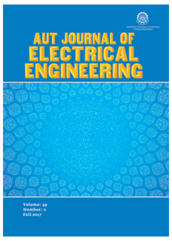


AUT Journal of Electrical Engineering

AUT J. Elec. Eng., 50(1)(2018)59-66
DOI: 10.22060/eej.2017.13284.5150

Reduction of Energy Consumption in Mobile Cloud Computing by Classifying of Demands and Executing in Different Data Centers

H. Yeganeh[1], A. Salahi[2*], M. A. Pourmina[3]

[1] Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2] Iran Telecommunication Research Center, Tehran, Iran
[3] Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

ABSTRACT: In recent years, mobile networks have faced with the increase of traffic demand. By emerging mobile applications and cloud computing, Mobile Cloud Computing (MCC) has been introduced. In this research, we focus on the 4th and 5th generation of mobile networks. Data Centers (DCs) are connected to each other by high-speed links in order to minimize delay and energy consumption. By considering a model of the geographical distribution of DCs which uses a wideband optical network, renewable energy and sharing resources for new generations of mobile networks, the real effect of issues on the consumed energy, cost, and profit in the mobile cloud computing are investigated. We derived a penalty function for cost and then by using Lyapunov optimization theorem; we designed an algorithm to minimize the average cost of energy consumption based on the online information in MCC. The time average cost is at most O(1/V) above the optimum target, while the average queue size is O(V). The parameter V can be tuned to make the time average cost as close to (or below) the optimum as desired. We designed three scenarios and two classes of applications to set up our simulation environment. The provided results illustrate the efficiency of our proposed scheme and validate the mathematical model.

Review History:
Received: 15 August 2017
Revised: 21 October 2017
Accepted: 6 November 2017
Available Online: 15 December 2017

Keywords:
Mobile Cloud Computing
Lyapunov Optimization
Energy Consumption
Penalty Function
Demand Classification

1- Introduction

In recent years, users of operators have preferred to receive the same services from a wireless environment as those received from fixed networks. The best solution is integrating wireless systems with the fixed ones. The main common goal of all researchers on the 4th and 5th generations of mobile technologies is an unlimited number of things that can communicate with each other with high communication capacity and quality of service [1]. In this paper, we focus on mobile cloud computing [2] in the 4th and 5th generations of mobile networks.

Furthermore, the DCs can be connected to each other with high-speed links. Among the problems related to DCs which are based on cloud computing, the cost of electricity is noteworthy. In the following, we will refer to cooling methods that can reduce the cost of cooling systems. DCs should take steps towards automation, and right decisions must be made in order to control the turning on and off of the servers [3]. To reduce the energy consumption of servers, the new online methods that do not need to predict the future conditions are used. In addition to these issues, the use of new energy sources such as wind and solar energy has a vital role in reducing the cost [4]. As a result, it would be better to transfer DCs to the cold climates or use new technologies to reduce the cost of cooling. In the proposed method, we divide the demands that are sent to the cloud in two classes, namely, real-time class and non-real-time class. Furthermore, since the quality of service in the 4th and the 5th generations is very important, the classification of demands and giving priority to them are also considered in the system model.

The outline of the paper is as follows. Section 2 discusses some related studies. In section 3, an overview of conventional and system architecture is given. Architecture is analyzed and compared with the state-of-the-art in section 4. The mathematical formulation for the two DCs is described throughout this section. The performance evaluation and simulation results are presented in section 5. Section 6 concludes the paper and propose possible future works.

2- Related Works

In recent years, many studies have been dedicated to the energy efficiency in the cloud computing [5]. In new generations of mobile, cloud computing in radio access networks has been proposed. The MCC provides computing resources for mobile devices in the cloud [6]. This architecture not only runs on available wireless networks but also is an essential part of the 4G and 5G networks [7]. In [8], minimizing energy consumption for MCC systems with off-loading computation has been considered. In this paper, a mobile user makes a decision on the amount of demands that should be transferred to the cloud to minimize the energy consumption of mobile devices. In [9], the leverage of cloud computing on poor resources of mobile devices is presented. In particular, mobile applications can be run on mobile devices or transferred to the cloud to save energy in a mobile device. In [10], efficient energy consumption is addressed and it has been suggested that energy reduction should be applied to all layers of DCs. Furthermore, the quality of service and reduction of energy consumption have been considered at the same time. In [11], the use of green energy (including renewable energy) instead of brown energy

The Corresponding Author; Email: salahi@itrc.ac.ir (energy production from fossil sources) is considered. In [12], power proportionality which means the ratio of turning servers on and off to the workloads dynamically is considered as a way to reduce energy consumption in DCs. In [13- 15] along with dynamic bandwidth allocation, geographically load balancing, and less response time have been introduced. The UPS and diesel generators for producing electricity are used to reduce costs in the DCs, this excess energy can be used in the peak time period when the cost of electricity is high [16]. Since a main part of the cost is spent on cooling equipment, relocation of DCs to the colder climates results in a reduction in power consumption [17]. In [18], it is stated that the power consumption is a significant part of operational costs in the DCs, and operators want to reduce their electricity bills as much as possible. The Lyapunov optimization technique is used to keep balancing explicitly between cost saving and stored energy. In [19, 20], an optical fiber for data communication has been proposed. The energy consumption of optical networks is much lower than the other transmission networks. Saving energy and reducing the environmental pollutions in the information technology industry are included in the green technology that has been addressed in [21].

Furthermore, dynamic pricing for many applications is used as a tool to improve the performance of resource management. In [22, 23], a dynamic pricing algorithm for the users of cloud computing is proposed to increase fairness by alignment and proper allocation of resources. An excellent example of dynamic pricing in wireless cloud computing has been used for the congestion control that has been investigated in [24]. In [3, 25], the pricing and the scheduling of workload in the mobile cloud computing are considered simultaneously, and Lyapunov optimization is used for the queue control.

