Energy-Optimized Bandwidth Allocation Strategy for Mobile Cloud Computing in LTE Networks

Xiang Sun, Student Member, IEEE, and Nirwan Ansari, *Fellow, IEEE*
Electrical & Computer Engineering Dept., New Jersey Institute of Technology, Newark, NJ 07102, USA
E-mail: xs47@njit.edu, nirwan.ansari@njit.edu

***Abstract*—This paper presents a mobile cloud computing application model and addresses how to minimize the energy consumption for uploading *L* size of data load within the *T* delay constraint. We propose a bandwidth allocation strategy for a LTE network with homogeneous sub-channel condition. Our objective is to allocate more bandwidth to each UE when its relative channel condition becomes better. We formulate the UE's objective function as the sum of two penalty functions: channel condition penalty function which incentivizes base stations to minimize the energy consumption for every UE and Service Level Agreement (SLA) demand penalty function which guarantees *L* size of data load that can be uploaded in time. In the network scenario, we formulate the EnerGy Optimized (EGO) bandwidth allocation strategy as a linear programming model and solve it by the Simplex Method. Simulation results show that EGO can save energy of up to 60% for each UE and decrease the SLA violation rate in the network of up to 30% in comparison with the existing bandwidth allocation strategy in the uplink of the LTE network.**

***Keywords*—Bandwidth allocation, mobile cloud computing, energy optimal, Service Level Agreement, linear programming, simplex method**

I. INTRODUCTION

Nowadays with increasingly more capable smartphones, different kinds of applications are emerging rapidly. However, due to the limitation of battery life, CPU and memory resources [1][2], it is still difficult to run computational intensive applications, such as image processing, mobile learning [3][4], mobile gaming [5][6], on the smartphones. Emergence of the Mobile Cloud Computing (MCC) technology has created a new commercial opportunity for smartphone application markets. By exploiting MCC, more complicated and computational intensive applications can be facilitated on the smartphones, as MCC may save energy for the mobile sets and reduce the execution latency [7][8][9]. Recently, several MCC platforms, such as MAUI [10], CloneCloud [11] and Cloudlet [12], have been proposed. The basic idea of these platforms is to associate each mobile set with one or more Virtual Machines (VMs) in the data centers to accomplish its MCC applications; this mechanism is referred to application offloading.

The process of offloading can be simply characterized as: different UEs upload the MCC application data load to the data center via an access network; a number of VMs execute the applications and return back the results to UEs. Each UE's data load can be depicted as <*L, T*> [7][13][14], where *L* is the data load size and *T* is the application completion deadline. However, not all the applications are suitable for execution in the cloud. Some of them may save energy and time by executing them locally rather than remotely. Therefore, research described in [13][14][15][16] focuses on designing a criterion to decide whether an application should be run locally or remotely. However, running some MCC applications, such as face identification and voice recognition, is still not a viable option since these applications require huge information which cannot be possibly and necessarily stored in the mobile set.

In this paper, we focus on how to decrease the energy consumption of the UE when the MCC applications must be executed remotely or UE decides to offload the applications while satisfying the application's Service Level Agreement (SLA) in terms of the time delay constraint *T.* In the MCC offloading process, the main energy consumption of UE is to upload the application's data load from UE to Base Station (BS). The total energy consumed by UE depends on how much bandwidth is allocated by BS in different time slots. We formulate the bandwidth allocation problem within the framework of linear programming and the solution is a policy specifying how much bandwidth is assigned to each UE in every time slot.

The rest of this paper is organized as follows. Section II provides an overview of related works. Section III presents an energy consumption model of UE in the LTE network and UE objective function for minimizing energy consumption and satisfying the application's SLA. We formulate the bandwidth allocation strategy in the LTE network as a linear programming model. Section IV presents the performance evaluation results. Section V summarizes the paper and provides future directions.

II. RELATED WORKS

Previous works [13][14][17] have investigated the concept of offloading computation to decrease the energy consumption of a single UE. Zhang *et al.* [13] developed an optimal data transmission scheduling of a single UE in order to minimize the energy consumption of UE for transmitting *L* bits within the time constraint *T.* The Gilbert-Elliott channel model is adopted in the paper and the channel state is determined by a discrete state space Markov model. Lei *et al.* [14] offered an energy saving model to make an offloading decision of UE. Huang *et al.* [16] proposed a dynamic offloading algorithm, i.e., an optimal strategy for migrating different offloadable components from UE to the cloud, to save more energy while meeting the

This work was supported in part by the National Science Foundation (NSF) under Grant Division of Computer and Network Systems (CNS)-1218181.

time requirement of the applications.

