

# VIRTUAL CPU SCALING FOR EFFICIENT SERVER POWER CONSUMPTION IN CLOUD DATA CENTERS

Noora N. Bhaya <sup>1</sup>, Rabah A. Ahmed <sup>2</sup>

<sup>1,2</sup> College of Information Engineering, Al-Nahrain University, Baghdad, Iraq  
noora.nahidh@gmail.com <sup>1</sup>, rabahalansary@coie-nahrain.edu.iq <sup>2</sup>

Received:13/1/2020, Accepted:22/4/2020

**Abstract-** Cloud computing is a fast growing technology used by major corporations these days because of the flexibility framework it provides to consumers. Cloud technology requires large data centers consisting of multiple IT equipment and servers. One main problem with these data centers is the vast amount of power consumed during servers operation. This reduces financial benefit and increases the need to produce more energy to cover the needs of operating the cloud infrastructure. This paper proposes an approach for managing the virtual central processing unit (vCPU) of a virtual machine to improve server power efficiency. A framework is used to study the proposed approach on a server while processing different types of workloads widely found in most general-purpose cloud computing applications. Results indicate an improvement in server power saving.

**keywords:** Virtualization, vCPU, Data center, Power consumption.

## I. INTRODUCTION

Nowadays technology have become a lot more popular allowing for easier access and use of internet technologies. These include storage technology and computing resources, they have become cheaper and a lot more influential. All these trends are placed under the broad term cloud computing. One of cloud computing characterizations is that it gives a flexible environment which increases the ability to extended the work [1] . Cloud computing has developed the communication and information technology industry as the growing demand increases. The fast growth in the distributed computing led to the creation of large scale data centers containing a huge number of complex servers. Data centers cooling and computing are two essential features as the energy is saved [2]. Cooling includes providing a specific technique to be able to reduce the heat generated and decrease the physical machine heat dissipation. Computing involves workload scheduling or virtual machines allocating to reduce energy consumption. The services of cloud computing have extensive use that expected to increase the consumed power through the rapid use of equipment in the environment of cloud computing. It has been reported by the United States in 2013 that the data centers consumed about ( 91 billion KW/ h/ year) [3]. This amount of electricity is the same amount generated by 34 large power plants in which each one generates ( 500 megawatts) . The predicted electricity consuming by the data centers is expected to grow to reach about (140 billion KW/ h/ year) in 2020 [3]. This increase in power consumption will led to an increase in the needed power plants to generate more electricity. That is not economically in cost and also will lead to an increase in pollution because of the carbon emissions generated by the power plants. The power consumed by data center server sorted into two types dynamic and static power. Dynamic power relies on transistors in the electronic devices while execution the workload. Static power associated to the power dissipation of powered- on servers [4]. The actual amount of utilization- energy varies, servers generally consume more than half of their peak load energy when in idle mode, and energy consumption increases with resource utilization. Energy consumed during idle mode is static energy consumption. The dynamic energy consumption is the additional energy consumed by

running processes in the Cloud. In this paper, a proposed approach used to reduce dynamic energy depending on vCPU of the server, where the CPUlimit tool were use to control the workload that over utilize the CPU while maintain the quality of experience of the consumer. The rest of paper is organize as follows: section 2 presents the related work with other papers. The concept of virtualization and vCPU will present in section 3, the methodology of Vcpu management will be presented in section 4, section 5 will discuss the evaluation of the proposed approach. Result will discuss in section 5 and section 6 and 7 is the conclusion and the future work.

## II. RELATED WORK

Researchers have been proposed many techniques for solving the power consumption issue. It can be summarized as follows:

