Optimal Scalable MEC Systems based on TDD-OFDMA

Wuyang Jiang\(^1\), Yunlong Gao\(^2\), Ying Cui\(^2a\), and Zhi Liu\(^3\)
\(^1\) Shanghai University of Engineering Science
333 Longteng Rd., Shanghai 201620, China
\(^2\) Department of Electronic Engineering, Shanghai Jiao Tong University
800 Dongchuan Rd., Shanghai 200240, China
\(^3\) School of Informatics and Engineering, The University of Electro-Communications
Tokyo 182-8585, Japan

\(^a\) cuying@sjtu.edu.cn

Abstract: This letter investigates the optimal tradeoff between user experience and communications and computation resource consumptions in designing efficient mobile edge computing (MEC) systems which are based on orthogonal frequency division multiple access (OFDMA) and operating in time division duplexing (TDD) mode. First, we establish a communication model for offloading scalable tasks in TDD-OFDMA systems. Then, we formulate the system utility maximization problem with respect to the service level selection and resource allocation. Finally, we propose two algorithms to obtain a globally optimal solution and a low-complexity suboptimal solution, respectively, using convex optimization and difference of convex (DC) programming.

Keywords: Mobile edge computing, OFDMA, TDD, resource allocation, service level selection, optimization.

Classification: Wireless communication technologies

References

[1] M. Li, S. Yang, Z. Zhang, J. Ren, and G. Yu, “Joint subcarrier and power allocation for OFDMA based mobile edge computing system,” in PIMRC, Oct 2017, pp. 1–6. DOI:10.1109/PIMRC.2017.8292207
[2] Y. Yu, J. Zhang, and K. B. Letaief, “Joint subcarrier and cpu time allocation for mobile edge computing,” in IEEE GLOBECOM, Dec 2016, pp. 1–6. DOI:10.1109/GLOCOM.2016.7841937
[3] J. Z. et al., “Energy-latency tradeoff for energy-aware offloading in mobile edge computing networks,” IEEE Internet Things J., vol. 5, no. 4, pp. 2633–2645, Aug 2018. DOI:10.1109/JIOT.2017.2786343
[4] X. Zhang, Y. Mao, J. Zhang, and K. B. Letaief, “Multi-objective resource allocation for mobile edge computing systems,” in PIMRC, Oct 2017, pp. 1–5. DOI:10.1109/PIMRC.2017.8292379
[5] Y. Gao, Y. Cui, X. Wang, and Z. Liu, “Optimal resource allocation for scalable mobile edge computing,” IEEE Commun. Lett., vol. 23, no. 7, pp. 1211–1214, July 2019. DOI:10.1109/LCOMM.2019.2916075
1 Introduction

Optimal communications and computation resource allocation for MEC systems based on OFDMA has received significant interest [1, 2, 3, 4]. Note that [2, 3, 4] assume that computation results are negligible and hence do not consider the transmission of computation results. Besides, [2, 1, 3, 4] assume that each computation task has only one quality level. However, for some applications such as object detection, there may exist multiple quality levels for a computation task associated with different computation workloads or/and computation results of possibly distinct sizes [5]. In this letter, we would like to address the above limitations. First, we establish a communication model for scalable tasks with non-negligible sizes of computation results and task execution durations in a TDD-OFDMA system. Then, we formulate system utility maximization problem with respect to (w.r.t.) the service level selection and resource allocation. Finally, we obtain a globally optimal solution and a low-complexity suboptimal solution using optimization techniques.

2 System Model

We consider a multi-user MEC system consisting of one serving node with powerful computing capability and $K$ users represented by $\mathcal{K} \triangleq \{1, 2, ..., K\}$. Each user $k \in \mathcal{K}$ has one computation task, i.e., task $k$. All tasks are assumed to be offloaded to the serving node for execution due to each user’s limited computing capability. All tasks are generated at time 0 and completed by deadline $T$ (in seconds) [5]. We adopt the scalable computation task model in our previous work [5], which enables a tradeoff between user experience and consumptions of communications resource and computation resource. In particular, each task $k$ has $L_k$ service levels that are denoted by $\mathcal{L}_k \triangleq \{1, 2, ..., L_k\}$. We characterize task $k$ at a certain level $l \in \mathcal{L}_k$ by four parameters: the size of the task before computation $N_{u,k,l}$ (in bits), workload $N_{e,k,l}$ (in number of CPU cycles), size of the computation result $N_{d,k,l}$ (in bits), and utility $U_{k,l}$ (which can reflect user experience).

