Reduction of Power Consumption in Cloud Data Centers via Dynamic Migration of Virtual Machines

Mona Arjmandi, Nik Mohammad Balouchzahi, Kaamran Raahemifar, Mahmood Fathy, and Ahmad Akbari

Abstract—**Cloud computing is the latest answer of technology to meet the computational requirements of users. The notable point in complicated computational works is energy consumption. The integration is one of the elements in the cloud system which can reduce the energy consumption and coordinate the software products. In this article, some solutions have been considered for determining the upper bound threshold of utilization for doing migration in order to reduce the power consumption. Also, it has been tried to use a solution to diagnose the overloaded host and eliminate it from the host as it can amend the efficiency as well as improve the host performance. All these will eventually reduce the power consumption. The results of evaluation show that the presented model has better efficiency in reducing the power consumption as well as decreasing the number of the migrations in comparison with the other models.**

Index Terms—**Cloud computing, overloaded host, power consumption, migration.**

I. Introduction

The cloud computing is a computational method which can present the IT's scalability and dynamic capabilities to the users via internet technologies [1]. The cloud computing model, similar to any other technology, has been created according to its own specific advantages while it involves some disadvantages. One of cloud computing disadvantages is producing a huge amount of CO2 due to the presence of very large data centers and high power consumption. The studies [2] regarding power consumption of server farms in 2005 show that the electric consumption by the wide world of servers (including cooling servers and their equipment) cost US $ 7.2. This study also shows that electricity consumption in that year, compared with consumption in 2000, has been doubled. However, the technology improvements have presented suitable hardware [3] including low-power processors, solid state drives and energy efficient monitors. A series of software solutions are optimally improving the energy efficiency. These two guidelines (hardware and software) should be evaluated as competitive supplements in order to incredibly increase their effects on the amount of energy consumption. The software solutions alongside software methods can improve the cloud computing efficiencies from the power consumption point of view. In this article one of the software factors, integration of virtual machines, will be evaluated. This factor reduces the server's power consumption. In this solution the low-load and overload servers will be diagnosed through a proper method and after migration of virtual machines, the low-load servers will be turned off. Moreover, those groups of the virtual machines which belong to the overload servers that ought to migrate are determined via the selective policy of the virtual machines. In the next step the selected machines will be transferred to the appropriate servers which had been determined on the destination host via virtual machines placement policies. The objective of this article is to improve the amount of power consumption in the data centers through presenting optimal solutions in order to determine the overloaded hosts. At the end, the applied method will be evaluated and compared with the other given policies which are related to this subject. The following part of the article will be organized as below. In Section II, the related works will be evaluated by considering the overloaded hosts. In Section III, the given solutions in this article will be discussed. The simulation results and efficiency evaluation regarding the proposed model will be presented in Section IV. We can find the conclusion in Section V.

II. Related Work

As it was mentioned in previous parts, the virtual machines dynamic integration had been divided in four phases. In this section the related works regarding phase 1, the diagnosing algorithm of the overloaded hosts, will be presented as follow:

Buyya and his colleagues [4] had evaluated hosts determination policies as well as selecting the virtual machines in the origin point. The statistical indicators and dispersion have been used in the considered policies in order to choose the overloaded host. Eventually, it has been shown that the effect of LR-MMT policy on number of migrations, energy consumption and SLA fault is higher than the other policies.

After evaluating overloaded host, Buyya [5] and colleagues came to the conclusion that the host's overload affects QoS directly. The reason is that if the source capacity used are fully utilized, most probably the operational programs face resource shortage and performance decline. Here, the Markov model has been used in order to diagnose overload of physical host. Evaluation of the presented mechanisms efficiency has been done according to the Planetlab workload.

Jung and colleagues [6] in their study have assessed the power distribution management in the virtual machines with

Manuscript received September 12, 2014; revised November 17, 2014.

Mona Arjmandi, Nik Mohammad Balouchzahi, Mahmood Fathy, and Ahmad Akbari are with the Iran University of Science and Technology, Tehran, Iran (e-mail: M.arjmandi27@gmail.com, Balouchzahi@iust.ac.ir, mahfathy@iust.ac.ir, Akbarii@iust.ac.ir).

Kaamran Raahemifar is with the Ryerson University, Toronto, Canada (e-mail: Kraahemi@ee.rayerson.ca).

DOI: 10.7763/IJIET.2016.V6.701

respect to the fixed high threshold in order to determine overloaded hosts. The results show that fixed high threshold is not suitable for the systems which have dynamic or unknown workload. The result indicates lack of proper efficiency of presented solution in the dynamic systems.