3- Overview of System Model

3- 1- System Model

Basically, the main goal of mobile service providers is increasing their long-term profit and the level of satisfaction of their customers. We prove that this could be achieved by optimizing the energy consumption of the DCs. However, as mentioned in [8], when the demands of mobile users are transferred to the cloud, the energy consumption of mobile devices and the DCs is optimized, and the level of satisfaction of users increases. Due to the use of fiber-optic network infrastructure, costs of data transmission and switching are low when one compares it with the traditional networks. The system model has been shown in Fig 1. In addition, a list of all parameters that are used in our paper is presented in Table1.

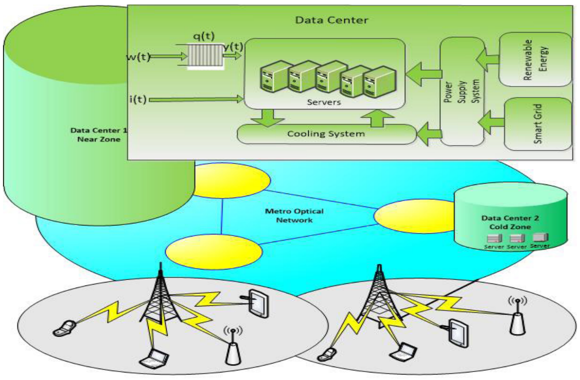


Fig. 1. System Model

In this model, two DCs are considered for mobile cloud computing. A group of servers are implemented in the regions which are closer to the users and mainly provide real-time applications; the other one is placed in cold regions and is used for non-real-time applications.

Table 1.List of all parameters and their definitions throughout the paper

$i(t)$	Number of real-time demands in a time slot
$w(t)$	Number of non-real-time demands in a time slot
$q(t)$	Length of queued non-real-time demands in a time slot
$y(t)$	Number of processed demands in a time slot
V	Trade-off parameter between a penalty and Lyapunov's drift
$P(t)$	Cost of energy consumption
$f(t)$	Normalized number of i(t) or w(t) in a time slot
μ, μ'	A fixed number that depends on the chiller's structure
E	A coefficient that maps f(t) on energy consumption
$\mathrm{e}_c(t)$	Energy consumption of cooling system in a time slot
$\mathrm{e}_p(t)$	Energy consumption for processing demands in a time slot
$\mathrm{e}_r(t)$	Renewable (solar and wind) energy in a time slot
$p_i(t)$	Service price for real-time demands in a time slot
$p_w(t)$	Service price for non-real-time demands in a time slot
$u(w(t), t)$	Utility function for each user in a time slot
$p_e(t)$	Price of electricity in a time slot
$C(t)$	Cost of energy in a time slot
$R(t)$	Profit of operator in a time slot
X	Capacity of data center
$e_{tr}(t)$	Energy consumption of transferring demands to DC2 in a time slot
θ	A coefficient that maps transmission cost to the processing cost
Ω	A coefficient that maps transmission delay to the processing delay
κ	A coefficient that maps queue length to delay
$\delta_k(t)$	The coefficient of demand status in each time slot
$\lambda_k(t)$	A relative factor for kth class of non-real-time demands
$e_{pi}(t)$	Consumed energy for processing real-time demands
$e_{py}(t)$	Consumed energy for processing non-real-time demands
$e_{cy}(t)$	Consumed energy for cooling real-time demands
$e_{ci}(t)$	Consumed energy for cooling non-real-time demands
$e_r(t)$	Amount of renewable energy in DC

The demands of users are classified into two categories. The first one is $i(t)$, that is used for real-time demands such as data mining, virtual searching, online audio and video services. The second one is $w(t)$ that is used for non-real time demand such as file transfer, remote login, and web services. We assume that a major part of the energy in a DC is consumed by the cooling system of server and storage (about 76% [26]), that keeps the servers at an appropriate temperature. In fact, to reduce the cost of energy consumption for cooling, we have placed another DC in a cold region. The demands that are not sensitive to delay, ($w(t)$ in Fig. 1) could be sent to the DC in the cold regions in order to be processed at a lower cost. Because the cost of transferring demands in the proposed infrastructure is less than that of cooling energy consumption.

3- 2- Lyapunov Optimization

The Lyapunov's drift is very critical in the optimal control of queues. Indeed, queue stability is achieved by optimizing performance-related objectives such as minimizing energy consumption and maximizing efficiency [27, 28].

If we consider N queues with length $q_k(t)$ where k=1,2,...,N at discrete time slot t ϵ{0, 1, 2, ...}, Lyapunov function of grade 2 is defined by

$$L(\mathrm{t}) = \frac{1}{2}\sum\nolimits_{k=1}^{N} q_k^2(t) \tag{1}$$

Lyapunov's drift in a time slot is defined as follows.

$$D(\mathrm{t}) = L(t+1) - L(t) \tag{2}$$

where

$$q_k(t+1) = \max(q_k(t) + w_k(t) - y_k(t), 0)$$

A control law should be designed to minimize the bound of the queue in a time slot. Adding a weighted penalty ($V.P(\mathrm{t})$) to the drift and minimizing the formula (3) lead to a drift-plus-penalty algorithm that is useful to achieve system stability and minimizes the penalty at the same time. The drift-plus-penalty algorithm is defined as:

$$DPP(\mathrm{t}) = D(t) + V.P(\mathrm{t}) \tag{3}$$

Since D(t) and P(t) do not have the same dimension, we need the constant V to make the second term have the same dimension as D(t) to be able to add them together, and we can control the penalty function, with the drift function. The goal is to keep queue stable by minimizing $P(t)$ in a time slot [29]. If we consider $R(t)$ as the profit, maximizing $R(t)$ is equivalent to minimizing $P(t)$. By considering that V> 0 and defining $P(t)$ as the negative of profit in each time slot, the drift plus penalty algorithm can be used to minimize the average energy consumption by restricting the queue size [30]. This algorithm would be appropriate for both flow control and network stability. The value of V can be adjusted in a way that the average of penalty is very close to optimum [31]. This idea has been used throughout the paper. By increasing the V, the delay grows up and as a result; more profit is provided for operators.

