Acta Technica *62*, No. 3B/2017, 63–74

A scheduling model of virtual machine based on time and energy efficiency in cloud computing environment [1]

Xin Sui[2,3], Li Li[2,4], Dan Liu[2], Huan Wang[2], Xu Di[2]

Abstract. The virtual machine scheduling strategy can effectively improve the resource utilization, reduce energy consumption and processing time. So far, many virtual machine scheduling strategies have been proposed, but most of the models are modeled on the virtual machine scheduling strategy from single factor of resource utilization, energy consumption and time, this paper proposes a new model of virtual machine scheduling based on time and energy, not only researching the time and energy of VM, but also taking into account the efficiency of resource utilization. The experimental results show that the model can shorten the processing time of VM, reduce the energy consumption of the cloud data center, and improve the utilization ratio of the resource.

Key words. Cloud computing, virtual machine, time-aware, energy efficiency.

1. Introduction

With the increasing size of the cloud data center, energy consumption has gradually increased, energy consumption has become a major concern of cloud service providers. In addition to power consumption and carbon emissions, power supply capacity has limits the scale of the cloud data center, but also has reduce the economic benefits of cloud service providers. On the other hand, unpredictable rental costs and tenants concerned about the quality of service problems make many potential tenants prohibitive. Therefore, in order to attract more tenants and maximize

[1]This work was supported in part by the project of the higher education in Jilin Province (2014), the science and technology project of "13th Five-Year" planning of the Education Department of Jilin Province (2016), the key project of "13th Five-Year" planning of the Education Science of Jilin Province (ZD16024), and Key Science and Technology Project of Jilin Province (20160204019GX).

[2]Changchun University of Science and Technology, College of Computer Science and Technology, 130022, China

[3]Jilin Provincial Institute of Education, ChangChun, 130022, China

[4]Corresponding author; e-mail: ll@cust.edu.cn

economic benefits, cloud service providers need to provide an effective way to reduce energy consumption for data centers.

The research results show that the energy consumed by the server and the cooling system is 70 % of the total energy consumption of the data center, and the remaining 30 % energy consumption is consumed by the network components. This has prompted researchers to study the energy consumption of servers and network components. At present, researchers have proposed an effective energy saving method based on virtualization technology, for example: energy aware virtual machine scheduling mechanism [1] (Dalvandi et al., 2015).

For tenants and cloud service providers, the execution time of task is also a very important factor. Cloud service providers allocate different resources for different tasks which helps to provide a more effective means of resource allocation, quality of service to meet the tenants at the same time to make the data center energy consumption to a minimum, so to achieve the goal of maximum economic efficiency.

Guangyu Du et al. [2] proposed a novel scheduling algorithm for heterogeneous virtual machines in virtualized environments to effectively reduce energy consumption and finish all tasks before a deadline. The new scheduling strategy is simulated using the CloudSim toolkit package. Experimental results show that our approach outperforms previous scheduling methods by a significant margin in terms of energy consumption. Aissan Dalvandi et al. [3] proposed a novel time-aware request model which enables tenants to specify an estimated required time-duration, in addition to their required server resources for Virtual Machines (VMs) and network bandwidth for their communication. Nguyen Quang-Hung et al. [4] proposed heuristic-based EM algorithm to solve the energy-aware VM allocation with fixed starting time and duration time. In addition, this work studied some heuristics for sorting the list of virtual machines to allocate VM. We evaluate the EM using CloudSim toolkit and jobs log-traces in the Feitelson's Parallel Workloads Archive. Simulation's results show that all of EM-ST, EM-LFT and EM-LDTF algorithms could reduce total energy consumption compared to state-of-the-art of power-aware VM allocation algorithms. Anh Quan Nguyen et al. [5] presented a new model to deal with VMMP that took into account the uncertainty of the VM execution time, hence allowing to obtain robust assignment solutions. The uncertainty of the VM execution time is modeled by (i) relying on a truncated normal distribution for constructing mapping instances, and (ii) by using the expected value of the generating truncated normal distribution. The proposed methods, for the optimization for the VMMP, are conducted on the Grid5000 in order to bring a detailed results comparison between the obtained results from the experimental study with different benchmarks. Rehana Begam et al. [6] proposed a time-sensitive resource (TSR) allocation scheme. A unified scheme that integrated the ideas of both EDF and TSR is also studied. By incorporating the classical mapping schemes when selecting the pool of resources, the proposed schemes were evaluated through extensive simulations using real-application trace data. The results showed that the proposed schemes could significantly out-perform the state-of-the-art deadline oblivious resource allocation scheme with up to 25 % more user requests being served and up to 5 % more benefit being achieved, especially for over-loaded cloud systems. To solve this BNP problem in polynomial time,

