ISSN: 2319-8753

International Journal of Innovative Research in Science,

Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

VM Assignment Algorithm Based Cost Effective Caching in Cloud Computing

Ramya.R[1]

Assistant Professor,Dept of Computer Applications, Bannari Amman Institute of Technology, Sathyamangalam, India[1]

Abstract : Cloud applications are evolving that offer data management services. The applications that involved in heavy I/O activities in the Cloud Technology, utilizes most of the services from Caching. Caching the data may be either spatial data or temporal data. The Local volatile memory might be an alternative support for Cache, but the capacity and the utilization of host machines reduces its usage. The existing system provided the Cache as a Service (CaaS) model as an additional service along with Infrastructure as a Service (IaaS). Particularly, the Cloud Provider sets a large collection of memory that can be dynamically separated and allocated to standard infrastructure services as Disk cache. An effective cache mechanism known as the elastic cache system provides the feasibility of CaaS using dedicated remote memory servers. The novel pricing scheme provides the maximization of cloud profit which gives the guarantee for user satisfaction. The performance degradation occurs due to increasing in the utilization of Energy. The proposed system utilizes Virtual Machines (VM) and doing server consolidation in a data center, by which a Cloud provider can reduce the total Energy consumption for servicing the clients as well as increasing the resource availability in the data center. Multiple copies of VMs are created and place these copies on the servers by using dynamic programming to reduce the total energy consumption. The communication parameters such as latency, bandwidth, and distance are considered in making the decision of assigning VMs to the servers. The algorithm is implemented with the help of the simulation tool (CloudSim) and the result obtained from this reduces the energy utilization and also increases the performance.

Keywords - CaaS, Remote Memory, Energy Efficiency, Virtual Machine, VM Assignment algorithm

I. INTRODUCTION

Cloud Computing is a technology that uses the internet and data centers to maintain data as well as applications. Cloud Computing allows business and consumers to access their personal files from any computer via internet. The “Cloud” simply denotes the default symbol of the internet in diagrams and the “Computing” encompasses the computation and storage. This technology allows more efficient Computing by centralized storage, processing, and bandwidth.

A simple example for this Cloud Computing is Gmail, yahoomail etc. There is no need of software or server to use them. Just an internet connection is needed to start sending an email. There is no need to worry about the internal processing. A cloud service provider is the responsible for all servers and e-mail management. The analogy is “If you want to stay in Chennai for one day, would you buy a house? The users get to use the software alone and enjoying the benefits of Cloud Computing.

For instance, a web server (ie.. a single computer)can run without a Cloud Computing. It means the computer can serve 500 pages per minute. If the website becomes popular, the audience will demand for more pages. At that time, the server became slow down and the audience loses their interest. For this, a server should move to the Cloud, you should rent computer power from the Cloud service provider who has thousands of servers, that all connected together allows sharing of work among each other. This solves the previous problems. It provides Pay-as-you-go model which denotes you have to pay for how much you use. We have to pay more rent for the extra usage. The ultimate goal of Cloud economy is to optimize 1) user satisfaction and 2) Cloud profit.

A. Cloud Storage:

Over the decades, the big internet based companies like Amazon and Google etc identified that only a small amount of data storage capacity is used. This leads to the renting out of space and storage of information on remote servers. Information is then cached on either desktop computers or mobile phones or other internet-linked devices. Amazon Elastic compute (EC2) and the simple storage solutions (s3) are the current best available facilities. In the cloud, there are three separate level on which you can cache. They are the server, load balancer and Content Delivery Network (CDN). In this paper, the data is cached on the server. The Energy Efficiency is low in Cache services. That can be improved with the help of Energy Efficient algorithms which has been described below.

II. RELATED WORKS

A. I/O Virtualization

Virtualization defines the separation of a resource or request for a service from the physical entity which is underlying. I/O Virtualization is a methodology to improve the performance of servers. Due to Virtualization overhead, I/O operations are more expensive than a native system. The significant I/O overhead with the page flipping technique can be replaced by the memcpy functions to avoid the overheads [10]. The I/O performance can be optimized with the help ofa Virtual Machine Monitorr (VMM). The I/O performance overhead can be tackled by doing full functional breakdown with the help of profiling tools [5].

