Green Fog Offloading Strategy for Heterogeneous Wireless Edge Networks

Yung-Lin Hsu*, Hung-Yu Wei*, and Mehdi Bennis†
* Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
† Centre for wireless communications, University of Oulu, Finland,
Email: * {d04942010,hywei}@ntu.edu.tw, †mehdi.bennis@oulu.fi

***Abstract*—Multi-access/Mobile Edge Computing (MEC) and fog computing are promising techniques to satisfy low latency requirements for emerging next generation applications. Moving computation entities closer to a user could reduce the overall serving latency. In terms of green communications, given the latency constraint, how to minimize the power consumption at the user equipment (UE) and the edge node (EN) sides is important. Considering several edge nodes, partially offloading a user's task to one or more edge nodes is key. In this paper, a multi-node partial task offloading MEC scenario is discussed, in which UEs locally compute the task and share the remainder with other edge nodes. In addition, a green task distribution algorithm which minimizes the system power consumption is proposed, considering queueing, transmitting and computing delay. The simulation results show that the proposed algorithm minimizes the power consumption while meeting the latency requirements, and the power saving efficiency outperforms a binary offloading strategy. Moreover, the coupling effects between the latency requirements, offloading signal strength within the edge nodes, computation capability of edge nodes and the number of sub-carriers used to transmit the offloading task are discussed.**

***Index Terms*—MEC, Fog node, Green communication**

I. Introduction

Nowadays, latency sensitive applications such as virtual reality (VR) and augmented reality (AR) have become one of the most prominent applications in 5G. Dealing with the stringent requirements for low latency is a critical challenge, whereby relying on centralized cloud computing is insufficient. In a heterogeneous network (HetNet), MEC and Fog Node systems are able to meet the latency requirement as network functionalities are pushed to the edge. In terms of HetNet design and deployment, the issues and methods are discussed in [1]. Fig. 1 briefly illustrates the concept of a MEC system in [2], [3]. More detailed discussion about 5G MEC and fog paradigm can be found in [4]. In a MEC system, the UE can offload the task wirelessly to transmission and reception points (TRPs). Assuming that two kinds of TRP are in the system, the fog node acts as an EN and the MEC node as a micro cell next generation node B (MCgNB). Typically, the MEC nodes is more powerful than the fog nodes regarding computational and other capabilities.

For delay sensitive applications, completing the task at the edge is more practical. Therefore, rather than forwarding the task to the cloud, this article focused on distributing the task to the TRP nodes with minimum power consumption. There are several MEC task offloading related works as follows.

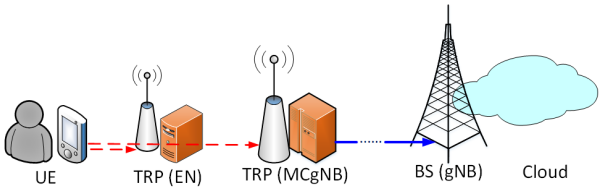


Fig. 1: Multi-access edge computing structure

An QoS-regarded task offloading scheme was proposed in [5]. [6] proposed a novel hybrid fog-cloud network and a task distribution algorithm to minimize the maximum computational latency of all the fog UEs. [7] proposed a novel framework aiming at the co-provisioning of radio access and computing services, and an algorithm is proposed to improve the system performance regarding latency reduction and energy efficiency. The authors in [8] proposed an energy-aware offloading scheme considering the number of the small cells and the CPU cycle frequency of UEs. In [9], the authors optimized the network energy efficiency subject to the network stability and proposed a joint computation allocation and resource management algorithm to achieve energy efficiency and delay tradeoff. A Boolean-domain task offloading algorithm is proposed to minimize the energy consumption and monetary cost from UE's perspective in [10]. This article not only investigated the offloading strategy but also considered the wireless resource allocation. The authors in [11] considered task offloading and computing optimization under a multi-antenna access point Time Division Multiple Access (TDMA) scenario. In addition, the influence of TDMA and Frequency Division Multiple Access (FDMA) on a MEC offloading scenario was discussed in [12]. [13] considered both edge computing and edge caching in fog networks, and [14] jointly considered energy harvesting, battery residual power and weather condition in task offloading. In [15], a power-delay tradeoff in the context of task offloading is studied. By using the tools from Lyapunov stochastic optimization, a novel network design is proposed in which latency and reliability constraints are taken into account.