Anton and Raj Kumar, in 2010 [5] propose dynamic virtual machines (VMs) reallocation depending on the presence of resources requirements, meanwhile maintain a reliable QoS. The reallocation is used to reduce the number of active physical servers allocated to serve the current workload, while the idle servers are shut down in order to reduce power consumption. Dynamic virtual machines reallocation is an independent technique that is not reliant on a certain type of workload; neither does it require information regarding any applications applied on VMs. Another technique of virtual machine deployment based on heuristic greedy algorithm has been proposed by Jinhai Wang and et al, in 2013 [6]. This technique developed to increase the utilization of resources and decrease power consumption by mapping the CPU- intensive and memory-intensive workloads to the same server. Patricia Arroba et al. , in 2016 [7] focused on physical CPU voltage and frequency scaling ( DVFS). They propose a policy referred to as DVFS- aware consolidation to improve server power consumption while mapping VMs to achieve accepted service quality. Managing the virtual central processing unit ( vCPU) of a virtual machine when processing vCPU- intensive workloads is studied by Aula Abdel Latief Dewan, in 2019 [8]. It has been found that controlling the processing of a vCPU can reduces VM live migration time and ultimately the accumulative VMs live migrations power consumption in data center. The related work shows that all the researchers used different techniques to reduce the dynamic power consumption. In which it can be notices that researchers [5] - [7] used live migration, VM placement and DVFS techniques on the physical CPU, while [8] used the live migration technique on the vCPU. Different vCPU workloads are available in most cloud computing applications, therefore placing the vCPU unit under large stress leading to maximum utilization at the processing level. The intensive workload will be used to evaluate the improvement of server power saving. From the results of the survey, it was noticed that more research projects are required to find ways to improve the saving percentage and expand the service to the consumer. Hence, this paper proposes an approach to control the vCPU of virtual machine to improve the efficiency of vCPU dynamic power consumption. The vCPU of the virtual machine would be adjusted to the workload needs of vCPU resource during runtime while saving the quality of experience of the user in order to decrease the dynamic power consumption.

## III. VIRTUALIZATION AND VCPU IN CLOUD COMPUTING

Virtualization refers to the operation of transforming a physical hardware resource to a virtual hardware resource. Many instances for the hardware resource can be virtualized such as server, storage, network. The creation of a virtual machine

started by allocating the physical hardware, then the operating system (OS). The OS of the virtual servers is called " Guest OS" , which is independent of the base OS that installed on the host server [9]. The Virtual Machine manager VMM works to manage every node and leads the operation of sharing the physical CPUs among the VMs that are defined in the node. Generally, the VMM defines the vCPU associations and it is able to force each vCPUs to run on the available physical CPUs. The available resources restrict the amount of VMs and consequently, each VM is supplied with a number of vCPUs based on the available physical CPUs [10].

#### IV. METHODOLOGY

##### A. VCPU management framework

A management framework is built to study the relation between vCPU usage of a VM and server power consumption. This framework comprises of two laptops, one is represents the server and the other for the client. The server machine consists of a 2.4 GHz Intel Core i5 processor, 8 GB of memory, and a 500 GB hard drive. As for the client side, the CPU is 2.5 GHz Intel Core i5 processor, 4 GB of memory, and a 500 GB hard disk. TP- Link TL- WR740N router switch was used to connect the two machines in star topology through Cat5 LAN cables. Server power consumption is measured using a current clamp, which allows to obtain real values during real-time while testing real scenarios [7]. The main reason for counter server AC input is that measuring only the DC current will eliminate the power consumed by the power supply, which is a main part of the server system. Consequently, the measurement would be inaccurate. Moreover, before starting the measurement process, the battery of the server laptop was removed, this is necessary to eliminate the current used to charge the battery. For the VM setup, Citrix Xen Server 6.5 is used as a hypervisor in server machine [11]. A single model of virtual machine with one 2.4 GHz vCPU, a disk capacity of 450 GB and a memory size of 6.0 GB is initialized and stored directly on the local storage of the server. This VM setup is applied using Citrix XenCenter VM manager [12]. XenCenter administrates the XenServer by connecting through the IP addresses. When the connection setup of the two sides is accomplished, new local storage is to holds the virtual machine. Linux Ubuntu 12.04 LTS operating systems is selected to run the VM [13]. The reason for using Linux Ubuntu was attributed due to the feature of open source in Linux operating system which made it flexible to use. The management of vCPU is carried out using one Linux package tool named CPUlimit. This tool uses a percentage value to limit the usage of a vCPU for a certain process. CPUlimit facilitates the batch jobs control of the system. This tool was used to limits the vCPU cycles of VM, which exceed process needs by preventing the running process to use more than the specified ratio. Also, this tool has the ability to adjust itself quickly and dynamically to the whole system load. CPUlimit controls the entire threads and child processes of a defined process to share the same percentage of vCPU.