$$DPP(t) = D(\mathrm{t}) - VR(t) \tag{4}$$

4- Problem Formulation

4- 1- Determining Cost, Profit and Service Pricing Functions

We assume that $e_c(t)$ and $e_p(t)$ increase linearly with the number of demands. This linear relationship is logically derived from experimental measurements [32, 33]. It is considered that the DCs, include servers that are similar and the servers have a normalized processing speed and energy consumption.

$$e_c(t) = E.\mu.f(\mathrm{t}) \tag{5}$$

$$e_p(t) = E.f(\mathrm{t}) \tag{6}$$

This model has been widely applied to the management of DCs [33, 34]. In order to reduce the complexity of computations, the costs of turning servers on/off are not considered. Hence the cost of energy is defined by $\mathrm{C}(t) = c(p_e(t), e(t))$ in which is equal to the purchased energy from electricity companies. It is assumed that $p_i(t) = \mathrm{p}_i$. The pricing is fulfilled for non-real-time demands and we have $p_w(t) \in [0, p_{\max}]$. Let the price be kept constant in each period. C(t) and R(t) are defined by

$$e_i(t) = e_{pi}(t) + e_{ci}(t) = E_i(i(t) + \mu i(t)) \tag{7}$$

$$e_y(t) = e_{py}(t) + e_{cy}(t) = E_y(\mathrm{y}(t) + \mu y(t)) \tag{8}$$

$$C(\mathrm{t}) = p_e(\mathrm{t})(\mathrm{e}_i(t) + e_y(\mathrm{t}) - e_r(t)) \tag{9}$$

$$\begin{gathered} R(\mathrm{t}) = p_i.i(\mathrm{t}) + p_w(\mathrm{t}).w(\mathrm{t}) - C(t) = \\ p_i.i(\mathrm{t}) + p_w(\mathrm{t}).w(\mathrm{t}) - p_e(\mathrm{t})(E_i(1+\mu)i(t) + E_y(1+\mu)y(t) - e_r(t)) \end{gathered} \tag{10}$$

It is assumed that the existing network connection between the DC and the base station (via a wideband backbone network) is not a data transmission bottleneck between the DC and users. The utility function is a time-varying function that increases as $w(t)$ increases. Here, $w(t)$ is a response to $p_w(t)$ and it is determined by the service provider. The optimal logarithmic utility function has been studied in prior studies [24].

$$u_k(w_k(\mathrm{t}), \mathrm{t}) = \delta_k(t)\log(1 + \sum_{k=1}^{N} \lambda_k w_k(t)) \tag{11}$$

The service provider pays a cost equal to $p_{wk}(t)$ for kth class, that results in a maximum utility equal to:

$$\max\,(\mathrm{u}_k(w_k(\mathrm{t}), \mathrm{t}) - \sum_{\mathrm{j}=1}^{n} \mathrm{p}_{\mathrm{wk}}(\mathrm{t}) w_k(\mathrm{t})) \tag{12}$$

4- 2- Two DCs with Classification

In this section, we assume that non-real-time demands have different QoS (Quality of Service) and the demands are divided into two categories, namely, Class 1 and Class 2. Class 1 has a higher priority and is more sensitive to the delay. The demands of class 2 are given the second priority for processing. Furthermore, it is necessary to consider how these two different classes can affect the average of service provider's profit.

$$e_y(t) = e_{py}(t) + e_{cy}(t) = E_y(1+\mu)\sum_{k=1}^{N} y_k(t) \tag{13}$$

$$\begin{gathered} C(\mathrm{t}) = p_e(\mathrm{t})(\mathrm{e}_i(t) + e_y(\mathrm{t}) - e_r(t)) = \\ p_e(t)(E_y(1+\mu)\sum_{k=1}^{N} y_k(t) + E_i(1+\mu)\mathrm{i}(t) - e_r(t)) \end{gathered} \tag{14}$$

$$\begin{gathered} R(t) = p_i.i(t) + p_w(t).w(t) \text{-} C(t) = \\ p_i i(t) + \sum_{k=1}^{N} p_{wk}(t) w_k(t) \text{-} p_e(t)(E_y(1+m)\sum_{k=1}^{N} y_k(t) + E_i(1+m)i(t) \text{-} e_r(t)) \end{gathered} \tag{15}$$

$$DPP(t) = D(\mathrm{q(t)}) - VR(t) \le M + \sum_{k=1}^{N} q_k(t)(\mathrm{w}_k(t) - \mathrm{y}_k(t)) - \mathrm{VR(t)} \tag{16}$$

Where, M is positive constant with upper bound: $M \ge \frac{1}{2}\sum_{k=1}^{N}(w_k(t) - y_k(t))^2$. Then,

$$\begin{gathered} DPP(t) \le M + \sum_{k=1}^{N} q_k(t)(\mathrm{w}_k(t) - \mathrm{y}_k(t)) \\ -\mathrm{V}(p_i\,\mathrm{i}(t) + \sum_{k=1}^{N} p_{wk}(t)\,\mathrm{w}_k(t) - p_e(t)(E_y(1+\mu)\sum_{k=1}^{N} y_k(t) + E_i(1+\mu)\mathrm{i}(t) - e_r(t)) \end{gathered} \tag{17}$$

DPP optimization problem that should be solved is summarized as

$$\min(\sum_{k=1}^{N} q_k(t)(w_k(t)-y_k(t)) - V(p_i i(t) + \sum_{k=1}^{N} p_{wk}(t)w_k(t)$$
$$-p_e(t)(E_y(1+\mu)\sum_{k=1}^{N} y_k(t)+E_i(1+\mu)i(t)-e_r(t)), \tag{18}$$
$$\text{s.t.}\quad e_r(t)£(1+\mu)(E_i i(t)+E_y\sum_{k=1}^{N} y_k(t)).$$