The fundamental issue tackled in this work is to minimize the system power consumption subject to latency constraints. Considering a MEC system, a UE is able to compute parts of the incoming task, thus the task could be partially offloaded to the TRPs such as EN and MCgNB. Moreover, the most

powerful TRP (e.g., MCgNB) is able to handle the rest of the offloading task from others with limited capacity (e.g., EN). To minimize the task offloading time, UEs need to directly transmit the task to the TRPs wirelessly. Under these assumptions, the main contributions of this paper are:

• A green multi-point task distribution MEC system is proposed. The task are partially transmitted to the TRPs wirelessly and computed.

• Both queueing systems and transmission rates are considered. With regard to transmission rate, signal strength is one of the most important factors.

• An efficient system power consumption algorithm is proposed, which yields the task distribution probability and the minimum system power.

The rest of this paper is organized as follows. Section II presents the MEC system model, computation delay and power consumption formulas. In section III, an objective function aiming to minimize the system power consumption is formulated. In section IV, an algorithm is proposed to solve the problem. The simulation results and discussion are carried out in section V. Finally, the conclusion is given in section VI.

II. System model

The MEC system scenario is as follows. As shown in Fig. 2, there are a few ENs and a MCgNB in the system with targeted UEs spread within the coverage of both EN and MCgNB. To simplify the analysis of UE offloading behavior, the UEs are divided into three groups, i.e., S_{C2EN}, S_{C2MC} and S_{FFMC}. Group S_{C2EN} represents the 40% of UEs which are the closest to EN, while group S_{C2MC} and S_{FFMC} are the half of the rest UEs which are closer to and farther from MCgNB, respectively. The behavior of the UEs in the three groups will be discussed in the simulation section. UEs could deal with part of the task by itself and offload the rest to the nodes such as EN and MCgNB. Compared with UE and edge node, the data processing latency of MCgNB is relatively tiny and could be ignored. Under their computational capabilities, EN and MCgNB could take the offloading requirement in case that they start to serve the UEs. In addition, the queueing system at UE, EN and MCgNB is an independent M/D/1 system. Orthogonal Frequency Division Multiple Access (OFDMA) is adopted in this scenario, where the sub-carriers and bandwidth are well organized for the UEs without interference, and the power consumption for each UE on one sub-carrier is fixed. In the following, the scenario will be focused on the UEs within one EN coverage.

A. Task distribution formulation

Assume there are U UEs in the MEC system scenario. For UE i, the rate of the task packets generated in a specific time is $X_u(i)$, and the expectational task distribution probability vector can be defined as

$$A(i) = [\alpha_u(i), \alpha_e(i), \alpha_m(i)], \quad \alpha_u(i) + \alpha_e(i) + \alpha_m(i) = 1. \tag{1}$$

where $\alpha_u(i), \alpha_e(i)$ and $\alpha_m(i)$ represent for the probabilities of the task distributed to the UE, EN and MCgNB side, the packet expectational arrival rate at UE, EN and MCgNB are $\lambda_u(i) = \alpha_u(i)X_u(i), \lambda_e(i) = \alpha_e(i)X_u(i)$ and $\lambda_m(i) = \alpha_m(i)X_u(i)$ as shown in Fig. 3. In such queueing system, even though "packet" is the smallest unit when distributing the task, $\lambda_u(i)$, $\lambda_e(i)$ and $\lambda_m(i)$ represent individual expectational packet arrival rate are unnecessary to be an integer.

B. Transmission power consumption and delay formulation

Since the link between UE and EN\MCgNB are wireless, by Shannon capacity theory, the transmission rate can be formulated as

$$r_{e\backslash m}(i) = B\log_2(1 + SNR_{e\backslash m}(i)). \tag{2}$$

Considering an OFDMA system, B represents for the bandwidth of one sub-carrier. $SNR_e(i) = \frac{G_e(i)P_u}{BN_0}$, which is the signal to noise ratio between UE i and EN. $G_e(i)$ is the channel gain related to the channel gain between UE and EN, and P_u is the transmission power on one sub-carrier of UE. The transmission power of UE to EN and MCgNB are assumed to be identical and fixed. N_0 is power spectrum density of AWGN. $SNR_m(i)$ is the signal to noise ratio between UE i and MCgNB, which is similar to that of EN.

In this system scenario, the link between UE-EN and UE-MCgNB are wireless and line-of-sight (LOS). For UE i, a path-loss (PL) formula in dB can be found in 3GPP TS 36.814 [16][1]. Having the channel gain and carrier frequency, the transmission rate could be obtained.