##### B. Workload selection

This framework uses multiple execution tasks during the evaluation. A set of vCPU intensive tasks has been selected to reflect vCPU activities to study the effect of controlling vCPU usage on server power consumption as well as the quality of Experience ( QoE) provided to consumers. This set of tasks includes MPEG- 4 video streaming, File Compression and Video Games. The characteristics of this set of vCPU intensive workloads will support the evaluation study of the efficient

vCPU usage that reduces server power consumption and maintains the QoE provided to the consumer. Different tasks can produce different vCPU activities and hence different impact on server's power and services. To account this issue, a number of each type of workloads has been used to represent the variety of workloads applied on vCPU. MPEG-4 videos are selected with a streaming time range from 60 to 90 minutes. These videos are processed while delivered directly by a local on-line service provider. These videos tested with different streaming quality of 720, 480, and 360 pixels. For the compression tasks, a collection of compression techniques include ( zip, tar.bz2, cbz, jar and tar.gz) are used on a file of 542MB. The file compression is differentiate from the video streaming task by the time factor consideration. As the compression does not have the graphic streaming as in the video. As for video gaming tasks, the activity of vCPU is tested using five different video games include ( Foobiliard, kiki, tennix, superkart, and motor x) video game. Gaming depends on the graphic motion and the interaction with the game panel without delay. This set of tasks can be representative cloud workload, since it can be found in most general purpose cloud computing applications. Since the scope of this study is vCPU activities, other components activities (memory, hard disk, network interface, etc.) are not exercised.

### C. Evaluation of quality of experience

Controlling vCPU usage may affect the quality Experience ( QoE) on the tasks ( workloads). QoE of a task represents the level of user satisfaction of using this task [14]. The goal is determining the minimum percentage of vCPU usage at which the consumer considers the service still to be accepted by using the subjective Metrics of QoE [15] . Thus, a questionnaire was used and filled out by 20 people ( 12 female, 8 male ) with age of (15 year- 55 year ). User include students, common computer users, pharmacists, and other people with different working fields. To evaluate the QoE of the video streaming, gaming and compression tasks. The questionnaire model used is simple, the user fills only if they are satisfied with the tested task or not. Fig. 1 show an example on the questionnaire. The criteria adopted are vary depending on the type of tasks under evaluation. For video streaming tasks, the criteria adopted are video quality and stream interruption [16]. From the user side, the quality of the experience can be achieved as long as video streams without quality degradation or interruptions. As for gaming tasks, the smooth interaction with the game panel immediately is the criterion adopted. The acceptance of the time taken to complete the compression process from the consumer perspective is the criterion adopted in the evaluation of compression.

## V. EVALUATION OF THE PROPOSED APPROACH

Allocation of vCPU resource to execute a process in VM must not exceeds the actual need to keep QoE. This section describes the evaluation of the proposed approach in real Xen server, which widely used in many public cloud data centers. Fig. 2 shows the schematic diagram of the framwork conducted in a laboratory environment with environment temperature of (30 Celsius). The framework described in IV- A has been used to evaluate the proposed approach and compare with Xen baseline model. The guest operating system of the virtual machines experienced the percentage control of vCPU for the workload process. During the evaluation, the identifier number (ID) of the process that represents the workload is predetermined. The CPUlimit tool is used as a command- line interface (CLI) in the terminal shell of the Ubuntu operating system as follow:

- First, use the top command in the shell to view the running system. In this case, it is used to show the ID of each running process and its CPU usage.

```
top
```

- The ability to use the CPU limiting tool can be achieved after knowing the ID and the amount of CPU usage. For example, the command used to control the process with the ID number (4510) and limiting the vCPU percentage to use 50% of the total CPU is:

```
cpulimit - p 4510 - 150%----
```

The ID of process is used as input to the CPULimit tool with the desired percentage of vCPU usage. Starting from the 100% of vCPU usage, the vCPU percentage was decreased in each scenario to determine the minimum percentage while the workload still valid to be used with an acceptable QoE for the user. For each set of workload, the procedure of CPULimit is carried on with operating the workload in the OS system on the client- side and starts to measure the current using the clamp meter that is located on the adapter on the server side and record the measurement of the current while operating each of the workload sets. To obtain the power consumed by each workload, the average server AC current consumption were multiplied by the AC voltage source. The procedure sequence during the experimental work is shown in Fig. 3 .

