JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN INFORMATION TECHNOLOGY

MANAGING UNDERLOADED HOST USING DYNAMIC VIRTUAL MACHINE CONSOLIDATION

[1]ASITHA MICHEAL, [2] VAISHALI GHATE, [3]JALPA MEHTA

Department of Information Technology
Shah & Anchor Kutchhi Engineering College, Mumbai

[1]asithamicheal@gmail.com, [2] vaishalighate@yahoo.co.in , [3]sakec.jalpam@gmail.com

ABSTRACT *- As Cloud Computing is growing rapidly and clients are demanding more services and better results, load balancing for the Cloud has become a very interesting and important research area. Due to increase in the demand in Cloud Computing model have led to the establishment of large scale virtualized data centres. Such data centre consumes enormous amounts of electrical energy resulting in high operating Costs. Energy efficiency has now become one of the major design constraints for current and future cloud data centre operators. One way to conserve energy is to transition idle servers into a lower power-state (e.g. suspend). Therefore, virtual machine (VM) placement and dynamic VM scheduling algorithms are proposed to facilitate the creation of idle times. Due to variability in the workloads experienced by modern application, the VM placement should be managed properly. Managing only the overloaded host will increase the availability of resource to the customer but it will increase the energy consumption.*

Energy Consumption can be reduced by switching off the idle server and migrating underutilized server. Dynamic Consolidation is an effective way to manage the underloaded Host. This paper proposes two algorithms to manage the underloaded host which will improve the resource utilization.

1. INTRODUCTION

Cloud computing is delivering software, storage & infrastructure as a provisioned service to end users. Due to its enhanced service there is tremendous increase in the usage of cloud computing each and every year. Whenever there is increase in demands Cloud vendors are based on automatic load balancing services, which allowed entities to increase the number of CPUs or memories for their resources to scale up according to their requirement. This service is optional and depends on the entity's business needs. Therefore load balancers served two important needs, primarily to promote availability of cloud resources and secondarily to promote performance.

Load Balancing is process of reassigning the total load to the individual nodes of the collective system to make resource utilization effective and to improve the response time of the job, simultaneously removing a condition in which some of the nodes are over loaded while some others are underloaded. Load balancing is a relatively new technique that facilitates networks and resources by providing a maximum throughput with minimum response time. Dividing the traffic between servers, data can be sent and received without major delay. Different kinds of algorithms are available that helps traffic loaded between available servers. A basic example of load balancing in our daily life can be related to

Websites. Without load balancing, users could experience delays, timeouts and possible long system responses. Load balancing solutions usually apply redundant servers which help a better distribution of the communication traffic so that the website availability is conclusively settled. There are different Techniques to manage the underloaded host. One well known technique to conserve energy besides improving the hardware is to virtualize the data centres and transition idle physical servers into a lower power-state (e.g. suspend) during periods of low utilization. That is to manage the underloaded Host is using dynamic VM Consolidation. VM Consolidation means placing the VM in the data centre so that all the resources in the data centre are used efficiently. Dynamic VM Consolidation means live migration of VM from least loaded host to another host which is not utilized fully. In this paper two algorithms ESWCT and modified ELMCT to make full usage of available resource in the data centre. Both algorithms investigate the problem of consolidating heterogeneous workloads. The algorithms try to execute all Virtual Machines (VMs) with the minimum amount of Physical Machines (PMs), and then power off unused machines to reduce power consumption. The two algorithms are based on the fact that heterogeneous workloads need a variety of resources in the data centre simultaneously.

2. REVIEW OF LITERATURE

Managing the Underloaded Host has become very important to reduce the energy consumption. This section focus on different Techniques used to manage the Underloaded Host and to reduce the power consumption.

Eugeen Feller, Louis Rillingy, Christine Morin [1] propose Ant Colony Optimization (ACO) based algorithm. This paper first accurately models the workload placement problem as an instance of the multi-dimensional binpacking (MDBP) problem. It then takes a nature-inspired approach derived from the behaviour of real ants and proposes a novel algorithm based on the Ant Colony Optimization (ACO) meta-heuristic to compute the placement dynamically according to the current load. This algorithm applied on a number of synthetic test instances and compares it with one frequently applied greedy algorithm (i.e., FFD). The results indicate that the ACO-based algorithm outperforms the evaluated greedy approach as it computes workload placements with superior energy gains through better resource utilization and requires fewer machines. Moreover by solving the model utilizing the IBM ILOG CPLEX solver, we show that the solutions computed by this approach are nearly optimal (i.e., small deviation of 1.1%). To the best of our knowledge this is the first work to: (1) apply ACO on the MDBP problem in the context of dynamic workload consolidation and (2) utilize ACO in order to conserve energy.