Regardless of the transmission queueing power consumption, having the transmission rates, the transmission time of offloaded packets to EN and MCgNB can be derived by $T_{e\backslash m}(i) = \frac{L\cdot 8\cdot\lambda_{e\backslash m}(i)}{r_{e\backslash m}(i)}$, L is the number of byte in one packet. Assuming that the transmission power on one sub-carrier is fixed regardless to CN or MCgNB, the overall power consumption $J_{tx,e\backslash m}(i)$ is proportional to the transmission time. That is

$$J_{tx,e\backslash m}(i) = P_u T_{e\backslash m}(i). \tag{3}$$

It should be noticed that when the probability of the offloading task is decided, given the single sub-carrier transmission rate to each node ($r_{e\backslash m}(i)$), the power consumption used to deliver the offloading tasks is determined however many sub-carriers are chosen. The only difference is that the more the sub-carriers are used, the faster the transmission process is completed.

The transmission queueing delay is also ignored, only the transmission time to CN and MCgNB is considered respectively. That is,

$$Q_{tx,e\backslash m}(i) = \frac{L\cdot 8\cdot\lambda_{e\backslash m}(i)}{N_{e\backslash m}(i)r_{e\backslash m}(i)}. \tag{4}$$

where $N_{e\backslash m}(i)$ is the number of the sub-carriers used to transmit the distributed task to EN and MCgNB, individually.

[1][16] In microcell scenarios, the LOS PL formula is $[16.9\log_{10}(d) + 32.8 + 20\log_{10}(f_c)]$. Where $3m < d < 100m$, f_c is carrier frequency given in GHz.

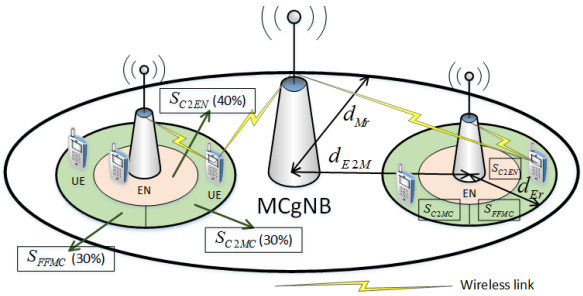


Fig. 2: Task offloading scenario in MEC network

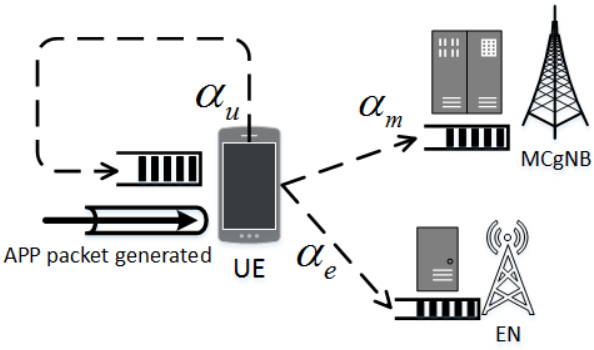


Fig. 3: Task offloading diagram

C. Computation power consumption and delay formulation

Regardless of the basic system supporting power, only computing power is considered. Refer to literature [11], both the computational power consumption of UE and EN are assumed to be linearly proportional to the bits of the task which needs to be computed. Thus, the power consumption formula is

$$J_{com,u\backslash e}(i) = \theta_{u\backslash e}\lambda_{u\backslash e}(i). \tag{5}$$

where $\theta_{u\backslash e}$ represents the computational power consumption per bit at UE\EN part. As for MCgNB, it is assumed to be always powered up by an electric grid and thus all of the power consumption could be ignored.

Next, the discussion of computation delay is as follows. Typically, the computation capability in MCgNB should outperform UE and EN. Thus, only the computation latency of UE and EN are considered in this scenario. Refer to literature [6], the delay formula is in the following.

$$Q_{com,u\backslash e}(i) = \frac{\lambda_{u\backslash e}(i)}{2\mu_{u\backslash e}(\mu_{u\backslash e} - \lambda_{u\backslash e}(i))} + \frac{1}{\mu_{u\backslash e}} + \kappa_{u\backslash e}\lambda_{u\backslash e}(i). \tag{6}$$

where $\mu_{u\backslash e}$ denotes the computation efficiency in UE\EN side. The computation efficiency of each UE (μ_u) is assumed to be the same and fixed, and each UE is fairly and independently assigned with computational resource of EN (μ_e.) The third term $\kappa_{u\backslash e}\lambda_{u\backslash e}(i)$ is the actual computation time of the given task, and $\kappa_{u\backslash e}$ is the computation time index incurred to the volume of the task.

III. Problem Formulation

Assuming the information of channel conditions and the computation capability of each nodes are well known in the system, given the maximum delay constraint, the appropriate task offloading strategies and system (UE and EN) power consumption could be found. In addition, regardless of the handover and mobility, the uplink and downlink transmission behaviors are quite similar. Therefore, only uplink and computation are considered in the following part.

