Green Energy Aware Avatar Migration Strategy in Green Cloudlet Networks

Xiang Sun, *Student Member*, *IEEE*, Nirwan Ansari, *Fellow, IEEE* and Qiang Fan, *Student Member*, *IEEE*

***Abstract*—We propose a Green Cloudlet Network (*GCN*) architecture to provide seamless Mobile Cloud Computing (*MCC*) services to User Equipments (*UEs*) with low latency in which each cloudlet is powered by both green and brown energy. Fully utilizing green energy can significantly reduce the operational cost of cloudlet providers. However, owing to the spatial dynamics of energy demand and green energy generation, the energy gap among different cloudlets in the network is unbalanced, i.e., some cloudlets' energy demands can be fully provided by green energy but others need to utilize on-grid energy (i.e., brown energy) to satisfy their energy demands. We propose a *Green-energy awarE Avatar migRation* (*GEAR*) strategy to minimize the on-grid energy consumption in GCN by redistributing the energy demands via Avatar migration among cloudlets according to cloudlets' green energy generation. Furthermore, GEAR ensures the Service Level Agreement (*SLA*) in terms of the maximum Avatar propagation delay by avoiding Avatars hosted in the remote cloudlets. We formulate the GEAR strategy as a mixed integer linear programming problem, which is NP-hard, and thus apply the Branch and Bound search to find its sub-optimal solution. Simulation results demonstrate that GEAR can save on-grid energy consumption significantly as compared to the Follow me AvataR (*FAR*) migration strategy, which aims to minimize the propagation delay between an UE and its Avatar.**

***Keywords*—Mobile cloud computing, cloudlet, live migration, energy optimization, Branch and Bound search**

I. Introduction

The emergence of Mobile Cloud Computing (*MCC*) is enabling execution of computation-intensive applications (e.g., augmented reality and speech recognition) in a user equipment (*UE*), i.e., the UE can offload some tasks to high performance Virtual Machines (*VMs*) in a data center and VMs can help the UE execute these tasks in order to improve the task execution time and reduce the UE's energy consumption. However, the existing MCC architecture suffers from the long communications latency between a UE and its VM in a remote data center as the communications link traverses the Wide Area Network (*WAN*) which does not guarantee any minimum QoS to the UE; it is also very hard to control the WAN latency [1]. According to a report [2], "Amazon famously claimed that every 100 millisecond reduction in delay led to a one percent increase in sales. Google also stated that for every half second delay, it saw a 20 percent reduction in traffic." Therefore, reducing the latency can bring a huge benefit to the application providers. The concept of cloudlets has thus been proposed to reduce the propagation delay between a UE and its VM [1]. A cloudlet is a tiny version of the data center [3], [4] and is located close to the UE, and so communications between the UE and its VM can be established via the local area network (*LAN*).

To reap benefits of cloudlets and make them sustainable, we propose the Green Cloudlet Network (*GCN*) architecture

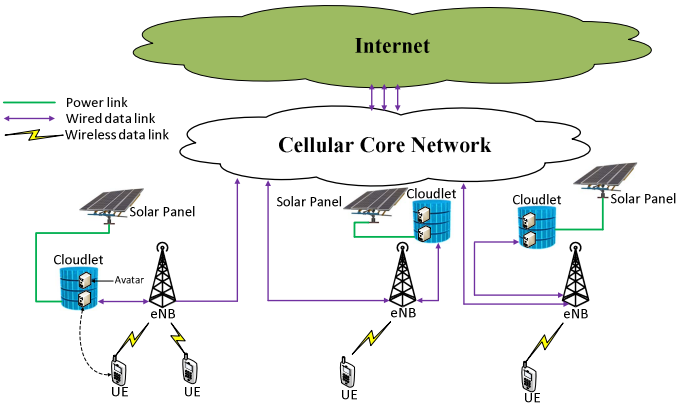

Fig. 1. Green cloudlet network architecture.

as shown in Fig. 1. Since the existing LTE network infrastructure can provide seamless connection between a UE and its eNB, each eNB is connected with a cloudlet via high speed fibers so that UEs can utilize the MCC technology everywhere. Meanwhile, the cloudlet is so close to the UE that the propagation delay is minimized. Each MCC UE subscribes one Avatar, a high performance VM in the cloudlet, to help run different tasks and provide extra storage space. Avatars are software clones of their UEs and always available to UEs when UEs are moving from one area to the others [5]. Moreover, in order to overcome the inefficient structure of the traditional Evolved Packet Core (*EPC*) network, Software Defined Network (*SDN*) based cellular core network [6], [7] has been proposed in the GCN architecture in order to decouple the control plane and data plane and provide efficient and flexible communications paths between Avatars in different cloudlets or between UEs in different eNBs.

GCN facilitates communications between a UE and its Avatar, but the distributed cloudlets increase the operational cost of cloudlet providers and CO_2 emission, i.e., a huge

This work was supported in part by the National Science Foundation under grant no. CNS-1320468. The authors are with the Advanced Networking Laboratory, Department of Electrical & Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA.
Emails: {xs47, nirwan.ansari, qf4}@njit.edu.

amount of on-grid energy (we assume energy from the power grid is brown energy, and thus the terms on-grid energy and brown energy are interchangeable for the rest of paper) will be consumed in order to maintain the GCN infrastructure. In order to reduce on-grid energy, "greening" is introduced in the GCN architecture, i.e., each cloudlet is powered by green energy generated from solar panels or other green energy collectors and uses on-grid energy as a backup. The power supply system of each cloudlet is shown in Fig. 2, in which the green energy collector absorbs energy from the green energy source (i.e., solar radiation) and converts it into electrical power, the charge controller regulates the electrical power from the green energy collector, and the electrical power is converted between AC and DC by the inverters. The smart meter records the electric energy from the power grid consumed by the cloudlet and eNB.

