Load Balancing in Cloud Computing

Pujesh Bilare, Vishal Gorle

Master of Computer Application,

ASM's Institute of Management & Computer Studies, University Of Mumbai.

Mumbai, India

ABSTRACT:

Cloud computing is an emerging technology in the computer and information technology industry where the computing and all other task are moved to a cloud platforms. Cloud computing is a flexible approach to enhance capacity and expand capabilities without the need for substantial investments in infrastructure, personnel training, or software licensing. The main objective of this paper is to address the challenges related to load balancing in cloud computing and propose techniques to mitigate waiting time and turnaround time. Load balancing is accomplished through the utilization of load balancers, which efficiently redirect incoming requests while ensuring transparency to the requesting client. Based on some of the predetermined parameters, such as availability,network bandwidth, payload capacity and current load, the load balancer uses various scheduling algorithms to choose which server should handle and forwards the request on to the selected server in the load balancing process.

Introduction:

Internet has been a major factor, and a contributor towards the various technologies that have been developed lately. In recent years, there has been a remarkable surge in the adoption of cloud computing, making it a prominent trend in the information technology landscape. The methodology of transitioning to cloud platforms has been driven by the many advantages it offers, such as cost reduction and data storage facilities for users and providers alike. Cloud computing can be described as a system of parallel and distributed computing, consisting of a network of interconnected computers over an network. These computers are dynamically made, allocated and presented as single one computing resources, governed by service-level agreements (SLAs) which is decided between the service provider and the consumer based on the consumers needs and negotiations between them. Load balancing is lately an emerging computational technology that enables the cloud's computing capabilities, and storage capacity and allowing data and services to be extensively hosted in the cloud and accessed from any internet-connected device. It is known as provider of dynamic services using very large salable and visualized resources over the internet. Load balancing is a technique in computer networking that involves the distribution of workload across various computer clusters, network links, or other available resources. The primary goal is to ensure optimal utilization of resources, maximize throughput, and minimize response time. and avoid overload. It is a mechanism where it tries to distributes the incoming dynamic local work load evenly across all the nodes in the whole cloud and the network and this is done to ensure and avoid a situation where some nodes in cloud network are heavily loaded while others are idle and nodes are being wasted as resources. Its goal is to improve the overall performance and the utility utilization the system

Challenges in Load Balancing in Cloud Computing:

Although cloud computing has been widely adopted.The field of cloud computing is still in its early stages of research, with many and vast number of challenges awaiting resolution by the scientific community. One such area that requires further exploration and research is load balancing, which has its own set of challenges. One key aspect of cloud computing is the ability to avail resources and services automatically, allowing for the dynamic allocation and release of serviecs and resources as per user needs. Virtual machine migration is a process enabled by virtualization technology, whereby an entire machine is represented as a file or a collection of files. This allows for the seamless transfer of a virtual machine from one physical machine to another, providing a means to alleviate the burden on heavily loaded physical machines.

- In the paper titled “Improved Max-Min Algorithm in cloud computing“, the author is trying out to allocate the task with maximum execution time to the resource with minimum completion time.

- In this approach, if we are having more no of tasks(lets say 10,000), then the average turn-around time of the tasks will be very high which will decrease the efficiency of the entire system.

- Now if the average turnarounds time will be high then the processing cost so as a result waiting time will also be increased.

Thus Load balancing mechanism helps in improving the performance by balancing the incoming dynamic load among the available resources like network links, CPU, disk,storage disk, and even on cloud and other storage devices.

Methodology:

- ➢ Initially we will try that all tasks will sorted according to their minimum execution length.
- ➢ Then we will calculate the expected completion time of each task on all resources.
- ➢ The Expected Completion time of task on a resource can be calculated as: CT(i,j)=ET(i,j)+ r(j), where ET(I,j) is the expected

□t(i) represent the execution time of task

□m(j) represents the machine

□r(j) represents the ready time of m(j) i.e. the time when m(j) becomes ready to

□Execute the task t(i).

- ➢ Now we will find minimum expected completion time of each task in MT(meta task table) and the resource that will obtain it.(tasks are collected into a set called meta task(MT)).
- ➢ Then we will arrange the resources in the descending order of MIPS(million instruction per second).
- ➢ And, then finally we arrange our tasks into the groups.

The groups would be

No. of groups = No. of Tasks/ Number of resources.

