Load Balancing with Tasks Subtraction

Ranjan Kumar Mondal[1]
Department of Computer Science & Engineering,
University of Kalyani, Kalyani, India

Payel Ray[2]
Department. Computer Science & Engineering,
University of Kalyani, WB, India

Enakshmi Nandi[3]
Department Computer Science & Engineering,
University of Kalyani, Kalyani, India

Debabrata Sarddar[4]
Assistant Professor, Department of Computer Science & Engineering,
University of Kalyani, Kalyani, India

Abstract.

Cloud computing has been becoming popular day by day to provide different type of web services and web resources and web applications to the web system. Cloud computing is working with web resources to execute applications and web resources. The objectives of Cloud computing is to share network resources and web services over the Internet of web nodes. In cloud computing, load balancing is one of the target issue. Load is a measure of the amount of works that a computation system performs which can be classified as CPU load and network load. Load balancing is the process of apportioning the load among various working nodes of a distributed system to improve both resource utilization and job response time while avoiding a situation where some of the nodes are heavily loaded while others are under loaded. Load balancing ensures that every node in the system does approximately equal amount of work as per their capacity at any instant of time.

In this paper, we propose a new scheduling algorithm in distributed system that chooses suitable nodes with their subtracting tasks. It is very easy way to select an appropriate node. This approach can provide efficient utilization of computing resources and maintain the load balancing in cloud computing environment.

Keywords: Cloud Computing, Load Balancing, Distributed System.

1. Introduction

Load balancing [1] is a new challenging role in distributed system in the cloud computing now. Always a distributed network is required because it is not cost efficient to maintain one or more idle servers as to fulfill the required demands. Jobs can't be assigned to suitable servers and clients individually for efficient load balancing as cloud is not being simple structure and components are present throughout a wide area.

Algorithms of load balancing in cloud computing are classified as static and dynamic algorithms. Static algorithms are suitable for homogeneous and stable environments and this type of algorithm can produce effectiveresults in these environments but they are usually not flexible and cannot match the dynamic changes to the attributes during the execution time. Whether dynamic algorithms are more flexible than static algorithm and it takes into consideration different types of

ISSN: 2278 – 7798

attributes in the system both prior to and during run-time.

Nowadays cloud computing is one of the major topics of discussion as it promises high levels of scalability and flexibility without making much efforts in setting up and running the large scale data centers and computing infrastructures. Cloud Computing is a paradigm which enables systems to scale up or scale down their use of resources of information technologies based on their prerequisites without making any efforts in infrastructures.

All resources in cloud computing are in a shared environment so that a resource management mechanism is required which can assign a task to processing units whenever necessary, handle requests, re-compute the responses that is returned by the processing units, manage the data availability to all the processing units etc. Load balancing is one of the most important part of the resource management, if load is not managed decently it can lead to prominent delays in execution. Different kinds of load balancing techniques will be discussed in this paper.

II. CLOUD COMPUTING

A. Overview

Cloud computing [2] is a on demand web based service in which shared resources work together to perform a particular task to get the results in minimum possible time by distribution of any dataset among all the connected processing units. Cloud computing is concerned to mention the network based services giving an illusion of providing a real server. Such virtual servers do not exist physically so they can be scaled up and down at any point of time.

Cloud computing is related to the processes to decomposing 'Big Data' tasks into small dataset and distributing them to several computing units, so the task can be performed in the minimum possible time. Three main characteristics are required by any cloud service system: resource utilization through a single service provider, ability to acquire transactional on demand and mechanisms to chargeto users on the basis of resource utilization.

B. Cloud Infrastructure

A cloud system consists of a group of processing unit together, so we can define that the basic unit of cloud is a processing units grouped together to achieve same goal. The processing units are connected to the master processing unit that is responsible for assigning the tasks to its slave units. The master system is again connected to the head node of the cloud which is responsible for receiving the tasks, dividing it into sub tasks and then assigning it to the masters system which further assigns the tasks to its slaves [3].

Data sharing is done by real time sharing over the network, since all the data is required to accessible to all the processing nodes for processing dataset.

C. Development models in Cloud Computing

There are mainly three different cloud computing models [4]:

- Private clouds
- Public Clouds
- Hybrid Clouds

Private cloud is a cloud where the computing resource, computing storage and computingsystem is owned by a private enterprise. The owner of the cloud is responsible for maintenance of the infrastructure.