Bo Wang et al. [7] proposed a heuristic algorithm. Its main idea was assigning the task closest to its deadline to current core until the core could not finish any task within its deadline. When there was no available core, the algorithm added an available PM with most capacity or rents a new VM with highest cost-performance ratio. Extensive experimental results showed that our heuristic algorithm saved 16.2 %–76 % rent cost and improved 47.3 %–182.8 % resource utilizations satisfying deadline constraints, compared with first fit decreasing algorithm. Jasmin James et al. [8] proposed a new VM load balancing algorithm and implemented for an IaaS framework in Simulated cloud computing environment, for the Datacenter to effectively load balance requests between the available virtual machines assigning a weight, in order to achieve better performance parameters such as response time and Data processing time. Shalom et al. [9] considered the following online scheduling problem in which the input consists of n jobs to be scheduled on identical machines of bounded capacity g (the maximum number of jobs that can be processed simultaneously on a single machine). Each job is associated with a release time and a completion time between which it is supposed to be processed. When a job is released, the online algorithm has to make decision without changing it afterwards. Tian et al. [10] considered online energy-efficient scheduling of virtual machines (VMs) for Cloud data centers. Each request is associated with a start-time, an end-time, a processing time and a capacity demand from a Physical Machine (PM). The goal is to schedule all of the requests non-preemptively in their start-time-endtime windows, subjecting to PM capacity constraints, such that the total busy time of all used PMs is minimized (called MinTBT-ON for abbreviation). Nguyen et al. [11] considered a virtual machine allocation problem. Each physical machine in cloud has a lot of virtual machines. Each job needs to use a number of virtual machines during a given and fixed period. The objective aims to minimize the cost induced by total execution time on each physical machine. This allocation problem is proved to be NP-hard. Additionally, three mixed integer linear mathematical models are constructed to represent and solve the problem. The performance comparison of the three proposed models is analyzed through some empirical results[11].

2. Methodology

2.1. Problem formulation

2.1.1. Energy problem. In the cloud data center, the element affecting server energy consumption are: CPU utilization, memory and store, and CPU is the major energy consuming equipment, energy consumption of physical server and CPU utilization are closely related. There is a certain relationship between the two factor, such as in formula (1). Let $P_i(u)$ be the power rate of server i, r_i represents the power ratio of server i when CPU is at minimum and maximum, $P_i^{\max}$ represents the maximum power rate of server i, u_i is the CPU utilization of server i at current time (Sui et al. [12]). Then

$$P_i(u) = r_i P_i^{\max} + (1 - r_i) P_i^{\max} u_i \,. \tag{1}$$

Due to the CPU utilization of server i is a changed value, so the energy consumption of server i can be expressed by formula (2). Let E_i represent energy consumption of server i between time t_1 and time t_n, $P(u_i(t_j))$ represents power rate of server i at time t_j, $u_i(t_j)$ represents the CPU utilization of server i at time t_j. Now

