Innovative Systems Design and Engineering
ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online)
Vol.7, No.5, 2016
www.iiste.org

A Model of Resource- Aware Load Balancing Scheme using Multi-objective Optimization in Cloud Environment

Kavita[1] Vikas Jhandu[2]
Department of Computer Science and Engineering
Swami Vivekanand Institute of Engineering & Technology, Banur, Punjab

Abstract
Cloud computing is a new class of network based computing that provides the customers with computing resources as a service over a network on their demand. The unique concept of cloud computing creates new opportunities for Business and IT enterprises to achieve their goals. In cloud computing, usually there are number of jobs that need to be executed with the available resources to achieve optimal performance, least possible total time for completion, shortest response time, and efficient utilization of resources etc. To accomplish these goals and achieve high performance, it is important to design and develop a multi objective scheduling algorithm. Hence it is most challenging to schedule the tasks along with satisfying the user's Quality of Service requirements. This paper proposes a multi- objective scheduling algorithm that considers wide variety of attributes in cloud environment. The paper aims to improve the performance of CPU, memory and network operations by reducing the load of a virtual machine (VM) by using Load Balancing Method. Finally, it optimizes the resource utilization by using Resource Aware Scheduling Algorithm.
Keywords: VM, QoS, Non- dominated sorting, Pareto optimal, Makespan, AHP

1. Introduction

In period of last few years, cloud computing has completely grabbed the IT market and majority of the IT industry is already using cloud computing. The word "cloud" in cloud computing refers as internet hence can be referred as Internet based computing or server based computing. It is a most recent new computing paradigm where high quality and low-cost information services are served by cloud service providers on pay-per-use basis. Cloud ensures the availability of resources to user at any time and from any location via internet. The resource demands for different jobs may fluctuate over time.

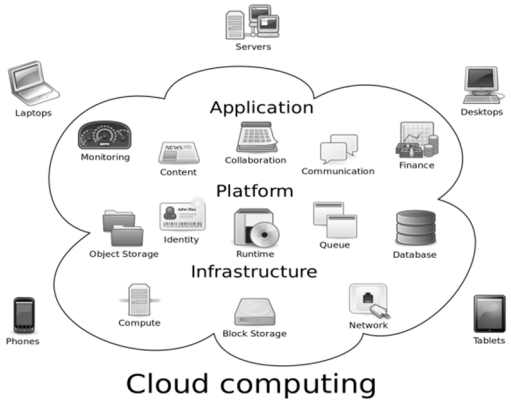


Figure 1 Cloud Computing

Cloud computing offers various service models to the client. These services include SaaS(Software as a service)where user is provided the application software and database access and is charged as per the use, IaaS (infrastructure as a Service) where user is provided the virtual machine by virtualising the physical machine encompassing storage, processing capability and other resources, PaaS(Platform as a Service) as the name itself suggests, a computing platform is provided which includes OS, programming language execution platform and web server [1]. In addition to providing these essential services, it is necessary to optimize the utilization of datacenter resources which are operating in most dynamic workload environments. A single datacenter generally contains a large number of virtual servers running at any instance of time, facilitating numerous tasks and in the meantime the cloud system continues acquiring the batches of task requests. During this context, one has to notice few target servers out many powered on servers, which can fulfill a batch of incoming tasks. So Task scheduling is an valuable issue which is greatly influences the performance of cloud service provider. It is crucial and extremely important to develop an efficient task scheduling algorithm to completely utilize the power of cloud computing and to enhance its performance and throughput.