## VI. RESULTS AND DISCUSSIONS

The experiments are conducted using the framework described previously. The results revealed an improvement in the efficiency of server power consumption when applying the proposed approach compared to the base system. Fig. 4 shows the server's power pattern when applying the proposed approach to a video stream workload of 720 and 480 pixels resolution. Results obtained for the experience of users under test indicated that CPU limit to 60% can be sufficient to provide acceptable quality of experience for the majority of users when experience video streaming workloads of 720 pixel, and 50% for 480- pixel. These results show that in an average 15% of server power was reduced by using the proposed approach compared to the basic system when processing a 720- pixel streaming workload, and 19% for the 480- pixel. As for the 360- pixel video streaming workload, the results of the questionnaire showed that the experience at 40% CPU limit is acceptable by the majority of uses as shown in Fig. 5, which may save 23% of server power consumption. As mention that the percentage of processor allocated by the base system (without the proposed approach) is not less than 96% for all the cases mentioned above. Furthermore, allocating 80%, 70% and 60% of processor time will be sufficient to obtain the acceptance of all users when experience a video streaming workload of 720, 480 and 360 pixel respectively. This decrease of vCPU will not affect the QoE because the decreased percentage of video workload is an over utilize percentage where each video would not need the full usage of the vCPU to work appropriately. For video gaming workloads, it was found

that the base system allocates more than 90% of processor time for processing such workloads. After applying the proposed approach, it appeared that 60% of the time of the vCPU is sufficient to reach the accepted experience of most of the tested users as shown in the Fig. 6. It is stated here that applying 70% CPU limit will be sufficient to meet the accepted experience for all the tested users. This is able to save energy at a rate of not less than 6% compared to the original system for the same working time. As for the file compression workloads, applying the propose approach can also reduce the direct power of the server. Fig. 7 describes the pattern of power consumption for different types of file compression methods, which shows a direct proportion to the percentage of CPU limit. Most of the users under test accept the compression experience when limiting vCPU by 40% and above. However, from the point of view of processing time, limiting of vCPU showed an adverse effect on the time taken to complete the compression process as shown in Fig. 8, as the solid line shows the mean time. This demonstrates a different effect of using the vCPU limit on file compression workloads compared to video streaming and video game workloads, as applying the vCPU limit will prolong the compression processing time. Fig. 9 depicts the energy consumed by the server during the processing time of different file compression methods as a relation to the percentage of CPUlimit. These results indicate that the proposed approach may not be beneficial for these types of time- sensitive workloads.

**CPU limitation evaluation**

Full name	Hilal M. Bahaya		Age	28
Gender	Male	Female	job	Engineer

1. Please choose minimum vCPU percentage that satisfies your experience in watching the following videos, considering video quality and stream interruptions:

Video	vCPU	70%	60%	50%	40%	30%
Video 1 resolution 360					✓	
Video 1 resolution 480				✓		
Video 1 resolution 720		✓				
Video 2 resolution 360					✓	
Video 2 resolution 480				✓		
Video 2 resolution 720		✓				
Video 3 resolution 360					✓	
Video 3 resolution 480				✓		
Video 3 resolution 720		✓				
Video 4 resolution 360					✓	
Video 4 resolution 480				✓		
Video 4 resolution 720		✓				
Video 5 resolution 360					✓	
Video 5 resolution 480				✓		
Video 5 resolution 720		✓				

2. Please choose minimum vCPU percentage that satisfies your experience to compress a file with the listed formats, consider compression time:

Format	vCPU	80%	70%	60%	50%	40%	30%	20%
Zip						✓		
Cbz						✓		
Jar					✓			
Tar.gz						✓		
Tar.bz2						✓		