In Green Cloud project of Buyya [7] and colleagues, they emphasized on optimal energy provision of cloud while the QoS necessities will be provided through SLA definitions and negotiation between providers and customers. This project will solve the energy- efficiency allocation issue of the virtual machines in the cloud data centers. The method will work on operational program services based on QoS necessities for the customers which include expiry date and budget constraints.

Guenter [8] and colleagues applied a dynamic aggregation system of energy-aware virtual machine based on the web operational applications. In that project the response time defines via SLA. They applied a weighted linear regression in order to foreseen the future workload and optimize the allocation of resources which has been carried out before. This regression which has called local regression and had been presented in the last works, will actively implement the resources placement on this algorithm. It will be used as a benchmark in this article.

Wang and colleagues [9] had evaluated the control circles for the allocation resources management based on the response time restraints of the QoS in server and cluster level. If the resource capacity of a server be insufficient for satisfying SLA of operational programs, then a virtual machine would be migrated from the server. All these works are similar to the discovery methods based on threshold which depend on the moment values of the performance characteristics. However, the history revision of the system states, which has been done to evaluate the future behavior of the system and optimize the min time, does not balance the efficiency.

III. Proposed Method

Due to the importance of overloaded hosts' determination in energy consumption of data centers, some authors have given solutions to diagnosing overloaded hosts. A summary of the most important works has been presented in table1. Most of the solution will reduce the power consumption of the data centers though; the other important factors in service quality of the cloud data centers such as SLA and the number of migration did not reduce by reducing power consumption. Therefore, in this article it has been tried to reduce all three parameters including power consumption, SLA fault and the number of migration by suggesting proper solutions for determining the overloaded hosts.

The given method is based on diagnosing the overloaded hosts and elimination of overload from those hosts. The issue consists of determining the overloaded hosts and then selecting the subjected virtual machines which should be migrated from the overloaded hosts to the selected hosts in order to eliminate the overload. The given method will be implemented in the first phase of data centers integration. Integration includes four phases as below:

1) The diagnosing overloaded host policies: all the hosts should be checked and in case of diagnosing any physical host with overload, some of its virtual machines ought to migrate to the other physical hosts. The server efficiency will be improved through that action and due to the better function after migration and reduction in overload; the amount of power consumption will reduce.
2) Diagnosing the low-load hosts: in this condition PABFD algorithm will be used. Hence, at the beginning the low-load hosts will be diagnosed and by transferring all the virtual machines on them, they will go to the sleep state in order to reduce the energy consumption.
3) Determining the virtual machines for migration: whenever a host is overloaded, some of its virtual machines should be migrated in order to reduce its overload. Therefore, some solutions will be applied in this phase to find the most appropriate machine on the host which is the best for migration.
4) Placement of the virtual machines: in case that the host is overloaded some specific number of its virtual machines and whenever the host is low-load all their virtual machines ought to be migrated. For the purpose of migration, the destination hosts should be determined precisely as by transferring the machines on them, they do not go to the overload state. The last phase of this process called placement of the virtual machines.

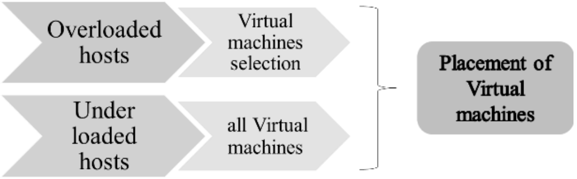

Fig. 1. The integration phases of virtual machines.

This article will discuss the first phase out of four integration phases. The overloaded host selection policies will be according to the high threshold determination. In this state if host productivity be higher than a certain threshold then that host would be overloaded therefore, some of its virtual machines ought to be migrated. Hence, on the works which have been done it was focused on the methods of determining the high productivity threshold. Also, in this article a method will be presented in order to determine the high productivity threshold as the amount of this threshold will cause a reduction in the number of migrations as well as reducing the power consumption of the cloud data centers. The most important policies of the selecting high threshold are: THR, MAD [10], IQR [11], LR [12], and LRR [13]. The overloaded hosts will be determined through the achieved thresholds. Therefore, during the next step it should be clear that which of the virtual machines should be selected from each hosts in order to migrate. The selective virtual machines policies in this article, in order to evaluate the efficiency of the presented method in compare to the other solutions, are: MMT, RS, MC.as it was shown in the previous solutions, the amount of SLA fault will be increased through power consumption reduction. In this article it has been tried to apply a solution that can reduce the number of migrations and SLA faults as well as energy consumption reduction. The median statistic index has been used in this state to determine productivity high threshold in order to diagnose the overloaded hosts. Cloudsim tool has been used to evaluate the efficiency of the applied method. The virtual machines determination policies are 1) MC [7], 2) MMT and 3) RS.