Moreover, some literatures have studied bandwidth allocation among different UEs on the uplink of the LTE network. Madan and Ray [18] formulated a convex optimization problem to maximize the total utility of all UEs' data rates by allocating bandwidth to different UEs on the uplink of the LTE network. Pfletschinger *et al.* [17] proposed a computationally efficient subcarrier allocation algorithm for a multiuser OFDM system for uplink and downlink transmission. Huang *et al.* [19] proposed a maximum weighted sum rate resource allocation strategy for the uplink OFDM system by considering the heterogeneity of sub-channel conditions and discrete nature of sub-channel assignments.

As compared to these previous efforts, this paper presents several enhancements. First, we apply the LTE technology as an access network for MCC application's offloading. The application profile is different from the traditional applications in the LTE network for bandwidth sharing, and thus results in different formulations. Second, we provide the theoretical framework of bandwidth allocation in order to reduce energy consumption of all UEs in the LTE network rather than of a single UE.

III. System Model and Problem Formulation

A. Mobile Cloud Computing Application Model

We consider the MCC environment as shown in Fig.1 [20] in which the data transmission between UE and the remote server in a data center is facilitated through wireline and wireless connections. Specifically, wireless connections are facilitated by the LTE network. The wireline network enables communications among BSs, the application server, and servers in a data center. The application server receives the MCC application requests from UEs and assigns specific servers to different MCC applications according to some policy [21][22][23]. Once the communication link is established, UE can upload the application data load to the servers in the data center through BS.

As mentioned earlier, every MCC application profile can be depicted as <*L, T*> where *L* is the application data load and *T* is the time constraint for completing the whole offloading process, which can be divided into four steps and involves time consumed for transmitting the data load from UE to BS, T_{upload}; time consumed for transmitting the data load from BS to data center, T_{trans}; time consumed for data load processing by servers in the data center, $T_{process}$; time consumed for returning back the results from servers to UE via BS, T_{back}. We suppose that the connection between BS and servers in the data center [24][25] is a non-blocking link or the congested link can be quickly resolved, and so T_{trans} can be approximated as zero [20]. Also, as compared to uplink data load size *L*, we assume the size of MCC application results is much smaller; and therefore, T_{back} can be negligible [13]. Hence, the time constraint of offloading *T* is primarily decided by T_{upload} and $T_{process}$. T_{upload} is determined by the amount of bandwidth resource that BS allocates to UE, i.e., if UE pays more for its data plan to the LTE network provider (e.g., AT&T or Verizon), it would incur less time to transmit the specific data load. On the other hand, $T_{process}$ is determined by the amount of computing resources allocated to UE, i.e., if UE pays more to the data center provider (e.g., Amazon or Microsoft) for renting more VMs [26], it would acquire higher computing speed in terms of small $T_{process}$. Therefore, the MCC application profile can be separated into two parts according to different steps of offloading process: <*L*, T_{upload} > and <*L*, $T_{process}$ >. In the paper, we do not consider the data load processing step, and so we modify the MCC application profile as <*L*, T_{upload} > in the uploading step.

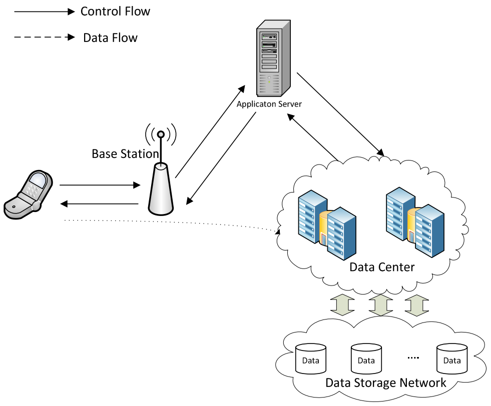


Fig. 1. Mobile cloud computing application model.

B. Energy consumption model of UE

The objective function of UE is to minimize the energy consumption for uploading *L* bits of data load while satisfying time constraint T_{upload}. We first discuss the energy consumption model of UE for the uplink transmission of the LTE network. If BS applies fractional power control [18] and we assume the sub-channel is homogenous, UE's transmission power P_{tx} in the Physical Shared Channel is [27]:

$$P_{tx} = \min\left(P_{\max}, P_0 + \alpha \times PL + 10\log_{10}\left(M\right)\right) \quad \text{[dBm]} \qquad (1)$$

where PL is the path loss from UE to BS, α is the path loss compensation factor, and P_0 is the power offset. Note that α and P_0 are the BS specific parameters, i.e., BS would broadcast the values to all the UEs for a time period. M is the number of Physical Resource Blocks (PRBs) allocated to UE. P_{max} is the maximum transmission power of UE. By converting the power units in Eq. (1) from dBm to mW, UE i's transmission power in each time slot becomes:

$$p_i^t = \frac{1}{Size_{PRB}} \times \varphi_i^t \times b_i^t \quad \text{[mW]}; \quad \forall t \,\&\, \mathrm{i}, \; p_i^t \le P_{\max} \qquad (2)$$

where p_i^t is the transmission power of UE i in time slot t, $Size_{PRB}$ is the amount of bandwidth in one PRB, b_i^t is the amount of bandwidth allocated to UE i, and $\varphi_i^t = 10^{\frac{P_0+\alpha\times PL_i^t}{10}}$. Here, we relax the constraint such that only an integer multiples of PRBs can be allocated to a UE and multiple users can share one sub-channel by time sharing [19].