3. Please choose the accepted percentage for each game considering graphic streaming and response to the controller :

Game	vCPU	70%	60%	50%	40%	30%
Game 1				✓		
Game 2		✓				
Game 3				✓		
Game 4				✓		
Game 5				✓		

Full name: Hilal M. Bahaya      Signature: 

Figure 1: Example on the QoE questionnaire

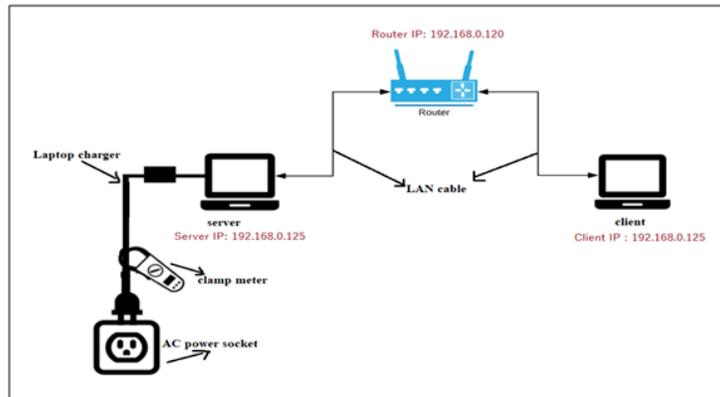


Figure 2: Schematic diagram for system structure

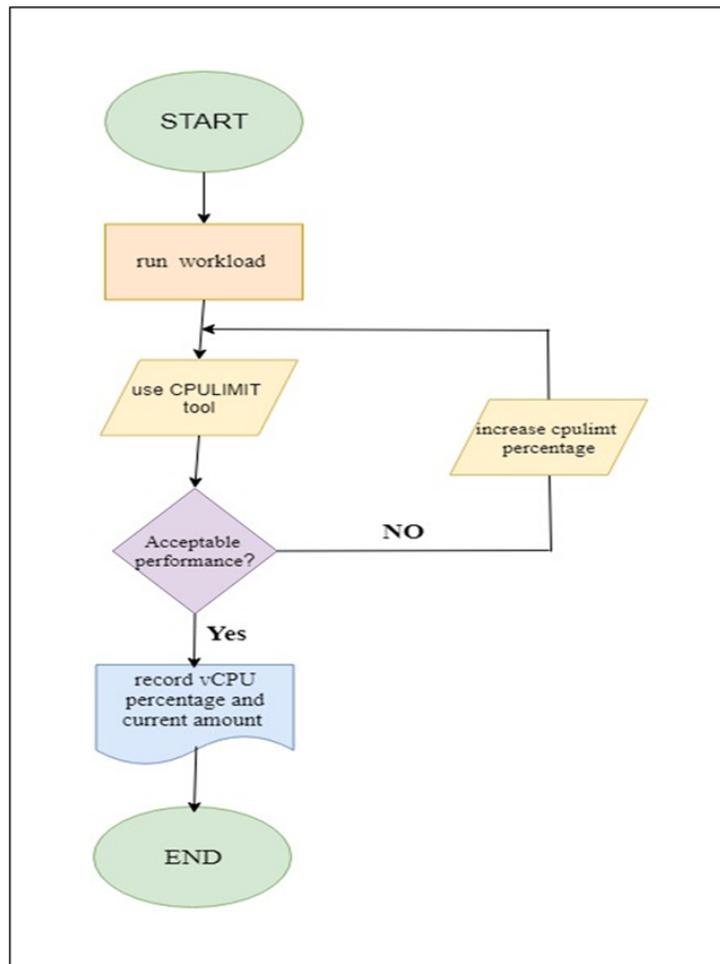


Figure 3: Flow chart of the proposed approach

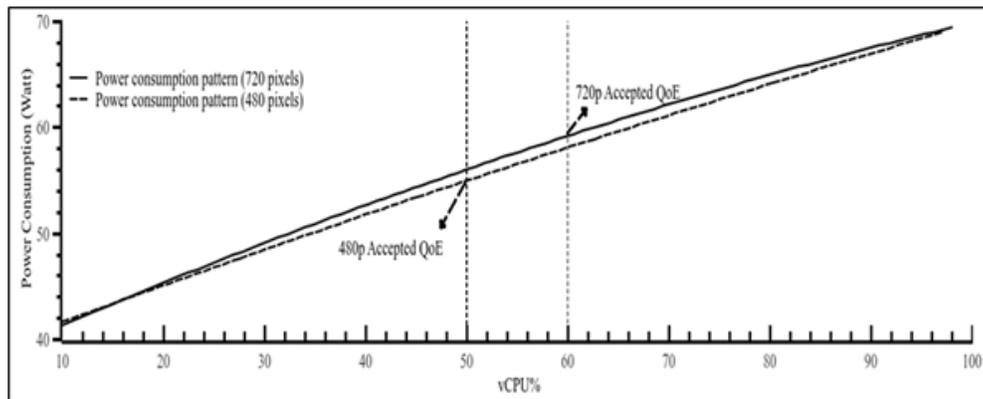


Figure 4: Performance of 720 p and 480 p videos