Let $s_{k,l}$ denote the service level selection variable for task $k$ and level $l$. Only one service level can be selected for each task. Thus, we have the following service level selection constraints:

$$s_{k,l} \in \{0, 1\}, \quad l \in \mathcal{L}_k, \quad k \in \mathcal{K},$$

(1)

$$\sum_{l \in \mathcal{L}_k} s_{k,l} = 1, \quad k \in \mathcal{K}.\quad (2)$$

$s_{k,l} = 1$ indicates that service level $l$ is selected for task $k$, and $s_{k,l} = 0$ otherwise. Given $s_k \triangleq (s_{k,l})_{l \in \mathcal{L}_k}$, the size before computation, workload, size of computation result and utility for task $k$ at the selected service level are given by $N_{u,k}(s_k) \triangleq \sum_{l \in \mathcal{L}_k} s_{k,l} N_{u,k,l}$, $N_{e,k}(s_k) \triangleq \sum_{l \in \mathcal{L}_k} s_{k,l} N_{e,k,l}$, $N_{d,k}(s_k) \triangleq \sum_{l \in \mathcal{L}_k} s_{k,l} N_{d,k,l}$ and $U_k(s_k) = \sum_{l \in \mathcal{L}_k} s_{k,l} U_{k,l}$, respectively. Given $s \triangleq (s_k)_{k \in \mathcal{K}}$, the total utility of the MEC system is $U(s) \triangleq \sum_{k \in \mathcal{K}} U_k(s_k) = \sum_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}_k} s_{k,l} U_{k,l}$.

1The results can be extended to the case with offloading scheduling or different task arrival times.
In contrast with our previous work [5], here we consider an OFDMA system in TDD mode. There are $N$ subcarriers, denoted by $\mathcal{N} \triangleq \{1, 2, ..., N\}$. Let $h_{k,n}$ denote the channel power on the $n$-th subcarrier for user $k$, assumed to be constant over time duration $[0, T]$. Define $\rho_{u,k,n}$ and $\rho_{d,k,n}$ as the uploading and downloading subcarrier assignment indicator variables for user $k$ and subcarrier $n$. One subcarrier is allocated to one user. Thus, we have the following subcarrier selection constraints:

$$\rho_{u,k,n} \in \{0, 1\}, \quad k \in \mathcal{K}, \quad n \in \mathcal{N},$$  \hspace{1cm} (3) $$\rho_{d,k,n} \in \{0, 1\}, \quad k \in \mathcal{K}, \quad n \in \mathcal{N},$$  \hspace{1cm} (4) $$\sum_{k \in \mathcal{K}} \rho_{u,k,n} = 1, \quad n \in \mathcal{N},$$  \hspace{1cm} (5) $$\sum_{k \in \mathcal{K}} \rho_{d,k,n} = 1, \quad n \in \mathcal{N}. $$  \hspace{1cm} (6)

$\rho_{o,k,n} = 1$ indicates that subcarrier $n$ is assigned to mobile $k$, and $\rho_{o,k,n} = 0$ otherwise, where $o \in \mathcal{O} \triangleq \{u, d\}$.

Let $r_{u,k,n}$ and $r_{d,k,n}$ denote the maximum transmission rates (in bit/s) for uploading and downloading on the $n$-th subcarrier for mobile $k$. Given $\rho_{o,k,n}$ and $r_{o,k,n}$, the transmission power (in W) on the $n$-th subcarrier for user $k$ can be expressed as $n_0 h_{k,n} \left(2^{\frac{r_{o,k,n}}{B \rho_{o,k,n}}} - 1\right)$, where $B$ (in Hz) and $n_0$ are the bandwidth and the complex additive white Gaussian channel noise power for each subcarrier, respectively. The transmission power constraints at user $k$ for uploading and at the serving node for downloading are given by:

$$\sum_{n \in \mathcal{N}} \rho_{u,k,n} n_0 h_{k,n} \left(2^{\frac{r_{u,k,n}}{B \rho_{u,k,n}}} - 1\right) \leq P_k, \quad k \in \mathcal{K},$$  \hspace{1cm} (7) $$\sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}} \rho_{d,k,n} n_0 h_{k,n} \left(2^{\frac{r_{d,k,n}}{B \rho_{d,k,n}}} - 1\right) \leq P_0. $$  \hspace{1cm} (8)

$P_k$ and $P_0$ denote the transmission power limits (representing the communications resource consumptions) for user $k$ and the serving node, respectively.