A. Overall power consumption model

The power at MCgNB side is assumed to be sufficient, thus only the power consumption of UE and EN are taken into account. The overall power consumption formulas at UE and EN part are shown as below.

$$\begin{cases} J_{all,u}(i) = J_{com,u}(t) + J_{tx,e}(i) + J_{tx,m}(i), \\ J_{all,e}(i) = J_{com,e}(i). \end{cases} \tag{7}$$

For UE i, $J_{all,u}(i)$ is the overall power consumption at the UE part, and this formula counts one computational power term and two task offloading transmission power terms. $J_{all,e}(i)$ is the power consumption at the EN part, in which only the computational power term is considered.

B. Overall delay model

The overall serving delay (transmission and/or computing delay) at UE, EN and MCgNB side are in the following.

$$\begin{cases} Q_{all,u}(t) = Q_{com,u}(t), \\ Q_{all,e}(t) = Q_{com,e}(t) + Q_{tx,e}(t) \\ Q_{all,m}(t) = Q_{tx,m}(t). \end{cases} \tag{8}$$

$Q_{all,u}(t)$ is the serving delay at UE part, only the computation delay is considered. At the EN part, computation delay and transmission delay are taken into account in $Q_{all,e}(t)$, and only the transmission delay are counted in $Q_{all,m}(t)$.

C. Objective function formulation

With the power consumption model and delay model in section III-A and III-B, the objective function aims to minimize the power consumption at UE and EN part can thus be formulated. In objective function Eq. 9, there are U UEs under both the coverage of MCgNB and EN, and the delay constraint is set to be D_{max}.

$$\begin{aligned} \text{min.} \quad & \sum_{i=1}^{U}(J_{all,u}(i) + J_{all,e}(i)) \\ \text{s.t.} \quad & C_1 : \alpha_u(i) + \alpha_e(i) + \alpha_m(i) = 1, \\ & C_2 : \alpha_u(i), \alpha_e(i), \alpha_m(i) \in [0,1], \\ & C_3 : Q_{all,u}(i) \le D_{max}, \\ & C_4 : Q_{all,e}(i) \le D_{max}, \\ & C_5 : Q_{all,m}(i) \le D_{max}, \\ & C_6 : \lambda_u(i) < \mu_u, \lambda_e(i) < \mu_e, \lambda_{u\backslash e\backslash m}(i) \in R^+, \\ & C_7 : N_{e\backslash m}(i) \in Z. \end{aligned} \tag{9}$$

C_1 and C_2 are the offloading probability constraints, and C_3, C_4 and C_5 are the delay constraints. In order not to be overwhelmed by the incoming task packets, C_6 is the constraint to ensure the transmitted packets would never be dropped by UE and EN. C_7 is the sub-carrier utilization condition, no further limit is set here.

After substituting Eq. 4, Eq. 3, Eq. 6 and Eq. 5 into the original objective function Eq. 9, the objective function is

$$\begin{aligned}
\text{min.} \quad & \sum_{i=1}^{U}(\theta_u\lambda_u(i) + P_uT_e(i) + \theta_e\lambda_e(i) + P_uT_m(i)) \\
\text{s.t.} \quad & C_1 : \alpha_u(i) + \alpha_e(i) + \alpha_m(i) = 1, \\
& C_2 : \alpha_u(i), \alpha_e(i), \alpha_m(i) \in [0,1], \\
& C_3 : \frac{\lambda_u(i)}{2\mu_u(\mu_u - \lambda_u(i))} + \frac{1}{\mu_u} + \kappa_u\lambda_u(i) \le D_{max}, \\
& C_4 : \frac{\lambda_e(i)}{2\mu_e(\mu_e - \lambda_e(i))} + \frac{1}{\mu_e} + \kappa_e\lambda_e(i) \\
& \quad + \frac{L\cdot 8\cdot\lambda_e(i)}{N_e(i)r_e(i)} \le D_{max}, \\
& C_5 : \frac{L\cdot 8\cdot\lambda_m(i)}{N_m(i)r_m(i)} \le D_{max}, \\
& C_6 : \lambda_u(i) < \mu_u, \lambda_e(i) < \mu_e, \lambda_{u\backslash e\backslash m}(i) \in R^+, \\
& C_7 : N_{e\backslash m}(i) \in Z.
\end{aligned} \tag{10}$$

IV. Problem solution

Since each UE is given with independent computation resources of EN and do not interfere with each other, minimizing system power consumption implies minimizing the power consumption of each UE. Consequently, finding out the optimal solution of Eq. 10 is equivalent to finding the minimum power consumption of each UE.

