# A Hybrid Method for Reduction of Energy Consumption in Cloud Networks

Mehran Tarahomi
Kish International Campus
Sharif University of Technology
Tehran, Iran
tarahomi@ce.sharif.edu

Mohammad Izadi
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
izadi@sharif.edu

***ABSTRACT:*** *Cloud computing is consisted of physical data centers each of which including hundreds or thousands of computers. A key technology which has enabled cloud computing feasible, is virtualization. Virtualization allows us to separate virtual machines in a way that each of these so-called virtualized machines can be configured on a number of hosts according to the type of application of the user. It is also possible to dynamically alter the allocated resources of a virtual machine. Different methods of energy saving in data centers can be divided to three general categories: 1) energy saving methods based on balancing the loads of resources; 2) scheduling by the help of hardware facilities; 3) reducing the consumption of energy through being aware of thermal characteristics of the environment.*

*One of the most important challenges in terms of cloud computing is to maintain an optimized load balance in cloud environment on servers because this issue is critical for energy consumption. Since load balance algorithms depend on the current behavior of the system, they act dynamically. Important factors which should be taken into account while developing such types of algorithms include load estimation, load ratio, adaptability of different systems, performance of the system, interaction between nods, scheduling, nature of the transferred service, selecting nodes and etc.*

*By taking a detailed look on previous methods and challenges of energy saving in cloud networks, we provide a hybrid method which enables us to save energy through finding a suitable configuration for placement of virtual machines and through being aware of special features of virtual environments for scheduling and balancing their dynamic loads by the help of live migration method.*

**Keywords:** Cloud Networks, Virtual Machines, Energy Saving, Scheduling, Load Balance, Live Migration, Cloudsim, DVFS

**Received:** 1 April 2019, Revised 9 June 2019, Accepted 8 July 2019

**DOI:** 10.6025/jnt/2019/10/3/92-103

## 1. Introduction

Providers of cloud services are motivated towards cloud computing because of the benefits they gain by providing such services; also customers and consumers of cloud services including organizations and economic firms are motivated towards cloud computing because of savings in costs related to maintenance of infrastructures of these services. Still, since the applications which the users run on these cloud services might be extremely sensitive, they need a guarantee from providers of these services. Such a guarantee is usually supplied through service level agreement (SLA). A key technology used in cloud computing is virtualization. This technology gives possibility to separation of virtualized machines in a way that each of these virtualized machines will then be able to be configured on a number of hosts according to the applications of the user [1]. However, still the specific features and attributes of the host on which the virtual machine is running are concealed from the user. Therefore, virtualized machines can be transferred to other hosts without interrupting services in order to increase the efficiency of applications run by users. It is also possible to dynamically alter the amount of resources allocated to a certain customer [8]. Nowadays, modern CPUs can run in different speeds.This is done through techniques including dynamic frequency scaling (DFS); Dynamic voltage scaling (DVS) or combined dynamic frequency and voltage scaling (DFVS). These techniques lead to formation of a non-linear power-frequency relation. According to [9], at each level of efficiency, the power-frequency relation can be approximated through a second-degree model. In cloud computing, for the amount of usage and services provided for users, there are costs that need to be paid. In addition, the work space is transferred to cloud servers through the personal computers of users [2, 3]. Nevertheless, users no more require high-end computer systems because the entire computing and storing processes are performed by well-equipped and advanced servers [12]. Currently, there are three different methods for providing cloud computing services: 1-Software as a service, 2-platform as a service, and 3-infrastructure as a service. With respect to their special needs and requirements, organizations and firms can select and make use of a single or a series of service strategies [4-7]. Service level agreement is an important aspect of cloud computing and it is also considered as a contract which includes the function of agreement regarding the service between the customer and the provider. However, nowadays with coming into play the large cloud service providers, most service level agreements are pre-determined and pre-standardized unless the customer is a major consumer of the services. The process of scheduling for creation of load balance and energy saving in cloud services depend on response-time of the machine. This load can be the load on the CPU, the load on Memory or the delay in the network. Whenever there is excessive load on the cloud network, load balance can inhibit problems in providing services for the consumers [8, 10].

A major challenge in terms of cloud computing is to maintain the load balance in the environment of the cloud on servers because this factor is highly effective on energy consumption. Investigating different ways of load balance with the objective of reduction of response-time in cloud environments is considered the most important counts of service agreement as well as a major anticipator of customer satisfaction [10-12]. Load balance is the process of distributing the load between nodes spread out through the system. In this way we can optimize the efficiency of resources, amount of consumed energy and response time. On the other hand, the situations in which some nodes are under heavy loads and some others are simultaneously almost free, are avoided. Load balancing is a method which facilitates maximum passage and minimum response time in networks and resources. By balancing the traffic between servers, data packets are sent and received with no major delays [14]. There are several different load balance algorithms which include water drop exploration algorithm, active clustering algorithm and algorithms of Max-Max; Min-Min and OLB+LBMM. But cloud computing covers different areas and therefore, we need methods which are suitable for most environments and reduce the costs simultaneously [12-14]. Since load balance algorithms depend on the current behavior of the system, they act dynamically. Important factors which should be considered while developing such algorithms include load estimation, load ratio, adaptability of different systems, efficiency of the system, interaction between nodes, scheduling, nature of the task that is being transferred, selection of nodes and etc. [9, 11].
In figure1, the following are assumed:

1- There are several applications being ran in a virtual machine and each of them require a specific amount of load,

2- Each host includes n virtual machines,

3- A component of the system which was correct at $T = 0$ and will remain that way until it encounters an error,

4- Failures are permanent and failures of virtual machines are independent of each other,

5- The confidence of entire similar virtual machines is similar.

