# Energy Cost and QoS based management of Data Centers Resources in Cloud Computing

Sahil Vashist[1]
Department of Computer Science and Engineering,
Chandigarh Engineering College,
Landran(Pb), India
Sahilvashist90@gmail.com[1]

Rajwinder Singh[2]
Department of Computer Science and Engineering,
Chandigarh Engineering College,
Landran(Pb), India
rwsingh@yahoo.com[2]

**Abstract- Cloud computing is the evolutionary step to transform a large part of the Information and Communication Technology industry. It is the result of the efforts to provide the opportunity to focus on hardware and software cost and the impending reduction of preservation. Cloud computing constitutes both business and an economic model which has been gaining popularity awareness in industry. Cloud computing services providers hope that the widespread adoption of the cloud will bring them more profit and they are actively promoting the technology. Energy Cost of the data-centers in the Cloud computing efficiency one of the main area which needs to be focus. In this paper we summarize previous offerings to the discussion of energy cost of cloud computing, provide a working definition of energy cost of cloud computing and discuss its importance. In this research work, we present an adaptive energy cost saving framework in the cloud to achieve and maintain best Service Level Agreements (SLA). We propose lightweight approach to accurately estimate the power usage of virtual machines and cloud servers at each geographical location. We consider both the system power usage and the SLA requirements, and influence the learning techniques to achieve optimal resource allocation and optimal power efficiency.**

**Keywords:** Energy Cost, Latency, Economic Model, Data centers, QoS

## I. INTRODUCTION

Cloud computing has rapidly emerged as a method for service delivery over TCP/IP networks such as the Internet. It dislocates the conventional IT computing environment. It provides organizations with an opportunity to subcontract the hosting and operations of their mission-critical business applications. The cloud computing represents a major step up in computing whereby shared computation resources are provided on demand. In such a scenario, applications and data can be hosted by various networked virtual machines (VMs). A request, in particular data-intensive applications, often necessitate communicating with data recurrently .Therefore, situation of virtual machines which is located in an application and movement of these virtual machines effects due to unexpected network latency or blockage which is critical to achieve and preserve the obedience routine. Based on the fundamental virtualization technologies, cloud computing has extremely altered the IT services delivery model as well as the hardware infrastructures. Thousands of equipment is composited as a pool of working out possessions of virtual machines (VMs) that are special consideration the end users' requests [1]. In such a surrounding, the VM that implement compliance is positioned on a physical machine in order to implement the required tasks. For a data-intensive request in cloud computing, the demand data might be spread in a number of immeasurably scattered data centers. As an application, especially a data-intensive application, often needs to communicate with related data frequently, the network I/O performance between the data centers that store the data and VMs that execute the applications could affect the performance of the applications significantly [3]. Current VM placement policy mainly focuses on the effectiveness and efficiency of the computing resources utilization [4][5], whereas the network aspects are largely ignored. This might make a VM that executes an application be placed on physical machines that are far away from the data centers that store the related data. As a result, the overall application performance and the system overhead would eventually deteriorate due to the costly data transfer time between the application and the data storage. Furthermore, the virtualization and processor distribution over physical machines frequently result in the unsteadiness of the communication within a cloud computing environment. For example, the TCP/UDP throughput sandwiched between the small instances in Amazon EC2 varies amid 1Gb/s and 2Gb/s frequently [6]. The unanticipated

network congestion and latency places another challenge to optimizing the data transfer time between VMs and the related data. This research addresses the above issues and proposes a policy to place the VM with consideration of the network I/O requirement. In addition, a VM relocation policy is presented to deal with the situation in which the unstable network connection deteriorates the application performance and likely to put in danger the existing concurrence amid the cloud service provider and the end user.