LI Hongyou, WANG Jiangyong& LIU Tang1 [2] delivers two algorithms called the Energy-aware Scheduling algorithm using Workload-aware Consolidation Technique (ESWCT) and the Energy aware Live Migration algorithm using Workload-aware Consolidation Technique (ELMWCT) to reduce the energy consumption in cloud Data Centre. The two algorithms are based on the fact that multiple resources (such as CPU, memory and network bandwidth) are shared by users concurrently in cloud data centre and heterogeneous workloads have different resource consumption characteristics. Both algorithms investigate the problem of consolidating heterogeneous workloads. They try to execute all Virtual Machines (VMs) with the minimum amount of Physical Machines (PMs), and then power off unused physical servers to reduce power consumption.

Michele Mazzucco, DmytroDyachuk and Ralph Deters [3] address the problem of maximizing the revenues of Cloud providers by trimming down their electricity costs. This paper proposes and evaluates energy-aware allocation policies that aim to maximize the average revenue received by the provider per unit time. This is achieved by improving the utilization of the server farm, i.e., by powering excess servers off. The policies proposed are based on (i) dynamic estimates of user demand, and (ii) models of system behaviour. The emphasis of the latter is on generality rather than analytical tractability. Thus, it uses some approximations to handle the resulting models. However, those approximations lead to algorithms that perform well under different traffic conditions and can be used in real systems. As a solution allocation policies which are based on the dynamic powering servers on and off are introduced and evaluated. The policies are based on (i) dynamic estimates of user demand, and (ii) models of system behaviour. The policies aim at satisfying the conflicting goals of maximizing the users' experience while minimizing the amount of consumed electricity.

Carlo Mastroianni& Giuseppe Papuzzo [4] presents ecoCloud, an approach for consolidating VMs on a single computing resource, i.e., the CPU. Here, the approach is extended to the multidimension problem, and is presented for the specific case in which VMs are consolidated with respect to two resources: CPU and RAM. With ecoCloud, VMs are consolidated using two types of probabilistic procedures, for the assignment and the migration of VMs. Both procedures aim at increasing the utilization of servers and consolidating the workload dynamically, with the twofold objective of saving electrical costs and respecting the Service Level Agreements stipulated with users. All this is done by demanding the key decisions to single servers, while the data centre manager is only requested to properly combine such local decisions. The approach is partly inspired by the ant algorithms and subsequently by a wide research community, to model the behaviour of ant colonies and solve many complex distributed problems. The characteristics inherited by such algorithms make ecoCloud novel and different from other solutions. Among such characteristics: 1) the use of the swarm intelligence paradigm, which allows a complex problem to be solved by combining simple operations performed by many autonomous actors (the single servers in our case); 2) the use of probabilistic procedures, inspired by those that model the operations of real ants; and 3) the self-organizing behaviour of the system, which ensures that the assignment of VMs to servers dynamically adapts to the varying workload. To evaluate the performance of ecoCloud it uses two complementary approaches. It first proposes a fluid mathematical model that derives the evolution of the system with time by assuming that the involved variables are continuous. The second approach consists of experiments performed on real data centres. The two approaches complement each other: the analytical model introduces some simplifying assumptions but allows for an easy exploration of a wide range of scenarios; conversely, the real experiments do not suffer from assumptions but are, somehow, less representative. Both the approaches show that ecoCloud achieves very good consolidation, and smoothly adapts to possible changes in the system condition

3. PROPOSED MODEL

The proposed system consists of 2 algorithm Energy-aware Scheduling algorithm using Workload-aware Consolidation Technique (ESWCT)& Modified Energy aware Live Migration algorithm using

Workload-aware Consolidation Technique (MELMCT). The proposed System can be explained using three models.

3.1CLOUD SYSTEM MODEL

In this project, as commented before, we focus on the IaaS cloud system. The target cloud system in our model is assumed to be deployed in a data centre which is composed by different hardware. With virtualization solutions, the users can use the resources in the form of leases safely. The IaaS cloud system uses virtualization technology to manage its resources and offers VMs to its customers.