The drift-plus-penalty algorithm is used in linear programming as well as convex optimization [29, 35]. Thus, solving (19) is based on linear programming by choosing $y_k(t)$ and $p_{wk}(t)$ in each time slot, independently. To calculate the cost and profit in the long term, a formula is required to subtract the entire cost from all revenues. For each $T \in \mathbb{Z}^+$ and $M \in \mathbb{Z}^+$ there is a $T_{end} = MT$ so that

$$\overline{C^*} = \frac{1}{T_{end}} \sum_{t=0}^{T_{end}-1} C^*(t) \tag{19}$$

$$\overline{R^*} = \frac{1}{T_{end}} \sum_{t=0}^{T_{end}-1} R^*(t) \tag{20}$$

In the above equation, $C^*(t)$ and $\mathrm{R}^*(t)$ are the optimal cost and profit through DPP algorithm. In order to process these categories properly, we focus on three different scenarios.

4- 2- 1- First Scenario

In this approach, the real-time demands and Class 1 demands (higher QoS of non-real-time demands) are processed in DC1 and the rest of non-real-time demands are processed in DC2. In this regard, adequate servers in DC1 are allocated to real-time applications and then through DPP algorithm, the amount of processing and the cost of services are determined. In DC2, only Class 2 demands are processed and the needed servers, as well as the cost of services, are optimized. Total costs can be computed according to the following equation.

$$e_{tr}(t) = \theta E_y\, \mathrm{y}_2(t) \tag{21}$$

$$\mathrm{C(t)} = p_e(t).[(\mathrm{E}_y((1+\mu)\,\mathrm{y}_1(t)+(1+\mu')\,\mathrm{y}_2(t)))+e_{tr}(t)+E_i(1+\mu)\mathrm{i}(t)-e_{r1}(t)-\mathrm{e}_{r2}(\mathrm{t})] \tag{22}$$

In the above equation, $e_{rk}(t)$ (k = 1, 2) represent the amount of renewable energy in the first and second DC. Because of locating DC2 in cold regions, energy consumption for cooling DC2 is lower ($\mu' < \mu$). To calculate the delay, the transferring time of non-real-time demands to DC2, is also considered. Finally, by calculating the amount of revenue resulted from the price of services and subtracting the total cost from it, the amount of profit is achieved.

$$R(\mathrm{t}) = p_i\,\mathrm{i}(t) + p_{w1}(t)\,\mathrm{w}_1(t) + p_{w2}(t)\,\mathrm{w}_2(t) -$$
$$p_e(t).[(\mathrm{E}_y((1+\mu)\,\mathrm{y}_1(t)+(1+\mu')\,\mathrm{y}_2(t))+\theta y_2(t))+E_i(1+\mu)\mathrm{i}(t)-e_{r1}(t)-\mathrm{e}_{r2}(\mathrm{t})] \tag{23}$$

The equation (25) shows the total delay is proportional to the queue length.

$$Delay(t) = \kappa(\sum_{i=1}^{t} q_1(\mathrm{i}) + \sum_{i=1}^{t} q_2(\mathrm{i}) + \omega\sum_{i=1}^{t} q_2(\mathrm{i})) \tag{24}$$

4- 2- 2- Second Scenario

In this approach, real-time demands are processed in DC1, and non-real-time demands of Class 1 and 2 are processed in DC2. As a result, applications that require real-time processing are not transferred, but applications that are less or non-sensitive delay are transferred to DC2 to impose a lower cooling cost. At first, the applications of Class 1 are processed and then the remaining capacity is allocated to Class 2 applications in DC2. The following equations show the total cost, profit, and delay in this scenario, respectively.

$$\mathrm{C(t)} = p_e(t).[(\mathrm{E}_y((1+\mu')\,\mathrm{y}_1(t)+(1+\mu')\,\mathrm{y}_2(t))+\theta(\mathrm{y}_1(t)+\mathrm{y}_2(t)))$$
$$+E_i(1+\mu)\mathrm{i}(t)-e_{r1}(t)-\mathrm{e}_{r2}(\mathrm{t})] \tag{25}$$

$$R(\mathrm{t}) = p_i\,\mathrm{i}(t) + p_{w1}(t)\,\mathrm{w}_1(t) + p_{w2}(t)\,\mathrm{w}_2(t) -$$
$$p_e(t).[(\mathrm{E}_y((1+\mu')\,\mathrm{y}_1(t)+(1+\mu')\,\mathrm{y}_2(t))+\theta(\mathrm{y}_1(t)+\mathrm{y}_2(t)))+E_i(1+\mu)\mathrm{i}(t)-e_{r1}(t)-\mathrm{e}_{r2}(\mathrm{t})] \tag{26}$$

$$Delay(t) = \kappa(\sum_{i=1}^{t} q_1(\mathrm{i}) + \sum_{i=1}^{t} q_2(\mathrm{i}) + \omega\sum_{i=1}^{t} q_1(\mathrm{i}) + \omega\sum_{i=1}^{t} q_2(\mathrm{i})) \tag{27}$$

4- 2- 3- Third Scenario

In this scenario, real-time demands are processed in DC1 and non-real-time demands of Class 1, and 2 are processed in DC2. The dedicated capacity of servers to Class 1 and Class 2 is optimized simultaneously. As a result, the equations of cost, profit and delay are similar to the second scenario, and the difference is the optimizing method of $y_1(t)$ and $y_2(t)$.

