Energy Management review in cloud

Rashmi K Gundakalli¹ & Dayananda Lal N²
¹Student, M.Tech Computer Science and Engineering, Acharya Institute of Technology, Bengaluru, India
²Assistant Professor, Computer Science and Engineering, Acharya Institute of Technology, Bengaluru, India

Abstract: Cloud computing is a much discussed technology. It transforms qualitative concepts into their corresponding quantitative expression. It is said to have restructured the distributed computing and has contributed much to business computing. Cloud computing is accompanied with the development of large data centers. The consolidation of resources causes large amount of energy requirement and dissipation of heat. The power consumption pattern in cloud computing, however differs from legacy storage oriented services. Datacenters require large amount of heat when they operate at their peak. Thus, calling for a considerable of investment coupled with the heat emission. This paper addresses such issues, its impact and efficient energy saving methods to reduce the huge energy consumption in the cloud datacenter.

1. Introduction

Cloud computing, as defined in [1], is a distributed type of computing resource consisting of physical machines with each hosting several virtual machines (VMs). These VMs can be provisioned and released dynamically, and are also presented to customers as processing and storage resources based on service level agreements (SLAs). Cloud Computing has already been proved to be one of the most important processing and storage resources for most applications and enterprises. Fast development in the field of cloud computing has made a great impact on Information Technology. It is now seen as an alternative to HPC clusters [2]. There are two main reasons: first, the clouds’ utility-based usage model allows users to pay as they use, in a manner similar to other public utilities services; second, there is relatively low investment needed for the end devices that access the cloud resources [2]. The development is accompanied by the ever growing data centers. To get more profit from the data centers, the brokers make complete use of the hardware. This has led to increased consumption of power in the data centers. Hence, a good resource scheduling and management schemes for every on demand resource allocation, performance optimization, load balancing and energy saving is required. However, much better energy saving methods in cloud systems can be achieved if they devoted to HPC and are augmented with global system-level optimizations. These optimizations focus on providing energy-efficient optimizations for the entire data center while taking each individual’s power behaviors as input parameters.

The computation oriented model is likely to widen the difference between the amount of energy required for peak periods and off-peak periods in a cloud data center due to [3]:

- CPU: Intensive computation requirement enlarges CPU utilization rate.
- DRAM: Increased computation demand brings up the required memory sizes. DRAM is one of the main consumers of energy in a computer.
- Hard Disk: Frequent computation requests raise the frequency of random disk accesses, which consume much more energy than the usually used sequential disk accesses for file transferring.

Mathematical methods have been developed for load forecasting the number of network requests is the key cause of the required load. To accurately predict power demand needs to forecast the number of user requests. Processing a user request involves several dynamic energy consuming activities that include CPU and disk actions. With the popularization of dynamic voltage and frequency scaling (DVFS), CPUs consume power in proportion to their utilization. On the other hand, power consumption of a hard disk is affected by the disk cache design, materials of storing media, etc.

The IT infrastructure provided by the datacenter owners/operators must meet various SLAs established with the clients. The SLAs may be resource related (e.g., amount of computing power, memory/storage space, network bandwidth), performance related (e.g., service time or throughput), or even quality of service related (e.g., 24-7 availability, data security, percentage of dropped requests).[4]
2. Problem Definition

Initial Assumption: Let us, throughout this paper consider that, each computing node consists of a processor, memory and a network adapter[ an energy]. The server at all nodes are considered to homogeneous and capacity of resource at each node is 1 unit. Energy consumed in transmission and switching are neglected. One or more VM’s could be run in server mode and each server has a VM monitor. Map-Reduce jobs are scheduled on these VM’s using VMwares and Xen virtualization [5]. Each VM further contains a Task Tracker and a DataNode, the computational and data components of the Map-Reduce framework. The job completion time can be calculated. The VM’s are processed using either batch processing method or online processing method. Let us further assume that there are servers can be put into sleep or hibernation mode, once all the jobs scheduled for a particular VM are done processing.