There should be two bottlenecks in Eq. 10, the computation delay and transmission bandwidth constraint. However, if the bandwidth constraint (e.g., the number of utilizing sub-carrier) is jointly considered, no optimal solution will be derived when the latency requirement is strict. In fact, the objective in this work is to minimize the system serving power, the sub-carrier utilization condition is less important. Thus, in section IV-A the constraints of sub-carrier utilization are released, and Eq. 10 becomes a linear first order box-constrained problem. In section IV-B, while the objective function is simplified, the offloading probability could be figured out. In section IV-C, the exact sub-carrier number could be calculated back. The three steps are discussed in detail next.

A. Step 1: Set the value of N_e and N_m to infinity

Free from transmission bandwidth constraint, N_e and N_m go to infinity first and the transmission delays become close to zero, C_3 and C_4 in Eq. 10 become unary quadratic inequalities and the objective function becomes a linear first order box-constrained problem such as Eq. 11. Theoretically, some optimal solutions will be obtained when the computation processing at EN side exactly takes all of the available time. However, it is impractical that the transmission time is absolute zero. Thus, an additional transmission guarantee gap ratio ε is given in C_4 Eq. 10, that is, $Q_{com,e} \le D_{max}(1-\varepsilon)$. The solution and the exact sub-carrier utilization number may change with ε. When ε is close to zero, the solution approaches optimal one but the sub-carrier utilization number may go large. In addition, in Eq. 10 constraint C_5 could be replaced by the default constraint ($0 \le \alpha_m(i) \le 1$), constraint C_2 and C_6 could also be merged into C_3 and C_4. As a result, for UE i, the objective function are as follows.

$$\begin{aligned}
\text{min.} \quad & K_u\alpha_u(i) + K_e\alpha_e(i) + K_m\alpha_m(i) \\
\text{s.t.} \quad & C_1 : \alpha_u(i) + \alpha_e(i) + \alpha_m(i) = 1, \\
& C_3 : 0 \le \alpha_u(i) \le \Phi_u, \\
& C_4 : 0 \le \alpha_e(i) \le \Phi_e, \\
& C_5 : 0 \le \alpha_m(i) \le \Phi_m.
\end{aligned} \tag{11}$$

where $K_u = \theta_u X_u$, $K_e = (\frac{P_u\cdot L\cdot 8}{r_e(i)} + \theta_e)X_u$ and $K_m = \frac{P_u\cdot L\cdot 8\cdot X_u}{r_m(i)}$. Let the feasible space of C_3 and C_4 in Eq. 10 are S_3 and S_4 respectively when N_e and N_m go to infinity, the upper bound of C_3 and C_4 in Eq. 11 can be defined as $\Phi_{u\backslash e} = \max\{0, \sup\{S_{3\backslash 4} \cap [0, \frac{\mu_{u\backslash e}}{X_u}]\}\}$. Where $\sup\{\cdot\}$ is the supremum[2] of $\{\cdot\}$. C_5 is the default constraint, $\Phi_m = 1$.

B. Step 2: find out the value of α_u, α_e and α_m

In Eq. 11, as a linear first order box-constrained problem, the optimal solution is at the boundary [17] and can be figured out by the following operation. First, set vector $G(i)$ to be the ascend-sorting(ASC) vector of K_u, K_e and K_m, ($G(i) = ASC\{K_u(i), K_e(i), K_m(i)\}$), and vector $\Phi(i)$ to be the constraint vector corresponding to vector $G(i)$, that is $\Phi(i) = \{\Phi_{j_1}(i), \Phi_{j_2}(i), \Phi_{j_3}(i)\}$. The relationship between $\{u, e, m\}$ and $\{j_1, j_2, j_3\}$ is defined as below

$$\begin{cases}
j_1 = \arg\min_{u,e,m}\{K_u, K_e, K_m\}, \\
j_2 = \arg\min_{u,e,m}\{K_u, K_e, K_m\}\backslash K_{j_1}, \\
j_3 = \arg_{u,e,m}\{K_u, K_e, K_m\}\backslash\{K_{j_1}, K_{j_2}\}.
\end{cases} \tag{12}$$

Then, the optimal solution of distributing probability vector $A(i)^* = \{\alpha_{j_1}, \alpha_{j_2}, \alpha_{j_3}\}$ is

$$\begin{cases}
\alpha_{j_1} = \Phi_{j_1}, \\
\alpha_{j_2} = \min\{(1-\alpha_{j_1}), \Phi_{j_2}\}, \\
\alpha_{j_3} = \max\{0, (1-\alpha_{j_1}-\alpha_{j_2})\}.
\end{cases} \tag{13}$$

Finally, the approximate minimum power consumption of UE i could be figured out by computing the inner product $P(i)^* = G(i)\cdot A(i)^*$. Once all the $P(i)^*$ are acquired, the system minimum offloading power consumption value can be calculated by $P^*_{all} = \sum_{i=1}^{U} P(i)^*$.