With respect to the aforementioned hypotheses, when a given amount of work is addressed to a machine of a cluster, if the accessible resources are used properly, then the load can be executed effectively. Therefore, there should be a mechanism for selecting those machines which have such resources. As you can see in figure 1, users who are connected to a cloud network can make use of the facilities of the network by sending requests to the cloud network. Since multiple requests from users forces the network to displace virtual machines, when the users of a network exceed a certain number, then the issue of load on cloud clusters turns into a serious debate [9].

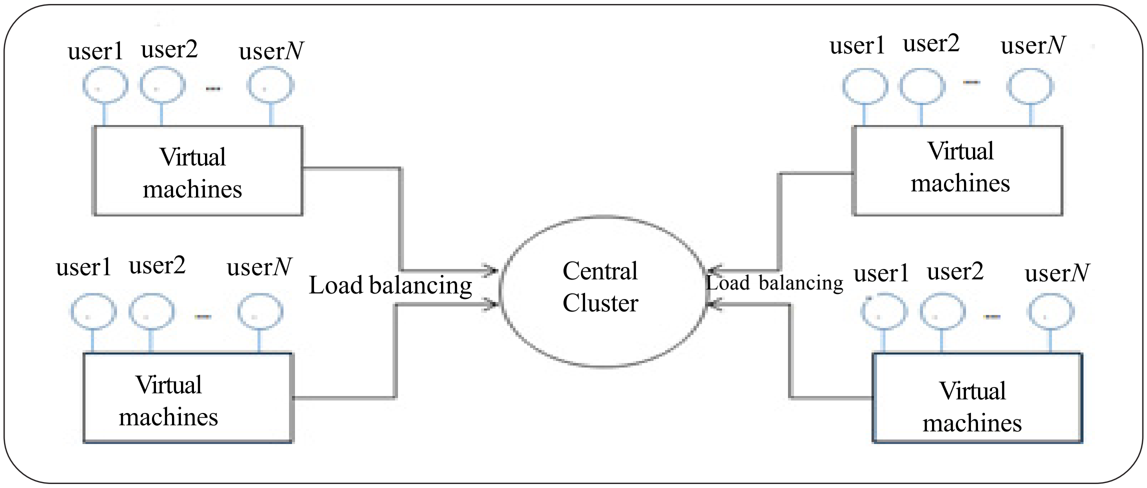

Figure 1. Addressing the Applied Load to Central Cluster

Virtualization technologies provide the processes and data of applications with the ability for being executed within the virtual component (virtual machines and virtual disks) without being dependent on physical resources. In other words, they make applications independent of machines. Virtualization may take place in several layers or levels in order to create a modern computer system. A level of virtualization which is applied to cloud computing data center is a systematic virtualization which enables virtual machines with different operating systems to be executed simultaneously in a real machine. The virtualization software which is called the virtual machine monitor (VMM) is tasked with management and controlling of consumption of hardware resources by each virtual machine. Xen [15] and VMware [16] are examples of virtualization methods which are applied on systems. In this paper, we are trying to suggest a method for the cloud computing data center which enables it to balance the distribution of load between hosts as well as reducing energy consumption. Load balancing is referred to a process of re-allocation of loads for single hosts within a data center in a way that efficiency of data centers are altered and response times are improved. Additionally, situations in which some hosts are under heavy loads while others are almost free or under small loads are avoided [9].

## 2. Objectives and Innovations of the Research

### 2.1 Objectives of this Study include

1- Reduction of consumption of energy in the infrastructure of virtualized data center,

2- Improvement of confidence and tolerance for failures in execution of applications,

3- Avoiding hotspots within the data center and balanced distribution of loads between the existing resources,

4- Avoiding the situation of thrashing among hosts of the data center,

5- Efficiency in all domains in order to have a cloud data center with high efficiency in terms of servicing.

### 2.2 Innovations of this Study Include

1- Developing a dynamic algorithm based on load balance and hardware techniques for optimization of energy consumption with considering for scheduling,

2- Monitoring the exploitation amount of service providing hosts,

3- Making use of the technology of live migration, 4- Lack of physical dependence on virtual machines for running the services in a place with a more inexpensive energy with respect to the proposed algorithm regarding navigation,

5- Variability of the scale of using processing resources according to demands of exploiters.

## 3. Related Works

Reducing energy consumption is feasible in static and dynamic views. In terms of the static method and from the perspective of hardware, energy management includes optimization methods [19] which are used at the time of designing in different levels including circuits and system architecture. In terms of optimization at circuit level, by designing complex gates and altering the size of transistors, the developers focus on reduction of energy consumption of switching in combined and sequential circuits. The purpose of optimization in logic level is to reduce the energy consumption of switching in combined and sequential circuits. Optimization methods at the architecture level include analysis of system design and application of energy consumption optimization techniques in the architecture of the system. Along the hardware optimization methods, the methods of management and allocation of resources to applications and optimization of execution of running applications are important in terms of management of energy consumption. In spite of complete and optimal design of the hardware, weak development of applications can led to reduction of efficiency and wastage energy.

Software methods of static energy consumption management include making use of parallel programming techniques, multi-thread algorithms and reduction of the size of programs which are all applied during the development of applications. Among these aforementioned software methods, some of static energy consumption management methods are unsuitable as a result of their static nature in data centers whose statuses usually change.

In addition, dynamic techniques of energy management can be divided into software and hardware as well. As a matter of fact, there lies difference between different hardware techniques of power management according to different hardware components. However, these methods are usually divided into two categories of dynamic scalability of efficiency including dynamic voltage scalability or complete turning off of components when the system is on stand-by. Software techniques of dynamic power management include mechanisms and policies which make use of knowledge of the current status of the system in order to accommodate the hardware with the current load of the system. These software techniques, with respect to their policies, make use hardware methods of dynamic power management.