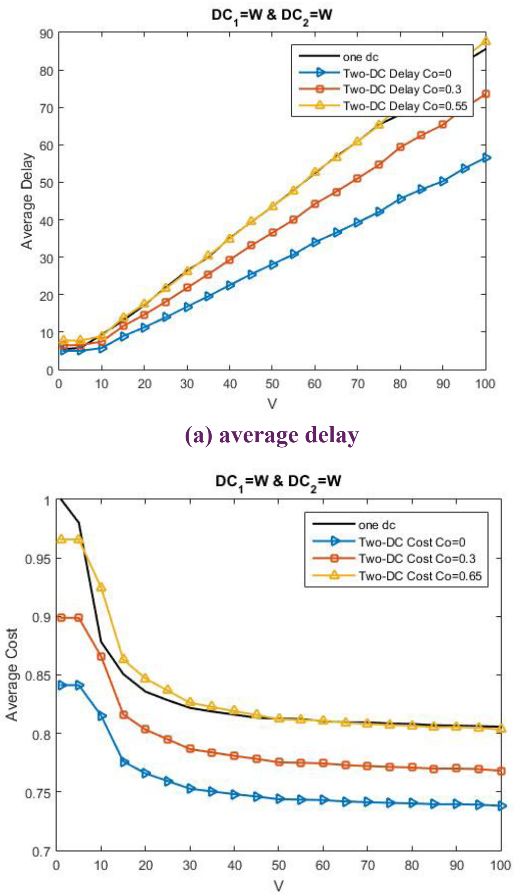
5- Simulation Results

In the beginning of each time slot, we use environmental information $p_e(t)$, *i(t)*, $e_r(t)$ and find the $w_k(t)$ from the utility function defined in (12). For time t = 0, we assume that the $p_{wk}(t) = P_m$ (maximum price), in order to find the $w_k(t)$ for the first time. We consider a time period up to 2000 time slots. In addition to the random data, we re-scale real data from California, America's resources for electricity and renewable-energy cost [36]. Suppose that the maximum number of servers *(W)* in each DC is 10. We assume that the cooling system uses 0.75 of the energy of servers. It is assumed that $\delta(t)$ is independent and distributed uniformly in the interval [0, 1] and *i(t)* is considered as a random variable with a uniform distribution at various intervals. It should be mentioned that this analysis can be proven with any other settings.

5- 1- Two DCs without classification

Since a part of the consumed energy of servers is used for cooling, placing a DC in the cold regions can reduce energy consumption. It cannot improve the profit of the service provider if delays and costs due to the transmission of non-real-time demands for DC2 exceed from a certain level. Therefore, it is necessary to compare the delay and cost of two-DCs regarding to one-DC. In the following, simulation results of delay and cost by changing the capacity of the first DC are presented. In all simulations, the capacity of one-DC approach is equal to the total capacity of the two-DC approach.

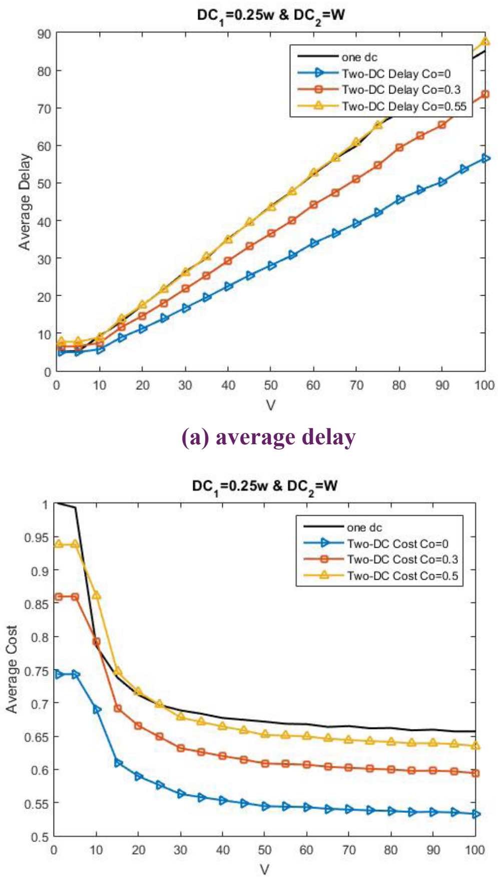
Fig. 2(a) shows the average delay for both one-DC and two-DC approaches with different delay coefficients (Delay Co). As can be seen, with using two DCs with small values of the coefficient, delay reduces. However, if transmission delay increases, using two DCs cannot lead to better results in delay. Fig 2(b) shows that when the cost of the transmission is less than a threshold, using two DCs reduces the cost, significantly. Fig. 3 shows the average delay and cost in terms of *V* for both approaches (X1=0.25X2) with different coefficients. In this case, because of low capacity for real-time demands and less revenue, we reach the same results with fewer network cost coefficients.



(a) average delay

(b) average cost

Fig. 2. The average delay and cost in terms of V, (X1=X2=W) for different coefficients



(a) average delay

(b) average cost

Fig. 3. The average delay and cost in terms of V, (X1=0.25X2) for different coefficients

5- 2- Two DCs with classification

In the previous section, it was supposed that the demands have the same priority for processing, and we considered the same QoS for non-real-time demands. In this section, we assume that the demands have different QoS and are divided into two categories, namely, Class 1 and Class 2. At first, by considering the DC, the effect of classification on the price of services as well as the delay of each class is investigated.In Fig. 4, the average price of service and average delay in two classes in the case of $i_{\max}=8$ have been compared. Clearly, the class 1 services with a lower delay have the higher price, compared to class 2. Fig. 5 and Fig. 6 show the average delay, cost and profit for $i_{\max}=2$ and $i_{\max}=8$ in different scenarios, respectively. It can be seen that initially (for small $i_{\max}$) the delay of the third scenario is larger than other scenarios', but it is less than the delay of one-DC method. In one-DC method, because all the processing is carried out inside a single DC and the electricity price is compared with a smaller threshold, the amount of processing is low. Thus, the amount of processing is low and, the delay is more than other scenarios. In the third scenario, $y_1(t)$ and $y_2(t)$ are optimized in the second DC, simultaneously. These values are determined so that the profit would be maximized. As a result, the profit of the third scenario is more than that of the other scenarios as well as single DC mode although, the third scenario imposes more delay, to increase the profit. In this regard, the third scenario has a higher delay compared to the other scenarios.

