DESIGN AND IMPLEMENTATION OF NEW VIRTUAL MACHINE SCHEDULING ALGORITHM

Neenu Joseph¹, Melody M Fernandes², Priyanka R Kamble³ and Mr. Santhosh B⁴

Abstract- For efficient operations in a distributed environment load balancing is an important factor. Load Balancing strategies are of huge significance in improving the reliability and performance of resources in data centres. One of the challenging problems in scheduling resources in a cloud data centre is the allocating and migrating of virtual machines, these virtual machines are reconfigurable. Therefore load balancing due to its challenges and importance for the cloud has become a major research area. Various algorithms were proposed to supply productive mechanisms for tending to the client's requests to available Cloud nodes. In this paper, we investigate the different multidimensional resource scheduling algorithms used to achieve scheduling of various task and balancing load in Cloud Computing.

Keywords –scheduling, integrated load balancing, imbalance value, cloud computing

I. INTRODUCTION

Cloud computing is one of the latest upcoming technologies, that is internet based with various advancement in virtualization, grid, web computing, utility computing and other related technologies. Many organizations use cloud computing for various purposes. Cloud computing allows us perform tasks with high quality, low cost, and uses pay-per-use model. It provides platform and applications on demand though the internet. The benefits of could computing are hiding and abstraction of complexity, virtualization of resources, and efficient use of distributed resources. Examples of could computing platforms are: Google App Engine, IBM blue cloud, Amazon EC2, Microsoft Azure.

A cloud datacenter is a distributed network structure that has many nodes. Resources like CPU, memory, and network bandwidth form a node. These different resources are the multi-dimensional resources. The traditional data centers contained physical machines. This configuration was expensive, with wasted energy and floor space, low resource utilization, and management overhead. The concept of virtual cloud data centers have the ability to be allocated form one set of resources to another in a proper manner. This allows us to use cloud more efficiently and usefully. We can handle large data centers with reduced maintenance cost. To use the cloud resources more efficiently scheduling of the resources plays an important role.

In this paper we have presented a modified version of the current LIF scheduling algorithm. This increases the efficiency of operations in a distributed environment.

A. Challenges in Load Balancing–

Modified resource allocation techniques and better improved strategies through efficient job scheduling are the main provisions of the load balancer. The load can be network load, CPU load,
memory capacity or delay. Load balancer performs the distribution of load among various nodes in a distributed system while also being able to avoid a situation where some of the nodes are idle or less loaded while others are overloaded and to improve both utilization of resources and response time of jobs. At any instant of time the load balancer ensures that equal amount of work is done by every node in the network or all the processors in the system. This is the important factor to be considered during the resource allocation but this has become more difficult especially in elastic cloud computing where the user can dynamically request for the resource. The performance prediction also plays an essential role in load balancing. But cloud environment is highly variable and unpredictable. Providers try to oversubscribe as many users to a shared infrastructure to increase resource utilization. This results in resource contention and interference. Heterogeneity within the same instance type and administrative action (e.g., eviction) to maintain the service level are some other factors that cause unpredictability of the environment.

II. PROBLEM STATEMENT

At a certain time period in a cloud data center there are $M$ Physical Machines (PMs), also called hosts, which configuration may be heterogeneous. Allocating and migrating virtual machines (VMs) which are reconfigurable and taking into consideration integrated features of hosting physical machine (PMs) are one of the challenging problems in scheduling resource in cloud data centres. This problem can also be defined as given a set of $n$ requests (VMs) and a set of $m$ identical machines (PMs) $PM_1, PM_2, ..., PM_m$, each request has a processing time, the objective of load-balance is to balance load on every machine while they are being assigned requests.

III. EXISTING ALGORITHMS

A. LIF Algorithm

Wenhong Tian et al [1], proposed an algorithm based on demands characteristics (for example, CPU intensive, high memory, high bandwidth requirements etc.), always selects lowest integrated imbalance value in different physical machines (as stated in equation (2-3)) and available resource to allocate virtual machines.

LIF algorithm considers imbalance values integrated CPU, memory and network bandwidth utilization, and the following parameters are considered: Average CPU utilization $CPU_{i}^{U}$ of a single server $i$: is averaged CPU utilization during observed period. For example, if the observed period is one minute and utilization of CPU is recorded every 10 seconds, then $CPU_{i}^{U}$ is the average of six recorded values of server $i$.

Average utilization of all CPUs in a Cloud data center. Let $CPU_{i}^{n}$ be the total number of CPUs of server $i$,

$$CPU_{i}^{A} = \frac{\sum_{i=1}^{N} CPU_{i}^{U} \times CPU_{i}^{n}}{\sum_{i=1}^{N} CPU_{i}^{n}}$$

(1)

Where $N$ is the total number of physical servers in a Cloud data center and $CPU_{i}^{n}$ represents the number of CPUs in physical server $i$. Similarly, average utilization of memory, network bandwidth of server $i$, all memories and all network bandwidth in a Cloud data center can be defined as $MEM_{i}^{A}, NET_{i}^{A}, MEM_{i}^{n}, NET_{i}^{n}$ respectively.

Datacenter-wide integrated imbalance value $ILBi$, of server $i$. In statistics variance is used as a measurement of how far a set of numbers are spread out from each other, therefore it is widely used. Using variance, an integrated load imbalance value ($ILBi$) of server $i$ is defined
Where

\[ \text{Average Utilization} = \frac{(\text{CPU} + \text{MEM} + \text{NET})}{3} \]  

is average utilization of multi-dimensional resource in a physical machine, also called integrated load in LIF.