This section divides the affairs related to applications of the technology of virtualization into three categories which include scheduling according to efficiency and load balance, making use of hardware facilities and awareness about thermal features of the environment. This categorization is shown in the following image and the tasks related to each of these three approaches are further explained in the text.

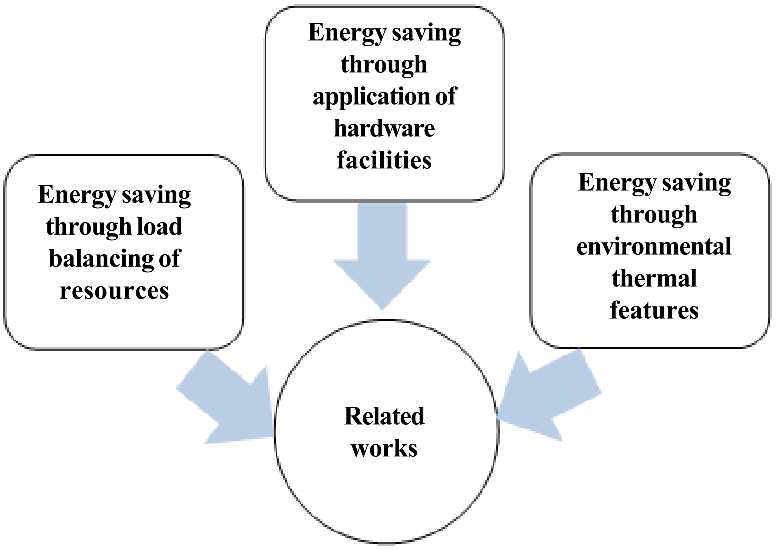
Energy saving through application of hardware facilities
Energy saving through load balancing of resources
Energy saving through environmental thermal features
Related works

Figure 2. Categorization of Related Works Regarding Energy Saving Through Virtualization

## 4. Energy Saving through Load Balancing of Resources

The RR algorithm is a method of forming load balance with the approach of optimization of energy consumption. This algorithm distributes tasks between the entire auxiliary processors. The entire tasks are allocated to auxiliary processes according to a RR order. This means that a selected processor is operated in an order and if the last processor is tasked, the loop will return to the first processor. Selection of processors takes place independent of other processors and the main advantage of the RR algorithm is that it doesn't require inter-process relations. RR is generally used in web-servers which generally have HTTP requests with similar natures. This defined algorithm is inconsistence with the principle of service level agreement [28].

Another way of achieving load balance is the application of random algorithm. This algorithm used random processor numbers for selection of auxiliary processors. Auxiliary processors in this fashion are produced with random numbers and are selected according to statistical distributions. This algorithm can have the very best performance for an application with a specific objective and this also is inconsistence with the principle of service level agreement [29].

Another algorithm for formation of load balance and reduction of energy consumption is the central manager model. In this model, the main processor selects a subsidiary or auxiliary processor for allocation of tasks in each step. This selected auxiliary processor, is a processor with minimum amount of load. The main processor is able to collect information from all other auxiliary processors; therefore the choice may be made according to this algorithm. The load manager makes load balance related decisions according to load information of system and allows the system to make the best choice while the process is running. This algorithm is expected to have a better performance than parallel applications, especially when different dynamic activities ate formed by different hosts [30].

Another algorithm which is investigated is the dynamic algorithm of load balance based on scheduling. In dynamic load balance algorithms, tasks are distributed among processors during execution. The main manager allocates the new process to an auxiliary processor according to new collected information. In a distributed system, dynamic load balance can be done in two ways: distributed and non-distributed (centralized). In distributed type, load balance algorithms are executed by the entire machines of the system and the responsibility of maintaining the load balance is shared between the entire engaged machines. The relation between the machines for obtaining load balance can be in either ways of collaborated or non-collaborated. For a general balance through the system, processors work together [13, 17].

On the other hand, in non-collaborated algorithms, each machine depends on its own resources' scheduling in which decisions are made independent of the rest of the system. On this basis, the machine may schedule or allocate tasks according to local performance. In collaborated algorithms, the tasks of energy consumption monitoring and maximization of profits are considered as ROI [14].

Dynamic load balance algorithms can create more messages compared to centralized algorithms; because each machine within the system needs to make contact with all other machines within the same system. An advantage of this is that if one or some machines of the system are failed, the total procedure of load balancing will not terminate and it can only affect the performance of system. However, in these algorithms the heuristic learning functions are used which is inconsistent with the service level agreement [14, 17, 18].

Currently, several methods and experiences are applied for optimization of efficiency of energy consumption. These methods include: optimization of algorithms of applications [31] and making use of hardware with low energy consumption [32].

By implementing the virtualization technology, we can make several virtual machines on a physical server. As a result, the amount of required hardware is reduces and also efficiency of resources increases. Cloud computing uses the virtualization technology for providing resources according to costumer demands. Making use of this technology in cloud computing naturally leads to optimized energy consumption which may be based on the following reasons:

- Increased efficiency of resources [36]

- Lack of physical and location dependence; virtual machines could be moved and executed in places where energy is cheaper [14]

- Alteration of the scale of implementation of resources; implementation of resources can be varied according to recent requirements[7]

In [33], reduced energy consumption is investigated in a virtualized cloud network data center. In this article, the authors suggest an efficient resource management system for cloud network data centers. This system reduces operational costs and improves service quality. This system's energy consumption saving is done through uninterrupted monitoring of virtual machines according to the current efficiency of resources. The main instrument that is applied here is making use of live migration for virtualized machines of the cloud network. Live migration is a feature for transferring virtual machines between different physical servers which have low overload and provide the cloud provider with the flexibility in which virtual machines can allocate resources dynamically and according to current and recent requirements.

