method, known as hybrid GA-PSO, allocates jobs to resources in the most effective manner possible. Utilizing genetic algorithm and particle swarm optimization, its objectives are attained. The authors suggest that using Max is less painful and less expensive while balancing the demand on cloud computing infrastructure.

Workflow scheduling was introduced by Nirmala SJ et al. [2] to ensure that crucial scientific procedures on IaaS clouds are planned. Therefore, workflow scheduling systems that utilize Catfish's particle swarm optimization (PSO) are more efficient and consume fewer resources than their competitors. When a large number of jobs are conducted concurrently, the execution of an algorithm consumes a significant amount of resources and takes a long time. These solutions do not account for the necessity of load balancing in the cloud work schedule, which continues to be a widespread problem.

Mishra et al. [3] devised the LB approach by modeling its structure after the characteristics of a flock of birds using the BSO-LB algorithm. Virtual machines (VMs) represent food particles in this scenario, while jobs represent birds. The authors received the datasets utilized for their measurements from the cloudlet-based GoCJ. The authors have drastically reduced the reaction time, allowing for an equal division of effort. The FCFS (First Come, First Serve), SJF (Shortest Job First), and RR (Round Robin) approaches are compared along with the proposed technique.

Muhammad Junaid et al. [4] used a support vector machine to classify the input request. Depending on the categorization of the assignment, it was then assigned to a hybrid metaheuristic approach that combined Ant colony optimization with file type formatting. According to them, the hybrid metaheuristic algorithm they developed can be used to keep cloud systems stable. Using criteria such as service level agreement violations, migration times, overhead times, throughput, and quality of service, they compared the efficacy of the suggested strategy.

A mutation that aids in workload distribution. When Awad AI et al. [5] proposed PSO, they aimed to increase load balancing and dependability while decreasing transmission costs, execution, transmission durations, make-spans, and round-trip times. In this method, each virtual machine will execute a proportional number of tasks to its load. If there are multiple significant projects, it is possible that the cloud system will not be able to correctly balance the load, causing the task to take longer than usual to complete. In addition, they did not consider the amount of time and money clients would need to implement their ideas. Arabnejad H et al. [6] found that although these algorithms achieve good outcomes, their great temporal complexity makes them less suitable for real-world computers. In their investigation, Shabnam et al. [7] utilized the bat algorithm. For load balancing and virtual machine optimization, a hybrid strategy is required. This swarm-based approach has been created and implemented to optimize the load on the virtual machine, ultimately resulting in the load on the actual computer being balanced. Yousef Fahim et al. [8] propose a metaheuristic bat technique for assigning virtual machine work.

In Kruekaew B et al. [9]'s implementation of a hybrid approach, an ABC algorithm and a heuristic technique are merged, which can be regarded as a hybrid approach. Considering the time required to develop and distribute the load is essential. It is possible to demonstrate that this method is effective in either homogeneous or heterogeneous contexts. Using this method, we could significantly minimize the number of conflicting factors previously investigated. This algorithm may have been evaluated against a dataset from the real world by assessing several different qualities of service criteria, such as resource consumption, reaction time, etc. Mala Yadav et al. [10] have developed a multi-criterion scheduling technique for multiprocessor computer systems. This application combines quantum computing and the gravitational search approach, both of which were inspired by nature. This study considers both homogeneous and heterogeneous conditions to determine whether the proposed strategy is beneficial. According to the findings, it produces better-than-anticipated results for the various scheduling objectives, such as load balancing, resource usage, and make-span. This method may have been enhanced in terms of consuming less energy and for workflow applications.

Mala Yadav and colleagues [11] conceived a hybrid metaheuristic algorithm. This system is composed of the genetic and particle swarm optimization algorithms. The primary objective of this study is to find a solution, or at least an approximation of a solution, to the problem of load balancing between virtual machines. The authors assert that the findings obtained from the tests were the best that could have been obtained under the conditions. Banerjee S et al. [12] designed load balancing to distribute cloudlets (tasks) around virtual machines (VMs) based on each machine's capability, hence decreasing task completion time as well as the makespan of VMs and hosts in the data center. This strategy assigns large workloads (those with a significant size) to VMs that are already available. However, prioritizing large tasks can significantly increase the delay for many minor actions (small jobs), resulting in a significantly longer overall completion time. The cost of execution and resource utilization have not been recognized as indicators of the essential quality of service by cloud providers and clients. Sequentially, many algorithms restrict themselves to a limited set of resources and job sizes to offer an optimal result (large and small tasks). By adopting a BPSO-based task scheduling technique, we provide an initial population and target function more suited to the work at hand and context.