Suppose the interference Power Spectrum Density (PSD) at the BS in every sub-channel is I_0 and noise PSD is N_0 (these two values can be measured by the BS periodically over unassigned resources [18]) and these two values are the same among different sub-channels since we assume sub-channels are homogenous. So, if UE transmits p_i^t of power over bandwidth b_i^t in time slot t, based on Shannon-Hartley theorem, it can achieve a data rate of:

$$r_i^t = b_i^t \log\left(1+\frac{G_i^t p_i^t}{b_i^t(\mathrm{N}_0+\mathrm{I}_0)}\right) = b_i^t \log\left(1+\frac{G_i^t \varphi_i^t}{\mathrm{Size}_{\mathrm{PRB}}\times\left(\mathrm{N}_0+\mathrm{I}_0\right)}\right), \quad (3)$$

where r_i^t is the data rate of UE i at time slot t and G_i^t is the channel gain of UE i at time slot t. The BS can measure G_i^t by decoding the Sounding Reference Signal (SRS) [28]. We define $c_i^t = \log\left(1+\frac{G_i^t \varphi_i^t}{Size_{PRB}\times(N_0+I_0)}\right)$ as the spectral efficiency of BS, and so UE i's data rate is a linear function of b_i^t:

$$r_i^t = c_i^t b_i^t \,. \quad (4)$$

C. Optimal transmission energy in LTE network

The objective function of a single UE is to minimize the energy consumption for transmitting L bits of data load and to satisfy SLA, i.e., meet the time delay constraint. In other words, for each time slot, more bandwidth should be allocated to UE if its channel condition is better or UE's SLA demand is much higher, i.e., transmitting more bits in less time slots. In order to achieve the goal, we define the channel condition penalty function and SLA demand penalty function.

1) Channel condition penalty function

From the energy saving aspect of view, good channel condition means UE can transmit more bits per unit of energy. So, we define the energy efficiency factor ζ_i^t in order to indicate the channel condition at each time slot for UE i as

$$\zeta_i^t = \frac{r_i^t \times \Delta T}{p_i^t \times \Delta T} = \frac{c_i^t \times Size_{PRB}}{\varphi_i^t}\,, \quad (5)$$

where ΔT is the duration of one time slot. Therefore, a large value of ζ_i^t means UE i's channel condition is good in comparison with other UEs at time slot t. However, a larger value of energy efficiency factor does not mean that UE is favored to obtain more bandwidth than others at time slot t. For instance, if UE i has better energy efficiency factor ζ_i^t than other UEs, but as compared with its average energy efficiency factor $\overline{\zeta_i^t}$ during its history window size, ζ_i^t is much smaller than $\overline{\zeta_i^t}$. In this case, UE i should be less favored to obtain bandwidth at this time slot since it has higher probability to get better channel condition in terms of energy efficiency factor in the future for transmitting data more efficiently. So, we define relative channel condition τ_i^t to measure the competence of obtaining bandwidth from the shared channels as

$$\tau_i^t = \frac{\overline{\zeta_{\mathrm{i}}^{\mathrm{t}}}}{\zeta_{\mathrm{i}}^{\mathrm{t}}}, \quad (6)$$

where τ_i^t is the relative channel condition factor of UE i at time slot t. Therefore, UE is more likely to get more bandwidth if it has better relative channel condition in terms of smaller value of τ_i^t, and vice versa.

If UE is assigned p_i^t of power to UE i at time slot t, then we define the channel condition penalty function g_i^t as follows:

$$g_i^t = \tau_i^t p_i^t = \frac{\varphi_i^t}{Size_{PRB}} \times \tau_i^t b_i^t \,. \quad (7)$$

Eq. (7) means if BS allocates more bandwidth to UE i with worse relative channel condition in terms of larger value of τ_i^t, BS would be penalized more, i.e., in order to decrease the penalty value, BS is incentivized to allocate more bandwidth to UE who has smaller value of τ_i^t. It is easy to see that g_i^t is a monotonically increasing function with respect to b_i^t, i.e., if we only consider minimizing energy consumption of UE, BS does not prefer to allocate any bandwidth to UE.