$$E_i = \sum_{t_1}^{t_n} Pu_i(t_j)) \,. \tag{2}$$

The energy consumption by network devices in the cloud data center is not negligible. According to the literature, the power rate of network device is related to the configuration and flow rate of the device. The configuration parameters of influencing power rate on network equipment including: backplane type, card type, card number and network interface configuration rate. Because of the energy consumed by the different network flow rate is almost identical, the power rate of network equipment is mainly affected by the network equipment configuration, which can be calculated by the formula (3). Among them, $P(C)$ represents the energy consumption of network equipment, C represents the configuration parameters of network equipment, $F(C)$ represents the sum of basic power rate and power rate of network line card, A is the network interface power rate, X is the count of network interface. Then

$$P(C) = F(C) + AX \,. \tag{3}$$

From the formula (2) and (3), the energy consumption of data center can be calculated, and expressed by formula (4). Among them, E_i represents the energy consumption of server i between time t_1 and time t_n, N is the number of running servers, and $P(C)$ is the power rate of network equipment. Now

$$E = \sum_{1}^{N} E_i + \sum_{t_1}^{t_n} P(C) \,. \tag{4}$$

2.1.2. Time problem. The previous virtual machine scheduling model considered the sum of all virtual machine completion time as parameters, without considering the response and processing time of each virtual machine, made response and processing time of some virtual machine longer. In this paper, the virtual machine scheduling model considered both the total completion time of virtual machines and the response and processing time of each virtual machine.

Because one server can run 1–m virtual machines, the response and processing time of each virtual machine is different. It is related to the number of CPU cores, the processing capacity of each CPU, the size of instructions executed by virtual machine and the scheduling mode of the virtual machines on the server.

The scheduling mode of virtual machine on the server include: time sharing and space sharing scheduling mode. Space sharing mode: there is only one virtual machine executing instructions on one server at the same time, as shown in Fig. 1. Time sharing mode: there are m virtual machines executing instructions on one server at the same time, the m virtual machines are in competition state for the

various resources, as shown in Fig. 2.

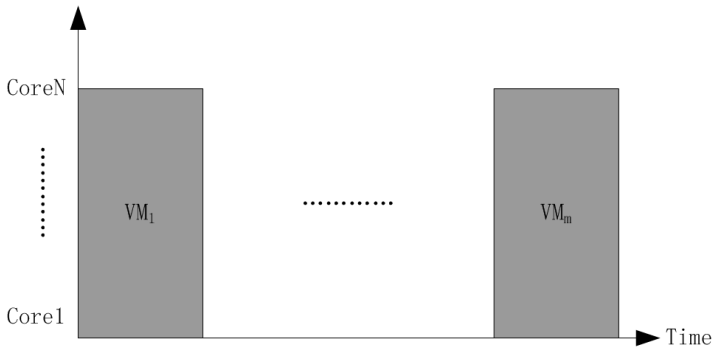

Fig. 1. Space sharing diagram

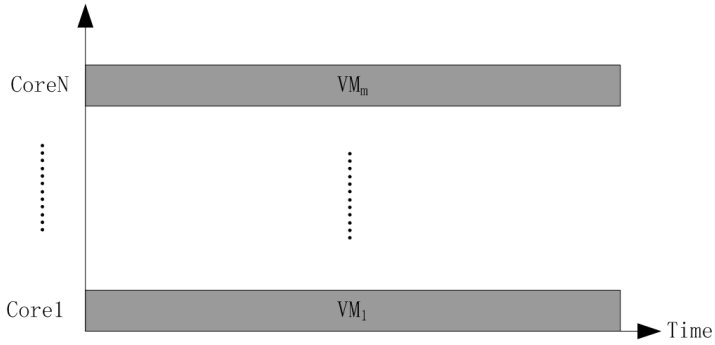

Fig. 2. Time sharing diagram

In space sharing mode, there was only one virtual machine at the executing state, the virtual machine executed sequentially on the physical server in a certain order, the response and processing time of each virtual machine can be calculated by formula (5) and formula (6). Among them, $T^S_{V_i}$ is the response time of virtual machine i, $T^R_{V_i}$ is the processing time of virtual machine i, $Instructions^{size}_i$ is the instruction size executed by the virtual machine i, $Core_{Capacity}$ represented the processing capacity of each CPU core; this paper assumed that the processing capacity of CPU core on one server is the same Chien et al. [13]).