In the proposed VM placement policy, our approach is to minimize the energy consumption cost and allocate the data access by placing a VM on the physical machine with the smallest data transfer time to the required data. In the proposed VM migration policy, VM migration is triggered when the data transfer time crosses a certain threshold due to the unstable network. Then the next optimized location is chosen according to the current network conditions and energy cost, and the VM is migrated to this particular physical machine for better performance without violating SLAs. Here, the entry can be determined by a time-related Service Level Agreement (SLA) between the cloud broker and the cloud user. Our experiments suggest that the switching to least cost data-center located in different geographical locations whose electricity charges are very less, also can lead to minimize overhead cost due to energy consumption. In terms of the proposed VM allocation policy, the task can access related data in a shorter time because the hosted VM is allocated on the physical machine that has better network connection status less energy cost. In terms of the proposed VM migration policy, the VM would be reallocated if the network connection were weaken to an unendurable extent so that the tasks can still running on the another VM with a shorter average achievement time.

## II. LITERATURE REVIEW

The data center has changed significantly as the development of information technology which has enabled it to become the critical nerve center of today's enterprise. As business difficulty increase, so does the number of data center facilities which house a rising amount of powerful IT equipment[7][8]. Data center managers around the world are running into resource limits related to power, cooling, and space, building the resource efficiency of data centers an important topic of debate. As a global consortium comprised of end-users, policy-makers, technology providers, facility architects, and utility companies, The Green Grid aims to address this significant topic[9][10].The services can be of any type e.g. Infrastructure as a Service (IaaS) e.g. Amazon[11][12], Platform as a Service (PaaS) and Software as a Service (SaaS) The major benefits of cloud data centers includes the tradition of financial system of scale to pay back the cost of ownership and the cost of system maintenance over a large number of machines. Customers will be able to access infrastructure and data from a cloud anywhere from the world. With the rapid growth of cloud data centers, the energy consumed by data centers is huge and straightforwardly associated to the number of hosted servers and their workload [13]. It has extremely increased over the past ten years. The power consumption of data centers has huge impacts on the environment[14]. The amount of electricity consumed by data centers worldwide dramatically grew also the electricity cost mostly in developing countries has already under a hike [15][16].

The scheduling in cloud is certainly tricky even when the user demand is absolutely predictable [17]. Authors in [17] make use of first-order estimation which heavily includes multi-service with dissimilar QoS parameters. Also, this work is extended to the model of cloud computing [18]. User demand is also main issue in cloud computing as the distribution of user is very unpredictable usually; the rough calculation of user demand mainly includes the literature that focuses on scheduling with different kinds of user demand uncertainty [19]. Also sample-based approximation are important in provided that a resourceful online scheduling system in cloud [20]. Most notably, some of the scheduling rules in the context of probability and discrete approximation have the desirable property of being independent of the user demand.

Authors in [21][22] have well thought-out the problem of energy-efficient supervision of homogeneous resources in hosting Data centers. The main challenge is to determine the resource request of each application at its current request load level and to allocate resources in the most efficient way. VMs' capability according to the power usage and users' service level agreement (SLA) requirements of the VMs. To accomplish this goal, a well d trade-off in between energy saving and system performance is compulsory. There exist some studies on resource provisioning in the cloud [23], [24]. Authors in [25] have put together the problem of power-aware dynamic placement of applications in virtualized heterogeneous systems as continuous optimization: at every time frame the placement of VMs is optimized to minimize power consumption and maximize performance. However, they do not either consider power and performance in decision making or cannot achieve a optimal solution efficiently by using heuristic method.

## III. SYSTEM MODEL

In this System model each Cloud provides services through its datacenter which houses physical servers. The unit of resource acquisition in the cloud is a Virtual Machine (VM) with many VM instances being instantiated on a physical server to cater to resource demand from end-users. This instantiation is typically done by the Hypervisor, which is a part of the cloud operating system and controls the virtualization aspects of cloud

computing. In this System Model each data center has a geographical location which can be different countries in the world. Also when a request comes to a cloud it is handled by the broker as shown in the figure 1. It is the responsibility of the Broker to submit the requests to the data centers as per internal policies of Cloud. As we can understand that each country has different electricity charges for the data centers, i.e. now in this model broker checks for the current electricity cost and gets the users location. Users' location is required in order to get the distance between request source and the request processor. As shown in Figure 2 the Pseudo-Code represents overall System Model of our frame work.