2) Service Level Agreement (SLA) demand penalty function

SLA demand means transmitting L bits of data load within the time delay constraint T. SLA demand changes at every time slots. It depends on how many bits UE has transmitted in the former time slots. Here, we use average data rate $\overline{\mathrm{R}_i^t} = \mathrm{L}_i^t/\mathrm{T}_i^t$ to measure the degree of SLA demand of UE i at time slot t. L_i^t refers to the number of bits remaining at the beginning of time slot t and T_i^t is the number of slots left at the beginning of time slot t. So, a larger value of $\overline{\mathrm{R}_i^t}$ means a higher degree of SLA demand. Suppose that UE i is assigned p_i^t of power at time slot t, we define the SLA demand penalty function Ω_i^t as follows:

$$\begin{aligned}\Omega_i^t &= \frac{\overline{R_i^{t+1}}}{r_i^t} p_i^t = \frac{L_i^t - \Delta T r_i^t}{\Delta T\left(T_i^t - 1 + \varepsilon\right) r_i^t} p_i^t \\ &= -\frac{\varphi_i^t}{Size_{PRB}} \times \frac{1}{T_i^t - 1 + \varepsilon} b_i^t + \frac{L_i^t \varphi_i^t}{\Delta T c_i^t \left(T_i^t - 1 + \varepsilon\right) \times Size_{PRB}}\end{aligned} \quad (8)$$

where $\overline{\mathrm{R}_i^{t+1}}$ is the average data rate of UE i in terms of SLA degree at time slot $t+1$ and ε is a very small value in order to make the equation meaningful when $\mathrm{T}_i^t = 1$. Eq. (8) means if UE i is assigned p_i^t of power and reaches the data rate of r_i^t at time slot t, then BS would be penalized more if UE i's SLA degree (i.e., $\overline{\mathrm{R}_i^{t+1}}$) becomes much higher at time slot $t+1$ as compared with the current data rate r_i^t. In other words, in order to decrease the penalty value, BS is incentivized to allocate more bandwidth to UE i in order to maximize the difference between the current date rate and SLA demand at the next time slot. Ω_i^t is a monotonically decreasing function with respect to b_i^t, i.e., if satisfying SLA demand is the only concern, BS prefers to allocate bandwidth to UE as much as possible.

3) EnerGy-Optimized (EGO) bandwidth allocation strategy in LTE network

There is a tradeoff between the two penalty functions, i.e., if BS wants to minimize the penalty of SLA demand, it would assign UE with enough bandwidth to transmit the whole data load, but on the other hand, UE would be penalized more for allocating more bandwidth in g_i^t. So, the optimal bandwidth allocation should minimize the sum of both penalty functions. Here, we define a penalty function f_i^t of UE i at time slot t as follows:

$$f_i^t = s_i \times g_i^t + \Omega_i^t \quad (9)$$

where s_i^t is a scalar to guarantee that the range of g_i^t and Ω_i^t is the same given the same increment value of b_i^t, i.e., g_i^t and Ω_i^t

should have the same influence on the penalty function f_i^t. So, we need to find the expression of s_i^t first.

If UE i's bandwidth is incremented by Δb_i^t , Ω_i^t is decremented by $\frac{\varphi_i^t}{Size_{PRB}} \times \frac{1}{T_i^t - 1 + \varepsilon} \Delta b_i^t$ and $s_i^t g_i^t$ is incremented by $\frac{\varphi_i^t}{Size_{PRB}} \times s_i^t \tau_i^t \Delta b_i^t$. So, we need to make sure $\frac{1}{T_i^t - 1 + \varepsilon}$ and $s_i^t \tau_i^t$ have the same value range. Since $T_i^t \epsilon [1, T_i^{initial}]$ where T_i^{inital} is UE i's initial time constraint, we have

$$\frac{1}{T_i^t - 1 + \varepsilon} \in \left[\frac{1}{T_i^{initial} - 1 + \varepsilon}, +\infty \right). \tag{10}$$

Assume ζ_i^{high} and ζ_i^{low} are the largest and smallest energy efficiency factor of UE i's historical record (BS records every UE's channel information with a sliding window), and so the expectation of the historical energy efficiency factor $\overline{\zeta_i^t}$ should be lower than ζ_i^{high}; we assume $\overline{\zeta_i^t} = \lambda_i^t \zeta_i^{high}$ where $\lambda_i^t \in [0,1]$ and based on Eq. (6), the range of τ_i^t is $\left[\frac{\overline{\zeta_i^t}}{\zeta_i^{high}}, \frac{\overline{\zeta_i^t}}{\zeta_i^{low}} \right]$. Plugging in the expression of $\overline{\zeta_i^t}$, we have

$$\tau_i^t \in \left[\lambda_i^t, \frac{\lambda_i^t \zeta_{\mathrm{i}}^{\mathrm{high}}}{\zeta_{\mathrm{i}}^{\mathrm{low}}} \right]. \tag{11}$$

Assume

$$s_i^t = \frac{1}{\lambda_i^t \left(T_i^{initial} - 1 + \varepsilon \right)}. \tag{12}$$

Then, we have:

$$s_i^t \tau_i^t \in \left[\frac{1}{T_i^{initial} - 1 + \varepsilon}, \frac{1}{T_i^{initial} - 1 + \varepsilon} \times \frac{\zeta_{\mathrm{i}}^{\mathrm{high}}}{\zeta_{\mathrm{i}}^{\mathrm{low}}} \right]. \tag{13}$$