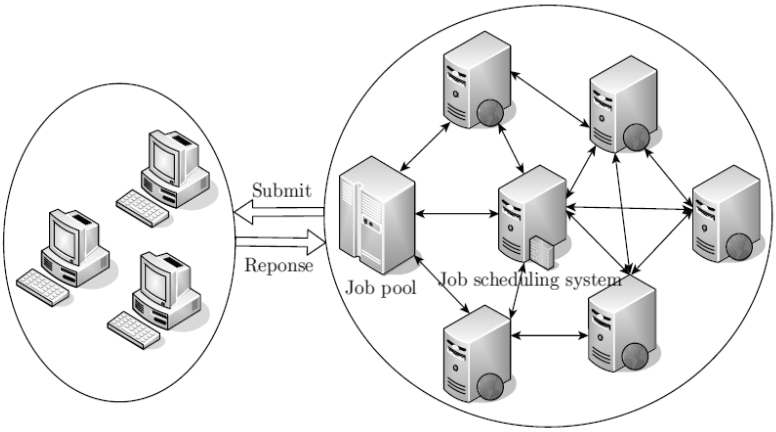


Figure 2 An illustration of job scheduling in cloud

Unlike other scheduling schemes which use include criterion, this paper is a way ahead and addresses multiple criteria including load balance, quality of service, economic principles, best running time and throughput by proposing a multi objective task scheduling scheme using Quality of Service parameters of resource nodes.

2. Related Work

Atul Vikas Lakra et.al [1] explained the underutilization of cloud resources due to poor scheduling of tasks in cloud and came up with a multi- objective task scheduling algorithm in this regard to optimize the throughput, cut the execution expenses at the same time keeping the SLAs intact. The proposed scheduling technique prioritizes the tasks according to the QoS such that low QoS valued tasks are assigned higher priority and vice versa. Cloud broker assigns QoS to VMs on receiving list of VMs from cloud service providers.

To minimize the energy consumption, a multi-objective genetic algorithm (MO-GA) was proposed by Jing Liu et.al [2] which focused on encoding rules, crossover operators, selection operators and the method of sorting Pareto solutions. It also maximizes the profit of services with the deadline constraints. It begins with proposing a job scheduling architecture in cloud computing containing various components to analyze the application along with allocating the appropriate resources to the applications hence improving the effectiveness and efficiency of the computing

Upendra Bhoi et.al [3] proposeda modification of Improved Max-min task scheduling algorithm. The proposed enhanced version of Max-min also includes the expected execution time as a selection basis just like the old improved Max-Min method but there is a slight difference in their working that differentiates them vividly. The old Improved Max-min algorithm assigns a largest task (having Maximum execution time) to a slowest resource (one producing Minimum completion time)whereas the Enhanced Max-min assigns the task having average execution time to the slowest resource (producing Minimum completion time).

To optimize the execution time and cut down the execution costs, Suraj Pandeyet.al [4] proposed a Particle Swarm Optimization technique to schedule the applications to the resources appropriately. PSO is a self-adaptive optimization technique that falls under population based genetic algorithms class and relies on the social behavior of the articles. The simulation results of the proposed technique compared with the existing Best Resource Selection (BRS) technique reveal a potential cost saving and achieved load balancing.

Lipsa Tripathy et.al [5] aimed to achieve the goals of of minimizing the switching time, improving the resource utilization; throughput and server performance by proposing a simple and novel scheduling scheme that mitigates the loopholes of the existing protocols. The paper explores out the drawbacks of the traditional scheduling techniques and improves the resource utilization by using Backfilling and minimizes the makespan by allocating the shortest distance resources to the jobs. Switching time is considered as it is more flexible and reliable.

Hong Sun et.al [6] emphasized on task scheduling algorithms based on comprehensive QoS. The paper considers the new Berger's Model under dynamic cloud computing environment and relates to benefit- fairness algorithm. New Berger game theory model apply the theory of social distribution and game theory on task scheduling in cloud computing environment. In cloud, task scheduling is achieved by assigning independent task to m virtual machine resource to fully utilize the resources in minimum finishing time. If FTi is the finish time of task i, then span is defined as

FT_{max}= max(FTi, i=1,2,3,…n)

The scheduled task is to find the optimal collection that make the spans FT_{max} and ΣFTI minimum in the 2m subset of the possible resource space.