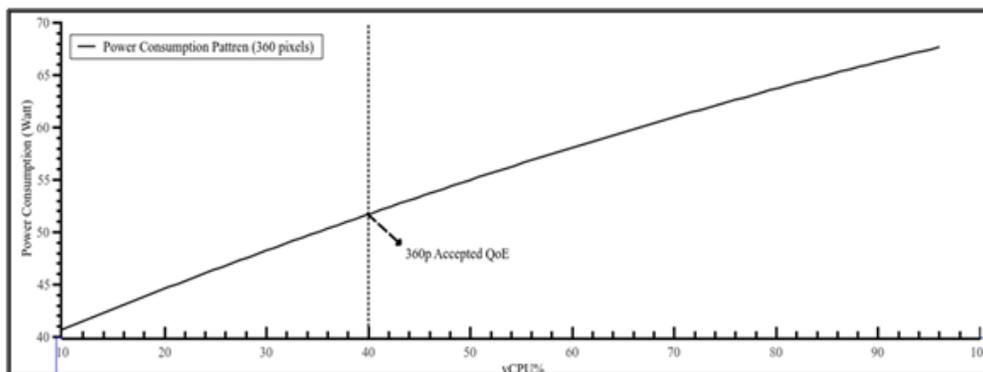


Figure 5: Performance of 360 p video

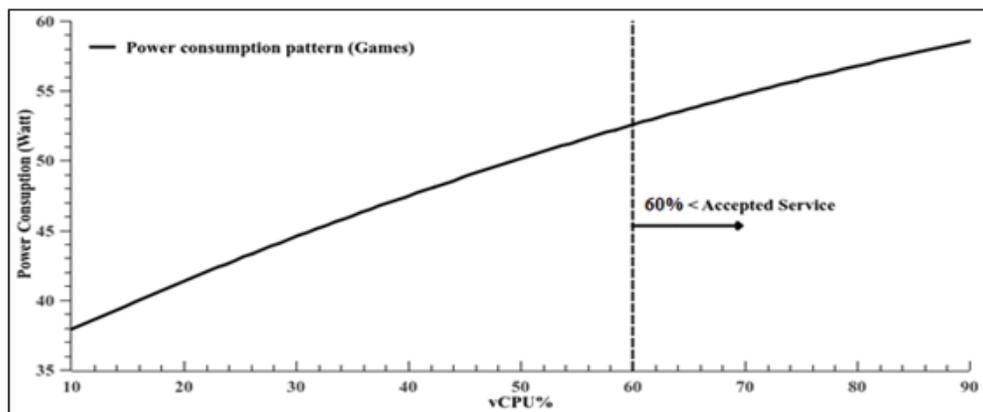


Figure 6: Performance of video gaming

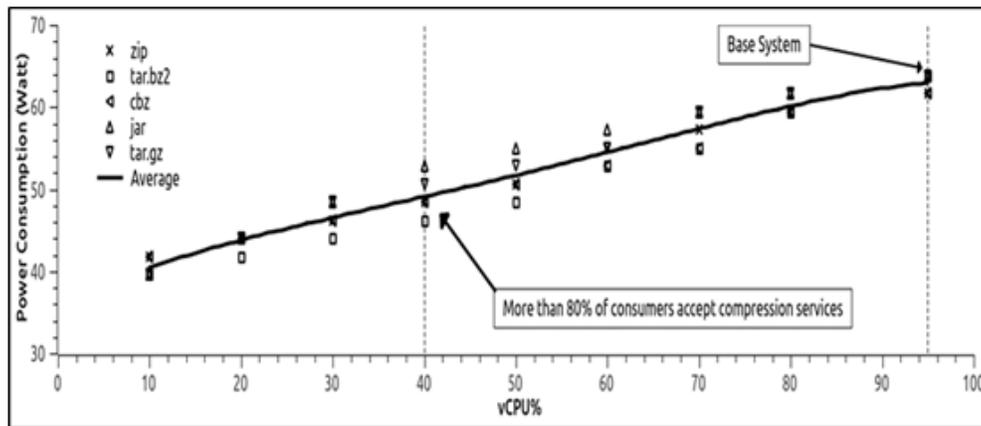


Figure 7: Compression performance

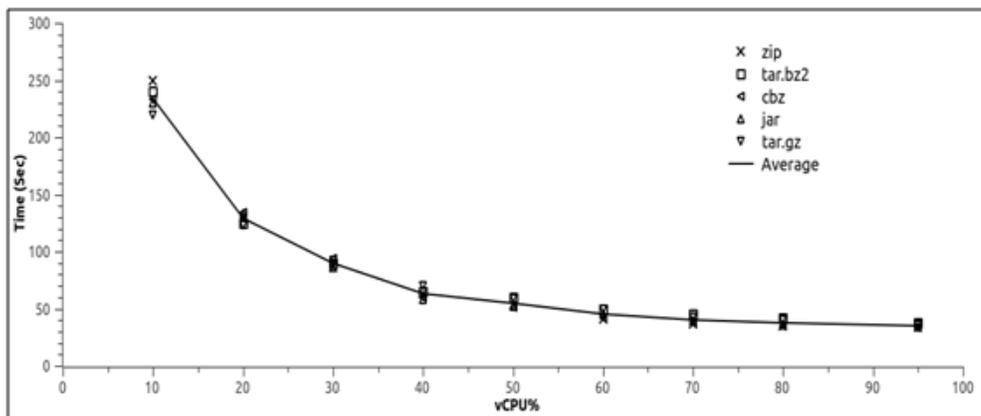


Figure 8: Compression consumption time

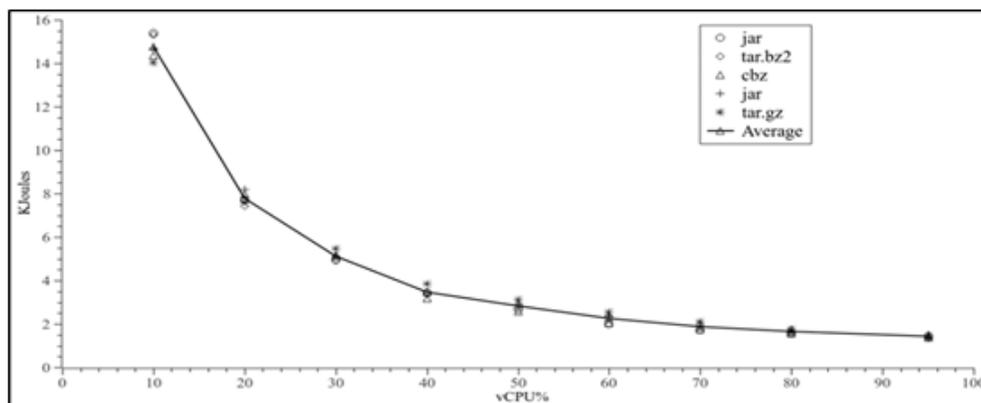


Figure 9: Energy consumed during the processing time

## VII. CONCLUSIONS

This paper proposes an approach to control the Central Process Unit of a virtual machine (vCPU) for improving the efficiency of a server power consumption in cloud data centers. The results obtained by this study revealed that while processing online video and game workloads, the range of vCPU usage could be reduce significantly while maintaining service quality of experience. Comparison results with the original system showed a significant reduction in server dynamic power consumption. However, this approach may adversely affect server dynamic power consumption when applied to the file compression process, where it may increase the execution time of the compression process.