In [34] an architecture is suggested which is aimed at management of energy in virtualized data centers. In this architecture, resources are managed under local and global levels. In terms of the local level, the system makes use of power management strategies embedded in guest operating systems. In other words, the same strategies which would have had been used if the system was running on a physical machine. Also in terms of the global level, for the purpose of allocation of virtualized machines, those general policies are used which make use of live migration technology.

## 5. Scheduling through Hardware Facilities

Dynamic voltage and frequency scaling is a hardware technique for dynamic management of power consumption which can be seen in some hardware pieces such as processors and CPUs. By making use of this capability, at the time of reduced loads, processors can enter operational statuses with lower power consumption. Therefore, the processing power of processors can be scaled according to current requirements of the system without having a flaw in efficiency of programs [21]. Anyway, in this technique processors are able to choose their operational voltage and frequency from a restricted and limited set of values [22]. In addition, by the use of this approach, energy consumption can only be reduced up to 10 to 20 percent [23]. As a result of existing limitations, many related tasks for reduction of energy consumptions make use of the approach of centralization of loads on a fewer number of brokers and turning off the idle brokers or putting them in sleep modes or hibernate modes or even a combination of the centralization method and dynamic frequency and voltage scaling method [20].

The problem of management of power consumption in brokers with web based applications and under agreements of service level is discussed in [24]. The main goal of the authors of this paper [24] was to reduce operational costs and reduction of downtimes resulting from excessive warming of physical brokers. For this purpose, they have made use of the approach of turning operational nodes on and off in addition to dynamic voltage and frequency scaling method for management of energy consumption. They evaluated their suggested policies through a number of simulations. Results of their simulations indicated that the policy of dynamic frequency and voltage scaling can reduce energy consumption up to 29 percent. The mechanism of turning operational nodes on and off can also reduce energy consumption up to 42 percent. This is while combination of these two methods can only save energy up to18 percent. Anyway, their research has failed to consider for many effective real parameters. For example, the required amount of time for turning brokers back on is not considered. In this approach, it is assumed that only one application is executed on a cluster at a time. They have neglected the changes of tasks' loads and this may lead to making inappropriate decisions. In addition, this research only investigates CPUs and neglects other units.

## 6. Reduction of Energy Consumption through Awareness about Thermal Features of the Environment

In another approach of reduction of energy consumption in virtual environments, the thermal features of the environment have been used. One of the most well-known efforts in this field belongs to Beloglazov and Buyya [25] in university of Melbourne. They have deigned a resource manager which dynamically reads thermal features of each node and tries to schedule virtual machines in a way that heavier loads are oriented towards machines which are cooler. Results of implementing this design on a simulated virtual context indicated that making use of this approach can significantly reduce the consumption of energy in virtual environments.

In two other researches in this field which are performed by Tang et al. [26,27], the emphasis has been put on the point that calculation nodes with high temperatures are not suitable choices for allocation of tasks and activities. On this basis, different methods have been proposed for scheduling activities according to thermal characteristics of working nods. Results of previous evaluations indicate that scheduling tasks with consideration for the temperature of the physical machine can have significant positive effects on optimization and improvement of accomplishment of tasks.

## 7. Comparison of Related Works and Challenges

With daily increase of demands for applications and increased complexity of calculation environments, distributional systems and data centers are beginning to consume more and more amounts of energy. Increased energy consumption in distributional systems such as cloud computing models has led to formation of several challenges. These challenges include emission of a large amount of carbon dioxide in environment, increased expenses, reduced efficiency and reduced confidence in systems. Therefore, nowadays, designing computer systems has moved towards increased efficiency and reduced energy consumption. With the recent advances in hardware technologies including development of processors with low energy consumption, the amount of consumed energy in data centers has significantly decreased. However, still the efficiency of energy consumption in data centers not only depends on the efficiency of physical hardware, but also lies in methods of management of resources and efficiency of programs being currently executed on physical systems. Nowadays, a large amount of research works have been done in the context of management of physical resources aimed at increased efficiency. With respect to the approach of the present paper, this section has mentioned a number of previous works regarding optimization of energy consumption and efficiency of programs and applications. As it can be seen in this section, different methods of reduction of energy consumption in data centers can be divided into three general categories including: 1) methods of reduction of energy consumption through balancing resources' loads; 2)scheduling through hardware facilities and 3) Reduction of energy consumption through awareness about thermal features of the environment.

By studying these methods, the challenges of reduction of energy consumption in cloud networks will be as follows:

- Finding a suitable configuration for placement of virtual machines,

- Being aware of special traits of virtual environments for dynamic scheduling of virtual machines.

## 8. The Proposed Hybrid Algorithm

In this section, we will propose an algorithm for load balancing. Figure 3, shows the flowchart of function of the proposed algorithm in the migration section. As you can see, the cloud network awaits receiving orders from users who are connected to the network. The central cluster is tasked with total management of the cloud network. This cluster investigates other connected clusters and uninterruptedly analyzes different types of routes, bandwidths, storage of attributes of cloud servers and etc. Tables shown in the third section of the flowchart show the entire routes according to step, IPs and address for accessing the cloud network. In order to signify the number of steps in a cloud network, navigation algorithms are used. On the other hand, in order to calculate the best route, the bandwidth of channel is also taken into account. Next, the algorithm elaborates on traffic of the cloud network. In this phase, it is checked whether the network is busy or not. If the answer was no, the network will keep its performance as before. Otherwise a more powerful is determined which is able to handle more user requests. If such a server was not proposed by the main cluster, the flowchart performs a migration to a new server.