On the other way**public cloud** is a cloudwhere the resources and system may or may not be owned by more than one organization but the resources are offered to external users.

A hybrid cloud is the category of cloud where a part of the cloud infrastructure is maintained by the administration itself with acquiring services from the public clouds.

These three different models of cloud have both advantages as well as disadvantages.

ISSN: 2278 – 7798

The main advantage of private cloud is the control over all the resources and infrastructure making it feasible for them to make changes in the infrastructure at any point of time according to their requirements. But the disadvantage of private cloud is the investment cost required to be put in establishment of the infrastructure, in the same way the a extra cost is required for the software and the maintenance activities. The advantage of public cloud is that an organizations itself does not have to take care of the cloud computing infrastructure and operational activities. The disadvantage of utilizing the services from a public cloud provider is that it is completely dependent upon another data

Head NodeMaster

ProcessingUnit

Data Sharing

Job Assignment to Processingunit

Job Distributionby host serverprocessing results tolead node business entity offering resources through public clouds.

D. Service models in cloud computing

Different types of services are offered for the use of a cloud infrastructure by the organizations. First of them is Software as a service that is called SaaS infrastructure which offers software applications for its customers. In this kind of service the customer is not allowed to make changes in the applications. Second is Platform as a Service that is PaaS infrastructure,the developers of the platform are provided with some API's in different languages which can be used to modify the application according to the use of the customer. Third is Infrastructure as a Service named is IaaS in which even infrastructure in terms of computing and storage resources is customizable by the users according to their preferences. This infrastructure is used to host applications. IaaS models often provide automatic support for on demand scalability of computing and storage resources [5].

III. LOAD BALANCING IN CLOUD COMPUTING

When anenquiry is requested by any node then it is distributed to all the slave nodes existing in that cloud. So the way the distribution is being done must get the response from all the slaves node at a time so that there should not be any waiting for any particular machine to reply before further processing could happen. But in the real time clouds heterogeneous computing devices exists and any process's execution time on the slave node is required to be estimated.

So the main featuresexisting in any load balancing [6] system is the asymmetric load distribution. A higher ratio of workload is required to be given to those with computation capabilities. But sometimes higher computation power cannot decidethat how much the task is required to be assigned to that system. This assignation of exact task to proper system in heterogeneous computing infrastructure is done by load balancing system.

Load balancer is also responsible for 'Priority Activation' meaning that when the number of slave computing nodes drops below a certain level the load balancer must wake some of the sleeping devices to maintain the computing performance for the customers.

A. Persistence Issue in Load Balancing

One of the main issues of load balancing system is when operating a load balancer service is 'how to handle information that must be kept across the multiple requests in a user's session'. If the data is stored only on system requesting server, it would not be accessible to other computing nodes and since subsequent requests takes space for this information this can lead to a performance issue.

The solutions to this problem is to send all the requests to the same computing systems containing this resources.

This is known as persistence. If the device containing the information goes down

ISSN: 2278 – 7798

the whole system gets down also the any session of the processes present on the node is also lost. So the problem is because of no backed up centralized system, one of the solution is using

PublicCloud
PrivateCloud
VirtualPrivate
CloudHybrid

CloudInfrastructure asa Service (IAAS)

AmazonEC2
IBM -Computing on Demand
EC2
VMware
vSphere

Platform asa Service (PAAS)

Google
App Engine
Yahoo Open
Strategy
EC2
Microsoft
Azure

Software asa Service (SAAS)

Google
Apps
Microsoft
Cloud Service
EC2
Salesforce
CRM

Development Models

Service Models are a backup system together but this leads to major performance issue. Next solution is related to backup system.

One other solution that can be used is by using database to store the resources, but this increases the loads on the applications. But databases provide solutions to some of the problems existing in systems. Databases can be backed up easily solving the problem of single node of failure, databases are highly scalable also. Since in a backed up system there are several nodes holding the same information the query load can also be distributed over them to get a better performance. All servers in the cloud system store information data on State Server and any server in the cloud system can retrieve the data.