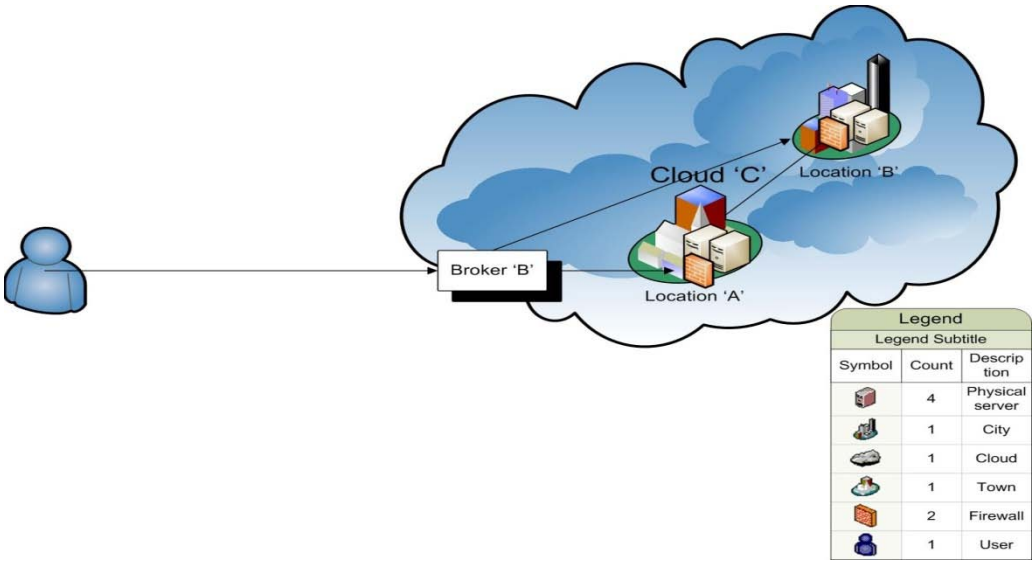



Figure 1. System Model.

## IV. SEQUENCE OF OPERATIONS

a. User interacts to a Cloud via Broker. User asks for VMs to fulfill its resource requirements. User negotiates with broker for VMs (resources) and when negotiation is done resources are handed over to users as shown in figure 1.
b. Users' location is very much important to cater latency issue. In this model broker finds out users location by pinging back to users request and allotted Data-center of the cloud accordingly.
c. If the data-center is located in the location 'A' as shown in the Figure 1 and another data-center is located in another location i.e. location 'B'. Then the distance between these two data-centers is find out and arranged accordingly.
d. In this case we will see the maximum flow in this flow network. For example in the Figure 2 each vertices is considered as data-center in different geographical location and the edges are representing the latency parameter, if we put Edmonds–Karp algorithm [7], in that it uses shortest augmenting paths (latency).

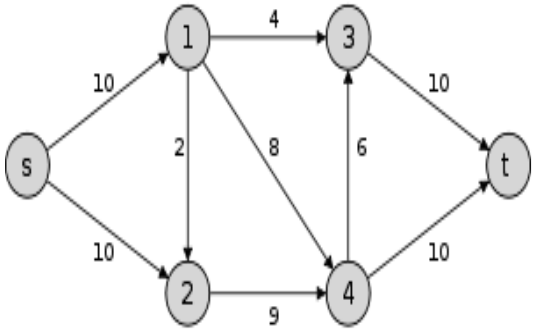
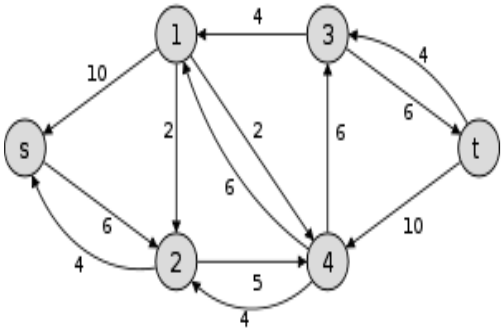



Figure 2. (a) With latency only




(b) Without energy cost

Since, {s, 1, 3, t} with 4 units of flow by comparing latency to energy cost at each geographical location we can put 2 parameters and then find out the best possible solution without violating SLA's. Therefore, the other graph can be representing as Figure 2 (b). Let $G = ((V, E), s, t)$ be a network with C(u,v) and f(u,v) the capacity and the flow of the edge (u,v) respectively.