References:

[1] Abdulhamid .S. M, Latiff .M. S. A and Idris .I (2015), "Tasks Scheduling Technique using League Championship Algorithm forMakespan Minimization in IaaS Cloud", ARPN Journal of Engineering and Applied Sciences, Vol. 9, pp. 2528-2533.

[2] Alexander .S (2014), "Efficient Cloud Storage Confidentiality to Ensure Data Security", IEEE CCSW 2014 International Conference on Computer Communication and Informatics, Vol. 03, No. 05.

[3] Amanpreet Kaur and Bikrampal Kaur (2018), "Load balancing optimization based on hybrid Heuristic-Metaheuristic techniques in cloud environment", Journal of King Saud University - Computer and Information Sciences.

[4] Amit Nathani, Sanjay Chaudhary and Gaurav Somani (2011), "Policy based resource allocation in IaaS Cloud" in ELSEVIER - Future Generation Computer Systems, Vol. 28, pp. 94-103.

[5] An Qingxian, WenYao, Ding Tao and Li Yongli(2018), "Resource sharing and payoff allocation in a three-stage system: Integrating network DEA with the Shapley value method", Elsevier, Journal of omega, Vol. 85, pp. 16-25.

[6] Anuradha .V.P and Sumathi .D (2014), "A Survey on Resource Allocation Strategies in Cloud Computing", ICICES - S.A.Engineering College, Chennai, Tamil Nadu, India.

[7] Armougum .A, Orriols .E, Gaston-Bellegarde .A, Marle .C. J.-L and Piolino .P (2018), "Virtual reality: A new method to investigate cognitive load during navigation", Journal of Environmental Psychology,101338. doi:10.1016/j.jenvp.2019.101338

[8] Kruekaew, B.; Kimpan, W. Enhancing of Artificial Bee Colony Algoithm for Virtual Machine Scheduling and Load Balancing Problem in Cloud Computing. *Int. J. Comput. Intell. Syst.* **2015**, *13*, 496–510. [Google Scholar] [CrossRef]

[9] Meng, X.B.; Gao, X.Z.; Lu, L.; Liu, Y.; Zhang, H. A new bio-inspired optimization algorithm: Bird Swarm Algorithm. *J. Exp. Theor. Artif. Intell.* **2016**, *28*, 673–687. [Google Scholar] [CrossRef]

[10] Yadav, M.; Gupta, S. Hybrid Meta-Heuristic VM Load Balancing Optimization Approach. *J. Inf. Optim. Sci.* **2017**, *41*, 577–586. [Google Scholar] [CrossRef]

[11] Banerjee, S.; Adhikari, M.; Kar, S.; Biswas, U. Development and analysis of a new cloudlet allocation strategy for QoS improvement in cloud. *Arab. J. Sci. Eng.* **2015**, *40*, 1409–1425. [Google Scholar] [CrossRef]

[12] Chaudhary, D.; Kumar, B. An analysis of the load scheduling algorithms in the cloud computing environment. In Proceedings of the IEEE 2014 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, India, 15–17 December 2014; pp. 1–6. [Google Scholar]

[13] Gupta, N.; Maashi, M.S.; Tanwar, S.; Badotra, S.; Aljebreen, M.; Bharany, S. A Comparative Study of Software Defined Networking Controllers Using Mininet. *Electronics* **2015**, *11*, 2715. [Google Scholar] [CrossRef]

[14] Devi, D.C.; Uthariaraj, V.R. Load balancing in cloud computing environment using improved weighted round robin algorithm for nonredemptive dependent tasks. *Sci. World J.* **2016**, *2016*, 3896065. [Google Scholar] [CrossRef]