After the importance of Virtualization is known, hardware level features become popular. It has been evaluated to seek for near Native performance. The Intel Virtualization technology is used to provide better I/O performance [7]. All these focuses only on Network I/O. The disk I/O will be considered in cache device to improve the low disk I/O performance.

To handle multiple applications, Virtualized servers require more network bandwidth connections to more networks and storage. In virtualized data centers, I/O performance problems are occurring by running multiple Virtual Machines (VMs) on one server. This can be overcome with the help of I/O Virtualization.

B. *Cache Device*

Cooperative Cache is a kind of Remote memory cache which improves the performance of a networked file system. It uses client memory as a cache. This caching scheme is effective because Remote Memory is faster than a local disk of the client who is requested. The advanced Buffer management technique for Cooperative caching has been introduced based on the degree of locality. This emphasizes the data that has high locality should be placed in the high-level cache and the data that has low locality should be placed in a low-level cache. The Cooperative caching system is also designed at the Virtualization layer reduces the disk I/O operations for shared working sets of virtual machines.

To improve the I/O performance of a local disk instead of a remote disk by using Remote Memory as a Cache in storage area network, Remote Direct Memory Access (RDMA) is used. The data management system utilizes either Solid State Drive (SSD) or the Hybrid Disk Drive (HDD) according to data usage patterns. However latency of an SSD is still higher than Remote Memory (RM). A new approach called RAM Cloud which reduces the latency by storing the data entirely in DRAM of distributed systems [9]. But RAMCloud incurs more cost and high usage of Energy.

The low disk I/O Performance can be enhanced with the help of Cache as a Service (CaaS) model. This is an additional service with IaaS.

C. Cache as a Service

Existing Cloud focus on the CaaS model which consists of two mechanisms: an elastic cache system and a service model with pricing scheme.

The elastic cache uses RM based cache at the block device level which is exported from dedicated memory servers. The elastic properties are On-demand allocation and reduction of storage and Computing resources. The elastic cache system can use any of the Cache replacement algorithms. VMs utilize RM to provide a necessary amount of cache on demand. The exported memory can be seen as available memory pool. The elastic cache uses this memory pool for VMs.

To deploy the elastic cache system, service components are necessary. The users can choose their cache service according to their cache requirement. The elastic cache system consists of two components i.e. VM and a cache server. A VM demands RM to use as a disk cache. A server can have several chunks. The chunk denotes the memory space. If VM wants to access RM, a VM should mark their rights on assigned chunks and then it uses that chunk as a cache. When multiple VMs try to mark their rights on the same chunk concurrently, the conflict can be eliminated with safe and Atomic chunk allocation method. It improves the performance and provides reliable environment. The effective use of capacity and utilization is not limited in this model.

The service model describes the modeling cache services and pricing model. This service model describes two CaaS types. They are High Performance (HP) which uses LM as a Cache and Best Value (BV) which uses RM as a cache. The goal of service model is to reduce the active number of physical machines.

The cost benefit of this CaaS model is Profit Maximization and Performance improvement. But it consumes more Energy to improve the performance effectiveness. The total cost of the system will be high as well as it gives high complexity.

D. Energy Efficient Algorithms

The major issues in Cloud computing are improving the Energy Efficiency. It can be done with the help of 1) Energy Aware Consolidation Technique 2) Dynamic VM Management Algorithm 3) Power and Migration cost aware Application placement and 4) Server Consolidation.

Energy Aware Consolidation

This technique is used to reduce the total Energy consumption in a Cloud Computing system. The server is modeled as a function of CPU and disk utilization. The performance can be determined only for small input size. It focuses only on the scalability of the system and it does not involve in the minimization of operational cost during the problem of assigning VMs on physical servers provides a major drawback of the system.

Dynamic VM Management Algorithm

This algorithm reduces the total power consumption with a restriction on SLA of each VM or minimizes the SLA violation rates by considering a fixed set of active servers.