By the use of this technology we are able to transfer a virtual machine from one server to another without interrupting the service. As a result, by the use of this technology and by transferring a specific amount of load to other servers we can perform the task of resource management in terms of balanced distribution of load between resources as well as optimization of energy consumption.

Nowadays, modern CPUs are able to take different speeds. This is done through making use of techniques such as DFS, DVS and DFVS. These techniques yield a non-linear power-frequency relation. According to [35] at each level of efficiency, the relation between power and frequency could be well approximated by a second type model. It means that we can write:

$$P(f) = P_{min} + \alpha (f - f_{min})^2 \tag{1}$$

In the upper relation, Power is estimated in watts and Frequency is estimated in GHz and also $\alpha$ is a coefficient based on $W/(GH)^2$. In our model, we have assumed that the CPUs of the hosts obey such a non-linear relation within the data center. The CPUs are run in different and limited values of frequency in a $[f_{min} - f_{max}]$ range. This leads to a tradeoff between efficiency and power related costs. In our model, all servers and hosts obey this relation. Also $\alpha$ is based on $W/(GH)^2$ because the highest amount of energy is consumed when CPU is run with the highest power. With respect to this content, $\alpha$ can be calculated as follows:

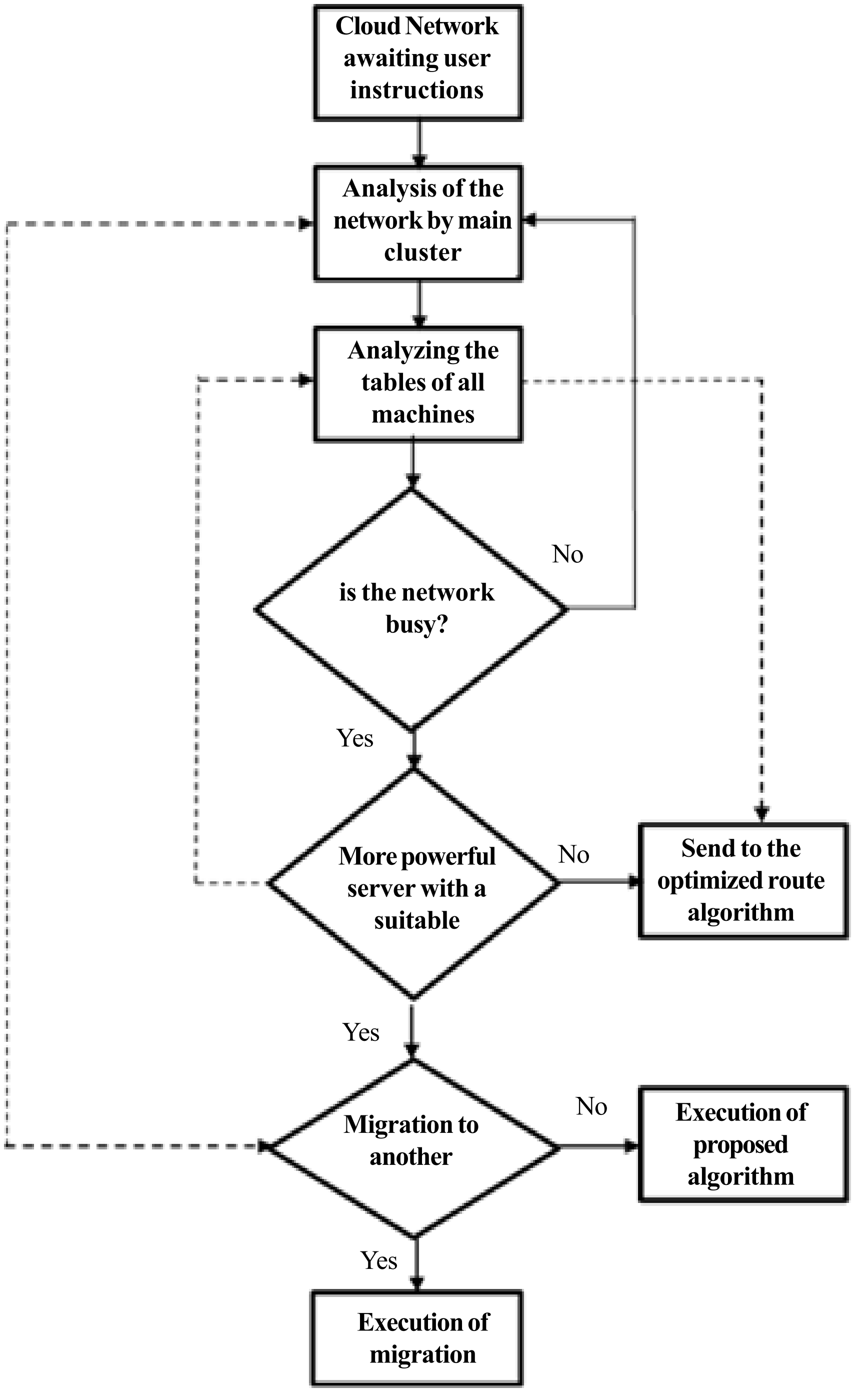
Cloud Network awaiting user instructions
Analysis of the network by main cluster
Analyzing the tables of all machines
is the network busy?
No
Yes
More powerful server with a suitable
No
Send to the optimized route algorithm
Yes
Migration to another
No
Execution of proposed algorithm
Yes
Execution of migration

Figure 3. Flowchart of Function of Proposed Algorithm in the Migration Section

$$\alpha = \frac{\rho_{max} - \rho_{min}}{(f_{max} - f_{min})^2} \qquad (2)$$

In addition, servers within the data center can go to stand-by if the amount of load on the data center was low. By taking a look at the upper relation, we can see that even if there were no loads on a server, still a minimum amount of power is required to keep the server in active mode. Therefore, putting the host into sleep or hibernate can have a significant effect on reduction of consumed energy.