Since $\frac{\zeta_i^{high}}{\zeta_i^{low}} \in [1, +\infty)$, $s_i^t \tau_i^t \in \left[\frac{1}{T_i^{initial} - 1 + \varepsilon}, +\infty \right)$. Substituting Eqs. (7), (8) and (12) into Eq. (9), we have:

$$f_i^t = \frac{\varphi_i^t}{Size_{PRB}} \left(\frac{\tau_i^t}{\lambda_i^t \left(T_i^{initial} - 1 + \varepsilon \right)} - \frac{1}{T_i^t - 1 + \varepsilon} \right) b_i^t + \frac{L_i^t \varphi_i^t}{\Delta T c_i^t \left(T_i^t - 1 + \varepsilon \right) \times Size_{PRB}} \tag{14}$$

where $\lambda_i^t = \overline{\zeta_i^t} / \zeta_i^{high}$. In the network scenario, suppose there are N UEs in the BS coverage area at time slot t and the total available bandwidth for BS is B. For each UE, the maximum transmission power is P_{max}. The objective of the system is to minimize the total penalty value of UEs in each time slot:

$$\arg\min_{b_i^t} \sum_{i=1}^{N} \left\{ \frac{\varphi_i^t}{Size_{PRB}} \left(\frac{\tau_i^t}{\lambda_i^t \left(T_i^{initial} - 1 + \varepsilon \right)} - \frac{1}{T_i^t - 1 + \varepsilon} \right) b_i^t + \frac{L_i^t \varphi_i^t}{\Delta T c_i^t \left(T_i^t - 1 + \varepsilon \right) \times Size_{PRB}} \right\}$$

$$s.t. \quad \sum_{i=1}^{N} b_i^t \le B; \tag{15}$$

$$\forall i, \quad b_i^t \times \varphi_i^t \le P_{\max}; \tag{16}$$

$$\forall i, \quad b_i^t \times c_i^t \times \Delta T \le L_i^t; \tag{17}$$

$$\forall i, \quad b_i^t \ge 0 \tag{18}$$

The first constraint means the total bandwidth allocated to UEs should be less than the system capacity. The second constraint means the transmission power for each UE should be less than the maximum transmission power [18]. The third constraint means the number of bits that UE uploads at this time slot should be less than UE's data load demand. It is a linear programming that can be solved by the Simplex Method [29].

IV. Simulation Results

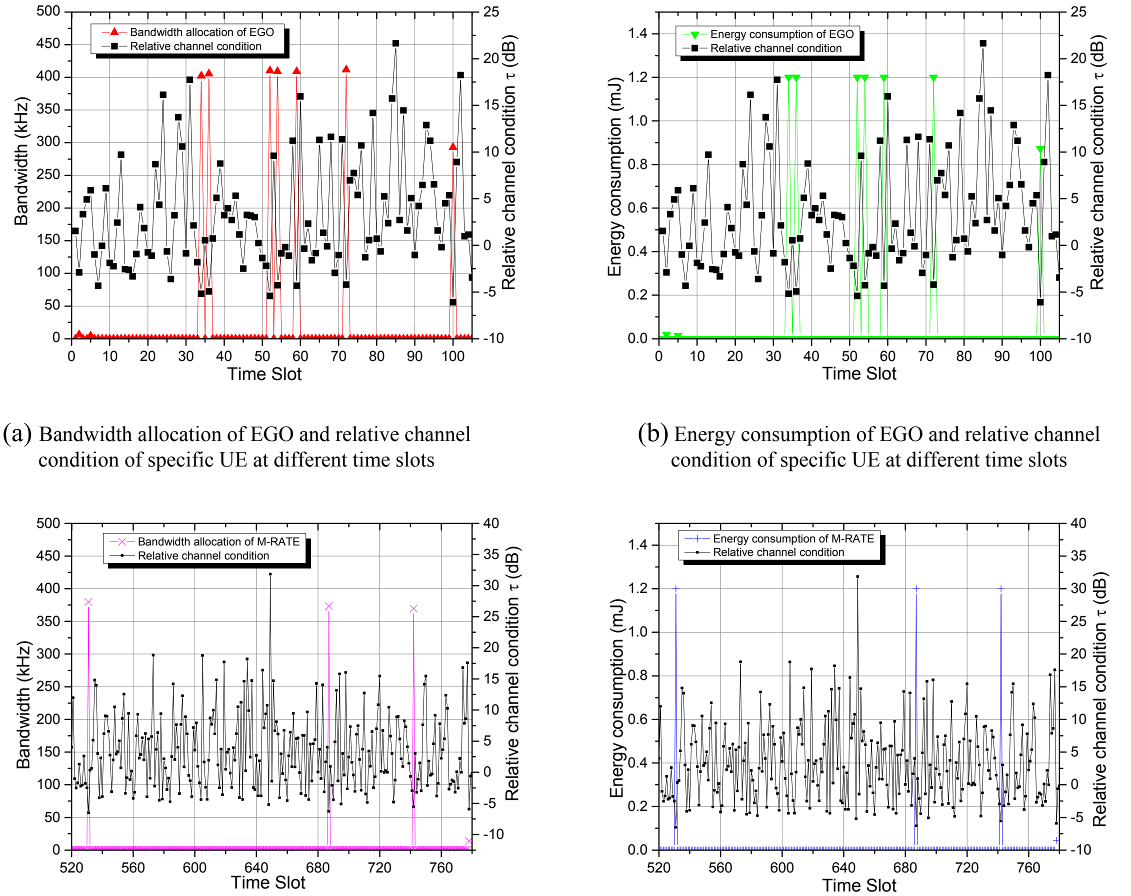
We simulate the proposed EGO strategy in the network scenario. For comparison, we select the bandwidth allocation strategy proposed in [18] whose goal is to Maximize the total UEs data RATE (M-RATE) at each time slot, i.e., $max \sum_{i=1}^{N} r_i^t$. Here, we assume the sub-channel condition is homogeneous.

