Effective Mechanism to Reduce Response Time in Public Cloud

K. Kiran kumar[1], R. Mangalagowri[2]

M.Tech[1], Assistant Professor[2]
Dept of CSE, SRM University, Chennai, India
kirankumarkajjayam@gmail.com[1] , mangalagowri.r@ktr.srmuniv.ac.in[2]

***Abstract*-Cloud computing is emerging as a key technology for sharing resources. Public cloud is an environment that contains collection of data centers, or cloud data storages distributed in many different locations and interconnected by high speed networks. Cloud computing is efficient and scalable but maintaining the stability of processing is very difficult. Public cloud has numerous numbers of nodes; to manage the public cloud effectively it has to be partitioned. Partitioning the public cloud is done based on the systems configuration like memory, processor, and network bandwidth which reduce the response time and increases the availability. All the partitions are kept under the control of a system called main controller. Initially tasks arrive at the main controller. The main controller schedules the task to the appropriate partition based on average execution time of tasks. The average execution time is calculated using Shortest Job First (SJF) scheduling algorithm which considers the weighted average of previous execution times. Distribution of tasks according to the execution time of the task reduces the waiting time and improves the response time and throughput.**

Keywords-public cloud, cloud partition, main controller, execution time.

I. INTRODUCTION

Clouds are large pool of easily usable and accessible virtualized resources such as hardware, development platforms and services. These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the infrastructure provider by means of customized service level agreements. Cloud computing has been coined as an umbrella term to describe a category of sophisticated on-demand computing services initially offered by commercial providers. It denotes a model on which a computing infrastructure is viewed as a “cloud”, from which businesses and individuals access applications from anywhere in the world on demand. The main principle behind this model is offering computing, storage, and software “as a service” [4].

Cloud computing has distributed technologies to satisfy a variety of applications and user needs. Sharing resources, software, information via internet are the main functions of cloud computing with an objective to reduced capital and operational cost, better performance in terms of response time and data processing time. The numbers of users accessing the cloud are rising day by day.

A. Types of cloud

Cloud computing technology and services can be implemented in different ways according to their purpose and characteristics. These different types of deployment of cloud are categorized in four ways as follows.

Private cloud

In this cloud, infrastructure is deployed and operated by an organization where all the resource scan be owned, maintained and controlled by it only

Community cloud

In this cloud, infrastructure of cloud is deployed and operated by several organizations in sharing that supports a specific community with common approaches.

Public cloud

In this cloud, infrastructure of cloud is available to the general public or large group of different kinds of organization. Client can access services without any control and at specific rent. Client’s services and data can be co-located with other users.

Hybrid cloud

In this cloud, infrastructure of cloud can be combination of private, community and public cloud infrastructure. This combination of two or more clouds is with unique characteristics, entities and benefits to the users. Cloud computing is efficient and scalable but maintaining the stability of processing is very difficult as much number of jobs arriving into the cloud.

The task arrival pattern is not predictable and the size of the each task is differs effective scheduling strategy is needed to schedule the tasks and to control the workload. Each task has to be processed without waiting to increase the throughput and system performance. The scheduling strategy used in this paper is aimed at public cloud which has numerous numbers of nodes with distributed computing resources in many geographical locations. Thus this model divides the public cloud into different partitions based on system characteristics like memory, CPU etc. The cloud has main controller which manages the partitions and chooses appropriate partition based on the task execution time.

II. RELATED WORK

There have been many studies of system performance for cloud environment. Many techniques and tools were introduced to improve the response time in the cloud. However, workload control in the cloud is still a new problem that needs new architectures to adapt too many changes.

There are many scheduling algorithms such as Round Robin, First come First Serve, priority scheduling etc. various studies on scheduling algorithms gave Shortest Job First (SJF) Scheduling algorithm minimizes the waiting time, reduces the response time and increases throughput these leads to enhancement of system performance.

III. SYSTEM MODEL

There are several cloud computing categories with this work focused on a public cloud. A public cloud is based on the standard cloud computing model, with service provided by a service provider. A large public cloud will include many nodes which are distributed in many locations. Cloud partitioning [5] is used to manage this large cloud. A cloud partition is a subarea of the public cloud with divisions based on the system configuration like memory, processor and network bandwidth. The systems with similar characteristics are considered as a partition. The cloud partitioning architecture is shown in Fig.1.