Power-Aware VM Placement

This algorithm is designed for heterogeneous servers to reduce the total Energy consumption. It does not consider multiple copies of VM and considers only one dimension of resource in the servers provides a major drawback of this algorithm. The proposed paper focuses on Server Consolidation to overcome all the drawbacks which are described above.

III. ENERGY EFFICIENT ASSIGNMENT ALGORITHM

The Energy consumption can be reduced with the help of an Energy Efficient Assignment algorithm and increases the resource availability in the data center.

A. Data Center Management

A data center consists of a number of heterogeneous servers from a well-known server types. The servers of a given type are designed by their processing capacity or CPU Cycles (C^c) or Memory Bandwidth (M^b). The Energy cost is related with Power utilization. The operational cost of the system is the total Energy cost for servicing all client requests. The Energy cost can be calculated by the server Energy consumption by the duration of time in seconds (T_S). The power cost of communication resources is also included in the data center power cost.

The client is assumed as VM. The amount of resources which is needed for each client is determined with the help of workload prediction. Each VM can be copied to different servers which imply the requests can be assigned to more than one server that is generated by a single client. Therefore upper bound L_b limit the maximum number of copies of VM in the data center. If multiple copies of VM have to be placed in different server means, it should satisfy the conditions which are given below

$$\sum_i \delta_{ij}^p C_j^p = c_i^p \text{.........}(1)$$

$$\delta_{ij}^m y_{ij} C_j^m = c_m^i \text{.........}(2)$$

Where δ_{ij}^p and δ_{ij}^m denotes the portion of the jth server CPU Cycles and Memory BW assigned to the VM which is related to the ith client.

c_i^p , c_m^i - Required total processing capacity and memory BW for the ith client

C_j^p , C_j^m - Total CPU Cycle and Memory BW of the jth server.

y_{ij} - a pseudo- Boolean factor to identify whether VM related to a client i is assigned to the server j or not.

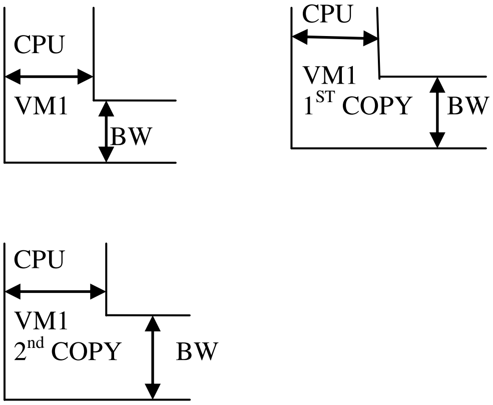

Fig 1.An example of multiple copies of VM

The Constraint (1) shows the summation of the reserved CPU cycles on the assigned servers to be equal to the required CPU cycles for a client *i*. The Constraint (2) shows the provided memory BW on assigned servers to be equal to the required memory BW for the original VM. This condition enforces not to give up the Quality of Service (Quos) for the clients.

B. Reducing Energy Cost of Data Center

Data center management is responsible for allowing the VMs in to the data center to reduce the Energy cost of the data center. The VM Controller (VMC) is responsible for identifying the resource requirements of the VMs and placing these on the servers as well as VM migration to minimize the performance overhead.

The VMC performs these operations based on two different optimization procedures. They are 1) semi-static optimization and 2) Dynamic optimization. Semi-Static optimization has to be done periodically whereas dynamic optimization can be done only whenever it is required.

Here, semi-static optimization procedure is focused. In this technique, the resource requirements for VMs are assumed to be identified based on the SLA specification for the next decision period. The Energy cost of this optimization can be done without considering the previous decision period [11].

The function of semi-static optimization in the VMC is to decide whether to create several copies of VMs on different servers and assign VMs to the servers. By considering the fixed payments by the client for Cloud services they use, the total Energy cost of active servers in data center gets reduced. This will increase the resource availability in the data center.

C. VM Migration

VM migration provides a major advantage in Cloud Computing via load balance in data centers. It is most beneficial in case of certain workload changes. VM migration is performed to minimize the workload changes in a Cloud Computing environment. VM migration reduces the migration cost with the help of semi-static optimization.