We have used a relation titled as Imbalance score for the load balance algorithm and calculation of the overload. This relation is as follows:

$$IBscore\,(\mu\,,\,T_{upper}) = e^{\,(\mu - T_{upper}\,)} \qquad (3)$$

$U$ is the utilization level of each of the resources of the server. For example utilization level of the CPU, RAM and etc. $T_{upper}$ is the upper threshold of server utilization which is used for initialization. The purpose of using this relation is to show the amount of load on each server and also the amount of each overload of server. If the utilization amount of the server is smaller than the threshold, then the value of $T_{upper}$ will be negative. As a result, as the amount of load on the server is reduced, the amount of the IB score value is reduced as well.

In addition, another concept titled as attractiveness is defined as follows. It's aimed at determining the attractiveness of a server for being selected as the destination of a virtual machine.

$$Attractiveness_{host} = IBscore_{host} \times LoadFraction_{vm,\ host} \qquad (4)$$

The proposed algorithm showed in figure 4 acts dynamically. It means that for a balanced distribution of loads on different resources, it only considers their current status. The purpose of using this algorithm is to avoid hotspots in servers within the data center because if the host is overloaded, the system is prone to reduction of efficiency and vulnerability. Though, this algorithm tries to keep the amount of load on each server lower than the threshold. Still, in special situations in which the systems in under heavy loads of work, we need to allocate the entire existing infrastructure resources to virtualized machines and makes as many virtual machines as possible on the infrastructure of our cloud computing network.Also in this situation, a large amount of load may be allocated to each virtualized machine. In this case, this algorithm is able to impose an overload on the model.

In our model, for the amount of costumer requests, virtual machines are able to be formed on the infrastructure of the cloud network. After that costumer requests make their to the data center, after allocation of virtualized machines according to specific policies of allocation, costumers' applications start to operate. We also investigate the efficiency of hosts in different time intervals. If the efficiency of a server exceeds its pre-defined threshold, a number of virtual machines are selected and then are transferred to a new host.

```
Load-Balance Algorithm
        For every power-on host
                If host-utilization > higher threshold-utilization then
                Add some VMs of host to migrate-set
                        Until host-utilization <=higher-Threshold-utilization
        For every VMs in migrate-set
                If(find target host) then
                Migrate VMs.
```

Figure 4. Proposed Load Balance Algorithm

As it was mentioned, machines are transferred lively. After selection of certain machines, the algorithm selects a number of hosts as destinations of virtualized machines. Among these servers, according to policies, the one is selected which has a smaller attractiveness function value. This algorithm also investigates if the selected overload of overload threshold of server is not exceeded after transferring the virtual machines. If so, another host or server is selected. In fact a host is selected which has a small value of attractiveness function and will not be overloaded after transfer. The former inhibits the situation of thrashing. After selection of the destination host, the process of migration starts and after this process, the execution of the programs continues by the new host and therefore resources of the source host are freed.

### 9. Evaluation of the Proposed Method

In order to evaluate the proposed algorithm, we have made use of the Cloudsim software. In fact we have expanded the core of this simulator for modeling the algorithm of load balance as well as modeling the effects of the DVFS technology in consumption of energy and calculation of added overload. In order to show the amount of overload, the service level agreement value is calculated in each time interval. Our simulated cloud computing data center is consisted of 100 servers which are able to be virtualized. In fact, it is been hypothesized that virtualization software such as Xen are installed on them. Each of the hosts is modeled as a single core host each of which has a power of 1000, 2000 or 3000 MIPS. In addition each server has 8 GB of RAM and 1 TB of H.D.D. we have created 220 virtualized machines on this center. One application is run in each machine and each application is also consisted of 1.500.000 instructions.