TABLE 1
System parameters

Parameter	Value
Cell layout	Circle grid,
Cell radius	500 m
The length of time slot ΔT	50 ms
Power offset P_0	-68dBm
Path loss compensation α	0.6
Noise Power Spectrum Density N_0	-174dBm/Hz (thermal) [31]
Interference Power Spectrum Density I_0	$8 \times N_0$ [31]
Path loss model (dB)	$34.5 + 35 \times log_{10}(d[m])$[30]
Shadow fading	Log-normal, 8 dB standard deviation[30]
Multipath fading	SCME [30]
Maximum transmission power P_{max}	24mW
Total available bandwidth for BS B	5MHz

The system parameters are listed in Table 1. Initially, UEs are randomly distributed in the BS's coverage area and the speed of UE is randomly chosen from 0 to 50m/s, meanwhile, the direction of UE's movement is randomly chosen from -180 to 180 degree. BS classifies the uploading services into 5 categories (5 service classes) with respect to the average data rate: 100 KB/s, 75 KB/s, 50 KB/s, 25 KB/s and 20 KB/s. High average data rate may cost UE more monthly payment. The average data rate corresponds to the relationship between the data load L and time constraint T, e.g., if the average data rate is 100 KB/s, then $T = \frac{L\ KB}{100\ KB/s}$. The data load of each UE is randomly selected from 10KB to 5MB.

In the first simulation, 1000 MCC UEs are uploading the data to BS, and they are uniformly assigned among the five service classes. We randomly choose one UE to trace its bandwidth allocation and energy consumption in different time slots with respect to the varying relative channel condition τ. The selected UE's initial profile is <880KB, 8.8s> in terms of transmitting 880KB data within 176 time slots. Figs. 2(a) and 2(b) show the bandwidth and energy consumption distribution among different time slots using the EGO strategy. In order to clearly show the bandwidth/energy consumption with respect to τ, we convert the unit of τ into the dB domain. In the figures, BS can select suitable time slot in terms of smaller value of τ to allocate enough bandwidth to UE in order to minimize the energy consumption. Meanwhile, SLA is satisfied (UE finished the uploading at the 100th time slot). Figs. 2(c) and 2(d) show the bandwidth and energy consumption distribution among different time slots using the M-RATE strategy We select the time range from520 to 770. M-RATE wants to maximize the total data rate at each time slot, and so in every time slot, BS would select the UEs who have larger values of spectral



(a) Bandwidth allocation of EGO and relative channel condition of specific UE at different time slots

(b) Energy consumption of EGO and relative channel condition of specific UE at different time slots

(c) Bandwidth allocation of M-RATE and relative channel condition of specific UE at different time slots

(d) Energy consumption of M-RATE and relative channel condition of specific UE at different time slots

Fig. 2. Bandwidth allocation and energy consumption of specific UE at different time slots with respect to relative channel condition τ.

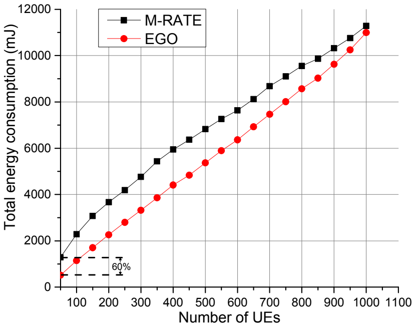


Fig. 3. Total energy consumption with respect to different number of UEs.