As it was mentioned, the median index will be used in order to determine the productivity high threshold. The median statistic index has been used in the high threshold equation. The aim is to determine the threshold with high accuracy as well as considering all the available values for threshold determination. . Moreover, the correct component function has been used in either sides of the given equation to round off the achieved decimal numbers. This will cause that the obtained numbers for *MAD`* will have got the smaller numbers. Hence, the obtained amount for the productivity high threshold in this state, will have got bigger numbers due to the equation of the *MAD`* high threshold equal with equation (1) .

$$MAD`= [\text{mean } (|x_i - \text{mean } x_j|)] \qquad (1)$$

$$Tu = 1 - s.\ MAD`\ .$$

IV. Evaluation and Simulation Results

The considered system is an infrastructure supplier cloud network which is named as Service (IaaS). The infrastructure includes a cloud data centre in a large scale and n numbers of heterogeneous hosts. Every node has got characteristics which contain processor, amount of memory and network bandwidth. The processor can be multi cores and its efficiency can be evaluated through MIPS. The servers have not got local disk and storage will be done in form of NAS to make the live migration possible. The environment has no knowledge regarding the workload. Several independent users request m numbers of heterogeneous virtual machines (with specific memory, bandwidth and efficiency). In fact those are the independent users who are managing the workload of virtual machines which depend on the combination of several virtual machines on a simple physical node. The combined workload has been consisted of various operational programs and web operational programs which are using the resources simultaneously. The usersdeliver SLAs with specific sending QoS. Suppliers will pay a large sum to users for breaching SLA. The local manager will install a module on each node to monitor virtual machines (VMM). The aim is to constantly monitor CPU productivity of nodes, change the size of virtual machine according to their needs and make decisions regarding when and which virtual machine ought to migrate from the node. The global managers have been located on the master nodes and they are gathering information from the local managers in order to control resources productivity. The global manager makes recommendations for optimal placement of VM. Virtual machine monitoring will make real changes and some changes will happen in nodes power by migration of virtual machines in order to reduce energy consumption. Whenever a virtual machine cannot use its available resources, the agreement breach will occur. There are 800 heterogeneous hosts for simulation that half of them are HP Proliant ML110G5 and other half are HP Proliant ML110G4. The server CPU frequency for each core based on MIPS is respectively MIPS1860 QND MIPS 2660. Each server needs a GB/S for bandwidth and the number of virtual machines and cloudlets is 1052. The productivity power of each virtual machine are rated at one of these numbers (2500, 2000, 1000, 500) MIPS, the amount of memory are equal to (0.85, 3.75, 1.7, 0.613) gigabytes and the maximum server consumption power is 250 watt. In this article a method has been applied to select overloaded physical host. The procedure is that, in the single threshold class, the utilization threshold which is a fixed number and equal 0.8 will be substituted by the new defined equation (1). The considered input for this new function, which has been added to single threshold class, is system productivity and its considered output is high threshold value as it be able to do migration of virtual machines from the physical host. This act will be done once for each host and whenever the value of the physical host threshold be more than the determined amount, that host would be recognised as overloaded host and some of its virtual machines ought to be migrated. In this article the cloudsim simulator has been used for simulation.

In Fig. 2, the applied solution has been compared with reference solutions from the view point of energy consumption and SLA violation. This comparison will be done through various policies in host selection as well as different policies for selecting virtual machines.

According to the Fig. 2 in which the energy has been calculated in terms of kilowatts per hour, the amount of MAD` energy consumption based on different policies of selecting virtual machine shows the lowest amount. It has been compared with the different policies of determining the productivity high threshold. However, it seems logical due to the increasing in the productivity high threshold of power consumption reduction.

The reason is that by reducing the number of migrations and as the off servers will be on later therefore, the obtained power consumption for this policy would be better than other policies.