We assume that the IaaS cloud system consists of a set N of n physical machines. N can be represented by N={pm1,pm2,…….pmn}. Each physical machine can be multiple compositions of CPU, memory and network cards. These heterogeneous physical machines in cloud data centres have different capabilities and speeds. Using virtualization technology, a set M of m virtual machines run on each physical machine. M can be represented by M = {vm1,vm2,……..,vmm}. The virtual machines on a physical machine can be restarted, paused and migrated to other physical machines in the cloud data centre. Because of the heterogeneity, we carefully consider the hardware resource heterogeneity (including CPU, memory, and network cards) when placing VMs. Here we use rcd to denote the computing capability of a dedicated reference server d. rcd can be measured by Millions of Instructions Per Second (MIPS). For an application i, it needs ricdv of the computing capability of the dedicated reference VM dv. Thus if the application i runs in another VM hv, the computing capability richv is

$r_{ichv} = \frac{r_{icdv} . r_{ch}}{r_{cd}}$……………………….Eq. (1)

Here rch denotes the computing capability of the server h. rch can also be measured by MIPS.

3.2 APPLICATION MODEL

In an IaaS cloud system, it is common for users to buy VMs to run their applications instead of physical machines. For example, the VMs in Amazon EC2 are leased to users at the price of ten cents per hour. Each VM offers 1.2 GHz computation power, 1.7 GB memory and 160 GB disk space. Cloud computing users usually pay for a statically configured VM size ignoring the actual resources needed by their applications. Therefore, it is very important to know the size of computing capability, memory capability and network capability for virtual machines so that customers can execute on demand. This is good for both cloud computing providers and cloud computing users. Accurate performance modelling of an application would help cloud computing providers in better VM sizing. Thus it is good for better scheduling. It is also good for users to size their VMs according to actual needs to reduce costs. However, it is really very difficult to create such accurate modelling of an application's performance. So in our study, we use a dedicated reference physical server. Thus we can obtain the actual size of the VM needed by the application. Here we use the demand vector rc, rm, rn to represent the requirements of computing capability, memory capability and network capability of an application.

3.3 ENERGY MODEL

The power consumption of physical machines in cloud data centres is mostly determined by processor, memory, disk storage and network interface controllers. The energy model in our study is based on the fact that processor utilization has a linear relationship with energy consumption. That is to say, the CPU consumes the main part of energy. The assumption has been used by many other works. We assume that physical machines in the cloud data centres work in three different modes: the idle mode, the active mode and the sleep mode. Many studies have shown that an idle sever consumes about 70% of the power consumed by the server running at the full speed on average. In order to reduce the total power consumption, we should switch idle servers to the sleep mode. We use the same energy model in our study as follows:

$P(u) = k \cdot P_{max} + (1- k) \cdot P_{max} \cdot u$…..Eq. (2)

In Eq. (2), Pmax is the power consumption at the peak load and k is the fraction of power consumed by the idle server. U is the CPU utilization.

3.4 PARAMETERS OF PROPOSED ALGORITHM

- Parameter 1:

Computing capability r_i^c, memorycapabilityr_i^m, and network capability$r_i^n$of a single physical machine i.

- Parameter 2:

The remaining computing capability r_{iR}^c , memory capabilityr_{iR}^m, and network capabilityr_{iR}^n of a single physical machine i.

- Parameter 3:

Computing capability rc, memory capability rm, and network capability rn of the VM which is going to be placed.

- Parameter 4:

average CPU utilization (CPU_i^U) of a single sever i, average memory utilization (MEM_i^U) of a single sever i , and average network bandwidth utilization (NET_i^U) of a single sever i.

- Parameter 5:

integrated resource utilization (IR_i^U) of a single server i is defined as

$\frac{CPU_i^U + MEM_i^U + (NET_i^U)}{3}$………………..Eq. (3)

- Parameter 6:

Imbalance Utilization Value (IUVi) among multi-dimensional resources of physical server i. Here we assume that the new VM is placed on the physical server i. And the current average CPU utilization, average memory utilization, average network bandwidth utilization are CPU_i^U,

MEM_i^U & NET_i^U respectively in Parameter 4. The imbalanced utilization value (IUVi) of sever i is defined as

$$\frac{(CPU_i^U - IR_i^U)^2 + (MEM_i^U - IR_i^U)^2 + (NET_i^U - IR_i^U)^2}{3} \text{……………………Eq. (4)}$$

In Eq. (4), when a new VM is going to be placed on the physical server i, we should make sure that $r_{iR}^c > r^c$, $r_{iR}^m > r^m$, $r_{iR}^n > r^n$

IUVi is used to match the VM and the server to get the best balanced utilizations between multi-dimensional resources.