C. Step 3: find out the exact value of N_e and N_m

Once the optimal distribution probabilities $\alpha^*_u(i), \alpha^*_e(i)$ and $\alpha^*_m(i)$ are obtained, the minimum value of carrier utilization N_e and N_m can be derived by the original two constraints C_4 and C_5 in Eq. 10.

The overall algorithm is presented as Algorithm 1.

[2] $\sup\{\varnothing\} \equiv -\infty$.

Algorithm 1 Green task offloading algorithm

Input:

The system settings[3].

The objective function Eq. 10.

Output:

for all the UE, $i = 1$ to U **do**

Let $N_e, N_m \to \infty$, give ε in Eq. 10.

obtain Eq. 11.

Derive the upper bound of C_3, C_4 and C_5 in Eq. 11

$\Phi_{u\backslash e} = \max\{0, \sup\{S_{3\backslash 4} \cap [0, \frac{\mu_{u\backslash e}}{X_u}]\}\}$, $\Phi_m = 1$.

Let $G = ASC\{K_u, K_e, K_m\}$

$K_u = \theta_u X_u$,

$K_e = (\frac{P_u \cdot L \cdot 8}{r_e(i)} + \theta_e) X_u$,

$K_m = \frac{P_u \cdot L \cdot 8 \cdot X_u}{r_m(i)}$.

Define the ASC index space $\{j_1, j_2, j_3\}$;

$j_1 = \arg\min_{u,e,m}\{K_u, K_e, K_m\}$,

$j_2 = \arg\min_{u,e,m}\{K_u, K_e, K_m\} \backslash K_{j_1}$,

$j_3 = \arg_{u,e,m}\{K_u, K_e, K_m\} \backslash \{K_{j_1}, K_{j_2}\}$.

Obtain the upper bound vector $\Phi = \{\Phi_{j_1}, \Phi_{j_2}, \Phi_{j_3}\}$;

Compute $A^* = \{\alpha_{j_1}, \alpha_{j_2}, \alpha_{j_3}\}$;

$\alpha_{j_1} = \Phi_{j_1}$,

$\alpha_{j_2} = \min\{(1 - \alpha_{j_1}), \Phi_{j_2}\}$,

$\alpha_{j_3} = \max\{0, (1 - \alpha_{j_1} - \alpha_{j_2})\}$.

Compute $P^* = G \cdot A^*$;

Substitute A^* into Eq. 10 C_3 and C_4

Compute N_e^*, N_m^*.

return A^*, P^*, N_e^*, N_m^*;

end for

$P_{all}^* = \sum_{i=1}^{U} P(i)^*$;

return P_{all}^*.

V. Simulation Results

The goal of this section is to figure out the theoretical performance of the proposed algorithm. The settings of the following simulations are as below[3]. The radius of EN(d_{Er}) and MCgNB(d_{Mr}) are 20 and 100 meters. All the UEs are uniformly distributed under the coverage of EN.

In Fig. 4, μ_e is fixed to be 12, the two ratios of EN to MCgNB distance(d_{E2M}) over EN radius(d_{Er}) are discussed, $R_{E2M,Er} = \frac{d_{E2M}}{d_{Er}}$. In Fig. 4a, the path loss is not that sever to the UEs in group S_{C2MC} when transmitting tasks to MCgNB. So, these UEs could offload more tasks to the MCgNB than others to save transmit power. However, when the value of $R_{E2M,Er}$ grows, the path loss between UEs and MCgNB goes poorly whatever groups the UEs are in. Therefore, the offloading probabilities are almost the same in Fig. 4b. In summary, these results show that the location of EN is essential especially in LOS scenario, it may affect the offloading distribution probability.