As can be seen, by increasing $i_{\max}$, the profit of the third scenario decreases gradually. In this case, due to a reduction in residual capacity to optimize profit, the allocation of all capacities to the Class 1 demands and, then, to the class 2 demands is more optimal than simultaneously optimizing two classes. Thus, in the case of $i_{\max}=8$, the second scenario has a higher profit for the mobile service provider than the other scenarios.

Fig. 7 and Fig. 8 compare the second and third scenarios in terms of demands, amount of processing and queue length for $i_{\max}=2$ and $i_{\max}=8$.

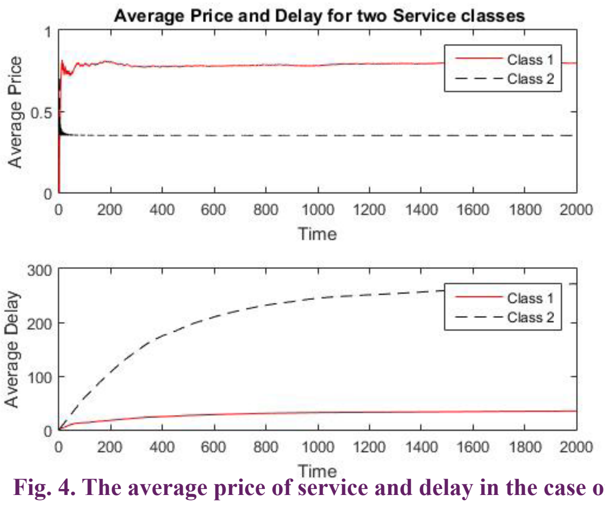


Fig. 4. The average price of service and delay in the case of classification with $i_{max}= 8$

6- Conclusion

In this paper, we found a new criteria for the 4th and 5th generation of mobile networks that adopt cloud computing for their uses. Furthermore, the delay, cost, and profit of service providers are investigated. A novel approach was proposed to transmit the demands for subscribers to the cloud. We focused on the processing of the real-time demands that are delay-sensitive in closer DC and also the processing of non-real-time demands that are less or non-delay-sensitive in the DC placed in cold regions. The mobile subscribers are modeled with their demands and, the demands are affected by deciding on price of service. To decrease the long-term cost and to increase the profit for service providers, we used drift-plus-penalty algorithm that can be implemented based on the online information. For real approach, the idea of classification of demands was used by using one and two DCs. We observed that the classification decreases the amount of cost and increases the profit substantially. Moreover, the results of using two DCs, with and without classification, to process non-real-time demands showed that when the amount of transmission delay is low, the cost of the service provider decreases significantly. It was shown that even by decreasing real-time demands as well as the capacity of near-zone DC, the transferring of non-real-time demands to the cold-zone can decrease delay and cost compared to that using one DC method. Extensive experimental results show that the proposed algorithm decreases the amount of average cost and increases the average profit substantially.