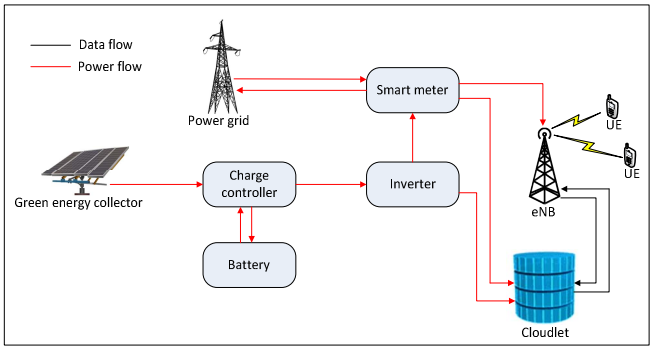

Fig. 2. The power supply system of the cloudlet.

Owing to the spatial dynamics of the distribution of UEs among different eNBs' coverage areas and the dynamics of application loads among Avatars, different cloudlets may require different energy demands for running the application loads of the hosting Avatars. Meanwhile, green energy generation also exhibits spatial dynamics. Therefore, some cloudlets, which have less energy demand and more green energy generated, would have excess of green energy. Conversely, some cloudlets, which have more energy demand and less green energy generated, would pull energy from the power grid (non-renewal). Such unbalanced energy gap (energy demand minus green energy generation) among different cloudlets increases the on-grid energy consumption. Therefore, fully utilizing green energy can tremendously reduce the on-grid energy consumption, and thus potentially decreases the operational cost of the cloudlet providers and CO_2 emission. In this paper, we propose the Green-energy awarE Avatar migRation (*GEAR*) strategy to minimize the on-grid energy consumption in GCN by redistributing the energy demands in terms of migrating Avatars among cloudlets according to cloudlets' green energy generation. Meanwhile, GEAR also guarantees the SLA, which is defined as the maximum Avatar propagation delay, i.e., the maximum propagation delay between the UE's eNB (the eNB which is serving the UE) and the UE's Avatar.

The rest of the paper is organized as follows. In Section II, we briefly review the related works. In Section III, we setup a power consumption model of a cloudlet. In Section IV, we formulate the avatar migration strategy in order to minimize the on-grid energy consumption. In Section V, we demonstrate the viability of GEAR via simulation results. The conclusion is presented in Section VI.

II. Related Works

Previous works [1], [8] have proved that cloudlets can significantly reduce the communications latency between UEs and VMs in the cloudlet. Tarik and Ksentini [9] introduced a follow-me cloud, i.e., a UE's service is continuously migrated to the data center which is much closer to the UE. Follow-Me Cloud tries to minimize the propagation delay between a UE and its VM in the data center, but it does not capitalize on green energy to optimize the energy consumption.

Rather than considering the communications latency between a UE and its VM, many works have focused on optimizing energy usage in Internet-scale Data Centers (*IDC*s). Studies [10]-[14] have aimed to minimize the electricity cost in IDCs which are only powered by brown energy, i.e., the workloads are migrated from high electricity cost IDCs to low electricity cost IDCs. Gkatzikis and Koutsopoulos [15] showed that introducing green energy in the cloud can significantly reduce the usage of brown energy, but there is a big challenge to match the dynamic green energy generation and dynamic energy demands of data centers in the green cloud network. Hatzopoulos *et al.* [16] assumed each task request is assigned to one VM and they tried to allocate running VMs into different data centers so that the total cost of power consumption in the green cloud network is reduced and the deadline of each request's response time is ensured. By considering the daily/seasonal effects of the green energy supplement in each geographical data center, Chen *et al* [17] proposed a holistic workload scheduling algorithm in order to minimize the brown energy consumption across all the data centers. Studies [18], [19] proposed the similar idea. Both of them design a profit maximization strategy which is to assign the incoming workloads among geo-distributed green data centers by considering the price of electricity, renewable power generation and SLA parameters.

As compared to the previous efforts, this paper presents several enhancements. First, GCN is implemented in order to provide seamless MCC services to UEs with lower and controllable communications latency. Different from traditional energy optimization in the green cloud network, SLA is defined as the maximum Avatar propagation delay (i.e., the propagation delay bound between a UE's eNB and UE's Avatar), and so our objective is not only to minimize the on-grid energy but also to guarantee the predefined SLA for

each UE in GCN. Ensuring the SLA is important for cloudlet providers, because each UE is moving over time and UEs' Avatars may migrate to the cloudlet with higher green energy generation and lower energy demands. Thus, some UEs' Avatars may be far away from themselves leading to the high propagation delay. As mentioned previously, the unbearable latency degrades the performance of MCC applications. Thus, guaranteeing the maximum propagation delay for each UE is an important factor to be considered when we design an optimal Avatar migration strategy. To the best of our knowledge, existing literature has not addressed this issue. Second, we propose a novel live Avatar migration strategy, i.e., GEAR, to achieve our objectives, and demonstrate the reduction of on-grid energy consumption without violating the SLA via extensive simulations.