D. Dynamic programming

A local search method is proposed to find out the number of copies for each VM and place these copies on servers to minimize the total cost in the system.

Initially the threshold can be set by the Cloud provider. The all servers with utilization less than the threshold means, the total Energy consumption will be reduced. The utilization of the server is defined as the maximum resource utilization in different magnitude.

The formula of the problem is given by

$$\sum_{i} \delta_{ij}^{p} C_{i}^{p} = c_{i}^{p} \text{.........(3)}$$

$$\delta_{ij}^{m} y_{ij} \leq L_{i} \text{............(4)}$$

Where L_i denotes the maximum number of servers which is allowed to serve the client i.

The constraint (3) describes the needed processing capacity is given. The constraint (4) guarantees that the number of copies of VM does not exceed the highest possible number of copies.

To identify the under-utilized servers, each of the servers is turned off one by one. With the help of dynamic programming method, the total Energy cost of utilization is determined by placing their VMs on other active servers. The dynamic programming is introduced to identify the number of copies of each VM and assign these VMs to the servers. This will reduce the total Energy consumption of a system in a Cloud Computing environment.

E. Server Consolidation

Server Consolidation is defined as the assignment of multiple VMs to a single physical server. In a Cloud Computing System, Server Consolidation is an efficient approach to minimize the total Energy consumption and provides the better utilization of resources. A single server is enough to consolidate VMs which is located in multiple under-utilized servers with the help of VM migration technology and the remaining servers can be set to the power-saving state i.e. by turning of the unused machines. Server Consolidation should consider the SLA constraints. The SLA constraints may be resource related (e.g. memory space, storage space, network bandwidth) or performance related (e.g. Throughput, reliability, scalability).

The steps involved in an Energy Efficient Assignment algorithm is

Step 1: Initially δ_j^p and δ_j^m for each server is set to zero.

Step 2: VMs are sorted based on their processing requirements in decreasing order.

Step 3: For every VM, a method based on DP is used to identify the number of copies placed on the server.

Step 4: The Energy cost can be calculated for assigning a copy of the ith VM to a server k is

$$C_{ik}(\alpha) = \delta_{ij}^p P_j^p + P_j^o c_i^m \big/ C_m^j \text{ ....(5)}$$

Where α denotes the size of the assigned VM to the server. The first term in (5) is the cost related to CPU utilization of the server. The second term denotes the replacement of the constant Energy cost of the active server.

The δ_{ij}^p can be calculated as

$$(\alpha u_i^p / L_i) \big/ C_p^j \text{ ....(6)}$$

Step 5: Find the active and inactive servers.

For active servers, value of cost is decremented by ε.

Step 6: Calculate cost for each assignment

$$Min \sum_{j \in p} y_{ij}^{\alpha} c_{ij}(\alpha) \text{............(7)}$$

w.r.to

$$\sum_{j \in p} \alpha y_{ij}^{\alpha} = L_i \text{..................(8)}$$

Where P denotes the server which is both active or inactive servers and y^{α}_{ik} denotes the assignment parameter. The communication resources like latency, bandwidth and distance parameters are considered here in decision making. Dynamic Programming is used to find the best assignment decision.