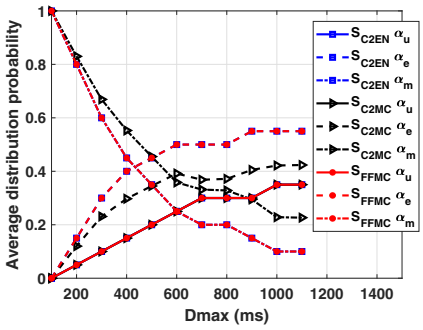
Next, the value of the $R_{E2M,Er}$ is fixed to 2.5 in Fig. 5 and Fig. 6. In Fig. 5, when the delay requirement is critical, most of the packets are unable to be handled by UEs and EN are offloaded to MCgNB. Transmitting such number of packets to the MCgNB is highly power-hungry, and this explains why the power consumption is large when the value of D_{max} is small. Obviously, the UEs in group S_{FFMC}, located the farthest away from the MCgNB consume the most of the power. However, it is not that intuitive that the power consumption of the UEs in group S_{C2EN} is greater than that of group S_{C2MC}. Theoretically, the UEs closer to EN should take less power to offload the task. As previously mentioned in IV-A, when the transmission constraint is released, the bottleneck of the system becomes only the computation capabilities of UEs and EN. Therefore, once EN is insufficiently powerful and a lot of packets are still await being transmitted to MCgNB, the UEs in group S_{C2MC}, which are the closest to the MCgNB, have greater advantage in packet transmitting. Furthermore, when the computation efficiency of EN is enhanced (e.g., μ_e from 12 to 16), the power consumption gap between all groups reduces faster. Then, considering the binary distribution strategy. In this simulation, the packet arrival rate is greater than computation efficiency of UE and EN ($X_u > \mu_{u\backslash e}$.) To avoid the traffic intensity ($\rho = \frac{X_u}{\mu_{u\backslash e}}$) exceeds one, all the tasks shall be forwarded to MCgNB whatever the delay tolerance is. Therefore, the power consumption without partial distribution should always be the same as the dash line in Fig. 5.

The sub-carrier utilization number is shown in Fig. 6. When the value of D_{max} is small, most of the packets need to be offloaded to MCgNB, thus the number of sub-carrier utilization is quite large. For all the UEs in group S_{C2EN}, S_{C2MC} and S_{FFMC}, the sub-carrier number used to transmit to MCgNB homogeneously decreases when D_{max} raises. Additionally, the sub-carrier number used to deliver packets to EN is relatively stationary. Limited by the computation capability, EN is unable to compute excessive packets per time unit. As a result, the ratio of the offloaded packets per time unit to EN is relatively fixed. Besides, there are some peaks appearing in the curves of EN (e.g., when $D_{max} = 900$ ms.) It is because EN is able to deal with more packets within the delay constraints, which will shorten the transmission time and thus UEs need to use more sub-carriers to compensate such condition. Finally, when adding a transmission constraint, some delay constraints will have no feasible solution. For instance, assuming the maximum sub-carrier utilization number is 1200 as the green line in Fig. 6, the UEs in group S_{FFMC} cannot make it until the delay constraint goes to 600 ms. However, the UEs in group S_{C2MC} could have a way to offload the tasks when D_{max} is just 300 ms.

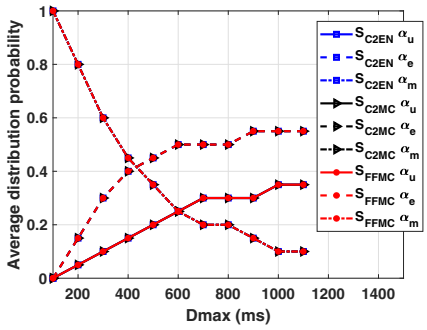
VI. Conclusions

Delay sensitive applications are key to the next generation communication systems, and MEC and fog computing are instrumental in enabling a HetNet system. In this paper, a multi-node partial task offloading MEC system is discussed and a green task offloading algorithm is proposed. Meeting the latency requirement of the application, the proposed algorithm could appropriately set the task distribution probability in order to yield the minimum system power consumption. The

[3] $U = 100$, $Pu = 20$ dBm, $Xu = 20$ packet/s, $B = 15$ kHz, $f_c = 2.4$ GHz, $L = 1500$ bytes, $\theta_u = 10^{-6}$ (J/bit), $\theta_e = 5 * 10^{-6}$ (J/bit), $\mu_u = 8, \kappa_u = 0.05, \kappa_e = 0.025, \varepsilon = 0.05$, $N_0 = -174$ dBm/Hz.