$$T^S_{V_{i+1}} = T^S_{V_i} + T^R_{V_i}\,, \tag{5}$$

$$T^R_{V_i} = \frac{Instructions^{size}_i}{\sum_1^n Core_{Capacity}}\,. \tag{6}$$

In time sharing mode, because it is a competitive relationship between virtual machines, the server resources obtained by each virtual machine are related to the number of CPU cores and the processing capacity of each CPU core, the processing time of each virtual machine can be calculated by formula (7). Here, $T_{V_i}^R$ is the processing time of virtual machine i, $Instructions_i^{size}$ is the instruction size executed by the virtual machine i, $Core_{Capacity}$ represented the processing capacity of each CPU core, and a is the number of CPU cores.

$$T_{V_i}^R = \frac{Instructions_i^{size}}{a * Core_{Capacity}} . \tag{7}$$

2.1.3. Virtual machine scheduling mode. Objective function:

$$F_1 = \min \sum_1^m T_{V_i}^R , \tag{8}$$

$$F_2 = \min \sum_1^m T_{V_i}^S , \tag{9}$$

$$F_3 = \min(E) . \tag{10}$$

In this paper, the goal of proposed virtual machine scheduling model is as follows: firstly, the total processing time of virtual machine is minimized, by formula (8). Secondly, the total response time of the virtual machine is minimized, by formula (9). Thirdly, ensuring that the energy consumption of cloud data center to a minimum, using formula (10).

Constraints of virtual machine placement:

$$\left\{ \begin{array}{c} \sum_{j=1}^m V_j^{CPU} P_{i,j} < C_i^{CPU} \\ \sum_{j=1}^m V_j^{MEM} P_{i,j} < C_i^{MEM} \\ \sum_{j=1}^m V_j^{STORE} P_{i,j} < C_i^{STORE} \end{array} \right\} , \tag{11}$$

$$\sum_{i=1}^n P_{i,j} = 1, \ P_{i,j} \in \{0,1\} . \tag{12}$$

The formula (11) represents that the sum of all the virtual machines capacity should not be greater than the server capacity, including CPU, memory, and store. Among them, m is the number of virtual machines and n is the number of servers. Quantities V_j^{CPU}, V_j^{MEM}, and V_j^{STORE} are. respectively. the requested size of CPU, memory, and store of virtual machines. Quaantities C_i^{CPU}, C_i^{MEM}, and C_i^{STORE} are, respectively, the owned size of CPU, memory, and store. The formula (12) indicates that one virtual machine can only be placed on one server, and $P_{i,j}$ represents virtual machine j is placed on server i or not.

3. Results

In this paper, the cloud computing simulation platform(CloudSim) is used to simulate the experiment. The TAE, IQR and MAD model are compared and analyzed in terms of processing time, resource utilization, energy consumption and so on.

3.1. Experiment parameter setting

This experiment ran in a cloud data center, which included 500 servers of two configurations and 500 virtual machines of four configurations, the server configuration parameters are shown in Table 1, the virtual machine configuration parameters are shown in Table 2.

Table 1. Server configuration parameter

Parameter	HOST1	HOST2
MIPS	1860	2660
PES (Number)	2	2
RAM (MB)	4096	4096
BW (Gbit/s)	1	1
STORAGE (TB)	1	1

Table 2. Vm configuration parameter

Parameter	VM1	VM2	VM3	VM4
MIPS	2000	1600	1000	500
PES (Number)	1	1	1	1
RAM (MB)	870	1740	1740	613
BW (Mbit/s)	100	100	100	100
STORAGE (GB)	2.5	2.5	2.5	2.5