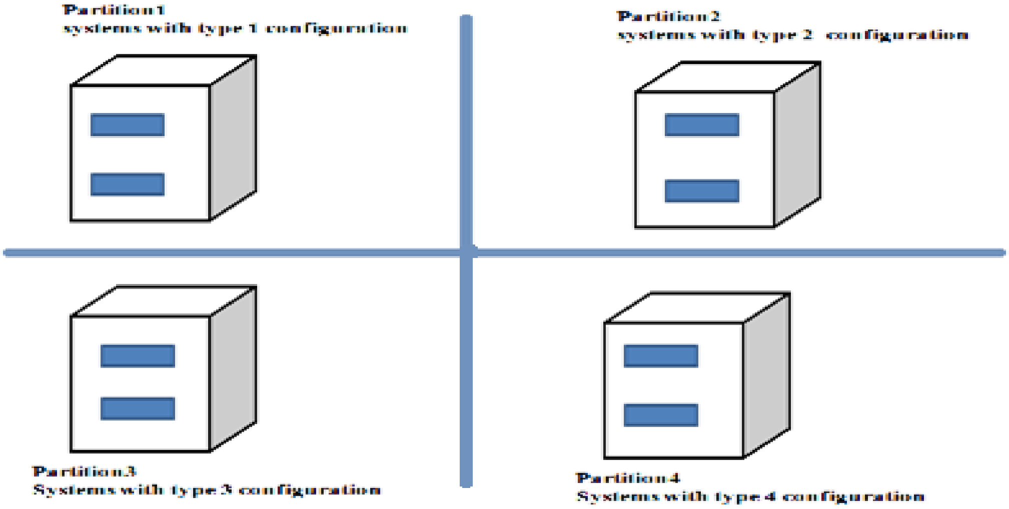
Partition1
systems with type 1 configuration
Partition2
systems with type 2 configuration
Partition3
Systems with type 3 configuration
Partition4
Systems with type 4 configuration

Fig. 1Typical cloud partitioning

The systems with similar configuration are comes under a partition. This technique is the efficient which controls the workload effectively and enhances the system performance. After creating the cloud partitions, the tasks given by the users are arrived at main controller which maintains the list of tasks given by the users within a certain time period and it schedules the jobs based on execution times to submit to appropriate partition.

A. Main Controller

The scheduling is done by the main controller. Main controller contains the configuration information of each partition and execution times of tasks submitted by users. For regular interval of time main controller buffers the new tasks. Main controller distributes the scheduled tasks to the appropriate cloud partition based on execution time of the task. Main controller contains the configuration information of each partition. Based on the task length the main controller decides to which partition the task has to be assigned to get the response in less time. The availability of systems is increased as tasks are distributed only to the appropriate partition.

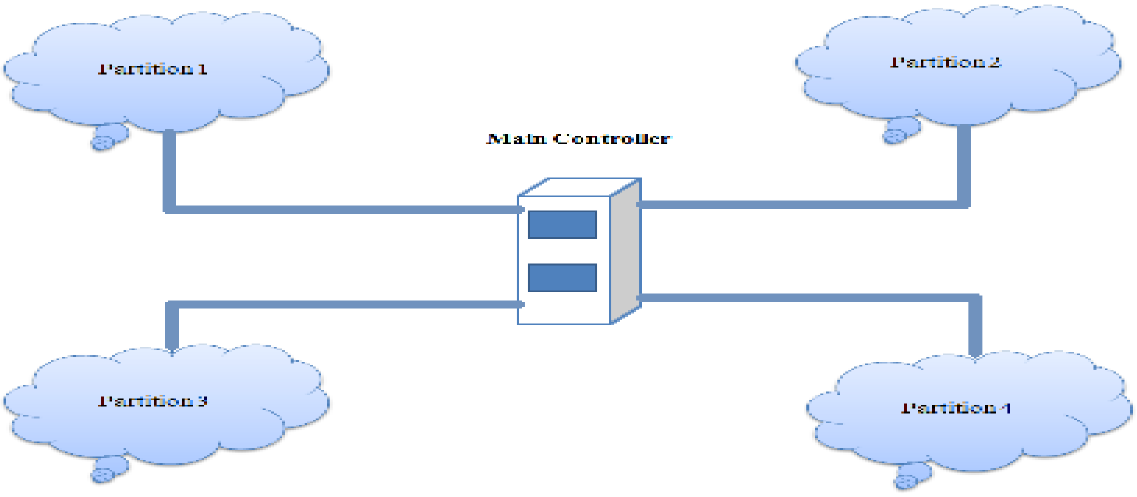

Fig. 2System architecture with main controller and partitions

B. Assigning jobs to the appropriate partition

When tasks arrive at public cloud initially they are allocated to main controller. Main controller maintains a list of tasks given by the users in a certain period of time and it finds out the execution time of all tasks in the list by using Shortest Job First (SJF) Scheduling Algorithm which uses the weighted averages of previous executions to calculate execution time of all tasks. After finding the execution time, it assigns the tasks to appropriate partition that has configuration which gives the response in less time.