B. Load Balancing Architecture

Both software and hardware are available in the market as a solution for the load balancing problem. Examples include the Apache web server's mod_proxy_balancer extension, nginx, Varnish, or the Pound reverse proxy and load balancer. Gearman can also be used to distribute appropriate computer tasks to multiple computers, so large tasks can be done more quickly.

Multiple layers of load balance system can also be employed to achieve more sophisticated systems that create a tree structure. The upper level load balance system distributes the task to lower level computing devices which uses their own load balancer to distribute the task further, this leveling in cloud system can go up to any number of levels creating very complex systems.

C. Scheduling Algorithms used in load balancing

There are a number of scheduling algorithms for load balancing for determining the computing device which should be sent the next computing task. One of the mostsimple is Round Robin algorithm. More sophisticated systems use additional factors to determine this such as server's reported load, number of active connections, geographic location recent response times, up/down status determined by a monitoring poll of some kind, capabilities or how much traffic it has recently been assigned.

D. MapReduce method for task distribution

MapReduce framework is fundamentally used for processing large datasets

ISSN: 2278 – 7798

across a huge number of computers i.e. a cluster of computer in the web. In this technique the processing is done in two way those are: Map and Reduce. Each has its own specific purpose that leads to the effect of computation. Both Map and Reduce functions are explained below.

"Map" Step: The process of converting a task to small sub-tasks and distributing it to the salve nodes in the node cluster is called as Map. There can be more distribution also that leads to multilevel tree structure. The slave computing devices are performing the processing and returning the result to the requesting device - Master Node.

"Reduce" Step : After receiving the response from the slave nodes the master node is required to perform the task of combining the results into one, which is done in this step.

MapReduce can also be performed in distributed systems. Since each mapping is independent to other and all mappings can be done in parallel, but this is limited by the number of independent nodes and number of CPU's near each source.

2. Problem Definition

Load balancing for a distributed system is one of the most important issuesthat has to be solved to enable the efficient utilization of the systems. This assignment can be compared to problems that arise in work distribution processes like that of scheduling all tasks that will be needed to construct a building. Several objectives have to be taken into consideration:

- The total work should be completed as early as possible.
- As employees are very expensive, thus they should be kept busy as possible. On the other way, the distribution can be done straight forward because of there is enough work to do.
- The work should be distributed fairly and regularly. Same amount of work should be assigned to every employee.
- There are precedence constraintsamongdissimilarworks. So we have to find aefficient processing order of the different tasks.

A dissimilar computing world is composed of computing resources where these resources can be a single PC, a workstations or a supercomputer. The main work of the task scheduling in the system is the efficiently allocating tasks to nodes. Tasks are originated from different users/applications, are independent.

3. The Proposed Method

There are several heterogeneous nodes in cloud computing system. All nodes have no capability to execute same tasks at a time; hence only consider the CPU remaining of the nodesare not enough when a node is chosen to execute a particular task. So, how to select an efficient node to execute that task is very important in a cloud computing.

4. Method:

Step 1: To subtract between maximum task and minimum task of all corresponding nodes.

Step 2: Choose maximum task value from all nodes.

Step 3: If there are more than one same tasks then select the node having its highest corresponding subtracting result

Step 4: Then assign node with minimum task value.

Step 5: If there are more than one same values, select the node having its highest corresponding summation result of particular tasks of all nodes.

Step 6: Then task is dispatched to the selected node for computation

Step 7: Select next higher task value and repeat step 3 to 5 until all task have been completed totally.

ISSN: 2278 – 7798

Example:

Table 1 shows an example

Task \ Node	C_{11}	C_{12}	C_{13}	C_{14}
t_1	12	13	10	14
t_2	16	24	13	25
t_3	26	31	12	33
t_4	17	24	18	31

Table 1

Table 2 shows all differences of all nodes

Task \ Node	C_{11}	C_{12}	C_{13}	C_{14}
t_1	12	13	10	14
t_2	16	24	13	25
t_3	26	31	12	33
t_4	17	24	18	31
Difference	**14**	**11**	**8**	**19**

Table 2

Table 3 selects highest subtracting value.