(a) $R_{E2M,Er} = 1.5$



(b) $R_{E2M,Er} = 2.5$

Fig. 4: Task offloading probability

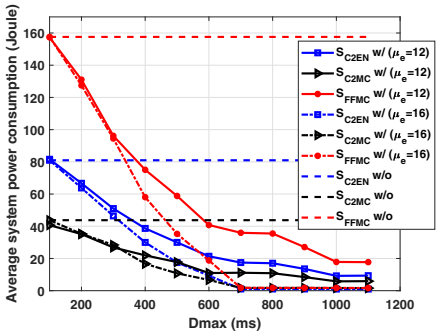


Fig. 5: System power consumption

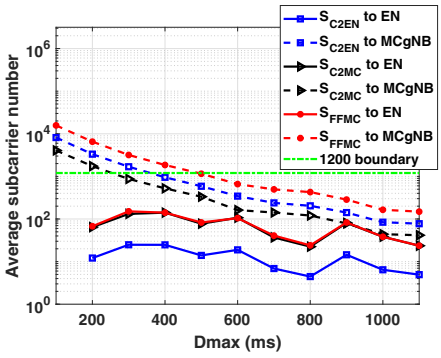


Fig. 6: System sub-carrier utilization

simulation results show that factors such as the location of TRP nodes, the wireless signal strength and nodes' computation capability may significantly impact the task offloading strategies and system power consumption. In addition, the merits of the partial task offloading are revealed in this paper. Compared with the binary offloading strategy, the partial offloading strategy outperforms thanks to its flexibility and power-saving ability. As a result, such partial offloading methodology should be taken into account when designing a MEC system strategy especially in capability limited systems and for green communication purposes. Future work will extend the current framework to incorporate latency and reliability [18].

ACKNOWLEDGEMENT

The research is partly funded by MOST of Taiwan under grant 106-2923-E-002-015-MY3, 107-2923-E-002-006-MY3, and 105-2221-E-002 -014-MY3, and by the EU Horizon 2020 grant No 761745.

REFERENCES

[1] *Design and Deployment of Small Cell Networks*. Cambridge University Press, 2015.

[2] ETSI MEC ISG, "Mobile edge computing (MEC); Technical Requirements," GS 002, Feb 2018, v 2.1.1.

[3] ——, "Mobile edge computing (MEC); Framework and Reference Architecture," GS 003, Nov 2017, v 2.1.1.

[4] Y. J. Ku, D. Y. Lin, C. F. Lee, P. J. Hsieh, H. Y. Wei, C. T. Chou, and A. C. Pang, "5g radio access network design with the fog paradigm: Confluence of communications and computing," *IEEE Communications Magazine*, vol. 55, no. 4, pp. 46–52, April 2017.

[5] T. Y. Kan, Y. Chiang, and H. Y. Wei, "Task offloading and resource allocation in mobile-edge computing system," in *2018 27th Wireless and Optical Communication Conference (WOCC)*, April 2018, pp. 1–4.

[6] G. Lee, W. Saad, and M. Bennis, "An online secretary framework for fog network formation with minimal latency," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[7] L. Chen, J. Xu, and S. Zhou, "Computation peer offloading in mobile edge computing with energy budgets," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[8] J. Zhang, X. Hu, Z. Ning, E. C. H. Ngai, L. Zhou, J. Wei, J. Cheng, and B. Hu, "Energy-latency trade-off for energy-aware offloading in mobile edge computing networks," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2017.

[9] S. Mao, S. Leng, K. Yang, Q. Zhao, and M. Liu, "Energy efficiency and delay tradeoff in multi-user wireless powered mobile-edge computing systems," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[10] J. Zhang, W. Xia, Y. Zhang, Q. Zou, B. Huang, F. Yan, and L. Shen, "Joint offloading and resource allocation optimization for mobile edge computing," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[11] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[12] H. Q. Le, H. Al-Shatri, and A. Klein, "Efficient resource allocation in mobile-edge computation offloading: Completion time minimization," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2513–2517.

[13] M. S. Elbamby, M. Bennis, and W. Saad, "Proactive edge computing in latency-constrained fog networks," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–6.

[14] J. Xu, L. Chen, and S. Ren, "Online learning for offloading and autoscaling in energy harvesting mobile edge computing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 3, pp. 361–373, Sept 2017.

[15] C. F. Liu, M. Bennis, and H. V. Poor, "Latency and reliability-aware task offloading and resource allocation for mobile edge computing," in *2017 IEEE Globecom Workshops (GC Wkshps)*, Dec 2017, pp. 1–7.

[16] 3GPP, "Evolved universal terrestrial radio access (e-utra); further advancements for e-utra physical layer aspects r9," TR 36.814, Mar 2017, v 9.2.0.

[17] C.-Y. Chi, W.-C. Li, and C.-H. Lin, *Convex Optimization for Signal Processing and Communications: From Fundamentals to Applications*, 2017.

[18] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *Proceedings of the IEEE*, Oct 2018, in press.