3.2. Experiment comparison results

3.2.1. Average processing time of VM in a certain number of tasks. The first experiment is to simulate the average processing time of VM in a certain number of tasks. The experimental results showed that when the virtual machine ID is 0–120, the average processing time of TAE model is reduced by 2.15 % and 0.3 %, respectively, compared with IQR and MAD models. When the virtual machine ID is 120–250, the average processing time of TAE model is reduced by 2.07 % and 0.22 %, respectively compared, with IQR and MAD models. When the virtual machine ID is 250–380, TAE model is reduced by 1.88 % than IQR model, increased by 0.0446 % than MAD model. When the virtual machine ID is 380–500, the TAE model is reduced by 0.775 % and 0.173 % than IQR, MAD model. When the virtual machine

ID is 0–500, the average processing time of TAE model is reduced by 1.43 % and 0.063 %, respectively, compared with IQR and MAD models. The experimental results showed that the TAE model is less than the IQR and MAD models in the virtual machine processing time. The result is depicted in Fig. 3.

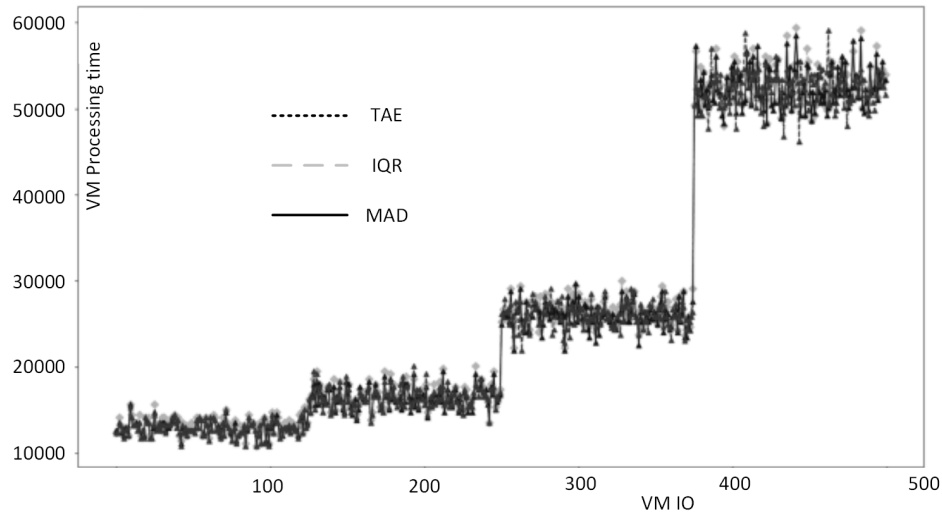

Fig. 3. Average processing time of VM in a certain number of tasks

3.2.2. Average processing time of VM at different number of tasks. The second experiment is to simulate the average processing time of VM in different tasks. The experimental results show that the average processing time of TAE model in this paper is the shortest, when the number of virtual machines is 50, the average processing time of TAE model is reduced by 2 % and 1.6 % than the IQR and MAD models. When the number of virtual machines increased to 500, the average processing time of TAE model is reduced by 1.45 % and 0.07 % than IQR and MAD models. The result is depicted in Fig. 4.

3.2.3. Average CPU utilization of data center at different number of tasks. The third experiment is to simulate the average CPU utilization of data center in different tasks. The experimental results showed that the CPU average utilization rate of TAE model had remain at between 53 %–60 %, IQR and MAD model was between 35 % and 40 %, so the TAE model was more effective than IQR and MAD model in improving average CPU utilization rate. The result is depicted in Fig. 5.

3.2.4. Average CPU utilization of data center at different time. The fourth experiment is to simulate the average CPU utilization of data center at different time. The experimental results showed that the proposed TAE model was higher than IQR and MAD models in average CPU utilization rate, maintained at between 30 % and 80 %. When the processing time was at 0–30000, the average CPU utilization rate