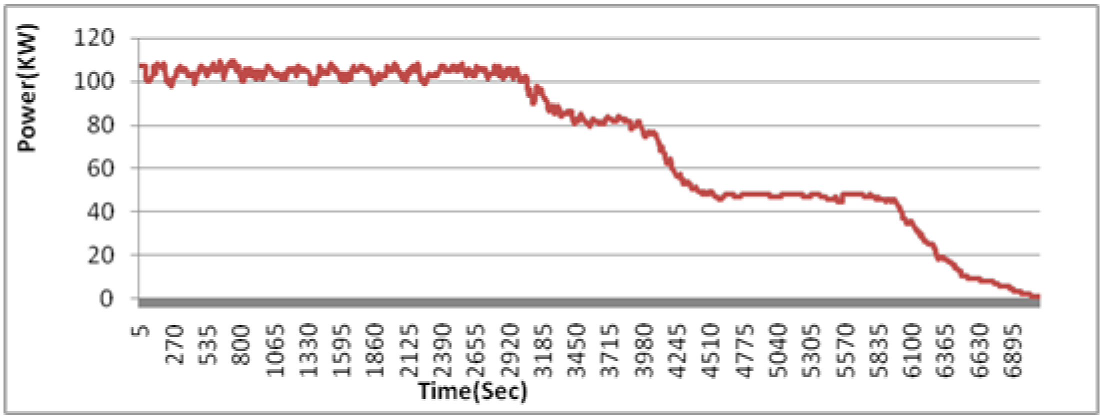

Figure 5. Energy Consumption Diagram before the Proposed Method

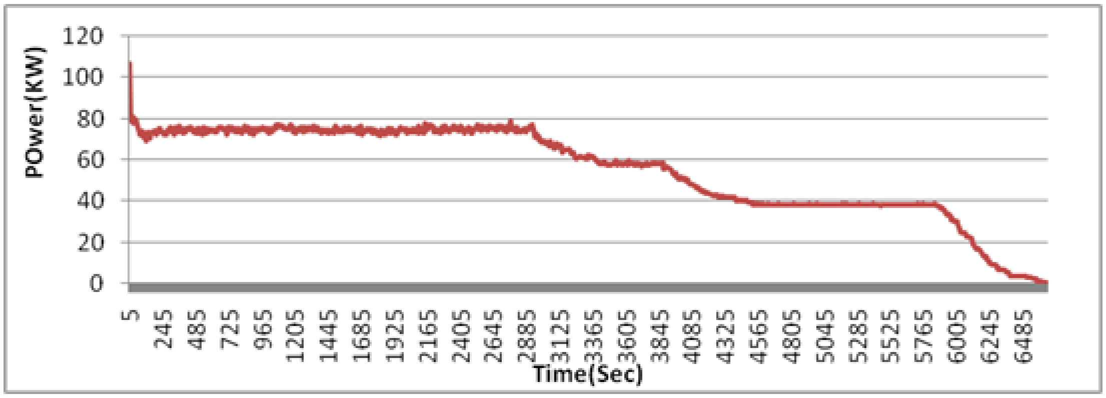

Figure 6. Energy Consumption Diagram after the Proposed Method

As we have also mentioned before, the purpose of implementing the algorithm of load balance is to have a balanced distribution of tasks between hosts which avoids accumulation of load on one host and also avoid formation of hotspots. However, implementing this algorithm violates the service level agreement. Figure 5 shows the output of simulation before using the proposed method for investigating violation of service level agreement. Figure 6 shows the amount of violation of service level agreement with a utilization threshold of 80 percent on each server during the execution of programs and applications. Since energy consumption has a large cost for providers of cloud based services, in our simulation we have used the DVFS method in conjunction with putting idle servers to standby.

## 9. Conclusions

Cloud computing is a novice concept and with respect to its advantages for both the provider and user, it is rapidly growing. However, in order to be able to realize the promises of the cloud computing, there are things and issues that require further investigation. In this paper, we have proposed a method for reduction of consumed energy based on load balance and hardware techniques in cloud computing data centers. Making use of the proposed algorithm in this research avoids formation of hotspots within the data center and formation of a balanced load distribution between the existing resources as well as reduction of consumed energy. For execution of each requested application, instead of running the request in a single virtualized machine, we run that request in three separate virtualized machines each of which are created on different servers. Also these servers may be selected from different clusters.

**References**

[1] Wang, L., Ranjan, R., Chen, J., Benatallah, B. (2011). Cloud Computing: Methodology, Systems, and Applications. CRC Press, 2011.

[2] Chandrasekaran, K . (2014). Essentials of Cloud Computing. Taylor & Francis, 2014.

[3] Schaefer, D. (2014). Cloud-Based Design and Manufacturing (CBDM): A Service-Oriented Product Development Paradigm for the 21st Century. Springer International Publishing, 2014.

[4] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., Brandic, I. (2009).Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, *Futur. Gener. Comput. Syst*., 25 (6) 17, 2009.

[5] Rittinghouse, J. W., Ransome, J. F. (2009). Cloud Computing: Implementation, Management, and Security. CRC Press.

[6] Keahey, K., Tsugawa, M., Matsunaga, A., and Fortes, J. (2009). Sky computing, *IEEE Internet Comput.*, 13, no. October, p. 43–51.

[7] Khalil, I., Khreishah, A., Azeem, M. (2014). Cloud Computing Security: A Survey, *Computers*, 3 (1) 1–35, Feb.

[8] Beloglazov, A., Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers, *Concurr. Comput. Pract. Exp*., V. 24, p. 1397–1420.

[9] Ding, Y., Qin, X., Liu, L., and Wang, T. (2015). Energy efficient scheduling of virtual machines in cloud with deadline constraint, *Futur. Gener. Comput. Syst*., vol. 50, p. 62–74, 2015.

[10] Aslazandeh, S., Chaczko, Z., Chiu, C. (2014). Cloud computing #x2014; The effect of generalized spring tensor algorithm on load balancing, *In:* Computer Aided System Engineering (APCASE), 2014 Asia-Pacific Conference on, 2014, p. 5–8.

[11] Chen, L., Shen, H., Sapra, K. (2014). RIAL: Resource Intensity Aware Load balancing in clouds, *In:* INFOCOM, 2014 *Proceedings IEEE*, 2014, p 1294–1302.

[12] Domanal, S. G., and Reddy, G. R. M. (2014). Optimal load balancing in cloud computing by efficient utilization of virtual machines, *In:* Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, 2014, p1–4.

[13] Mashaly, M., and Kuhn, P. J. (2012). Load balancing in cloud-based content delivery networks using adaptive server activation/deactivation, *In:* Engineering and Technology (ICET), 2012 International Conference on, 2012, p. 1–6.

[14] Ajit, M., Vidya, G. (2013). VM level load balancing in cloud environment, *In:* Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, 2013, p. 1–5.

[15] Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., Warfield, A. (2003). Xen and the art of virtualization, *ACM SIGOPS Oper. Syst. Rev*., vol. 37, p. 164, 2003.

[16] Calheiros, R. N. (2009). Automated Emulation of Distributed Systems through System Management and Virtualization, PhD Thesis, Pontif. Cathol. Univ. Rio Gd. do Sul Porto Alegre, Brazil, 2009.

[17] Shaw, S. B., Singh, A. K. (2014). A survey on scheduling and load balancing techniques in cloud computing environment, *In:* Computer and Communication Technology (ICCCT), *In:* 2014 International Conference on, 2014, p. 87–95.