VM Assignment Algorithm

The algorithm shows the assignment solution for each VM.

```
Inputs:
$C_j^m, C_j^p, P_j^o, P_j^p, c_i^m, c_i^p, L_i, c_i^b, c_i^d, c_i^t, C_j^b, C_j^d, C_j^t, B_j^o, D_j^o,$
$T_j^o, B_j^b, D_j^b, T_j^b$
Outputs:
$\phi_{ij}^p, \phi_{ij}^m, \phi_{ij}^b, \phi_{ij}^d, \phi_{ij}^d$ (i is constant in this alg)
P= { }, B= { }, D= { }, T= { }
For (k = 1 to number of server types)
ON=0; OFF=0;
For ($\alpha$ = 1 to $L_i$)
$\phi_{ij}^p = (\alpha c_i^p / L_i) / C_j^p$
$\phi_{ij}^b = (\alpha c_i^b / L_i) / C_j^b$
$\phi_{ij}^d = (\alpha c_i^d / L_i) / C_j^d$
$\phi_{ij}^t = (\alpha c_i^t / L_i) / C_j^t$
```

```
$C_{ik}(\alpha) = (\phi_{ij}^{p} P_j^{p} + P_j^{0} c_i^{m} / C_j^{m}) + (\phi_{ij}^{b} B_j^{p} + B_j^{0} c_i^{b} / C_j^{b}) + (\phi_{ij}^{d} D_j^{p} + D_j^{0} c_i^{d} / C_j^{d}) + (\phi_{ij}^{t} T_j^{p} + T_j^{0} c_i^{t} / C_j^{t})$
End
$J^{ON} = \{j \in s_k \mid (1- \phi_j^{m}) \geq c_i^{m} / C_j^{m} ; (1- \phi_j^{b}) \geq c_i^{b} /$
$C_j^{b}; (1- \phi_j^{d}) \geq c_i^{d} / C_j^{d}; (1- \phi_j^{t}) \geq c_i^{t} / C_j^{t} \}$
$J^{OFF} = \{j \in s_k \mid \phi_j^{p}=0, (1- \phi_j^{m}) \geq c_i^{m} / C_j^{m} ; \phi_j^{b}=0, (1- \phi_j^{b}) \geq c_i^{b} / C_j^{b} ; \phi_j^{d}=0, (1- \phi_j^{d}) \geq c_i^{d} / C_j^{d} ; \phi_j^{t}=0,(1- \phi_j^{t}) \geq c_i^{t} / C_j^{t} \}$
Foreach ($j \in s_k$)
If ($j \in J^{ON}$ & ON< $L_i$)
$P = P \cup \{j\}$, ON++, $cost_{ij}(\alpha) = c_{ik}(\alpha) - \varepsilon$
$B = B \cup \{j\}$, ON++, $cost_{ij}(\alpha) = c_{ik}(\alpha) - \varepsilon$
$D = D \cup \{j\}$, ON++, $cost_{ij}(\alpha) = c_{ik}(\alpha) - \varepsilon$
$T = T \cup \{j\}$, ON++, $cost_{ij}(\alpha) = c_{ik}(\alpha) - \varepsilon$
Else if ($j \in J^{OFF}$& OFF< $L_i$)
$P = P \cup \{j\}$, OFF++, $cost_{ij}(\alpha) = c_{ik}(\alpha)$
$B = B \cup \{j\}$, OFF++, $cost_{ij}(\alpha) = c_{ik}(\alpha)$
$D = D \cup \{j\}$, OFF++, $cost_{ij}(\alpha) = c_{ik}(\alpha)$
$T = T \cup \{j\}$, OFF++, $cost_{ij}(\alpha) = c_{ik}(\alpha)$
End
End
$X = L_i$, and Y = maxsize (P, B, D, T)
Foreach ($j \in$ (P, B, D, T))
For (x = 1 to X)
D[x,y]= infinity; //Auxiliary X×Y matrix used for DP
For (z=1 to x)
D[x,y]=min(D[x,y],D[x-1,y-z]+ costij(z))
D[x,y]=min(D[x,y], D[x-1,y])
End
End
Back-track and update $\phi_j$'s
```

IV. SIMULATION RESULTS

In cloud computing, Simulation can be done with the help of CloudSim. Compare with the traditional VM Placement Algorithm, the proposed model of the VM Assignment algorithm is efficient. The communication parameters such as bandwidth, distance, and latency are considered in making decision of assigning VMs to servers. This will improve the performance of the system by reducing the energy cost of the server.