Task \ Node	C_{11}	C_{12}	C_{13}	C_{14}
t_1	12	13	10	14
t_2	16	24	13	25
t_3	26	31	12	33
t_4	17	24	18	31
Difference	14	11	8	**19**

Table 3

Table 4 selects next highest value

Task \ Node	C_{11}	C_{12}	C_{13}	C_{14}
t_1	12	13	10	14
t_2	16	24	13	25
t_3	26	31	12	33
t_4	17	24	18	31
Difference	**14**	11	8	19

Table 4

Table 5 selects next highest value

Task \ Node	C_{11}	C_{12}	C_{13}	C_{14}
t_1	12	13	10	14
t_2	16	24	13	25
t_3	26	31	12	33
t_4	17	24	18	31
Difference	14	**11**	8	19

Table 5

ISSN: 2278 – 7798

Table 6 selects last value

Task \ Node	C_{11}	C_{12}	C_{13}	C_{14}
t_1	12	13	10	14
t_2	16	24	13	25
t_3	26	31	12	33
t_4	17	24	18	31
Difference	14	11	**8**	19

Table 6

5. Result Analysis:

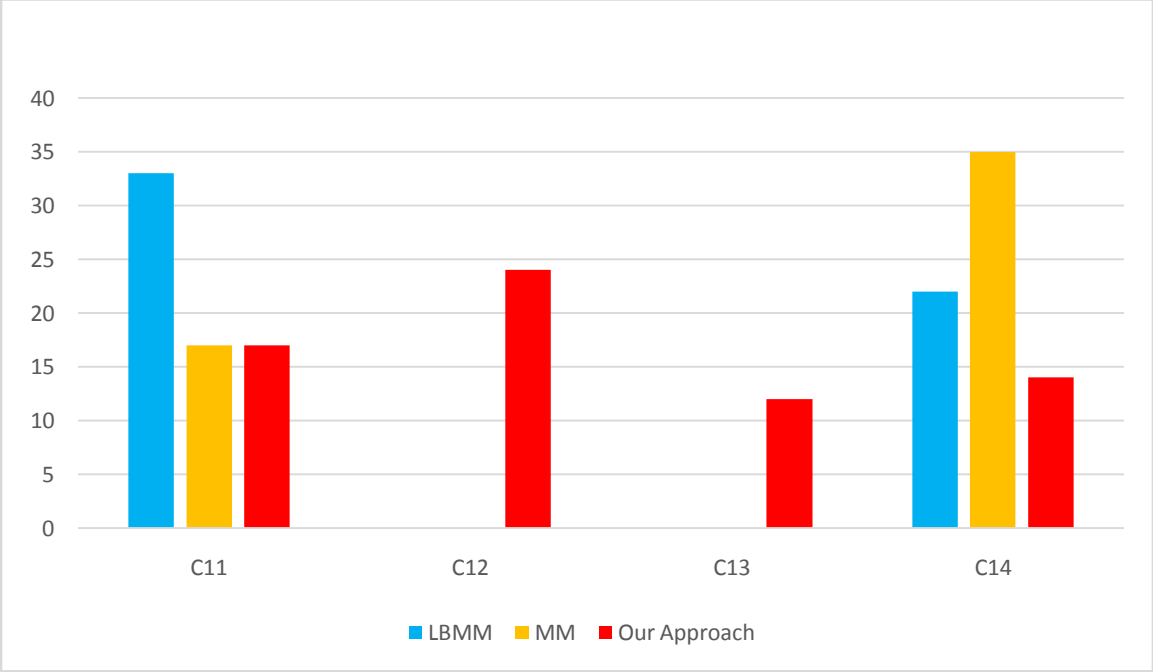

Fig 1.The comparison of completion time of tasks at different nodes.

6. Conclusion

In this paper, we proposed an efficient scheduling algorithm, LBST, for the cloud computing network to assign tasks to computing nodes according to their resource capability. Similarly, our approach can achieve better load balancing and performance than other algorithms, such as MM and LBMM from the case study.

In this paper, we have presented a new scheduling algorithm for scheduling. The goal of the scheduler in this paper is minimizing make-span and maximizes resources utilization.

1. References

[1]. Hu, Y. F., and R. J. Blake. "An optimal dynamic load balancing algorithm." Daresbury Laboratory. 1995.

[2]. Wang, S. C., Yan, K. Q., Liao, W. P., Wang, S. S.: Towards a Load Balancing in a threelevel cloud computing network. In: Computer Science and Information Technology, pp. 108—113, (2010).