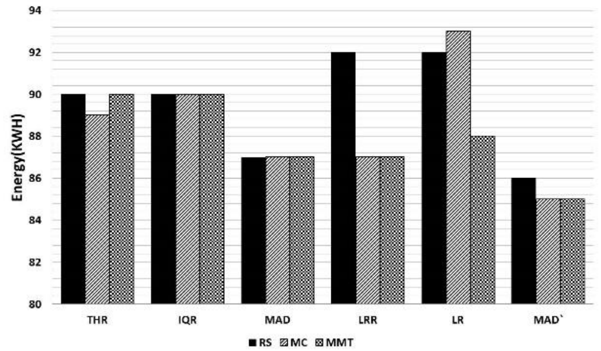

Fig. 2. The amount of energy consumption regarding various policies.

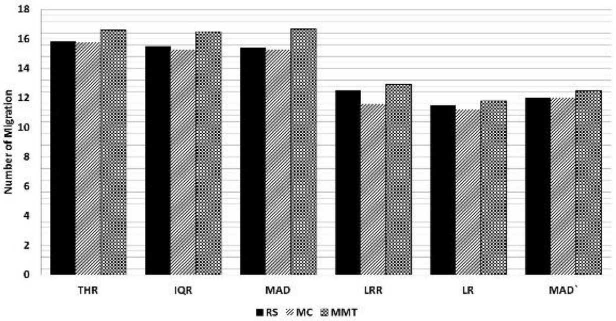

Fig. 3. The number of migrations according to the different policies.

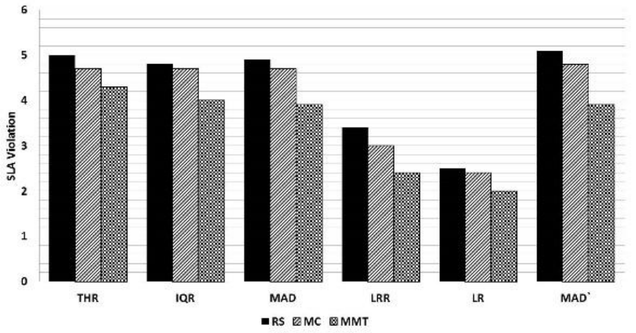

Fig. 4. The SLA violation values regarding the different policies.

Also, in Fig. 3 which shows the number ofvirtual machines migrations, *MAD*` policy has got the least numbers of migrations in comparison with the other high threshold determination policies. It is because of increasing in the high threshold for the migration which reduced the number of migrations. At the end, the SLA breach will be discussed that may slightly increase due to reduce in number of migrations and increase the migration threshold. Also, it has been shown in Fig.4 the SLA breach had slightly increased. This policy can be used wherever the SLA breach is not very important and has not disadvantages for the users. Hence, if we divided users in any of the service models (SaaS, PaaS, IaaS) into three subdivisions 1) important users 2) medium users and 3) less important users (the categorizing would be their own decision) and the amount of allocated values to each group have been considered differently, then, this policy can be used for the less important users that company will pay smaller amount for them. In the other words, if the supplier losses becomes less than energy consumption profits through SLA breach, then it would be beneficial for them.

V. Conclusion

As it has been discussed in the evaluation part, the new policy has been greatly beneficial for reduction of power consumption and reduction of migrations as well as energy consumption, though, the SLA breach has slightly increased. However, it would be a logical point because, as much as migration threshold increases and the number of migration reduce, the possibility of SLA breach will consequently increase. By increasing the SLA breach in the future, it is planned to find a host selection policy as it can decrease the SLA breach optimally or to choose a specific policy for selecting the virtual machines as it cause reduction of SLA breach. Generally, another solutions can be used to reduce power consumption such as special hardware equipment, suitable switching power supply and some other methods which can considerably reduce the power consumption and eventually the aim is to transfer this new technology to a green technology.

References

[1] A. Schurr, “Cloud computing confusion leads to opportunity,” *Network World*, September 11, 2009.

[2] J. G. Koomey, *Estimating Total Power Consumption by Servers in the US and the World*, Oakland, CA: Analytics Press, February 15, 2007.

[3] G. Koch “Extending the benefits of Moore’s law,” *Technology Discovering Multi-Core*, p. 1, 2005.

[4] A. Beloglazov and R. Buyaa, “Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers,” 2012.

[5] A. Beloglazov and R. Buyaa, “Managing overload host for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints,” *IEEE Transaction On Parallel and Distributed Systems*, vol. 24, No. 7, July 2013.

[6] VMware Inc., “VMware distributed power management concepts and use,” *Information Guide*, 2010.