A good scheduler is one which adapts the scheduling strategy according to the type of task and the changing environment. With this concept, Dr. Amit Agarwal et.al [7] proposed a Generalized Priority algorithm

for efficient execution of task. It first detailed down the basic scheduling strategies FCFS and Round Robin scheduling and identified their disadvantages. The proposed technique addresses those loopholes by prioritizing the tasks according to their size and ranking VMs according to their MIPS. The simulation results proved GPA to be better than the other two.

Shamsollah Ghanbari et.al [8] came up with a Job scheduling algorithm based on priority concept in cloud computing including the theory of Analytical Hierarchy Process (AHP). In scheduling, priority of jobs is an essential issue since some jobs are needed to be addressed sooner than other as those jobs are not able to stay for quite a while in a system. The proposed algorithm involves three levels of priorities including: scheduling level, resource level and job level.

Ekta S. Mathukiya et.al [9] introduces a multi-objective task scheduling algorithm that additionally performs non-dominated sorting for ordering of tasks. The proposed algorithm considers task size, makespan, and deadline as the criteria. The task's priority here is allocated in accordance with the QoS value and QoS are assigned to the VMs on the basis of their Millions Instructions Per Second (MIPS) value. The list of VMs possessed by cloud broker is updated after fixed time interval. Based on MIPS the list of VMs is sorted in descending order starting from high QoS VM to low QoS VM and non- dominated sorting is performed on the list to generate non-dominated task's set. These tasks are bounded to VMs sequentially and the process of allocation is repeated for all tasks.

3. Job Scheduling Model and Preliminaries

3.1 Job Scheduling in Cloud

Though cloud computing has received considerable attention and has emerged out as a promising approach and a commercial reality in the information technology domain, however there are still many issues that need more attention. Job scheduling is one such fundamental issue that needs to be addressed to enhance the cloud performance. Job scheduling is actually the efficient allocation of resources to the required jobs under the constraint of the Service Level Agreements (SLAs). It is one of the most prominent activities executed in the cloud computing environment to get maximum profit.

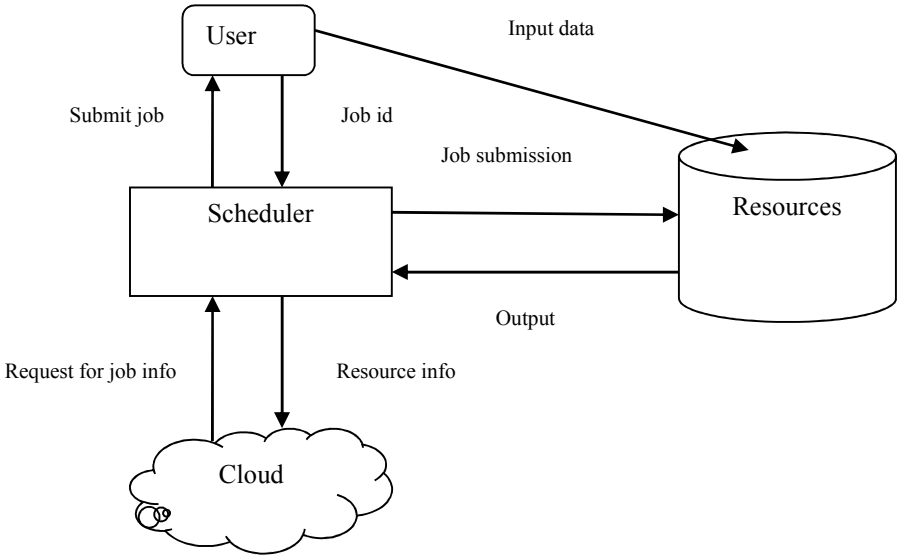


Figure 3 Job scheduling in cloud

The main motive of task scheduling is to attain better cloud performance in terms of better throughput, load balancing among resources, quality of service (QoS), economic feasibility and the optimal operation time. Task scheduling can be viewed from two directions- from the cloud resources users' view, users have to identify which cloud computing resource can meet their job QoS requirements for computing and how much amount to be paid for the cloud computing resources. While, from the cloud computing service providers view, its concern is to gain the maximum profits by offering cloud computing resources, apart from meeting the CCU's job QoS requirements.