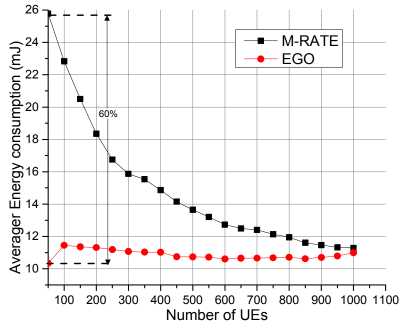


Fig. 4. Average energy consumption with respect to different number of UEs.

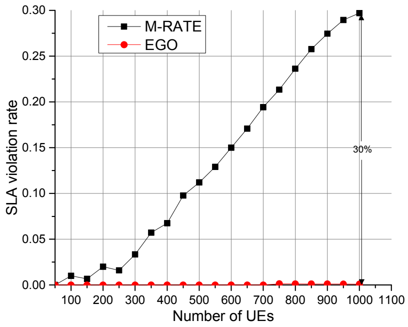


Fig. 5. SLA violation rate with respect to different number of UEs.

efficiency c_i^t and allocate enough bandwidth to them. The selected UE is completing its uploading process until the 744th time slot at which its SLA is violated. This is because the selected UE has the smaller spectral efficiency c_i^t than the other competitors (it is probably far away from the BS), and so it cannot obtain bandwidth at earlier time slots until the UEs who are closer to the BS in terms of better spectral efficiency finish the uploading process.

In the second simulation, we vary the number of UEs from 50 to 1000 in the network and record the total energy consumption, average energy consumption in terms of $\frac{\text{total energy consumption}}{\textit{number of UEs}}$ and SLA violation rate in terms of $\frac{\text{number of UEs volate the SLA demand}}{\textit{number of UEs}}$. As shown in Figs. 3 and 4, EGO can save the total energy consumption and the average energy consumption of up to 60% when the number of UEs is small. This is because BS only considers the UE i's current spectral efficiency c_i^t, i.e., if it is larger than others, BS would allocate enough bandwidth to it. However, UE i's current c_i^t cannot indicate whether it has better channel condition at the current time slot as compared to its history, i.e., even if UE i's current channel condition is worse than those of the past time slots, it also has the probability $q_i^t$to get enough bandwidth for transmission data load, and q_i^t gets small as the number of competitors increases. Therefore, we can see that when the number of UEs in the network increases, the two strategies' total energy consumption and average energy consumption are getting closer. In Fig. 5, the M-RATE's SLA violation rate increases (up to 30%) as the number of UEs in the network increases. This is because M-RATE does not consider SLA demand in its objective function, and so UE who has worse spectral efficiency always gets lower priority to acquire bandwidth no matter how high its SLA demand degree is. On the other hands, the SLA violation rate of EGO is very low (around 0.1%). Since EGO considers SLA demands as part of its objective function, when the UE's SLA demand degree becomes higher, BS is incentivized to allocate more bandwidth to it even if its relative channel condition is a little bit worse.

V. CONCLUSION

In this paper, we have investigated the issue of saving energy for resource-constrained mobile devices during the uploading process. The MCC application's profile is characterized as <*L*, *T*>, and we have proposed a bandwidth allocation strategy to decrease the energy consumption of UEs for uploading *L* data load while satisfying the SLA demand. The objective function of each UE is characterized by two penalty functions, which reflect the relative channel condition and the level of SLA demand. Simulation results have demonstrated that the proposed bandwidth allocation strategy, EGO, can save energy of up to 60% for each UE as well as decrease the SLA violation of up to 30% as compared with the existing bandwidth allocation strategy M-RATE.

In the future, we will apply the method for heterogeneous sub-channel condition of the uplink channel and evaluate the performance by simulating the strategy in different scenarios.

REFERENCES

[1] S. Robinson, "Cellphone energy gap: Desperately seeking solutions," Strategy Analytics, Acceesible from: www.strategyanalytics.com/default.aspx?mod= reportabstractviewer&a0=4645, 2009.

[2] Y. Wen, *et al.*, "Energy-optimal execution policy for a cloud-assisted mobile application platform," Tech. Rep., Sep 2011.

[3] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," Wireless Communications and Mobile Computing, vol. 13, pp. 1587-1611, 2013.

[4] H. Gao and Y. Zhai, "System design of cloud computing based on Mobile Learning," in Knowledge Acquisition and Modeling (KAM), 2010 3rd International Symposium on, pp. 239-242, 2010

[5] Z. Li, C. Wang, and R. Xu, "Computation offloading to save energy on handheld devices: a partition scheme," in Proceedings of the 2001 international conference on compilers, architecture, and synthesis for embedded systems, pp. 238-246, 2001

[6] S. Wang and S. Dey, "Rendering adaptation to address communication and computation constraints in cloud mobile gaming," in Global Telecommunications Conference , Miami, FL, Dec. 6-10, 2010, pp. 1-6.

[7] Y. Xu and S. Mao, "A survey of mobile cloud computing for rich media applications," IEEE Wireless Commun., vol. 20, pp. 46-53, 2013.