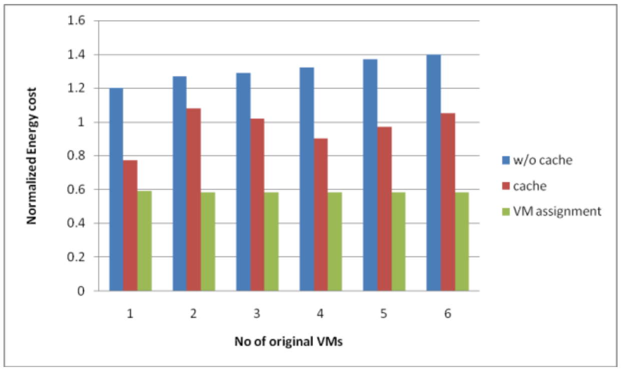

Fig 2. Normalized total energy cost of VM assignment techniques for different servers

Fig 2. describes the no of original VMs versus the Energy cost which shows the reduction in energy cost up to 10-15% than previous approaches.

V. CONCLUSION AND FUTURE WORK

In this paper, an algorithm is proposed to minimize the total Energy cost by considering the communication resources parameters such as latency, bandwidth and distance while making decision in assigning VMs to servers. This method also increases the resource availability in the data center.
Cloud provider can decide how to service VMs with big processing resource requirements and how to distribute the client request. For future work, other resources such as secondary storage can also be considered in this decision making. Moreover, different methods can be provisioned for consistency between VM copies and failure recovery.

REFERENCES

[1] Hyuck Han, Young Choon Lee, Woong Shin, Hyungsoo Jung, Heon Y. Yeom and Albert Y. Zomaya, Fellow, “Cashing in on the Cache in the Cloud,” VOL. 23, NO. 8, August 2012.

[2] M.D. Dahlin, R.Y. Wang, T.E. Anderson, and D.A. Patterson, “Cooperative Caching: Using Remote Client Memory to Improve File System Performance,” Proc. First USENIX Conf. Operating Systems Design and Implementation (OSDI ’94), 1994.

[3] T.E. Anderson, M.D. Dahlin, J.M. Neefe, D.A. Patterson, D.S.Roselli, and R.Y. Wang, “Serverless Network File Systems,” ACM Trans. Computer Systems, vol. 14, pp. 41-79, Feb. 1996.

[4] H. Kim, H. Jo, and J. Lee, “XHive: Efficient Cooperative Caching for Virtual Machines,” IEEE Trans. Computers, vol. 60, no. 1,pp. 106-119, Jan. 2011.

[5] A. Menon, J.R. Santos, Y. Turner, G.J. Janakiraman, and W.Zwaenepoel, “Diagnosing Performance Overheads in the Xen Virtual Machine Environment,” Proc. First ACM/USENIX Int’l Conf. Virtual Execution Environments (VEE ’05), 2005.

[6] L. Cherkasova and R. Gardner, “Measuring CPU Overhead for I/O Processing in the Xen Virtual Machine Monitor,” Proc. Ann. Conf.USENIX Ann. Technical Conf. (ATC ’05), 2005.

[7] X. Zhang and Y. Dong, “Optimizing Xen VMM Based on Intel Virtualization Technology,” Proc. IEEE Int’l Conf. Internet Computing in Science and Eng. (ICICSE ’08), 2008.

[8] J. Liu, W. Huang, B. Abali, and D.K. Panda, “High Performance VMMBypass I/O in Virtual Machines,” Proc. Ann. Conf. USENIX Ann. Technical Conf. (ATC ’06), 2006.

[9] J. Ousterhout, P. Agrawal, D. Erickson, C. Kozyrakis, J. Leverich, D. Mazie`res, S. Mitra, A. Narayanan, G. Parulkar, M. Rosenblum, S.M. Rumble, E. Stratmann, and R. Stutsman, "The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM," ACM SIGOPS Operating Systems Rev., vol. 43, pp. 92-105, Jan. 2010.

[10] E.R. Reid, "Drupal Performance Improvement via SSD Technology," technical report, Sun Microsystems, Inc., 2009

[11] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for Cloud Computing," In proc. of the 2008 conference on Power aware Computing and systems (HotPower'08). 2008.

[12] A. Verrna, P. Ahuja and A. Neogi, "pMapper: Power and migration cost aware application placement in virtualized systems," In proc. Of the 9th ACM/IFIP/USENIX International Middleware Conference.2008.