The mapping defined as,

$$c_f(u,v) = c(u,v) - f(u,v) \text{ if } (u,v) \in E$$
$$c_f(v,u) = f(u,v)$$
$$c_f(u,v) = 0 \text{ Otherwise.}$$

An augmenting path is an (s-t) path in the residual graph $G_f$.

Also, $\{s,1,3,t\}$ with 4 units of flow in figure 2(a) and 5 units of flow in figure 2(b).

e. Selection and re-negotiation is done by minimizing the energy-cost versus latency keeping SLA violation under consideration.

f. After negotiation

```
SendRRQuery ()

        //propagates the query containing the specified number of requirements from users

1:      query= new query();

2:      query.add(this); // Adds self to enable responses to be directed to it

3:      query. add(numServersRequired); // Adds required servers required to  broker

4:      sendQuery(query) ; //  Sends "query" to all data centers

ProcessQuery (query) // implemented at each data center

        //if maximum number of hops specified in the query is exceeded drop it

1:      IF (query.numHops > query.maxHops)// this will check for the availability of the VMs in a
        //data center

2:      dropQuery (query); //when data center foud

3:      END IF

        //if this data center meets the selection criteria

5:      IF (this.getAvailableresources () >=query. numInstanceRequired)

6:              check cost and ping for latecy

7:      IF (this.energy_cost <= desired && latecy <= desired  && SLA== Non Voilated)

8:  Allot the datacenter
```

Figure. 2 pseudo code for resource acquisition

## V. SIMULATION

*Assumptions*

a. We consider only one datacenter per cloud service provider, although it can be easily extended to include multiple datacenters.
b. For simplicity, we assume that the physical servers in datacenters are of similar pattern and have the capability of running the same number of VM instances.
c. The resource requests are in terms of VM instances, although cloud users consume services ranging from Infrastructure-as-a-Service (in terms of physical server instances) to Software-as-a-Service (higher order applications and services). We assume that the "Broker" will translate such requests into VM instances required to provide those services.

*The following parameters were considered during simulation:*

- Number of cloud service providers: 50
- Number of physical servers per datacenter: 100~300
- Maximum Virtual Machines per server: 5
- Resource request quantum: 10~50 vms per request
- Resource request frequency: 2~5 per minute
- Duration of resource usage: 30~60 minutes
- Flash-crowd scenario frequency: once every 3 hours
- Flash-crowd scenario duration: 10 minutes
- Flash-crowd resource request frequency: 15~20 per minut

## EXPERIMENTAL RESULTS.

CloudSim is considered for this purpose, and we have implemented energy cost as one of the parameter and associated each data center with a cost flag and then evaluated the cost each data center is executing for energy consumption then we execute our model which has associated cost and latency and migration an allocation of VMs are done. In the figure 3 we can see that our model has almost linear cost and it is not fluctuating this is because each request is almost send to the data center whose energy cost is very little when compared to other without violating SLAs.