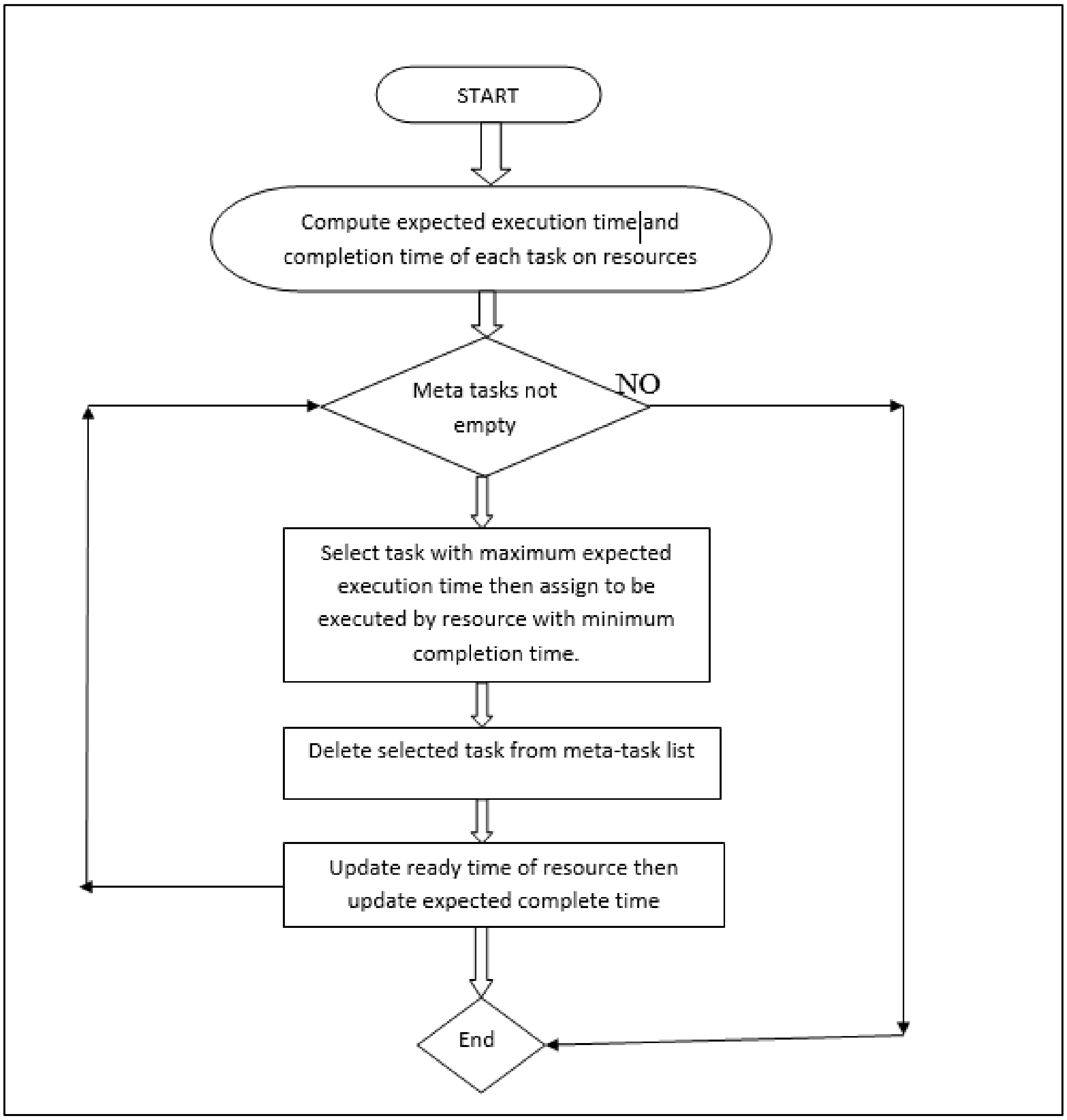
So, the choice of cloudlet in the group.

Task: Size = more/max.

Task: Size = less/ min.

In the given scenario, T1, T2, T3, T4, T5, and T6 represent the tasks, while R1, R2, and R3 represent the available resources. By dividing the total number of tasks by the number of resources, we can determine the number of groups:

6/3= 2, 12/3 =4


START
Compute expected execution time and completion time of each task on resources
Meta tasks not empty
NO
Select task with maximum expected execution time then assign to be executed by resource with minimum completion time.
Delete selected task from meta-task list
Update ready time of resource then update expected complete time
End

Conclusion:

This paper is based on cloud computing technology which has a very vast potential and is still unexplored. The capabilities of cloud computing are endless. Cloud computing can be used as Saas (Software As a Service), IaaS (Infrastructure As a Service),PaaS (Platform as a Service) and other needs as computing and storage to are provided. One Major issues which is a set back in it is the ability to handle the dynamic incoming load which if not handled properly it might hamper many factors and slow down the performance. So there is always a requirement of efficient load balancing algorithm and efficient utilization of resources to manage and handle this incoming data. Our paper focuses on the load balancing algorithms and their applicability in cloud computing environment to handle the load and keep the system up and running.

References:

1. O.M. Elzeki . “Improved Max-Min Algorithm in Cloud Computing”. International Journal of Computer Applications(0975-8887) Volume 50-No.12,july 2012.
2. Amandeep Kaur Sidhu. “Analysis of load balancing techniques in cloud computing”. International Journal of computers & technology volume 4 No. 2, March-April, 2013, ISSN 2277-3061.
3. Ektemal Al-Rayis. “Performance Analysis of load balancing Architectures in Cloud computing” 2013 European Modeling Symposium. 978-1-4799-2578-0/13$31.00@2013 IEEE.