III. SYSTEM MODEL

Cloudlets are distributed in the network powered by both on-grid and green energy. We assume that the servers in GCN are homogeneous, i.e., the configuration of each server is the same. Every UE's Avatar is also homogeneous, but the application loads of different UEs' Avatars are different. So, each server can host a fixed number of Avatars τ, but the application loads on different servers vary. The cloudlet provider plays the role of an Infrastructure as a Service (*IaaS*) provider, i.e., the cloudlet provider supplies virtualized computing resources in terms of Avatars to UEs. Although the provider does not impose any SLA of applications onto Avatars, it does ensure the SLA to UEs, i.e., the maximum Avatar propagation delay ϵ.

As mentioned earlier, Avatar migration is enabled to adjust the energy demand among the cloudlets. However, live migration may cost several seconds or minutes to activate an Avatar moving from the source server in one cloudlet to the destination server in another cloudlet, and so proactive migration decision should be made, i.e., the decision maker should determine each Avatar's location (i.e., in which cloudlet) in the next time slot based on the prediction of the energy demand and the green energy generation of each cloudlet. The hourly solar energy generation can be estimated by using typical annual meteorological weather data for a given geolocation [20]. Meanwhile, the estimation of the cloudlet's energy demand can be calculated by means of forecasting the power consumption of each active server in the cloudlet as will be discussed later by setting up the active server power consumption model.

In order to identify the location of different Avatars, two indicator functions $\delta(i,k)$ and $\eta_i(j,k)$ are introduced where i is the index of cloudlets in the network, j is the index of servers in one cloudlet, and k is the index of Avatars in the network. So, $\delta(i,k)=1$ implies that Avatar k is located in cloudlet i. Meanwhile, $\eta_i(j,k)=1$ indicates that Avatar k is in cloudlet i's server j. Therefore,

$$\delta(i,k)=\sum_{j=1}^{n_i}\eta_i(\mathrm{j},k), \tag{1}$$

where n_i is the number of active servers in cloudlet i and it is a function of $\delta(i,k)$:

$$n_i=\left\lceil\frac{\sum_k\delta(i,k)}{\tau}\right\rceil, \tag{2}$$

where τ is the number of Avatars hosted in the server.

A. Server Power Consumption Model

In this section, we model the power consumption of active servers in a cloudlet. The power consumption of active server j in cloudlet i can be characterized as follows [21]:

$$P_{i,j}=P^s+P_{i,j}^{vir}+\alpha\times u_{i,j}^{app}, \tag{3}$$

where $P_{i,j}$ is the total power consumption of active server j in cloudlet i; P^s is the power consumption of the server when it is in the standby mode, i.e., when the server's CPU load is zero; $P_{i,j}^{vir}$ is the power consumption of server j for doing virtualization and we will discuss it in the next paragraph; $\alpha\times u_{i,j}^{app}$ is the power consumption of server j for running different applications on its hosting Avatars, where $u_{i,j}^{app}$ is server j's CPU usage for running Avatar i's application load and α is the coefficient that maps the CPU usage into the power consumption. So, if the CPU usage for running Avatar k's application load is u_k^{app}, then the server j's CPU usage can be considered as a function of $\eta_i(j,k)$:

$$u_{i,j}^{app}=\sum_k\eta_i(j,k)\times u_k^{app}. \tag{4}$$

As mentioned earlier, $P_{i,j}^{vir}$ is the power consumption of server j for performing virtualization, and it includes two parts [21]:

$$P_{i,j}^{vir}=P_{i,j}^{hyper}+P_{i,j}^{idle}, \tag{5}$$

where $P_{i,j}^{hyper}$ is the power consumption of server j in cloudlet i for running a hypervisor without any Avatar load (i.e., the hypervisor only manages the configuration of different Avatars in the server). In order to determine the power consumption of an idle hypervisor, Warkozek *et al.* [21] showed that $P_{i,j}^{hyper}$ is proportional to the number of Avatars hosted in the server:

$$P_{i,j}^{hyper}=\beta\times\sum_k\eta_i(j,k), \tag{6}$$

where $\sum_k\eta_i(j,k)$ indicates the total number of Avatars in sever j and β is the Avatar power coefficient, which is the power cost of the hypervisor for maintaining one Avatar.

$P_{i,j}^{idle}$ in Eq. (5) is the power consumption of Avatars in the idle mode (Avatar does not take any application load from UE, but runs basic system operation instances) in server j of cloudlet i. $P_{i,j}^{idle}$ is determined by the number of Avatars in server j and the amount of CPU usage for running the operating system (*OS*) kernel instances for each idle Avatar.