[7] A. Verma, G. Dasgupta, T. K. Nayak, P. De, and R. Kothari, “Server workload analysis for power minimization using consolidation,” in *Proc. the 2009 Usenix Annual Technical Conference*, San Diego, CA, USA, 2009, pp. 28–28.

[8] B. Guenter, N. Jain, and C. Williams, “Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning,” in *Proc. of the 30st Annual IEEE Intl. Conf. on Computer Communications (INFOCOM)*, 2011, pp. 1332–1340.

[9] X. Wang and Y. Wang, “Coordinating power control and performance management for virtualized server clusters,” *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, vol. 22, no. 2, pp. 245–259, 2011.

[10] A. Beloglazov, J. Abawajy, and R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing,” *Future Generation Computer Systems*, 2011.

[11] G. Upton and I. Cook, *Understanding Statistics*, Oxford University Press, p. 55, 1996.

[12] W. S. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” *Journal of the American Statistical Association*, 1979, vol. 74, no. 368, pp. 829–836.

[13] W. S. Cleveland, *Visualizing Data*, New Jersey: Hobart Press, 1993.

Mona Arjomandi received her B.S. degree in electronics engineering from Islamic Azad University, Zahedan Branch, in 2009 and M.S. degree in information technology (IT) from Iran University of Science and Technology, Tehran, Iran, in 2014 under supervision of Dr. Mahmood Fathy and Dr. Kaamran Raahemifar. Currently, she is working in Bank Melli Iran.

Nik Mohammad Balouchzahi received the B.S. degree in computer engineering from Yazd University, Iran, in 2001 and M.S. degree in computer engineering from Iran University of Science and Technology, Tehran, Iran, in 2004. Since 2004 he has been a faculty member of Electrical and Computer Engineering Department at University of Sistan and Baluchestan. Currently, he is a PhD candidate in computer engineering at Iran University of Science and Technology under supervision of Dr. Mahmood Fathy and Dr. Ahmad Akbari. His main research interests include computer networks, vehicular communications and intelligent transportation systems.

Kaamran Raahemifar received his B.Sc. degree (1985-1988) in electrical engineering from Sharif University of Technology, Tehran, Iran, his MASc. degree (1991-1993) from Electrical and Computer Engineering Dept., Waterloo University, Waterloo, Ontario, Canada, and his Ph.D. degree (1996-1999) from Windsor University, Ontario, Canada. He was the chief scientist (1999-2000), in Electronic Workbench, Toronto, Ontario, Canada. He joined Ryerson University in Sept. 1999 and was tenured in 2001. Since 2011, he has been a professor with the Department of Electrical and Computer Engineering, Ryerson University. He is the recipient of ELCE-GSA professor of the Year Award (Elected by Graduate Student's body, 2010), Faculty of Engineering, Architecture, and Science Best Teaching Award (Apr. 2011), and Department of Electrical and Computer Engineering Best Teaching Award (Dec. 2011). He has been awarded more than $2.3M external research fund during his time at Ryerson. His research interests include: 1) optimization in engineering: theory and application, which includes grid optimization and net-zero communities, as well as biomedical signal processing, 2) big data analysis (dictionary/sparse representations, interpolation, predictions), 3) modelling, simulation, design, and testing, and 4) time-based operational circuit designs.

Mahmood Fathy received the B.S. degree in electronics from Iran University of Science and Technology, Tehran, Iran, in 1984, the M.S. degree in computer architecture from Bradford University, West Yorkshire, U.K., in 1987, and the Ph.D. degree in image processing and its architecture from the University of Manchester Institute of Science and Technology, Manchester, U.K., in 1991. Since 1991, he has been an academic member with the Department of Computer Engineering, Iran University of Science and Technology. His research interests include the quality of service in computer networks, including video and image transmission over Internet, the applications of vehicular ad hoc networks in intelligent transportation systems, real-time image processing, with particular interest in traffic engineering and remote health monitoring.

Ahmad Akbari received the BSc. degree in electronics engineering and the MSc. degree in communications engineering from Isfahan University of technology (IUT) in 1986 and 1989 respectively. He received the DEA and Ph.D. degrees in signal processing and telecommunications from university of Rennes 1, Rennes, France in 1992 and 1995 respectively. In 1996 he joined the Computer Engineering Department at Iran University of Science and Technology (IUST) as an assistant professor, where he is now the director of Computer Engineering Department. His research interests include computer networking, network security, acoustic modeling of speech, robust speech recognition, speech enhancement, implementation of signal processing algorithms, voice applications and interfaces and web technologies.