[3]. Hung, Che-Lun, Hsiao-hsi Wang, and Yu-Chen Hu. "Efficient Load Balancing Algorithm for Cloud Computing Network." In International Conference on Information Science and Technology (IST 2012), April, pp. 28-30. 2012.

ISSN: 2278 – 7798

[4]. Armstrong, R., Hensgen, D., Kidd, T.: The relative performance of various mapping algorithms is independent of sizable variances in run-time predictions. In: 7th IEEE Heterogeneous Computing Workshop, pp. 79—87, (1998)

[5]. Freund, R., Gherrity, M., Ambrosius, S., Campbell, M., Halderman, M., Hensgen, D., Keith, E., Kidd, T., Kussow, M., Lima, J., Mirabile, F., Moore, L., Rust, B., Siegel, H.: Scheduling resources in multi-user, heterogeneous, computing environments with SmartNet. In: 7th IEEE Heterogeneous Computing Workshop, pp. 184—199, (1998)

[6]. Ritchie, G., Levine, J.: A Fast, Effective Local Search for Scheduling Independent Jobs in Heterogeneous Computing Environments. Journal of Computer Applications, vol. 25, pp. 1190—1192, (2005)

[7]. Braun, T. D., Siegel, H. J., Beck, N., Bölöni, L. L., Maheswaran, M., Reuther, A. I., Robertson, J. P., Theys, M. D., Yao, B., Hensgen, D., Freund, R. F.: A Comparison of Eleven Static Heuristics for Mapping a Class of Independent Tasks onto Heterogeneous Distributed Computing Systems. Journal of Parallel and Distributed Computing, vol. 61, pp. 810—837, (2001).

[8]. Ranjan Kumar Mondal, Debabrata Sarddar "Load Balancing with Task Subtraction of Same Nodes". International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 3, Issue 4, pp: (162-166), Month: October - December 2015.

[9]. Ranjan Kumar Mondal, Enakshmi Nandi, and Debabrata Sarddar. "Load Balancing Scheduling with Shortest Load First." *International Journal of Grid and Distributed Computing* 8.4 (2015): 171-178.

[10]. Ranjan Kumar Mondal, Payel Ray, Debabrata Sarddar "Load Balancing with Task Division and Addition". International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882, Volume 5, Issue 1, January 2016: 15-19.

Authors Profile

Ranjan Kumar Mondal received his M.Tech in Computer Science and Engineering from University of Kalyani, Kalyani, Nadia; and B.Tech in Computer Science and Engineering from Government College of Engineering and Textile Technology, Berhampore, Murshidabad, West Bengal under West Bengal University of Technology, West Bengal, India. At present, he is a Ph.D research scholar in Computer Science and Engineering from University of Kalyani. His research interests include Cloud Computing, Wireless and Mobile Communication Systems.

Payel Ray received her M.Tech in Computer Science and Engineering from Jadavpur University, Jadavpur, India; and B.Tech in Computer Science & Engineering from Murshidabad College of Engineering and Technology, Berhampore, Murshidabad, West Bengal under West Bengal University of Technology, West Bengal, India. At present, she is a Ph.D research scholar in Computer Science and Engineering from University of Kalyani. Her research interests include Cloud Computing, Wireless Adhoc and Sensor Network and Mobile Communication Systems.

Enakshmi Nandi received her M.Tech in VLSI and Micro-electronics from Techno India, Salt Lake, West Bengal and B.Tech in Electronics and Communication Engineering from JIS College of Engineering, West Bengal under West Bengal

ISSN: 2278 – 7798

University of Technology, West Bengal, India. At present, she is Research scholar in Computer Science and Engineering from University of Kalyani. Her research interests include Cloud Computing, Mobile Communication system, Device and Nanotechnology.

Debabrata Sarddar is an Assistant Professor at the Department of Computer Science and Engineering from University of Kalyani, Kalyani, Nadia, West Bengal, India. He completed his PhD from Jadavpur University. He did his M. Tech in Computer Science & Engineering from DAVV, Indore in 2006, and his B.E in Computer Science & Engineering from NIT, Durgapur in 2001. He has published more than 75 research papers in different journals and conferences. His research interests include Cloud Computing, Wireless and Mobile Communication Systems.

ISSN: 2278 – 7798