[8] X. Ma, Y. Zhao, L. Zhang, H. Wang, and L. Peng, "When mobile terminals meet the cloud: computation offloading as the bridge," IEEE Network, vol. 27, pp. 28-33, 2013.

[9] K. Kumar and Y. Lu, "Cloud computing for mobile users: Can offloading computation save energy?," Computer, vol. 43, pp. 51-56, 2010.

[10] E. Cuervo, *et al.*, "MAUI: making smartphones last longer with code offload," in Proceedings of the 8th international conference on Mobile systems, applications, and services, New York, NY, 2010, pp. 49-62.

[11] B. G. Chun, *et al.*, "Clonecloud: elastic execution between mobile device and cloud," in Proceedings of the sixth conference on Computer systems, New York, NY, 2011, pp. 301-314.

[12] S. Ibrahim, H. Jin, B. Cheng, H. Cao, S. Wu, and L. Qi, "CLOUDLET: towards mapreduce implementation on virtual machines," in Proceedings of the 18th ACM international symposium on High performance distributed computing, New York, NY, 2009, pp. 65-66.

[13] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, D. Wu, "Energy-Optimal Mobile Cloud Computing under Stochastic Wireless Channel," IEEE Transactions on Wireless Communications, vol. 12, pp. 4569-4581, 2013.

[14] Y. Wen, W. Zhang, and H. Luo, "Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones," in INFOCOM, 2012 Proceedings IEEE, Orlando, FL, Mar. 25-30, 2012, pp. 2716-2720.

[15] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on wireless heterogeneous networks for mobile cloud computing," Wireless Communications, IEEE, vol. 20, pp. 34-44, 2013.

[16] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," Wireless Communications, IEEE Transactions on, vol. 11, pp. 1991-1995, 2012.

[17] Munz, G.; Pfletschinger, S.; Speidel, J., "An efficient waterfilling algorithm for multiple access OFDM," Global Telecommunications Conference, Taipei, Taiwan, Nov. 17-21, 2002, vol.1, pp.681-685.

[18] R. Madan and S. Ray, "Uplink Resource Allocation for Frequency Selective Channels and Fractional Power Control in LTE," in Communications (ICC), IEEE International Conference on, Kyoto, Japan, Jun. 5-9, 2011, pp. 1-5.

[19] J. Huang, *et al.*, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," Selected Areas in Communications, IEEE Journal on , vol.27, pp.226,234, Feb. 2009.

[20] R. Kaewpuang, *et al.*, "A Framework for Cooperative Resource Management in Mobile Cloud Computing," Selected Areas in Communications, IEEE Journal on, vol. 31, pp. 2685-2700, 2013.

[21] C. Papagianni, A., *et al.*, "On the optimal allocation of virtual resources in cloud computing networks," Computers, IEEE Transactions on, vol. 62, pp. 1060-1071, 2013.

[22] K. Le, *et al.*, "Capping the brown energy consumption of internet services at low cost," in Green Computing Conference, Chicago, IL, Aug. 15-18, 2010, pp. 3-14.

[23] C. Chen, B. He, and X. Tang, "Green-aware workload scheduling in geographically distributed data centers," in Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, Taipei, Taiwan, Dec. 3-6, 2012, pp. 82-89.

[24] Y. Zhang and N. Ansari, "Fair Quantized Congestion Notification in Data Center Networks," IEEE Transactions on Communications, vol. 61, no.11, pp. 4690-4699, Nov. 2013.

[25] Y. Zhang and N. Ansari, "On Architecture Design, Congestion Notification, TCP Incast and Power Consumption in Data Centers," IEEE Communications Surveys and Tutorials, vol. 15, no. 1, pp. 39-64, 2013.

[26] X. Sun and N. Ansari, "Improving Bandwidth Efficiency and Fairness in Cloud Computing," Proc. IEEE Global Communications Conference (GLOBECOM 2013), Atlanta, GA, Dec. 9-13, 2013, pp. 2313-2318.

[27] 3GPP, "E-UTRA, Physical layer procedures," ed: TS 36.202 V8.10.0, 2010.

[28] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, 3G evolution: HSPA and LTE for mobile broadband: Academic press, 2010.

[29] S. Boyd and L. Vandenberghe, Convex optimization: Cambridge university press, 2009.

[30] D. S. Baum, J. Hansen, and J. Salo, "An interim channel model for beyond-3G systems: extending the 3GPP spatial channel model (SCM)," in Vehicular Technology Conference, VTC 2005-Spring, 2005 IEEE 61st, Dallas, TX, May 30-Jun. 1, 2005, vol. 5, pp. 3132-3136.

[31] M. Lauridsen, A. R. Jensen, and P. Mogensen, "Reducing LTE uplink transmission energy by allocating resources," in Vehicular Technology Conference (VTC Fall), 2011 IEEE, San Francisco, CA, Sep. 5-8, 2011, pp. 1-5.