Assuming that all UEs' Avatars use the same OS, thereby CPU usage for running the OS kernel instances u^{idle} is the same for all Avatars. Therefore, we have

$$P_{i,j}^{idle} = \alpha \times u^{idle} \times \sum_k \eta_i(j,k). \quad (7)$$

Substituting Eqs. (4)-(7) into Eq. (3) yields the power consumption of server j in cloudlet i as follows:

$$P_{i,j} = P^s + \beta \times \sum_k \eta_i(j,k) + \sum_k \left[\eta_i(j,k) \times \alpha \left(u^{idle} + u_k^{app} \right) \right]. \quad (8)$$

Since the servers in the cloudlet network are homogeneous, P^s and α, which are constants, can be pre-determined. Meanwhile, if all servers are installed with the same type of hypervisor, such as Hyper-V, ESX or Xen, then β is the same for all servers. We define $u_k = u^{idle} + u_k^{app}$ as the total CPU usage (including OS kernel CPU usage and application load CPU usage) for running Avatar k in the server, and so Eq. (8) can be expressed as:

$$P_{i,j} = P^s + \beta \times \sum_k \eta_i(j,k) + \sum_k \left[\eta_i(j,k) \times \alpha u_k \right]. \quad (9)$$

B. Cloudlet Power Consumption Model

Aside from running the cooling system and cloudlet network equipment, the major power consumption of a data center is the power consumed by the active servers. However, a cloudlet is a tiny version of the data center that does not need to maintain a powerful cooling system and plenty of switches, and so we assume all the power consumption of a cloudlet is contributed by the computing equipments such as servers, and we calculate cloudlet i's power consumption as the sum of the power consumption of the active servers:

$$P_i = \sum_{j=1}^{n_i} P_{i,j} = n_i P^s + \beta \times \sum_{j=1}^{n_i} \sum_k \eta_i(j,k) + \sum_{j=1}^{n_i} \sum_k \left[\eta_i(j,k) \times \alpha u_k \right] \quad (10)$$

By approximating Eq. (2) into $n_i \approx \frac{\sum_k \delta(i,k)}{\tau}$ and substituting Eq. (1) and $n_i \approx \frac{\sum_k \delta(i,k)}{\tau}$ into Eq. (10), we have the power consumption of cloudlet i:

$$P_i \approx \sum_k \left[\delta(i,k) \times \left(\frac{P_s}{\tau} + \beta + \alpha u_k \right) \right]. \quad (11)$$

C. Avatar Propagation Delay Model

Usually, a UE and its Avatar may not associate with the same eNB. As mentioned previously, UEs are moving over time and Avatars migrate to the cloudlet with more green energy generation and less energy demands. Thus, the communications between a UE and its Avatar might traverse SDN based cellular core network. As shown in Fig. 3, if UE 1 tries to communicate with its Avatar (i.e., Avatar 1 in cloudlet B), the communications path should traverse eNB 1, openflow switches, eNB 2 and cloudlet B. Thus, the communications delay between a UE and its Avatar comprises three parts: wireless communications delay between the UE and its eNB, propagation delay between the eNB and the cloudlet where the UE's Avatar is located, and the propagation delay within the cloudlet. However, the

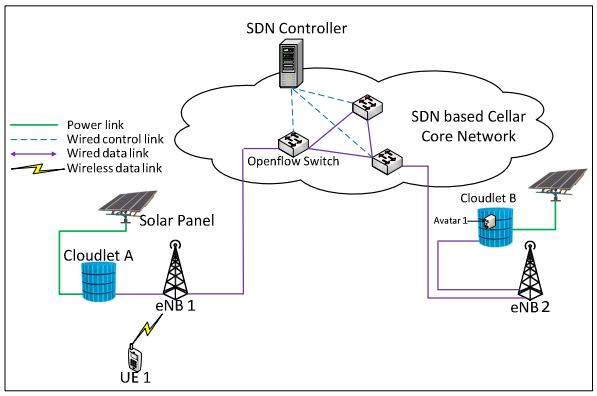

Fig. 3. The communications between a UE and its Avatar.

wireless communications delay is determined by the UE's billing plan and the LTE network provider's bandwidth allocation strategy which are not controlled by the cloudlet provider. Meanwhile, we assume the propagation delay within the cloudlet is negligible. So, we define Avatar propagation delay as the latency for propagating one packet between the UE's eNB and its cloudlet. Cloudlet providers only need to guarantee the SLA in terms of maximum Avatar propagation delay for each UE.

If the UE's Avatar is located at cloudlet i and the UE is associated with eNB e (e is the index of eNB in GCN), we can express the Avatar propagation delay as $T_{i,e}$, which comprises two parts: $T_{i,e}^{prop}$, i.e., the propagation delay for transmitting a packet between cloudlet i in which the UE's Avatar resides and the UE's eNB e; $T_{i,e}^{proc}$, i.e., the total processing delay for all the openflow switches on the routing path in processing one packet. Assume $T_{i,e}^{prop}$ is proportional to the distance $d_{i,e}$ [22], i.e., the distance between cloudlet i in which the UE's Avatar resides and the UE's eNB e, i.e., $T_{i,e}^{prop} \propto d_{i,e}$. Meanwhile, if we assume that the number of openflow switches on the routing path is also proportional to $d_{i,e}$ (i.e., the longer distance between Avatar's cloudlet and the UE's eNB, the more openflow switches the packet needs to traverse) and the average packet processing time on every openflow switch is the same, then $T_{i,e}^{proc}$ is also proportional to $d_{i,e}$, i.e., $T_{i,e}^{proc} \propto d_{i,e}$. So, we conclude that:

$$T_{i,e} = \sigma \times d_{i,e} \times \delta(i,k), \quad (12)$$

where σ is the coefficient that maps the distance to the time delay.