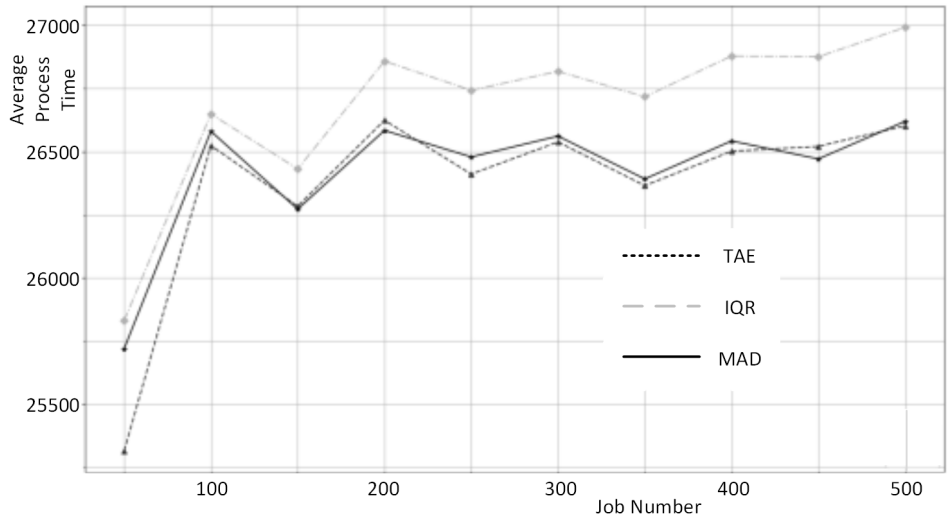

Fig. 4. Average processing time of VM at different number of tasks

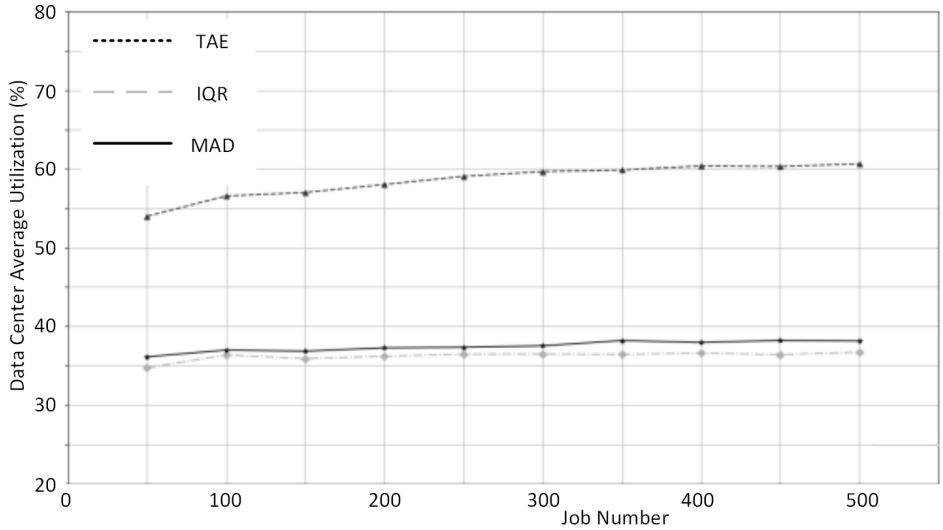

Fig. 5. Average CPU utilization of data center at different number of tasks

is relatively stable at 45 %–65 %. When the processing time was at 30000–55000, the average CPU utilization rate was fluctuated, maintained at between 30 %–80 %, illustrated that a part of cloud tasks had been completed, caused the server CPU utilization increased or decreased. When the processing time was at 55000–60000, the average CPU utilization rate decreased rapidly, that was because all the tasks had been completed. The result is depicted in Fig. 6.