3. Problem Formulation

3.1 VM Displacement based on SLA

The goal is to solve semi-static VM placement problem and to minimize the operational cost in datacenter. This includes power and mitigation costs and penalty of violating response time as well. It is a NP hard problem. It can be modeled as follows:

\[
\text{Min } C_p \sum_{i} [x_i \rho_i^p + \sum_{j} \phi_{ij} T_0 + \sum_{i} z_i \text{cost}_i^m + T_e \sum_{i} f_i \lambda_i \sum_{j} z_{ij} \text{cost}_i^m]
\]

The first term in this formula represents the energy cost which is inclusive of idle server cost and the server utilization costs. The second term in the formula capture the migration costs of the VM in the server and the last term signifies the penalty that might occur whenever the SLA is being violated. There is no efficient solution for this problem. Since, the algorithm has to scale for large number of client and servers. There are two disadvantages:

a. SLA doesn’t specify the number of resources required. This means that there has to be some way to translate the SLA constraints to number of resources required for each client.

b. Considering the fixed number of resources might degrade the performance guarantee for others. [6]

3.2 Dependencies between the network and VM Displacement

Let S×V matrix capture the placement of the VM on the physical servers [7]. Let it denoted by f, where let \(f_{old}\) be the old placement of the VM and the new VM placement is denoted by \(f_{new}\). Let the overall CPU utilization be denoted by \(CPU_x\). Further, let \(\text{Req}_x\) denote the required computing capacity for the VM \(V_x\).

\[U_x(F,t) = \sum_{i} f_{ij} \times (\text{Req}_x \times CPU_i) / CPU_x\]

\[\sum_{i} f_{ij} \times (\text{Req}_x \times CPU_i) / CPU_x \leq 1\]

\[P_x(F,t) = 0.7P_x^{\text{max}} + 0.3P_x^{\text{max}} \times U_x(F,t)\]

\[\text{Energy}(F,t) = \sum_{i=1}^{n} \left[ \sum_{t=1}^{m} \left( \sum_{j} \right) \right]\]

\[\text{comm}(F) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{t} c_{ij}^m f_{ij} d_{ij}\]

Where, \(S_k\) represents the server, \(P_k^{\text{max}}\) represents the power of the server \(S_k\). Eqn 1, \(U_x(F,t)\), signifies the placement of VM F at time t and utilization of the server \(S_k\). Eqn 2, explains that the overall usage of the server cannot be more than 1. The third equation is about the power consumption of the VM F at time t on server \(S_k\). The overall energy that is being consumed is given by Eqn 4. Eqn 5 signifies the load on the network incurred from \(t_1\) to \(t_2\) when VM F is being placed. The last equation explains about the entire implementation costs for F.

4. Problem Statement

Minimization Algorithms

4.1 Placing VM based on SLA

To reduce the cost of energy in the data centers, a simple two step algorithm is being used. There are two algorithms that are being considered-dynamic programming and convex optimization methods for step 1. In step 1, based on the resource requirement and the based on the previous epoch decisions, customers are being ordered and hence the VM is allotted [6]. In step2, to cater the resource utilization by the under-utilized and under competing servers, local searches are executed. Only a small number of servers from each of the server groups can be considered every time for the calculations. To categorise between the active and inactive servers, their respective energy slopes are being considered. Further, to improvise on the allocation of the resources and to service only the active servers, local search methods are used.