IV. Problem Formulation

Owing to the spatial dynamics of energy demand and green energy generation among different cloudlets, energy demand of some cloudlets can be met by green energy, but some cannot and need to consume on-grid energy. Meanwhile, owing to the disadvantages of "banking" green energy [23], we assume that green energy is disinclined to be stored for each cloudlet [24], [25]. Therefore, green energy should be

fully utilized in each time slot so that the on-gird energy can be minimized. Denote ρ_i as the on-grid energy consumption of cloudlet i, i.e., $\rho_i = max[0, \Delta T(P_i - G_i)]$, where ΔT is the length of one time slot; P_i and G_i are the power demand and the green power generation of cloudlet i. The objective of the GEAR strategy is to minimize the on-grid energy consumption of GCN in each time slot. So, we formulate GEAR as follows:

$$\min_{\delta(i,k)} \sum_i \rho_i \tag{13}$$

$$s.t. \quad \forall i, \ \rho_i \geq \sum_k \left[\delta(i,k) \times \left(\frac{P_s}{\tau} + \beta + \alpha u_k \right) \right] - G_i, \tag{14}$$

$$\forall i, \ \rho_i \geq 0, \tag{15}$$

$$\forall k, \ \sigma d_{i,e} \delta(i,k) \leq \varepsilon, \tag{16}$$

$$\forall i, \ \frac{1}{\tau} \sum_k \delta(i,k) \leq m_i, \tag{17}$$

$$\forall k, \ \sum_i \delta(i,k) = 1, \tag{18}$$

where $\delta(i,k)$ is a binary variable indicating the location of Avatars, ϵ is the SLA provided by the cloudlet provider, and m_i is the capacity of cloudlet i, i.e., the total number of servers owned by cloudlet i. Constraints (14) and (15) indicate $\rho_i = max(0, P_i - G_i)$. Constraint (16) means the cloudlet provider should guarantee the SLA for all UEs. Constraint (17) implies that the total number of Avatars assigned to the cloudlet should not exceed the cloudlet's capacity and Constraint (18) means each Avatar should be assigned to no more than one cloudlet.

To solve GEAR (which is proven to be an NP-hard problem [6]), we use the Branch and Bound search method [26] to find the sub-optimal solution to the problem. Therefore, in each time slot, each Avatar estimates its average CPU utilization for the next time slot by adopting the CPU workload prediction model [27], [28], acquires the location of its UE, and reports the information to the GCN manager. The GCN manager decides the location of all Avatars by solving the above optimization problem.

TABLE 1
SYSTEM PARAMETERS

Parameter	Value
The length of time slot, ΔT	15 *mins*
Capacity of server, τ	16 *Avatars*
Power consumption of standby server, P^s	80 *Watts*
CPU usage to power mapping coefficient, α	0.2 *%CPU/Watts*
Avatar to power mapping coefficient, β	0.3 *Watts/Avatar*
Distance to delay mapping coefficient, σ	3.33 *ms/km*
SLA, ε	10 *ms*

V. SIMULATION RESULTS

We simulate the proposed GEAR strategy in GCN. For comparisons, we select the other Avatar migration strategy, i.e., Follow me AvataR (*FAR*) migration strategy. The idea of FAR is similar to the previous work [9], which tries to minimize the propagation delay between a UE and its VM in the cloud. Similarly, FAR does not minimize the on-grid energy consumption but minimizes the propagation delay between a UE and its Avatar by selecting the nearest cloudlet as the host of the UE's Avatar. Some system parameters are listed in Table 1.

To demonstrate the viability of GEAR, we set up a network with the topology shown in Fig. 4, which includes 16 cloudlet-eNB combinations (4×4) in a square area of 64 km^2. The coverage of each eNB is a square area of 4 km^2. The whole area is divided into 2 parts, i.e., urban and rural areas. Initially, each cloudlet's capacity m_i in terms of the number of servers is randomly chosen between 10 and 30, and UEs are uniformly distributed in the network. The location of each Avatar is initially chosen to be its nearest cloudlet. Each Avatar's CPU is assigned by one physical core in the server and the CPU utilization of each Avatar is randomly chosen between 10% and 100% in each time slot (we assume the OS kernel instances cost 10% of the Avatar's CPU utilization).

A. Spatial Dynamics of Energy Demand

Energy demands of different cloudlets in GCN exhibit spatial dynamics, and so we setup the simulation scenario as follows: UE mobility adopts the modified random waypoint model, i.e., each UE randomly selects a speed between 0 and 10 *m/s* in every time slot and moves toward its destination, and the locations of UEs' destinations (i.e., the values of *x* and *y* coordinates) are randomly selected according to a normal distribution$\sim N(4km, 1.4km)$, which implies that UEs more likely move toward the center of each urban area (i.e., based on the characteristics of the normal distribution, UEs more likely select their destinations which are close to the center of the network). For the green energy generation, we use the local daily solar radiation data trace (Millbrook, NY in Jan. 1st, 2015) from National Climatic Data Center [29], as shown in Fig. 5, where each point indicates the average solar radiation within the current hour. Assume the size of the solar cell equipped in each cloudlet is 5 m^2 and the efficiency for converting solar radiation into electricity is 46% [30]. Also, suppose the green energy generated in different cloudlets is the same in the same time slot. Fig. 6 shows the total on-grid energy consumption of two Avatar migration strategies in different time slots. When there is no or little green energy provision in GCN (from 8 a.m. to 9 a.m.), there is no difference between the two live migration strategies since all the cloudlets are powered by on-grid energy. However, when more green energy is generated at each cloudlet, GEAR can save more on-grid energy than FAR, since GEAR can migrate Avatars from the cloudlets with higher energy demand to the cloudlets with lower energy demand so that green energy can be fully utilized. Fig. 7 shows the total on-grid energy consumption in GCN in the whole day.