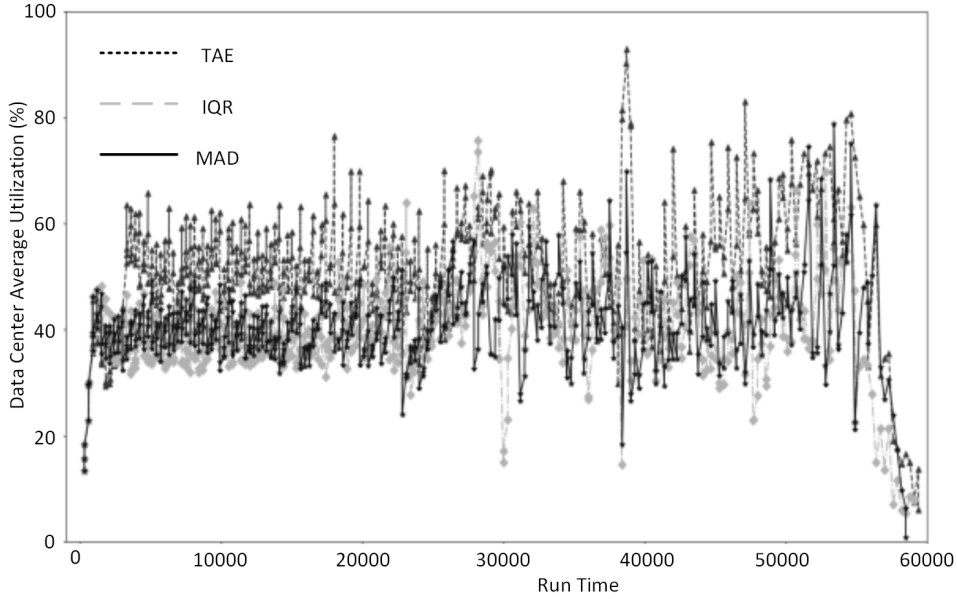

Fig. 6. Average CPU utilization of data center at different time

3.2.5. Energy consumption of data center at different number of tasks. The fifth experiment is to simulate the energy consumption of data center in different tasks. The experimental results showed that when the task number is 50, the energy consumption of IQR and MAD models is 9.08 and 9.16, and the TAE model was 7.47, decreased by 17.7 % and 18.5 %. When the task number is 500, IQR and MAD models are 89.46 and 88.32, and the TAE model was 73.37, respectively, decreased by 17.99 % and 16.93 %. Therefore, the proposed TAE model is more energy efficient than IQR and MAD models.

4. Discussion

Compared with the IQR and MAD models, the TAE model not only takes the energy consumption problem as a model parameter, but also takes the time and space sharing problem as parameter. The experimental results show that the processing time of virtual machine, CPU utilization, energy consumption are better than the other two algorithms, which is due to the factors of the model considering the energy consumption, the CPU utilization rate of server and processing time of virtual machine, rather than from one or two to study, but the average processing time in a very few number of tasks is almost the same as MAD model, optimization space still exists.

