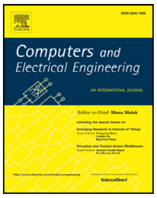
Computers and Electrical Engineering 000 (2017) 1–17

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Deadline constrained based dynamic load balancing algorithm with elasticity in cloud environment☆

Mohit Kumar*, S.C. Sharma

IIT Roorkee, India

ARTICLE INFO

Article history:
Received 22 April 2017
Revised 12 November 2017
Accepted 14 November 2017
Available online xxx

Keywords:
Cloud service provider
Task scheduler
Virtual machine
Load balancing
Makespan time
Elasticity

ABSTRACT

The most challenging problem for a cloud service provider is maintaining the quality of service parameters like reliability, elasticity, keeping the deadline and minimizing the makespan time as also the task rejection ratio. Therefore, the cloud service provider needs a dynamic task scheduling algorithm that reduces the makespan time while increasing the utilization ratio of cloud resources and meeting the user defined QoS parameters. In this paper, we have developed a dynamic scheduling algorithm that balances the workload among all the virtual machines with elastic resource provisioning and deprovisioning based on the last optimal k-interval. Further, the algorithm has been tested on variable number of tasks (10 to 30) to achieve better scalability. The computational results (Figs. 5–10) show that the developed algorithm decreases the makespan time and increases the task to meet the deadline ratio compared with the min-min, the first come-first-serve and the shortest-job-first algorithms in all conditions.

1. Introduction

Cloud computing provides the services either in the form of software application or hardware infrastructure on the basis of pay per use over the internet. It is the collection of heterogeneous resources that contain the characteristics of on demand self-service, scalability (scale-out and scale-up), resource pooling, broad network access, rapid elasticity and higher availability. It provides the different types of services like infrastructure as a service (IaaS), storage as a service (SaaS), software as a service (SaaS), platform as a service (PaaS) etc. These types of services are useful in scientific, business and industrial applications. User can send the request at any time for the resource to cloud service provider and cloud resource broker (cloud service provider) selects the best resource within user-defined deadline and budget. Cloud resource broker provides the on-demand service to the user. The number of users and applications are increasing gradually in cloud environment and in turn there is increase in the workload and traffic at the web applications which are deployed in the virtual machine (cloud resource). Therefore cloud resource broker needs an efficient algorithm that distributes the task fairly in all the running virtual machines and reduces the task rejection ratio so that the entire user task can be executed.

The main objective of the load balancing is to utilize the cloud resource in a manner that improves the average resource utilization ratio, response time and scalability of the web application. Efficient load balancing gives the minimum makespan time of tasks and increases the performance of the system. It also prevents bottleneck of the system which may occur due to load imbalance. Load balancing is one of the challenging research areas in the field of cloud computing whose

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. L. Bittencourt.

* Corresponding author.

E-mail addresses: mohit05cs33@gmail.com (M. Kumar), scs60fpt@iitr.ac.in (S.C. Sharma).

https://doi.org/10.1016/j.compeleceng.2017.11.018
0045-7906/© 2017 Elsevier Ltd. All rights reserved.

Please cite this article as: M. Kumar, S.C. Sharma, Deadline constrained based dynamic load balancing algorithm with elasticity in cloud environment, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.018

objective is to balance the workload among the virtual machines. There are two steps of achieving the load balancing in cloud environment; first one is task scheduling and monitoring the virtual machine. Task scheduling is one of the best known optimization problems (NP Complete) in the field of computer science because cloud has heterogeneous resource (different host and virtual machine configuration) and very rapid on demand request change. Therefore it is difficult to predict and calculate all possible task-resource mapping in cloud environment. So we need an efficient task scheduling algorithm which can distribute the task in effective manner, so that less number of virtual machines faces overload or under-load condition. Second one is to monitor the virtual machine continuously and perform the load balancing operation (either task migration or virtual machine migration). Task migration approach has several advantages like over the virtual machine migration therefore we used the task migration approach in this paper. Cloud resource broker (CRB) monitor the virtual machine continuously in cloud environment. If any virtual machine is in overload or under-load condition after the task scheduling then cloud resource broker start the load balancing operation on the virtual machine and migrate the task from overloaded virtual machine to under loaded virtual machine.