TABLE 1 . COMPARATIVE RESULTS OF SYSTEM MODEL WITH TRADITIONAL APPROACH FOR ENERGY COST.

| Data Center | VMS | Energy Cost before optimal allocation/hr ($) | Energy Cost after optimal allocation/hr ($) |
|---|---|---|---|
| Data_Center0 | 200 | 5 | 3 |
| Data_Center1 | 300 | 8 | 5 |
| Data_Center2 | 400 | 5 | 2 |
| Data_Center3 | 300 | 6 | 3 |
| Data_Center4 | 300 | 5 | 2 |

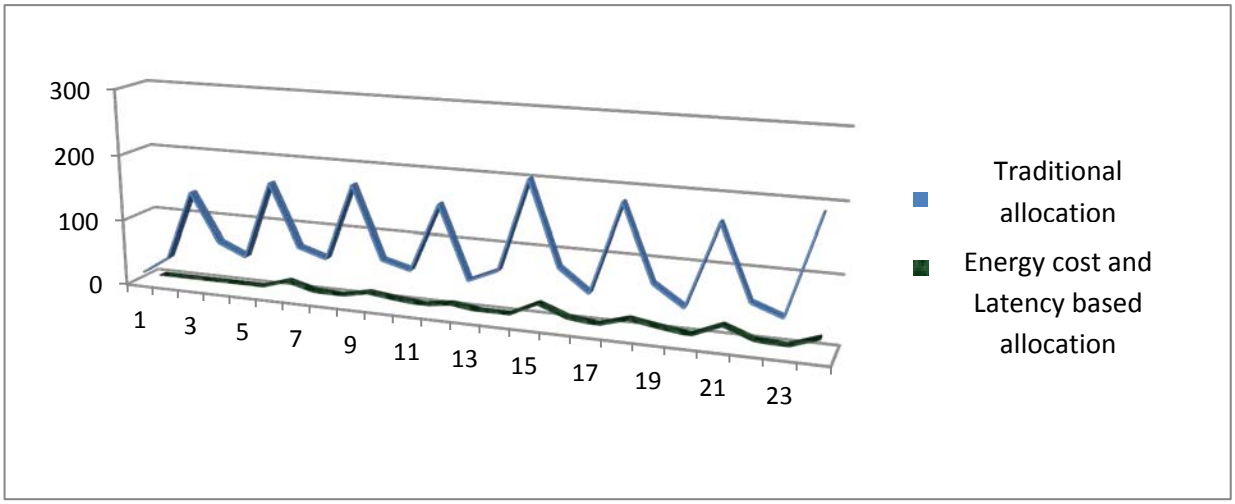



Figure 3 . Simulation Results

## CONCLUSION AND FUTURE WORK

Migration of virtual machines is a well-organized system used to implement cost saving and load balancing in virtualized cloud computing data center. In this paper, we study the request allocation of multiple virtual machines from experimental perspective and investigate different resource reservation methods in the energy cost saving process as well as other complex migration strategies such as parallel migration and workload-aware migration. Experimental results show that: (1) Migration of virtual machine brings some performance overheads. (2) Resource reservation in target machine is necessary to avoid the migration failures and performance cost. (3) The energy cost-aware migration strategy can efficiently improve the cost benefit of a cloud.

Future work will include designing and implementing smart allocation mechanism to improve the energy cost of the cloud and studying the migration strategies as an optimization problem using mathematical modeling methods.

## REFERENCES

[1] H. Wei, I. L. Yeb and B. Thuraisingham, Dynamic Service and Data Migration in the Clouds. Computer Software and Applications Conference, 2009. COMPSAC '09. 33rd Annual IEEE International conference.

[2] K. Sato, H. Sato, and S. Matsuoka. A Model-Based Algorithm for Optimizing I/O Intensive Applications in Clouds Using VM-Based Migration. Cluster Computing and the Grid, 2009. CCGRID '09. 9th IEEE/ACM International Symposium.

[3] S. K. Bose, and S. Sundarrajan, Optimizing Migration of Virtual Machines across Data-Centers. Parallel Processing Workshops, 2009. ICPPW '09. International Conference.

[4] C. Tang, M. Steinder, M. Spreitzer, and G. Pacifici. A Scalable Application Placement Controller for Enterprise Data Centers. Proceedings of the 16th international conference on World Wide Web, 331 – 340, Banff, Alberta, Canada, (May) 2007.