Several algorithms & protocols have been proposed till date regarding the scheduling mechanism of the cloud computing. Most of the traditional scheduling strategies were single- objective which could focus on one particular issue. The main examples of traditional scheduling algorithms are FCFS, Round-Robin, Min-Min algorithm, Max-Min algorithm and meta- heuristic algorithms (ACO, GA, Simulated annealing, PSO, Tabu search and many more). In last few years, multi- objective strategies have been explored and developed. Multiple criteria include improving the various Quality of Services and at the same time maintaining the efficiency and fairness

among the jobs.

3.2 Job Scheduling Goals

The job scheduling in Cloud computing is a mapping mechanism from users' tasks to the appropriate selection of resources and its execution. Specific goals of job scheduling include load balance, quality of service (QoS), economic principle, and optimal operation time and system throughput.

- Load Balancing: It is an important measure in cloud and is closely linked to job scheduling. It helps to distribute all loads between all the nodes and ensures that every computing resource is distributed efficiently and fairly. It helps in preventing bottlenecks of the system which may occur due to load imbalance. It is a relatively new technique that provides high resource utilization and better response time.
- Quality of Service: Cloud mainly aims to provide users with computing and cloud services and resources to the users so as to achieve quality of service.
- Throughput: Optimized performance with task scheduling can be measured by throughput. Task scheduling aims to achieve high throughput.

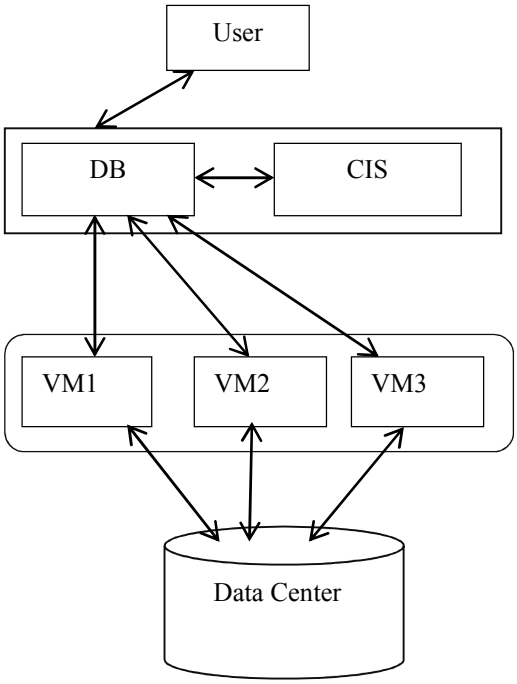


Figure 4 Stages of scheduling

3.3 Contributions

The paper aims to study and analyze the processing time of low level scheduling algorithms and to develop a multi- objective task scheduling using quality of service parameters of resource nodes. It also attempts to improve the performance of CPU, memory and network operations by reducing the load of a virtual machine (VM) by using Load Balancing Method. Finally, it optimizes the resource utilization by using Resource Aware Scheduling Algorithm.

4. Proposed Work

The research involves exploring various low level algorithms in cloud computing like FIFO, SJF and Max-min and analyzes their processing times.

The paper proposes a multi objective task scheduling using QOS parameters for virtual machine. In this, tasks are sorted using non-dominating sort and virtual machines are sorted using MIPS and granularity size. Then the tasks are mapped to virtual machines using some grouping factor in order to optimize the processing and average waiting time.

4.1 Multi Objective Task Scheduling

There are number of datacenters in cloud computing to serve customers demands. Datacenters further consists of number of servers and subsequently each server runs number of VMs. The execution capacity of different VMs varies to execute tasks with different QoS parameter.