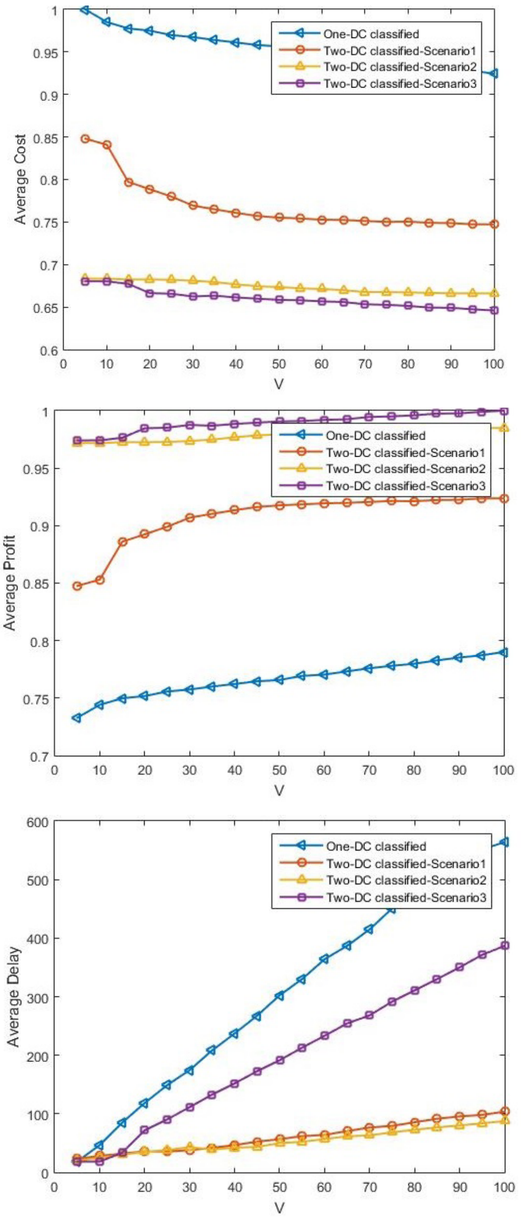


Fig. 5. The average delay, cost and profit for i_{max}= 2 in different scenarios

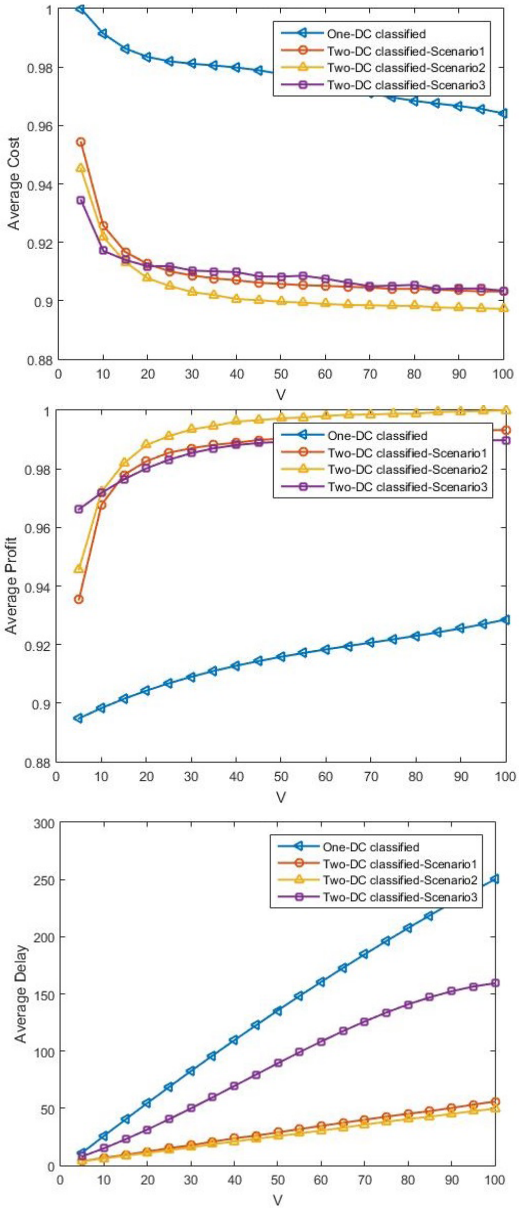
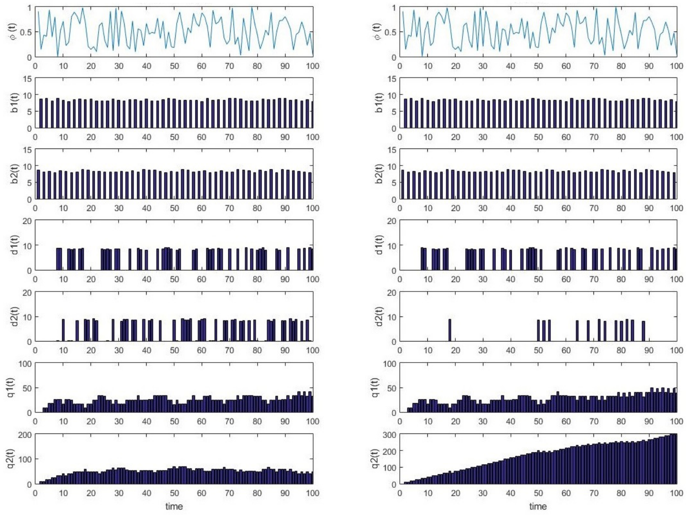
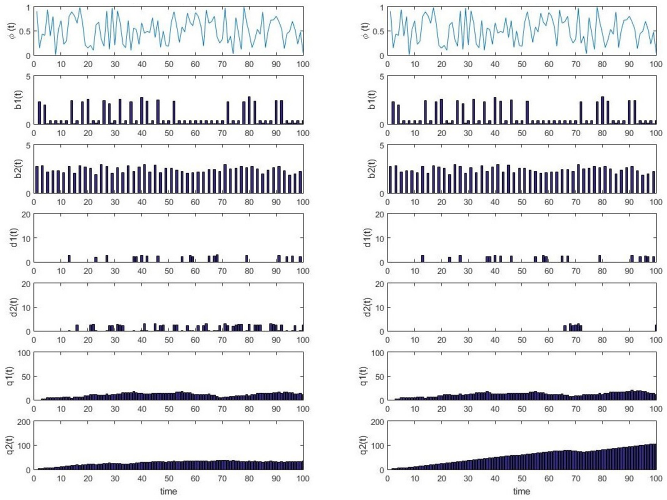


Fig. 6. The average delay, cost and profit for i_{max}= 8 in different scenarios



(a) Second Scenario (b) Third Scenario

Fig. 7. A comparison between the second and third scenarios for $i_{max}= 2$



(a) Second Scenario (b) Third Scenario

Fig. 8. A comparison between the second and third scenarios for $i_{max}= 8$

References

[1] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini, T. Braun, Network slices toward 5G communications: Slicing the LTE network, IEEE Communications Magazine, 55(8) (2017) 146-154.

[2] M. Hogan, F. Liu, A. Sokol, J. Tong, Nist cloud computing standards roadmap, NIST Special Publication, 35 (2011) 6-11.

[3] S. Ren, M. van der Schaar, Dynamic scheduling and pricing in wireless cloud computing, IEEE Transactions on Mobile Computing, 13(10) (2014) 2283-2292.

[4] I.B. Software, Workload automation: Helping Cloud Computing Take Flight, in, http://documents.bmc.com/products/documents/62/56/286256/286256.pdf,04.02.2017

[5] A. Hameed, A. Khoshkbarforoushha, R. Ranjan, P.P. Jayaraman, J. Kolodziej, P. Balaji, S. Zeadally, Q.M. Malluhi, N. Tziritas, A. Vishnu, A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems, Computing, 98(7) (2016) 751-774.

[6] B.A. Bjerke, LTE-advanced and the evolution of LTE deployments, IEEE Wireless Communications, 18(5) (2011).

[7] I.F. Akyildiz, D.M. Gutierrez-Estevez, E.C. Reyes, The evolution to 4G cellular systems: LTE-Advanced, Physical communication, 3(4) (2010) 217-244.

[8] R. Ferzli, I. Khalife, Mobile cloud computing educational tool for image/video processing algorithms, in: Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE, IEEE, 2011, pp. 529-533.

[9] Y. Wen, W. Zhang, H. Luo, Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones, in: INFOCOM, 2012 Proceedings IEEE, IEEE, 2012, pp. 2716-2720.