[5] H. N. Van, F. D. Tran, and J-M, Menaud. Autonomic Virtual Resource Management for Service Hosting Platforms. Software Engineering Challenges of Cloud Computing, CLOUD'09. ICSE Workshop on, Vancouver, Canada, 1-8, (May) 2009.

[6] W. Guohui, and T. S. E. NG, The Impact of Virtualization on Network Performance of Amazon EC2 Data Center. INFOCOM, 2010 Proceedings IEEE.

[7] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman,L. Youseff, and D. agorodnov, "The eucalyptus open-source cloud-computing system," in Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid-Volume 00, pp. 124–131,2009.

[8] B. Sotomayor, R. Montero, I. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," *IEEE Internet Computing*, pp. 14–22, 2009.

[9] The green grid consortium, 2011. URL: http://www.thegreengrid.org.

[10] K. Ye, X. Jiang, D. Ye, and D. Huang, "Two Optimization Mechanisms to Improve the Isolation Property of Server Consolidation in Virtualized Multi-core Server," in *Proceedings of 12th IEEE International Conference on High Performance Computing and Communications*, pp. 281–288, 2010.

[11] Amazon Elastic Computing Cloud, aws.amazon.com/ec2

[12] Amazon Web Services, aws.amazon.com

[13] R. Nathuji, K. Schwan, Virtualpower: coordinated power management in virtualized enterprise systems, ACM SIGOPS Operating Systems Review 41 (6),pp. 265–278,2007.

[14] R. Buyya, A. Beloglazov, J. Abawajy, Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges, in: Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 2010,Las Vegas, USA, 2010.

[15] E. Elnozahy, M. Kistler, R. Rajamony, Energy-efficient server clusters, Power-Aware Computer Systems ,pp. 179–197,2003.

[16] E. Pinheiro, R. Bianchini, E.V. Carrera, T. Heath, Load balancing and unbalancing for power and performancee in cluster-based systems, in: Proceedings of the Workshop on Compilers and Operating Systems for Low Power, pp.182–195, 2001.

[17] W. Yi, M. B. Blake, "Service-Oriented Computing and Cloud Computing: Challenges and Opportunities," IEEE Internet Computing, vol. 14, no. 6, pp. 72-75, 2010.

[18] S. Yeo, H.-H. Lee, "Using Mathematical Modeling in Provisioning a Heterogeneous Cloud Computing Environment," Computer, vol. 44, no. 8, pp. 55-62, 2011.
[19] L. Zhou, Y. Zhang, K. Song, W. Jing, and A. V. Vasilakos, "Distributed Media-Service Scheme for P2P-based Vehicular Networks," IEEE Transactions on Vehicular Technology, vol. 60, no. 2, pp. 692-703, 2011.
[20] X. Song, B.-P. Paris, "Measuring the size of the Internet via importance sampling," IEEE Journal on Selected Areas in Communications, vol. 21, no. 6, pp. 922-933, 2003.
[21] J.S. Chase, D.C. Anderson, P.N. Thakar, A.M. Vahdat, R.P. Doyle, Managing energy and server resources in hosting centers, in: Proceedings of the 18th ACM Symposium on Operating Systems Principles, ACM, New York, NY, USA, pp. 103–116, 2001.
[22] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, X. Zhu, No ''power'' struggles:coordinated multi-level power management for the data center,SIGARCH Computer Architecture News 36 (1) ,pp. 48–59,2008.
[23] D. Gupta, L. Cherkasova, R. Gardner, and A. Vahdat. Enforcing performance isolation across virtual machines in xen. In Proceedings of the ACM/USENIX International Conference on Middleware (Middleware), 2006.
[24] D. Ongaro, A. L. Cox, and S. Rixner. Scheduling I/O in virtual machine monitors. In Proceedings of ACM International Conference on Virtual Execution Environments (VEE), 2008.
[25] A. Verma, P. Ahuja, A. Neogi, pMapper: power and migration cost aware application placement in virtualized systems, in: Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, Springer, pp. 243–264, 2008.