- Parameter 7:

average utilization of all CPUs(CPU_u^A) in a cloud data centre. Let NUM_i^{cpu} be the total number of CPUs of server i,

$$CPU_u^A = \frac{\sum_i^N CPU_i^u NUM_i^{cpu}}{\sum_i^N NUM_i^{cpu}} \text{…….. Eq. (5)}$$

where N is the total number of physical servers in a cloud data centre. Similarly, MEM_u^A and NET_u^A can be defined.

- Parameter 8:

integrated resource utilization IR_A^U of a cloud data centre, is defined as

$$\frac{CPU_u^A + MEM_u^A + NET_u^A}{3} \text{………………Eq. (6)}$$

3.5 ESWCT ALGORITHM

When an application arrives, the cloud data centre rents a VM to the application. The ESWCT algorithm shows where to place the VM to get a better balance utilization of resource components among a physical server considering resource consumption. Algorithm ESWCT includes three fundamental steps: 1) compute every component capability of each physical sever; 2) get every component capability of the VM; 3) get the smallest value of IUVi and assign the VM to that sever. ESWCT algorithm is only the first part of VM allocation. The algorithm deals with the admission of new applications for VM provisioning and places the VM on the server. In order to use all the provided resources in cloud data centres, heterogeneous workloads should be consolidated to the minimum number of physical machines. Thus, idle servers can be switched to the sleep mode to reduce energy consumption. And the second parts of VM allocation are the optimization of the current VM allocation and dealing with where to migrate VMs. We can dynamically migrate running VMs using ELMWCT from underused physical machines to others which are mostly fully used to reduce power consumption.

Algorithm 1 ESWCT (vj, N)

Input: A new VM vj and a set N of n active physical servers in a cloud data centre

Output: A VM-sever match

```
{
1) Let n∗ = 1 and IUV ∗ =1;
2) for∀ni∈ N
3) compute $r_{nR}^c$, $r_{nR}^m$, $r_{nR}^n$ of nith physical server
4) get the results rc, rm and rn of the VM vj according to CCM, MCM, NCM;
5) calculate IUVi in Eq. (7);
6) if ( IUVi< IUV ∗ )
7) Let n∗ = ni ;
8)          IUVi = IUV ∗
9) end if
10)end for
11)if ( IUV 1 ∗ == )
12) Randomly turn on a new physical server rand n which is in sleep mode to active mode and place theVM to the sever;
13) N = N + {nrand} ;
14) else
15) Assign vj to n∗ ;
16) end if
}
```

3.6 MELMCT ALGORITHM

MELMCT algorithm is the second part of VM allocation. The algorithm MELMCT is the optimization of the current VM allocation. It is carried out in two steps. At the first step, the algorithm MELMCT chooses the VMs which are needed to be migrated. To determine which VMs should be migrated, integrated resource utilization is introduced. If the utilization of CPU, memory and network of a single physical machine is below the given integrated resource utilization, all the VMs on the physical machine have to be migrated to other physical machines. At the second step, the chosen VMs are allocated to other physical machines using the ESWCT algorithm. After all the VMs migrated to other machines, the physical machine has to be switched to the sleep mode to reduce power consumption.

Algorithm 2 MELMWCT

Input: A set N of n active physical machines in a cloud data centre

Output: VMs on underused physical servers to those which are mostly fully used

```
{
1) for∀ni∈ N
2) compute $CPU_n^U$,$MEM_n^U$, $NET_n^U$  of the n physical server;
3) Calculate $IR_n^U$  of the n physical Server
4) Sort $IR_n^U$of the n physical server in ascending order
5) Select the least $IR_n^U$ server
6) for∀mj∈ni -th server
5) ESWCT (Vm, N –{ni});
6) end for
7) Turn n -th server to sleep mode;
8) N=N-{ni};
9) end for
}
```

REQUIREMENT ANALYSIS OF PROPOSED ALGORITHM

Table shows the different configuration given as a input to the project. Table 4.1 shows the Configuration detail of Virtual Machine. It shows that 8 Virtual Machine with different configuration given as an input. It consists of Configuration of CPU measured in MIPS, Memory measured in GB & Bandwidth in Mbps. For each and every Virtual Machine different values are provided as a input due to heterogeneous workload. Similarly table 4.2 shows the Configuration detail of Physical Machine. The table shows that 4 physical machine of different configuration is gives as a input.