Cloud broker is an important entity in cloud computing as it sends request to the cloud service provider

for the QoS of requested task and for list of VMs. The task's priority is allocated as per the QoS value of task in such a way that the lower QoS valued task is given high priority and vice versa. To assign QoS to VMs, Millions of instructions per second (MIPS) of VM are used in such a way that the VM with high MIPS is declared high QoS VM and similar is with low MIPS VM.

The list of VMs possessed by cloud broker is updated after fixed time interval. Based on MIPS the list of VMs is sorted in descending order starting from high QoS VM to low QoS VM and non- dominated sorting is performed on the list to generate non-dominated task's set. These tasks are bounded to VMs sequentially and the process of allocation is repeated for all tasks.

Multiple Objective optimization is a mathematical optimization with more than one objective functions to be achieved and mathematically can be expressed as[9]:

Find X = {x1,x2,……xn}

Which Min(f1(x), f2(x),……fk(x))

where k>=2 represents the number of objectives and X is the set of feasible solutions. A multi-objective optimization does not normally provide a sufficient solution that minimizes all objective functions at the same time. As objectives are conflicting to each other so Pareto optimal solutions are considered where such solutions cannot be enhanced for any objectives without degrading atleast one of the other objectives. Hence, an important concept in multi objective scheduling is domination. In Dominated solution there exists at least one feasible solution better than it in all objectives. In mathematical form, a feasible solution x1 dominates another solution x2 if:

1. fi(x1) ≤ fi(x2) for all objectives i Є {1,2,……,k } and
2. fj(x1) <fj(x2) for at least one objective j Є {1,2,……,k }

Here solution x1 is a Pareto optimal solution as there does not exist better solution that dominates it.

4.2 Load Balancing

Load balancing can be achieved by first mapping the tasks to VM's and further all the VMs to host resources by the means of Task-Based System Load Balancing method. This algorithm achieves the system load balancing by transferring only extra tasks from an overloaded VM rather than migrating the entire overloaded VM. The loads are formulated as:

VMload = Bandwidth+ Ram+ Mips

4.3 Resource Utilization

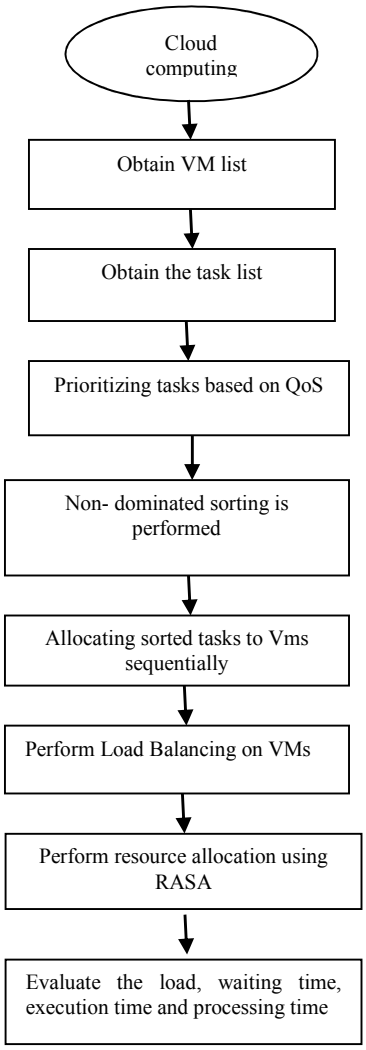
RASA, a combined approach of Max-Min and Min-Min algorithms, is used to achieve efficient resource utilization, further optimizing the resources in terms of accuracy and efficiency. To realize this optimization, it first estimates the completion time of the tasks on each of the available cloud resources and then applies the Max-Min and Min-Min algorithms, alternatively. Small tasks are executed by using Min-Min strategy before the large ones. Max-Min strategy comes into play to avoid delays in the execution of large tasks, to support concurrency in the execution of large and small tasks.

The key objective of the proposed algorithm is to obtain better results than existing algorithm of Shortest Job First (SJF)to reduce the execution time of jobs.