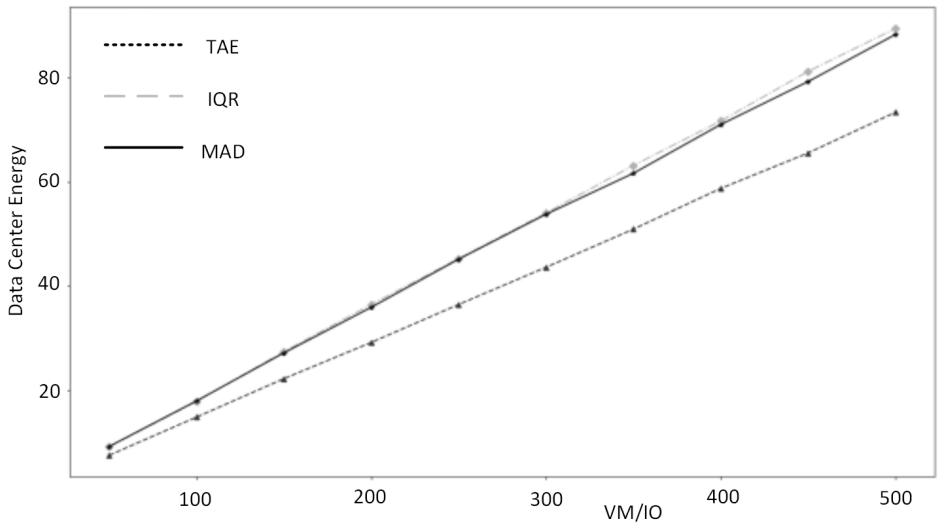

Fig. 7. Energy consumption of data center at different number of tasks

5. Conclusion

This paper studied the relationship problem of energy consumption and CPU utilization of server, the response and processing time of virtual machine based on time and space sharing, according to the constraints of virtual machine placement, proposed a virtual machine placement model (TAE) based on time-aware and energy-aware. During the experiment, the TAE model is compared with IQR and MAD models in the virtual machine processing time, resource utilization and energy consumption by using the cloud computing simulation platform(CloudSim). The experimental results showed that the TAE model is superior to the IQR and MAD models in the above three aspects.

References

[1] A. Dalvandi, M. Gurusamy, K. C. Chua: *Power-efficient resource-guaranteed VM placement and routing for time-aware data center applications.* The International Journal of Computer and Telecommunications Networking *88* (2015), No. C, 249–268.

[2] G. Y. Du, H. He, Q. G. Meng Q.G: *Energy-efficient scheduling for tasks with deadline in virtualized environments.* Mathematical Problems in Engineering (2014), No. 5, ID 496843.

[3] A. Dalvandi, M. Gurusamy, K. C. Chua: *Time-aware VMFlow placement, routing, and migration for power efficiency in data centers.* IEEE Transactions on Network and Service Management *12* (2015), No. 3, 349–362.

[4] Q. H. Nguyen, T. Nam: *Minimizing total busy time for energy-aware virtual machine allocation problems.* International Symposium on Information and Communication Technology (SoICT), 3–4 Decenber 2015, Hue City, Vietnam, ACM New York, USA 900–905.

[5] A. Q. Nguyen, P. Bouvry, E. G. Talbi: *A new model for VMMP dealing with execution time uncertainty in a multi-clouds system.* Proc. IEEE International Conference on Cloud Networking (CloudNet), 5–7 October 2015, Niagara Falls, ON, Canada, IEEE Conference Publications 165–167.

[6] R. Begam, D. Zhu: *Time-sensitive virtual machines provisioning and resource allocation in clouds.* Proc. IEEE 17th International Conference on High Performance Computing and Communications, 24–26 August 2015, New York, USA, 660–665

[7] B. Wang, Y. Song, Y. Sun, J. Liu: *Managing deadline-constrained ag-of-tasks jobs on hybrid clouds.* Proc. 24th High Performance Computing Symposium 3–5 April 2016, San Diego, CA, USA, paper 22.

[8] J. James, B. Verma: *Efficient VM load balancing algorithm for a cloud computing environment.* International Journal on Computer Science and Engineering, *4* (2012), No. 9, 1658–1663.

[9] M. Shalom, A. Voloshin, P. W. H. Wong, F. C. C. Yung, S. Zaks: *Online optimization of busy time on parallel machines.* Theoretical Computer Science *560, Part 2* (2014), 190–206.

[10] W. H. Tian, Q. Xiong, J. Cao: *An online parallel scheduling method with application to energy-efficiency in cloud computing.* The Journal of Supercomputing, *66* (2013), No. 3, 1773–1790.

[11] Q. T. Nguyen, Q. H. Nguyen, H. T. Nguyen, H. T. Van, T. Nam: *Virtual machine allocation in cloud computing for minimizing total execution time on each machine.* Proc. IEEE International Conference on Computing, Management and Telecommunications (ComManTel), 21–24 January 2013, Ho Chi Minh City, Vietnam, 241 to 245.

[12] X. Sui, L. Li, D. Liu, H. W. Yang, X. Di: *A loading balance model of virtual machine live migration in cloud computing environment.* Acta Technica, *61* (2016), No. 4B, 261 to 270.

[13] N. K. Chien, N. H. Son, H. D. Loc: *Load balancing algorithm based on estimating finish time of services in cloud computing.* Proc. 18th International Conference on Advanced Communication Technology (ICACT), 31 January–3 February 2016, Pyeongchang, South Korea, 228–233.

Received April 30, 2017