Types	CPU/ MIPS	Memory/ GB	Network/ (Mbits)
1	125	1	2
2	250	1	5
3	250	2	8
4	500	2	8
5	500	4	10
6	1000	4	10
7	1000	6	12
8	1000	10	20

Table 4.1 Types of Virtual Machine

Types	CPU/ MIPS	Memory/ GB	Network/ (Mbits)
1	800	11	100
2	1000	12	30
3	2000	16	51
4	4000	33	120

Table 4.2 Types of Physical Machines

4. SIMULATION RESULTS

The table shown below shows the Comparison of Power Consumption of ELMCT & MELMCT algorithm. The Table consist of measurement of Power Consumption (kw) with respect to time (minutes). It shows the record of each and every 10 minutes. The Average Power Consumption of ELMCT Algorithm is 65 kw while MELMCT Algorithm is 61. Thus the above result proves that power consumption of MELMCT Algorithm is less than ELMCT Algorithm.

	Power (KW)	
Time(minutes)	ELMCT Algorithm	MELMCT Agorithm
10	70	65
20	60	55
30	80	55
40	55	65
50	60	65
Average	65	61

Table 4.1 Comparison of Power Consumption of ELMCT & MELMCT Algorithm

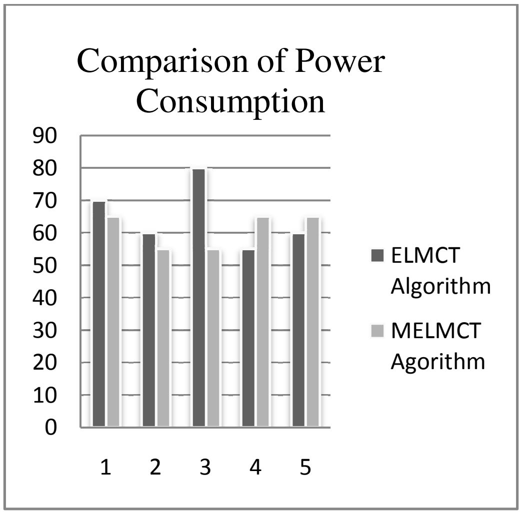

Fig 5.1: Comparison of Power Consumption

5. CONCLUSION

In this project two algorithms had been implemented to reduce the power consumption by considering the different workload and resource. It migrate the VM and switch off the idle server which is one of the efficient energy saving process. This project is used to maintain only underloaded host. ESWCT Algorithm properly places the VM in Data Centre considering the all 3 resources such as CPU, RAM, BW. Each and every placement of VM is done by calculating the current utilization and computing capability of VM and Data centre while the MELMCT controls the migration and switching the idle to sleep mode. This algorithm selects the less utilized host for live migration, so that unnecessary live migration can be avoided and eliminated. MELMCT selects the underutilized server by calculating average utilization of the entire server in the data centre.

Implementation of managing overloaded host will increase the advantage of this project. Adding more objectives into our model and then implementing the algorithms in a real cloud data centre constitute our future work.

6. REFERENCES

[1] Eugen Feller, Louis Rillingy, Christine Morin, “Energy-Aware Ant Colony Based Workload Placement in Cloudsim”, 12th IEEE/ACM International Conference, 2011.

[2] LI Hongyou, WANG Jiangyong, PENG Jian, WANG Junfeng, LIU Tang, “Energy-Aware Scheduling Scheme Using Workload-Aware Consolidation Technique in Cloud Data Centres” in China Communications, December 2013.

[3] Michele Mazzucco and DmytroDyachukand Ralph Detersy, “Maximizing Cloud Providers Revenues via Energy Aware Allocation Policies”, IEEE/ACM International Conference, Feburary 2011.

[4]Carlo Mastroianni, MichelaMeo, and Giuseppe Papuzzo, “Probabilistic Consolidation of Virtual

Machines in Self-Organizing Cloud Data Centres", IEEE transactions on cloud computing, vol. 1, no. 2, july-december2013.

[5]Anton Beloglazov and RajkumarBuyya, "Managing Overloaded Hosts for Dynamic Consolidation of Virtual Machines in Cloud Data Centres under Quality of Service Constraints", IEEE transactions on parallel and distributed systems, vol. 24, no. 7, july 2013.

[6]Mayank Mishra, Anwesha Das, PurushottamKulkarni, and AnirudhaSahoo, "Dynamic Resource Management Using Virtual Machine Migrations", IEEE Communications Magazine September 2012

[7]SeyedSaeidMasoumzadeh and Helmut Hlavacs, "Integrating VM Selection Criteria in Distributed Dynamic VM Consolidation Using Fuzzy Q-Learning" 9thCNSM: workshop SVM 2013

[8]Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, Cesar A. F. De Rose and RajkumarBuyya, "CloudSim: a toolkit for 3d modelling and simulation of cloud computing environments and evaluation of resource provisioning algorithms" 24 August 2010 in Wiley Online Library