In this paper we have analysed scalability (elasticity) aspect, which is the ability of a system to fit in a problem such that if scope of the problem increases (number of request increase, length of request vary randomly etc.). The ability of auto scaling on upcoming demands in cloud computing is biggest advantage for services provider as well as for user [1]. Auto scaling can reduce the risk, which is associated with request/load overflow causing server failure. Two types of auto scaling approaches are available in cloud environment i.e. reactive and proactive. Reactive approach gives their response to event after they have occurred. Reactive approaches are good for short term services but expensive for long term service. Proactive approach tries to eliminate the problem before they have chance to occur. A proactive scaling system that enables the services providers to schedule dynamic changes in the capacity that should be match with expected changes in application demand. To perform proactive scaling, firstly understand the expected change in workload i.e. one should try to understand that how much workload changes are possible from the excepted workload. Main drawback of proactive approach is that if predictive model is fails then resources will be in overutilization or underutilization mode and violates of the service level agreement (SLA). Proactive approach can be classified into three categories, cyclic, event, and prediction based. In this paper we are using prediction based approach that predicts the future demands on the basis of past history.

Scalability can be classified into two type's horizontal scalability (scale out) and vertical scalability (scale up). Vertical scalability can be achieved by making changes in the existing resources such as memory, hard drives, CPU's, etc. Vertical scalability is not generally used in cloud environment because common operating systems don't support these changes without rebooting on existing resources like memory, CPU's. Adding or releasing of one or more machine instance or computing node of same type is called horizontal scaling. Adding the IT resource in horizontally is called scale-out and releasing the IT resource horizontally scale-in. horizontal scaling is better than vertical scaling in cloud environment because it is less expensive and not limited by hardware capacity.

The reminder of this paper is organized as follows: Section 2 describes the related work in which we will discuss existing load balancing algorithm and technique in cloud environment that is related to present research work, Section 3 holds the problem formulation, Section 4 describes the proposed architecture section 5 presents the dynamic load balancing algorithm with elasticity, Section 6 shows the Analysis and comparison of results and Section 7 conclusion and future scope of the paper work.

2. Related work

Several static [2,5–6] and dynamic algorithms [7,12–19] for load balancing have been proposed in last decade. Static algorithm needs advanced information about the upcoming number of request (task) and information about available cloud running virtual machine (cloud resources). There is no need to monitor the cloud resources continuously because this type of algorithm works well when node has low variation in workload. It is difficult to predict the job execution time in cloud environment therefore job scheduler must be dynamic in nature. Existing job scheduling algorithms like conservative backfill, EASY etc. are unable to fill the resource gap efficiently. The work done by [2] focus at improving the backfill algorithm (IBA) not only improves the processing time of jobs but also provide the guarantee of quality of services in cloud environment. In this paper author's improve the IBA using balanced spiral (BS) method but this algorithm do not provide better processing time when job requests enter randomly in cloud environment. Dubey et al. [3] proposed an algorithm for metascheduler to solve the job scheduling problem in cloud computing and removed the limitation of IBA algorithm. In this paper, authors improved the processing time of upcoming jobs and resource utilization ratio of cloud resources considering priority of job as quality of service parameter. Sahoo et al. [4] proposed an algorithm based upon the greedy technique that reduces the makespan time and execution time of the tasks without using task migration or virtual machine migration approach for load balancing; proposed algorithm does not provide better results in real environment. First come first serve (FCFS) and shortest job first (SJF) [5,6] are static algorithms that are suitable for batch system. Static algorithm gives better results when there is low variation in workload. Literature review on dynamic based algorithm is reported in Table 1. Chen et al. [7] proposed a user guided min-min load balancing algorithm that not only minimize the execution time of the tasks but also remove the drawback of min-min algorithm (load is not properly balanced at each node). Proposed algorithm schedule n different length tasks to m heterogeneous cloud resource where scheduler chooses the minimum size task T_i and calculate the processing time at all the running virtual machine (resources) and assigns that virtual machine who can execute the task in minimum time. Task T_i is removed from the set S after the assignment and same procedure is repeated until all

Please cite this article as: M. Kumar, S.C. Sharma, Deadline constrained based dynamic load balancing algorithm with elasticity in cloud environment, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.018