The evaluation parameters considered for performance evaluation are:

a) Load
b) Waiting Time, Execution Time,
c) Processing time

4.4 Flowchart



5. Conclusion and Future Scope

In this paper, Multi- objective task scheduling is emphasized and to achieve the same a multi- objective scheduling algorithm is proposed considering the user's Quality of Service requirements. The task sorting is done using Non-dominated sorting. In addition to this, the paper aims to achieve load balancing on VMs and improved resource utilization using RASA. The evaluation parameters considered in the paper include waiting time, execution time and processing time.

The work can be further extended in future aiming to achieve more efficient performance results. The proposed work is using RASA for resource allocation. In future work, resource allocation can be performed with the improved versions and evolutionary algorithms. In addition to this, the proposed system can be further implemented in real time scenario.

References

[1] Lakra.A.V,Yadav. D.K," Multi-Objective Tasks Scheduling Algorithm for Cloud Computing Throughput Optimization ," International Conference on Intelligent Computing , Communication & Convergence , 2015

[2] Liu J, Luo X. G, Zhang X.M, Zhang F and Li B.N, "Job Scheduling Model for Cloud Computing Based on Multi-Objective Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 3, January 2013

[3] Bhoi U, Ramanuj P.N, "Enhanced Max-min Task Scheduling Algorithm in Cloud Computing", International Journal of Application or Innovation in Engineering and Management, Volume 2, Issue 4, April 2013

[4] Pandey S, Wu L, Guru S M, Buyya1R, "A Particle Swarm Optimization-based Heuristic for Scheduling Workflow Applications in Cloud Computing Environments"

[5] Tripathy L, Patra R.R, "Scheduling In Cloud Computing", International Journal on Cloud Computing: Services

and Architecture (IJCCSA) Vol. 4, No. 5, October 2014
[6] Sun H, Chen S.P, Jin C, Guo K, "Research and Simulation of Task Scheduling Algorithm in Cloud Computing", TELKOMNIKA, Vol.11, No.11, November 2013, pp. 6664~6672e-ISSN: 2087-278X
[7] Agarwal A, Jain S, "Efficient Optimal Algorithm of Task Scheduling in Cloud Computing Environment", International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 7– Mar 2014
[8] Ghanbaria S, Othman M, "A Priority based Job Scheduling Algorithm in Cloud Computing", International Conference on Advances Science and Contemporary Engineering 2012(ICASCE 2012)
[9] Mathukiya E.S, Gohel P.V, "Efficient Qos Based Tasks Scheduling using Multi-Objective Optimization for Cloud Computing", International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 8, August 2015
[10] Chen H, Wang F, Dr Helian N, Akanmu G, "User-Priority Guided Min-Min Scheduling Algorithm for Load Balancing in Cloud Computing"
[11] Vijayalakshmi M, Muthusamy K, "An Efficient Study of Job Scheduling Algorithms with ACO in Cloud Computing Environment", Volume 3, Special Issue 3, March 20142014 International Conference on Innovations in Engineering and Technology
[12] Tarek Z, Zakria M,.Omara F A, "PSO Optimization algorithm for Task Scheduling on The Cloud Computing Environment", ISSN 2277-3061, International Journal of Computers And Technology Vol. 13, No. 9
[13] B. Kanani and B. Maniyar," Review on Max-min Task Scheduling Algorithm for Cloud Computing ," Journal of Emerging Technologies and Innovative Research ,vol. 2 , pp. 781-784 ,2015
[14] A.Bhatia and R.Sharma , " An Analysis Report of Workflow Scheduling Algorithm for Cloud Environment ," International Journal of Computer Applications, vol.119, pp. 21-25 ,2015
[15] Kowsik , K. RajaKumari ," A Comparative Study on Various Scheduling Algorithms in Cloud Environment , " International Journal of Innovative Research in Computer and Communication Engineering , vol.2 , 2014