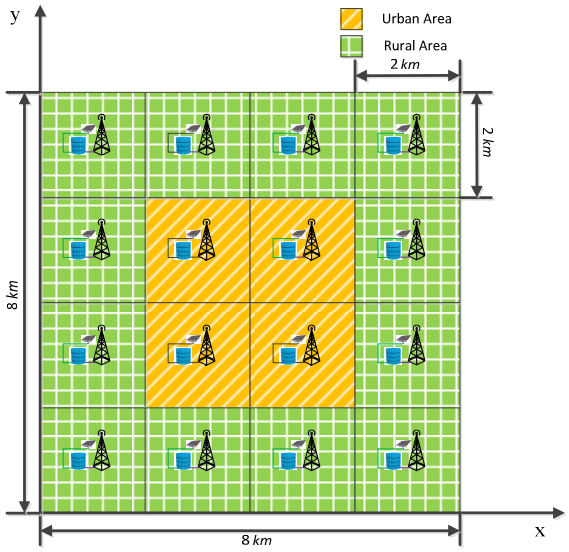

Fig. 4. Network topology.

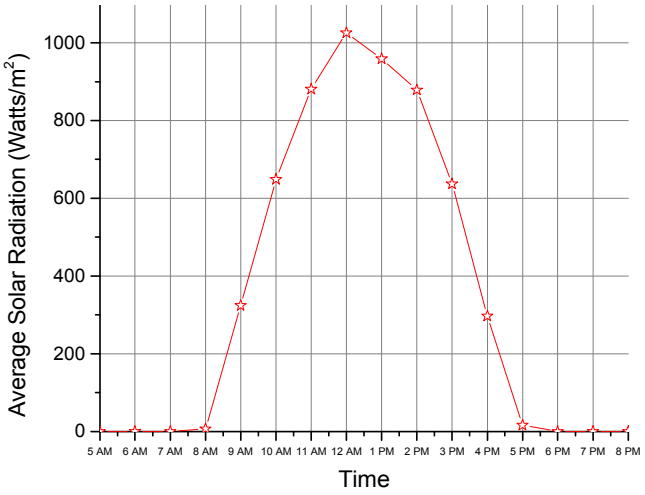

Fig. 5. Average solar radiation generated at different time.

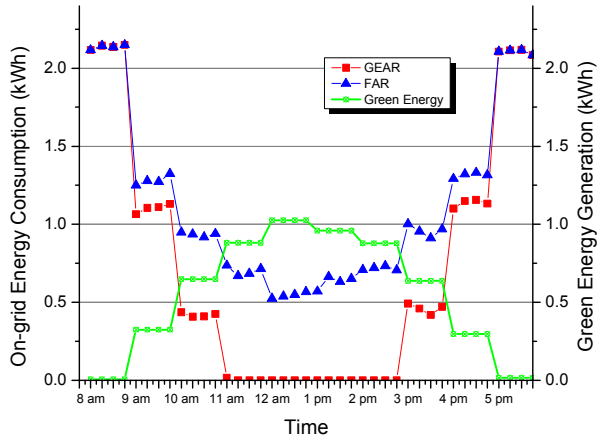

Fig. 6. On-grid energy consumption at different time.

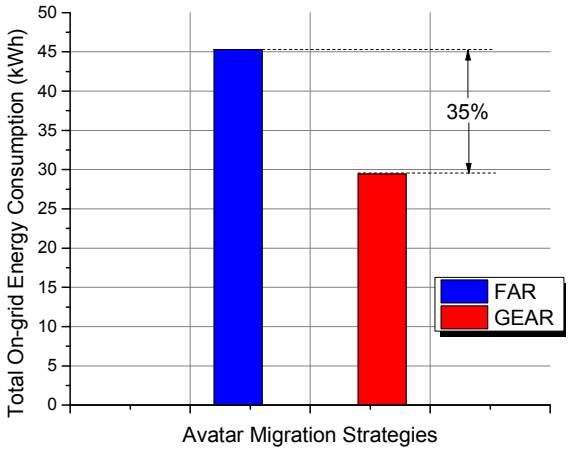

Fig. 7. Total on-grid energy consumption in one day.

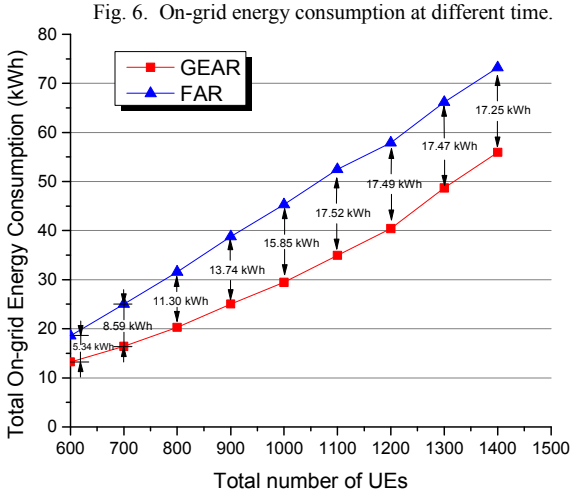

Fig. 8. Total on-grid energy consumption over the number of UEs in GCN.