B. **ZHJZ Algorithm**

Wood et al. [2], introduced a few virtual machine migration techniques. One integrated load balance measurement is applied as follows:

\[ V = \frac{1}{(1-CPU_u)(1-MEM_u)(1-NET_u)} \]  

---

C. **ZHJZ Algorithm**

Zheng et al [3], proposed integrated load-balancing measurement as following:

\[ B = \frac{aN1iCi}{N1mCm} + \frac{bN2iMi}{N2mMm} + \frac{cN3iDi}{N3mDm} + \frac{dN4iNeti}{Netm} \]  

---

D. **Rand Algorithm**

Randomly assigns requests (virtual machines) to physical machines which have available resource.

E. **Round Robin (RR)**

The round-robin is one of the most used algorithm for scheduling (for example by Amazon EC2 and Eucalyptus [4]), in which PM’s are allocated VM’s in turns. Simplicity in implementation is the advantage of this algorithm.

IV. **PROPOSED ALGORITHM:** Extended_LIF
LIF algorithm considers imbalance values integrated CPU, memory and network bandwidth utilization. It always selects lowest integrated imbalance value in different physical machines and available resource to allocate virtual machines.

Extended_LIF algorithm considers hard disk utilization along with CPU, memory and network bandwidth utilization. The following parameters are considered:

- Average CPU utilization $C_{PU}^i$ of a single server $i$: is averaged CPU utilization during observed period. For example, if the observed period is one minute and utilization of CPU is recorded every 10 seconds, then $C_{PU}^i$ is the average of six recorded values of server $i$.

- Average utilization of all CPUs in a Cloud data center. Let $C_{PU}^n$ be the total number of CPUs of server $i$.

\[
C_{PU}^i = \frac{\sum_{l=1}^{N} C_{PU}^l C_{PU}^n}{\sum_{l=1}^{N} C_{PU}^l}
\]  

Where $N$ is the total number of physical servers in a Cloud data center and $C_{PU}^n$ represents the number of CPUs in physical server $i$. Similarly, average utilization of memory, network bandwidth and hard disk of server $i$, all memories, all network bandwidth and all hard disks in a Cloud data center can be defined as $M_{MEM}^i$, $N_{NET}^i$, $M_{NET}^A$, $N_{NET}^A$, $H_{DISK}^i$, $H_{DISK}^A$ respectively.

Datacenter-wide integrated imbalance value $ILBi$, of server $i$. In statistics variance is used as a measurement of how far a set of numbers are spread out from each other, therefore it is widely used. Using variance, an integrated load imbalance value ($ILBi$) of server $i$ is defined

\[
ILBi = \frac{(Avg_{i} - CPU^A)^2 + (Avg_{i} - MEM^A)^2 + (Avg_{i} - NET^A)^2 + (Avg_{i} - HD^A)^2}{4}
\]  

Where

\[
Avg_{i} = \frac{(C_{PU}^i + M_{MEM}^i + N_{NET}^i + H_{DISK}^i)}{4}
\]

is average utilization of multi-dimensional resource in a physical machine, also called integrated load in LIF

V. EXPERIMENTAL RESULTS

Algorithms are compared using the cloud simulator – cloudsched. Similar to Amazon eight EC2 instances with high CPU, high memory and standard configuration but not exactly the same, eight types of virtual machines with equal probability of requests are generated. Cloudsched simulator generates different requests as follows: the total numbers of arrivals (requests) can be randomly set, all requests follow Poisson arrival process and have exponential length distribution. Therefore to test the algorithm, it is executed six times and its average has been taken. Experiment is conducted using the following data center characteristics.

![Figure 1. DataCenter characteristics_1](image-url)
PM Type 1: CPU 16.0 GHz Memory 58.0 G Bandwidth 3380.0 M Hard disk 14 G= 50
PM Type 2: CPU 52.0 GHz Memory 136.8 G Bandwidth 3380.0 M Hard disk 14 G= 25
PM Type 3: CPU 40.0 GHz Memory 14.0 G Bandwidth 3380.0 M Hard disk 14 G= 25

Table - 1: Research utilization and unbalanced degree of data center

<table>
<thead>
<tr>
<th></th>
<th>LIF</th>
<th>Extended_LIF</th>
<th>Random</th>
<th>Round Robin</th>
<th>ZHJZ</th>
<th>ZHJZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbalanced degree of data center (Variance)</td>
<td>0.17%</td>
<td>0.14%</td>
<td>0.24%</td>
<td>0.20%</td>
<td>0.21%</td>
<td>0.19%</td>
</tr>
</tbody>
</table>

Table 1 shows average imbalance value of a cloud data center, it can be seen that Extended_LIF algorithm has lowest imbalance value of a cloud data center.

VI. CONCLUSION

Essential requirements of a dynamic resource scheduler is to reduce imbalance level of the datacenter and to make all the resources are equally loaded. Simulation results shows that proposed Extended_LIF algorithm has low unbalance degree of data centre. It makes load on resources are equally distributed.

VII. ACKNOWLEDGEMENT

The authors would like to thank the dedicated SIG-research group in the area of Cloud & Grid Computing, wireless networks at the Dept. of Computer Science, AIMIT, St Aloysius College,, Mangalore, India, for many simulating discussions for further improvement and future aspects of the Paper. Lastly but not least the author would like to thank everyone, including the anonymous reviewers.
REFERENCES