[18] Abrahamsson, P., Helmer, S., Phaphoom, N., Nicolodi, L., Preda, N., Miori, L., Angriman, M., Rikkila, J., Wang, X., Hamily, K., Bugoloni, S. (2013).Affordable and Energy-Efficient Cloud Computing Clusters: The Bolzano Raspberry Pi Cloud Cluster Experiment," Cloud Computing Technology and Science (CloudCom), *In:* 2013 IEEE 5th International Conference on, vol. 2. p 170–175, 2013.

[19] Devadas, S., Malik, S. (1995). A survey of optimization techniques targeting lowpower VLSI circuits, *In:* 32nd Annual ACM/ IEEE Design AutomationConference, p. 242-247, Spain.

[20] Petrucci, V., Loques, O., Mosse, D. (2010).Dynamic optimization of power andperformance for virtualized server clusters, *In:* Proceedings of the 2010 ACMSymposium on Applied Computing, p. 263-264, USA, 2010.

[21] von Laszewski, G., Wang, L., Younge, A. J., He, X. (2009). Power-awarescheduling of virtual machines in dvfs-enabled clusters, *In:* ClusterComputing and Workshops, IEEE International Conference on ClusterComputing, p. 1-10, 2009.

[22] Rotem, E., Naveh, A., Moffie, M., Mendelson, A. (2004). Analysis of thermalmonitor features of the Intel Pentium processor, TACS Workshop,Hongkong,.

[23] Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., and Jiang, G. (2009). Powerand performance management of virtualized computing environments vialookahead control, *Cluster computing*, vol. 12, p. 1-15, 2009.

[24] Elnozahy, E., Kistler, M., Rajamony, R. (2003). Energy-efficient server clusters, *In:* Workshop on Power-Aware Computer Systems, p. 179-197, Italy, 2003.

[25] Beloglazov, A., Buyya, R. (2010). Energy efficient resource management invirtualized cloud data centers, IEEE/ACM International Conference onCluster, *Cloud and Grid Computing*, p. 826-831, Sydney, 2010.

[26] Tang, Q., Gupta, S., and Varsamopoulos, G. (2007). Thermal-aware task schedulingfor data centers through minimizing heat recirculation, *IEEE InternationalConference on Cluster Computing*, p. 129-138, USA, 2007.

[27] Tang, Q.,Gupta, S. K. S., Varsamopoulos, G. (2008). Energy-efficient thermalawaretask scheduling for homogeneous high-performance computing datacenters: A cyber-physical approach, *Parallel and Distributed Systems, IEEETransactions on*, vol. 19, p. 1458-1472.

[28] Banerjee, S., Adhikari, M., Biswas, U. (2014). Development of a smart job allocation model for a Cloud Service Provider, *In:* Business and Information Management (ICBIM), *In:* 2014 2nd International Conference on, 2014, p. 114–119.

[29] Chang, F., Viswanathan, R., Wood, T. L. (2012). Placement in Clouds for Application-Level Latency Requirements, *In:* Cloud Computing (CLOUD), *In:* 2012 IEEE 5th International Conference on, 2012, p. 327–335.

[30] Alakeel, A. M. (2010). A Guide to Dynamic Load Balancing in Distributed Computer Systems, *Int. J. Comput. Sci. Inf. Secur*., 10 ( 6) 153–160, 2010.

[31] Yan, S., Wang, X., Razo, M., Tacca, M., and Fumagalli, A. (2014). Data center selection: A probability based approach, Transparent Optical Networks (ICTON), *In:*2014 16th International Conference on. p. 1–5, 2014.

[32] Wang, H., Bergman, K. (2012). Optically interconnected data center architecture for bandwidth intensive energy efficient networking, Transparent Optical Networks (ICTON), 2012 14th International Conference on. p. 1–4, 2012.

[33] Beloglazov, A., and Buyya, R. (2010). Energy Efficient Resource Management in Virtualized Cloud Data Centers, *Cluster, Cloud and Grid Computing* (CCGrid), *In:* 2010 10th IEEE/ACM International Conference on. p. 826–831, 2010.

[34] Ghosh, T. K., Goswami, R., Bera S., and Barman, S.(2012). Load balanced static grid scheduling using Max-Min heuristic," *In:*Parallel Distributed and Grid Computing (PDGC), 2012 2nd IEEE International Conference on, 2012, p. 419–423.

[35] Urgaonkar, R., Kozat, U. C., Igarashi, K., Neely, M. J. (2010). Dynamic resource allocation and power management in virtualized data centers, *Netw. Oper. Manag. Symp*. NOMS 2010 *IEEE*, no. Vm, p. 479–486, 2010.

[36] Adhikari, J., Patil, S. (2013). Double threshold energy aware load balancing in cloud computing, in Computing, *In: Communications and Networking Technologies* (ICCCNT),2013 Fourth International Conference on, 2013, p 1–6.

## Authors

**Mehran Tarahomi** received his Bachelor in Computer Engineering and Master of Software Engineering at during the years 2002 and 2005 respectively from Central Branch of Azad University in Tehran and Shiraz University. He is currently PhD candidate in Software Engineering at Kish International Campus, Sharif University of Technology. He has many papers published in national and international conferences and journals. His research interests include software engineering and energy-aware techniques specifically in the area of cloud computing. His email address is: tarahomi@ce.sharif.edu

**Mohammad Izadi** is an Assistant Professor and the vice-chairman of educational affairs in the department of computer engineering at Sharif University of Technology, Tehran, Iran. He received a PhD degree incomputer science from Leiden University, the Netherlands and another PhD Degree in software engineering from Sharif University of Technology, respectively in 2011 and 2008. He also has a master degree in Philosophyof science and another one in computer engineering and has BSc degree allfrom Sharif University of Technology. His fields of research include distributed models and algorithms, game theory, verification of component based computing systems, computational complexity, and mathematical logic. His email address is: Izadi@sharif.edu