## VIII. FUTURE WORKS

To extend this work in the future, several suggestions were made such as using different workloads set; use more than vCPU for XenServer can be used to study the influence of the power consumption, and use different operating system instead of using Linux.

## REFERENCES

- [1] Puthal Deepak, Bibhudutta PS Sahoo, Sambit Mishra, and Satyabrata Swain, " Cloud computing features, issues, and challenges: a big picture." International Conference on Computational Intelligence and Networks, pp. 116- 123, IEEE, 2015.
- [2] Kaur Amanpreet, V. P. Singh, and Sukhpal Singh Gill, " The future of cloud computing: opportunities, challenges and research trends" , 2nd International Conference on I- SMAC ( IoT in Social, Mobile, Analytics and Cloud) ( I- SMAC) I- SMAC ( IoT in Social, Mobile, Analytics and Cloud (I- SMAC), 2nd International Conference, IEEE, 2018.
- [3] Delforge Pierre , Whitney Jos, " Data Center Efficiency Assessment Scaling Up Energy Efficiency Across the Data Center Industry: Evaluating Key Drivers and Barriers" , IP:14- 0- a ,Kyoto University,2014.
- [4] Gandhi A, Harchol- Balter M, Das R, Lefurgy C, " Optimal power allocation in server farms" , ACM SIGMETRICS Perform Eval Rev, 37( 1) : 157- 168, 2009.
- [5] Beloglazov Anton, and Rajkumar Buyya, " Energy efficient allocation of virtual machines in cloud data centers" , 10th IEEE/ ACM International Conference on Cluster, Cloud and Grid Computing, IEEE, 2010.
- [6] Wang Jinhai, Chuanhe Huang, Kai He, Xiaomao Wang, Xi Chen, and Kuangyu Qin, "An energy- aware resource allocation heuristics for VM scheduling in cloud" , 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, IEEE, 2013.
- [7] Arroba Patricia, Jose M. Moya, Jose L. Ayala, and Rajkumar Buyya, " Dynamic Voltage and Frequency Scaling- aware dynamic consolidation of virtual machines for energy efficient cloud data centers" , Concurrency and Computation: Practice and Experience, 2016.
- [8] Aula A Dewan. Rabah A.Ahmed, " Enhancing the performance of pre copy live migration method for virtual machines in cloud environment" , ( master dissertation) , Alnahrain University,collage of information engineering, Baghdad, Iraq, 2019.
- [9] Lee Hyungro, " Virtualization basics: Understanding techniques and fundamentals." School of Informatics and Computing, Indiana University, 2014.
- [10] Buyya Rajkumar, James Broberg, and Andrzej M. Goscinski, " Cloud computing: Principles and paradigms" , Vol. 87, John Wiley & Sons, 2010.
- [11] Citrix, " XenServer6.5 Service Pack 1 Installation Guid" , Edition 1.0 , United States of America, 2015.
- [12] Citrix, " Xencenter Documentation" , United States of America, 2019.
- [13] Ubuntu, " Ubuntu manua" , <http://manpages.ubuntu.com/manpages/trusty/man1/cpulimit.1.html>, 2019.
- [14] Laghari, Asif Ali, Hui He, Imtiaz A. Halepoto, M. Sulleman Memon, and Sajida Parveen, " Analysis of quality of experience frameworks for cloud computing" , IJCSNS 17, no. 12 ,2017.
- [15] Juluri, Parikshit, Venkatesh Tamarapalli, and Deep Medhi, " Measurement of quality of experience of video- on- demand services: A survey" , IEEE Communications Surveys & Tutorials 18, no. 1 , 2015.
- [16] Laghari, Asif Ali, Hui He, Muhammad Shafiq, and Asiya Khan, " Assessing effect of Cloud distance on end user's Quality of Experience (QoE)" , In 2016 2nd IEEE international conference on computer and communications ( ICC) , pp. 500- 505. IEEE, 2016.