[10] P.J. Havinga, G.J. Smit, Energy-efficient wireless networking for multimedia applications, Wireless communications and mobile computing, 1(2) (2001) 165-184.

[11] Z. Liu, M. Lin, A. Wierman, S.H. Low, L.L. Andrew, Greening geographical load balancing, in: Proceedings

of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems, ACM, 2011, pp. 233-244.

[12] B. Guenter, N. Jain, C. Williams, Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning, in: INFOCOM, 2011 Proceedings IEEE, IEEE, 2011, pp. 1332-1340.

[13] S.W. Paper, The Seven Standards of Cloud Computing Service Delivery, in, https://www.salesforce.com/assets/pdf/datasheets/SevenStandards.pdf,07.02.2017.

[14] I.T.C.C. Standards, Cloud Computing Standards: Overview and ITU-T positioning, in, https://www.itu.int/dms_pub/itu-t/oth/06/5B/T065B00001C0043PDFE.pdf,10.09.2016.

[15] M. Lin, Z. Liu, A. Wierman, L.L. Andrew, Online algorithms for geographical load balancing, in: Green Computing Conference (IGCC), 2012 International, IEEE, 2012, pp. 1-10.

[16] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, B. Maggs, Cutting the electric bill for internet-scale systems, in: ACM SIGCOMM computer communication review, ACM, 2009, pp. 123-134.

[17] H. L. Barroso, U, The Data Center as a Computer, Morgan & Claypool 2009.

[18] Y. Guo, Z. Ding, Y. Fang, D. Wu, Cutting down electricity cost in internet data centers by using energy storage, in: Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE, Ieee, 2011, pp. 1-5.

[19] M. Anastasopoulos, A. Tzanakaki, G. Zervas, B.R. Rofoee, R. Nejabati, D. Simeonidou, Virtualization over converged wireless, optical and IT elements in support of resilient cloud and mobile cloud services, in: European Conference and Exhibition on Optical Communication, Optical Society of America, 2012, pp. P5. 15.

[20] A. Tzanakaki, M.P. Anastasopoulos, G.S. Zervas, B.R. Rofoee, R. Nejabati, D. Simeonidou, Virtualization of heterogeneous wireless-optical network and IT infrastructures in support of cloud and mobile cloud services, IEEE Communications Magazine, 51(8) (2013) 155-161.

[21] E. Skejić, O. Dšindo, D. Demirović, Virtualization of hardware resources as a method of power savings in data center, in: MIPRO, 2010 Proceedings of the 33rd International Convention, IEEE, 2010, pp. 636-640.

[22] S. Irani, S. Shukla, R. Gupta, Online strategies for dynamic power management in systems with multiple power-saving states, ACM Transactions on Embedded Computing Systems (TECS), 2(3) (2003) 325-346.

[23] D.B. Rawat, C. Bajracharya, Software defined networking for reducing energy consumption and carbon emission, in: SoutheastCon, 2016, IEEE, 2016, pp. 1-2.

[24] P. Hande, M. Chiang, R. Calderbank, S. Rangan, Network pricing and rate allocation with content provider participation, in: INFOCOM 2009, IEEE, IEEE, 2009, pp. 990-998.

[25] C. Joe-Wong, S. Sen, S. Ha, M. Chiang, Optimized day-ahead pricing for smart grids with device-specific scheduling flexibility, IEEE Journal on Selected Areas in Communications, 30(6) (2012) 1075-1085.

[26] M. Dayarathna, Y. Wen, R. Fan, Data center energy consumption modeling: A survey, IEEE Communications Surveys & Tutorials, 18(1) (2016) 732-794.

[27] L. Tassiulas, A. Ephremides, Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks, IEEE transactions on automatic control, 37(12) (1992) 1936-1948.

[28] L. Tassiulas, A. Ephremides, Dynamic server allocation to parallel queues with randomly varying connectivity, IEEE Transactions on Information Theory, 39(2) (1993) 466-478.

[29] M.J. Neely, Stochastic network optimization with application to communication and queueing systems, Synthesis Lectures on Communication Networks, 3(1) (2010) 1-211.

[30] M.J. Neely, Energy optimal control for time-varying wireless networks, IEEE transactions on Information Theory, 52(7) (2006) 2915-2934.

[31] M.J. Neely, L. Huang, Dynamic product assembly and inventory control for maximum profit, in: Decision and Control (CDC), 2010 49th IEEE Conference on, IEEE, 2010, pp. 2805-2812.

[32] C.D. Patel, R.K. Sharma, C.E. Bash, M.H. Beitelmal, Energy flow in the information technology stack: Introducing the coefficient of performance of the ensemble, in: ASME 2006 International Mechanical Engineering Congress and Exposition, American Society of Mechanical Engineers, 2006, pp. 233-241.

[33] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, C. Hyser, Renewable and cooling aware workload management for sustainable data centers, in: ACM SIGMETRICS Performance Evaluation Review, ACM, 2012, pp. 175-186.

[34] M. Lin, A. Wierman, L.L. Andrew, E. Thereska, Dynamic right-sizing for power-proportional data centers, IEEE/ACM Transactions on Networking, 21(5) (2013) 1378-1391.

[35] M.J. Neely, Distributed and secure computation of convex programs over a network of connected processors, in: DCDIS Conf., Guelph, Ontario, Citeseer, 2005, pp. 285-307.

[36] California ISO- Todays Outlook in, http://www.caiso.com/Pages/TodaysOutlook.aspx , 04.05.2017.

Please cite this article using:

H. Yeganeh, A. Salahi, M. A. Pourmina, Reduction of Energy Consumption in Mobile Cloud Computing by Classifying of Demands and Executing in Different Data Centers, *AUT J. Elec. Eng.*, 50(1)(2018) 59-66.

DOI: 10.22060/eej.2017.13284.5150