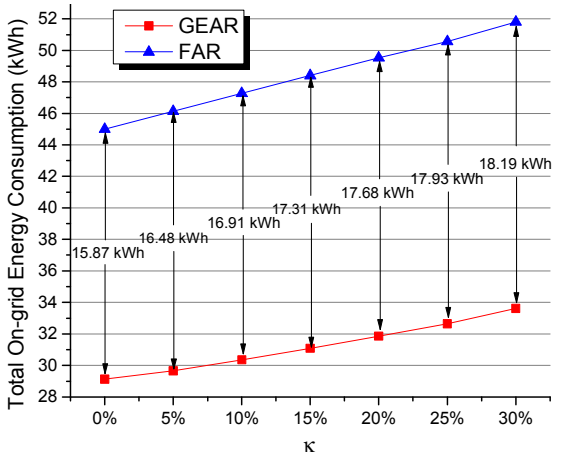

Fig. 9. Total on-grid energy consumption over different values of κ.

We next examine the effect of the density of UEs by increasing the number of UEs from 600 to 1400. More UEs in the network result in more out-of-balanced energy demand between the urban area and the rural area because all the UEs prefer to go to the urban rather than the rural area. Fig. 8 shows the total on-grid energy consumption in one day with respect to different numbers of UEs in the network. We can see that when the number of UEs increases from 600 to 1100,

the difference of the total on-grid energy consumption between GEAR and FAR is increasing, i.e., if the energy demand is more unbalanced between areas, GEAR can save more on-grid energy as compared to FAR by balancing the energy gap among the cloudlets. However, as the number of UEs exceeds 1100, the difference of the total on-grid energy consumption between GEAR and FAR remains static because green energy has already been fully utilized by GEAR when the number of UEs reaches 1100, and if the energy demands are still increasing, on-grid energy has to be tapped.

B. Spatial Dynamics of Green Energy Provision and Energy Demand

In the real environment, not only the energy demand but also the green energy provision may exhibit spatial dynamics, i.e., the solar cell in different cloudlets may generate different amount of green energy because of the position of the sun, the spatial dynamics of atmospheric conditions, etc. Also, evidence shows that the solar radiation of the rural area is greater than the urban area [31]. Therefore, we setup the simulation scenario as follows: the UE's parameters are the same as in the previous simulation scenario and the hourly average solar radiation in the rural area still follows the data trace as shown in Fig. 5. However, the hourly average solar radiation in the urban area is reduced by κ percentage. Fig. 9 shows the total on-grid energy consumption of the network in one day with respect to different values of κ. Note that as the value of κ increases (i.e., hourly green energy generation is getting less in the urban area and the energy gap between the two areas is getting larger), FAR consumes more on-grid energy because the energy gap of the urban area is getting larger, and GEAR can save more on-grid energy as compared to FAR by balancing the energy gap between the two areas.

VI. CONCLUSION

In this paper, we have proposed the GCN architecture to provide seamless and lower latency MCC services to UEs, i.e., UEs can offload tasks to their powerful Avatars with shorter propagation delay. However, owing to the spatial dynamics of energy demand and green energy provisioning, a significant amount of green energy is wasted, thus resulting in more grid energy consumption. Therefore, we have proposed the GEAR strategy to redistribute the energy demand by migrating Avatars among cloudlets according to cloudlets' green energy generation and to guarantee the maximum Avatar propagation delay. Simulation results have demonstrated that GEAR can significantly save on-grid energy as compared to the FAR strategy.

REFERENCES

[1] S. Mahadev, P. Bahl, R. Caceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14-23, Oct.-Dec. 2009.

[2] *Latency Considerations in LTE*, Sep. 2014. [Online]. Available: http://mavenir.com/files/doc_downloads/Stoke_Documents/130-0029-001_LTELatencyConsiderations_Final.pdf.

[3] Y. Zhang, and N. Ansari, *"*HERO: Hierarchical Energy Optimization for Data Center Networks*," IEEE Systems Journal,* vol. 2, no. 9, pp. 406-415, June 2015.

[4] 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, First Quarter, 2013.

[5] C. Borcea, *et al*., "Avatar: Mobile Distributed Computing in the Cloud," in *3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud '15),* San Francisco, CA, Mar. 30-Apr. 3, 2015, pp. 151-157.

[6] X. Sun, N. Ansari, and Q. Fan, *Green Energy Aware Avatar Migration Strategy in Green Cloudlet Networks*, NJIT Advanced Networking Lab., Tech. Rep. TR-ANL-2015-006; also archived in Computing Research Repository (CoRR), arXiv:1509.03603, 2015.

[7] X. Jin, L. E. Li, L. Vanbever, and J. Rexford, "Softcell: Scalable and flexible cellular core network architecture," in *Proceedings of the 9th ACM conference on Emerging networking experiments and technologies*, Santa Barbara, CA, Dec. 09-12, 2013, pp. 163-174.