IV. SCHEDULING STRATEGY

A. Motivation

Good scheduling will improves the performance of the entire cloud. However, there is no effective strategy to schedule the tasks. Various strategies have been developed; each of them has some drawbacks. The current model is very simple and effective one to improve the system performance.

B. Shortest Job First (SJF) Scheduling Algorithm

Shortest job first scheduling algorithm that assigns to each process the length of its next CPU burst/execution time. CPU is then given to the process with the minimal CPU burst from the waiting queue. SJF is provably optimal, in that for a given set of processes and their CPU bursts/Execution Times it gives average waiting time for a process is given by CPU is allocated to the process with least CPU-burst time amongst the tasks in the ready queue. CPU is always assigned to the process with least CPU burst requirement. If two processes having the same length, next CPU burst then First Come First Serve Scheduling is used i.e. one which arrives first, will be taken up first by the CPU.

Algorithm:

Step 1: Start the process.
Step 2: select process which has shortest burst time among all processes will execute first.
Step 3: if processes have same burst time length then FCFS (First come First Serve) scheduling algorithm is used.
Step 4: make average waiting time length of next process.
Step 5: start with first process, selection as above and other processes are to be in queue.
Step 6: calculate the total number of burst time.
Step 7: stop the process.

Predicting the time the process will use on its next schedule:

$t(n+1) = w*t(n)+(1-w)*T(n)$

Here

$t(n+1)$	is time of next burst.
$t(n)$	is time of current burst.
$T(n)$	is average of all previous bursts.
W	is a weighting factor emphasizing current or previous bursts.

V. CONCLUSION

The proposed scheduling algorithm uses effective scheduling strategy which leads to minimum waiting time, minimum response time and maximum throughput. Hence the system performance is increased.

VI. FUTURE ENHANCEMENT

1. Cloud partitioning is not a simple task; effective strategies must be needed to partition the cloud.

2. Grouping of tasks based on execution times is not only the strategy, tasks may be grouped based on length or some other factors. More studies are needed to develop effective grouping strategies.
3. Prediction of execution times of jobs based on weighted average of previous executions is difficult; it may not produce exact values. More work is needed to develop effective techniques to know execution times based on other parameters.

REFERENCES

[1] Mangal Nath Tiwari * Kamalendra Kumar Gautam Dr Rakesh Kumar Katare, “Analysis of Public Cloud Load Balancing using Partitioning Method and Game Theory”, IJARCSSE Volume 4, Issue 2, February 2014.

[2] Monica Choudhary, Sateesh Kumar, Peddoju “A Dynamic Optimization Algorithm for Task Scheduling in Cloud Environment”, IJERA Vol. 2, Issue 3, May-Jun 20.

[3] Pinky Rosemarry, Ravinder Singh, Payal Singhal and Dilip Sisodia, “grouping based job scheduling algorithm using priority queue and hybrid algorithm in grid computing”, IJGCA 2012.

[4] panagiotis kalagiakos, panagiotis karampelas “Cloud Computing Learning” IEEE , 2011.

[5] Gaochao Xu, Junjie Pang, and Xiaodong Fu, “A-Load- Balancing-Model-Based-on-Cloud-Partitioningfor-the-Public-Cloud”, IEEE 2013.

[6] M. Vijayalakshmi, V.Venkatesa Kumar “Investigations on Job Scheduling Algorithms in Cloud Computing” 2014.

[7] Vijindra and Sudhir Shenai. A, “Survey of Scheduling Issues in Cloud Computing”, 2012, Elsevier Ltd.

[8] Jeongseob Ahn, Changdae Kim, Jaeung Han, Young-ri Choi†, and Jaehyuk Huh,“Dynamic Virtual Machine Scheduling in Clouds for Architectural Shared Resources”, 2011.

[9] Zheng Hu, Kaijun Wu, Jinsong Huang,“An Utility-Based Job Scheduling Algorithm for Current Computing Cloud Considering Reliability Factor”,2012.

[10] Xiaoli Wang, Yuping Wang, Hai Zhu,“Energy-Efficient Multi-Job Scheduling Model for Cloud Computing and Its Genetic Algorithm”, Hindawi Publishing Corporation Mathematical Problems in Engineering, Volume 2012.