4.2 Communication Aware Algorithm to minimize the network load

This algorithm tries to reschedule the load iteratively to minimize the network traffic. Let \(V_K\)
denote the VM to be hosted on server $S_X$. Let $\deg(V_K)$ denote the number of adjacent VM’s to $V_K$. There are at the most $V_K$ destinations for the placement of one $V_K$. The algorithm computes the migration of the VMs to the servers which will serve them. This migration is said to feasible migration as the destination server has enough resources to serve the needs of the migrating VM. Let $t_1$ and $t_2$ determine the time interval and let $f$ determine the placement. The total load incurred by the network is given by eqn 7(below). Eqn 8 gives the state of the VM after migration ($f^{old} \rightarrow f^{new}$). Eqn 9 is used to maximize the potential benefits of VM migration.

$$L(v_1, f, t_1, t_2) = \sum_{m=1}^{F} \sum_{n=1}^{F} C_{m,n} f_{m} f_{n} d_{m}$$ Eq.7

$$MC(v_1, F^{old}, F^{new}) = \sum_{m=1}^{S} \sum_{n=1}^{S} S(v_1) f_{m}^{old} f_{n}^{new} d_{m}$$ Eq.8

$$B(v_1, F^{old}, F^{new}, t_1, t_2) = L(v_1, F^{old}, t_1, t_2) - MC(v_1, F^{old}, F^{new})$$ Eq.9

5. Energy Consumption Pattern

The study of the energy consumption is important so as to improvise it to reduce the energy consumption. Servers are one of the important parts of the cloud environment. The energy being consumed by the servers forms large part of the energy in the cloud and varies according to its utilization. The varying consumption of energy is also dependent on the type of the computation being run on the server which may be data reading or writing [8].

Networking equipment’s, lightening and pumps consume 6% of the total energy, which means these instruments consume significantly less energy.

5.1 Power Consumption Pattern

To understand the measurement metric, we consider the hardware monitoring tool or software application. This in turn will help us understand the power consumption pattern in the system.[2]The power consumption pattern can further be studied using the benchmarks. The following table describes the same.

<table>
<thead>
<tr>
<th>Table 1. Definitions of Benchmark Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
</tr>
<tr>
<td>Performance</td>
</tr>
<tr>
<td>Power consumption</td>
</tr>
<tr>
<td>Execution time</td>
</tr>
<tr>
<td>Power-efficiency</td>
</tr>
<tr>
<td>Energy consumption</td>
</tr>
</tbody>
</table>

The energy that is consumed by the cloud can be presented by the following equation:

$$E_{Cloud} = E_{Node} + E_{Switch} + E_{Storage} + E_{Others}$$ Eq.10

Where, $E_{node}$ represents the node’s energy consumption, $E_{switch}$ represents the energy consumption of the switching equipment’s, the $E_{storage}$ represents[9] the energy consumption of the storage device and $E_{others}$ represents the energy being consumed by other factors such as fan, current conversion loss etc. The energy consumption model of the cloud can also be modeled as follows:

$$E_{Cloud} = \sum_{m=1}^{M} \sum_{n=1}^{N} P(mm, f_{CPU}, f_{Memory}, mode_{Data}, speed_{Data}, SpeedFactor_{Data})$$ Eq.11

To remove the heat so produced while maintaining the required humidity and air quality, cooling systems are used. Cooling starts with when the dissipated heat is channelized to the chilled water cooling loop while cold air is being supplied to the facility continuously. This process contributes significantly to the energy required. However, its requirement increases with both cloud environment thermal load and outside temperature.

5.1.1 Energy Modeling

The energy calculation depends on various factors such as type, diversity and complexity of the data. The interaction between the sub system (server, storage, network etc) also affects the energy consumption pattern because of the non-linear communication between them. The energy that is consumed can be divided into two parts:

i. Fixed Energy Consumption (server idle time and server cooling time energy details)

ii. Dynamic Energy Consumption( Cloud task energy consumption)

The power consumed by cloud applications which are interactive in nature, can be formulated as follows:
\[ P_l = P_d + E_d B + (N_e E_c + N_t E_d + E_{Init} + E_{Run}) B + E_d B + P_d \]

Eq. 12

6. Conclusion

Energy consumption and its management has been a prime issue for over a decade now. The methods mentioned in this paper will prove to be useful in this regard. However, it should be borne in mind that, each of these algorithms or the methods are inefficient in one or other way. Hence, hunt for better algorithm to reduce the energy consumed is always on.

7. References