[8] W. Liu, J. Cao, X. Qiu, and J. Li, "Improving Performance of Mobile Interactive Data-Streaming Applications with Multiple Cloudlets," in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom)*, Singapore, Dec. 15-18, 2014, pp. 46-53.

[9] T. Taleb and A. Ksentini, "Follow me cloud: interworking federated clouds and distributed mobile networks," *IEEE Network*, vol. 27, no. 5, pp.12-19, Sep.-Oct. 2013.

[10] D. Xu, X. Liu, and B. Fan, "Minimizing energy cost for Internet-scale datacenters with dynamic traffic," in *2011 IEEE 19th International Workshop on Quality of Service (IWQoS)*, San Jose, CA, Jun. 6-7, 2011, pp. 1-2.

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

[12] N. Buchbinder, N. Jain, and I. Menache, "Online job-migration for reducing the electricity bill in the cloud," in *2011 NETWORKING*, Valencia, Spain, May. 9-13, 2011, pp. 172-185.

[13] A. Qureshi, *et al*., "Cutting the electric bill for internet-scale systems." in *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp. 123-134, 2009.

[14] B. Aksanli, J. Venkatesh, T. Rosing, and I. Monga, "A comprehensive approach to reduce the energy cost of network of datacenters," in *2013 IEEE Symposium on Computers and Communications (ISCC'13)*, Split, Croatia, Jul. 7-10, 2013, pp. 275-280.

[15] L. Gkatzikis and I. Koutsopoulos, "Migrate or not? Exploiting dynamic task migration in mobile cloud computing systems," *IEEE Wireless Communications*, vol. 20, no. 3, pp. 24-32, Jun. 2013.

[16] D. Hatzopoulos, I. Koutsopoulos, G. Koutitas, and W. V. Heddeghem, "Dynamic virtual machine allocation in cloud server facility systems with renewable energy sources," in *Proceedings of IEEE International Conference on Communications (ICC'13)*, Budapest, Hungary, Jun. 9-13, 2013, pp. 4217-4221.

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

[18] M. Ghamkhari. and H. Mohsenian-Rad, "Energy and performance management of green data centers: A profit maximization approach," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 1017-1025, 2013.

[19] A. Kiani, and N. Ansari, "Toward Low-Cost Workload Distribution for Integrated Green Data Centers," *IEEE Communications Letters*, vol. 19, no. 1, pp. 26-29, Jan. 2015.

[20] T. Han and N. Ansari, "On Optimizing Green Energy Utilization for Cellular Networks with Hybrid Energy Supplies," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 3872-3882, August 2013.

[21] G. Warkozek, E. Drayer, V. Debusschere, and S. Bacha, "A new approach to model energy consumption of servers in data centers," in *2012 IEEE International Conference on Industrial Technology*, Athens, Greece, Mar. 19-21, 2012, pp. 211-216.

[22] Cisco System, Inc. (2007). "Design Best Practices for Latency Optimization," Financial Services Technical Decision Maker White Paper. [Online]. Available: https://www.cisco.com/application/pdf/en/us/guest/netsol/ns407/c654/ccmigration_09186a008091d542.pdf

[23] Í. Goiri, *et al.*, "GreenHadoop: leveraging green energy in data-processing frameworks," in *Proceedings of the 7th ACM european conference on Computer Systems*, Bern, Switzerland, Apr. 10-13, 2012, pp. 57-70.

[24] T. Han, and N. Ansari, "Green-energy Aware and Latency Aware user associations in heterogeneous cellular networks," in *Proceedings of IEEE Global Communications Conference (GLOBECOM'13)*, Atlanta, GA, Dec. 9-13, 2013, pp. 4946-4951.

[25] T. Han, and N. Ansari, "A Traffic Load Balancing Framework for Software-Defined Radio Access Networks Powered by Hybrid Energy Sources," *IEEE Transactions on Networking*, DOI: 10.1109/TNET.2015.2404576, early access, 2015.

[26] G. Mitra, "Investigation of some branch and bound strategies for the solution of mixed integer linear programs." *Mathematical Programming*, vol. 4, no. 1, pp. 155-170, 1973.

[27] K. Qazi, Y. Li, and A. Sohn, "Workload Prediction of Virtual Machines for Harnessing Data Center Resources," in *2014 IEEE 7th International Conference on Cloud Computing*, Anchorage, AK, Jun. 27-Jul. 2, 2014, pp. 522-529.

[28] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1107-1117, 2013.

[29] *Daily solar radiation data trace from National Climatic Data Center*. [Online]. Available: http://www1.ncdc.noaa.gov/pub/data/uscrn/products/hourly02/2015/CRNH0203-2015-NY_Millbrook_3_W.txt

[30] M. K. Islam, T. Ahammad, E. H. Pathan, A. M. Haque, et al., "Analysis of Maximum Possible Utilization of Solar Radiation on a Solar Photovoltaic Cell with a Proposed Model," *International Journal of Modeling and Optimization*, vol. 1, no. 1, pp. 66-69, Jan. 2011.

[31] G. Codato, A. P. Oliveira, and J. F. Escobedo, "Comparative study of solar radiation in urban and rural areas," in *Anais do XIII Congresso Brasileiro de Meteorologia*, Fortaleza, Brazil, 2004.