Given $s_k$ and $r_{o,k} \triangleq (r_{o,k,n})_{n \in \mathcal{N}}$, the corresponding uploading, downloading and executing durations (in seconds) for task $k$ are given by $t_{o,k}(s_k, r_{o,k}) = \frac{N_{o,k}(s_k)}{\sum_{n \in \mathcal{N}} r_{o,k,n}}$, $o \in \mathcal{O}$ and $t_{e,k}(s_k) = \frac{N_{e,k}(s_k)}{F}$. Here, $F$ represents the CPU frequency (in number of CPU cycles per second) of the MEC server at the serving node, which can be interpreted as the consumption of computation resource at the server. In the TDD-OFDMA system, for ease of implementation, assume that the uploading operations of all $K$ tasks are completed before the executing operation of any task and that the executing operations of all $K$ tasks are completed before the downloading operation of any task. Thus, there are three phases, i.e., the uploading phase of time duration $t_u$, the executing phase of time duration $\sum_{k \in \mathcal{K}} t_{e,k}(s_k)$, and the downloading phase of time duration $t_d$. We have the time allocation constraints:

$$t_{u,k}(s_k, r_{u,k}) = \frac{\sum_{l \in \mathcal{L}_k} s_{k,l} N_{u,k,l}}{\sum_{n \in \mathcal{N}} r_{u,k,n}} \leq t_u, \quad k \in \mathcal{K}, $$  \hspace{1cm} (9)
\[ t_{d,k}(s_k, r_{d,k}) = \frac{\sum_{l \in L_k} s_{k,l} N_{d,k,l}}{\sum_{n \in N} r_{d,k,n}} \leq t_d, \quad k \in K, \quad (10) \]
\[ t_u + \sum_{k \in K} t_{e,k}(s_k) + t_d = t_u + \sum_{k \in K} \sum_{l \in L_k} s_{k,l} N_{e,k,l} \leq T. \quad (11) \]

3 Problem Formulation

In this section, the service level selection \( s \), subcarrier allocation \( \rho \triangleq (\rho_{o,k,n})_{o \in O, k \in K, n \in N} \), transmission rate allocation \( r \triangleq (r_{o,k,n})_{o \in O, k \in K, n \in N} \) and transmission time duration allocation \( t \triangleq (t_o)_{o \in O} \) are optimized to maximize the total utility \( U(s) \) under the service level selection constraints, subcarrier selection constraints, transmission power constraints, and time allocation constraints.\(^2\)

Problem 1 (Total Utility Maximization)

\[
U^* \triangleq \max_{r \succeq 0, s, \rho, t} U(s) \\
\text{s.t.} \quad (1), (2), (3), (4), (5), (6), (7), (8), (9), (10), (11).
\]

Let \((s^*, \rho^*, r^*, t^*)\) denote an optimal solution. \((U^*, P, F)\), where \( P \triangleq (P_0, P_1, ..., P_K) \), gives the optimal tradeoff between user experience and communications and computation resource consumptions. Problem 1 is a challenging mixed discrete-continuous problem. It is crucial to reduce the computational complexity caused by the discrete variables.

4 Optimal Solution

In this section, an algorithm is proposed to obtain an optimal solution (under a mild condition). Define \( S \triangleq \{ s : (1), (2) \} \). First, consider the problem:

Problem 2 (Equivalent Problem of Problem 1)

\[
\max_{s \in S} U(s) \\
\text{s.t.} \quad t_u^*(s) + \sum_{k \in K} \sum_{l \in L_k} s_{k,l} N_{e,k,l} \leq T,
\]
where \( t_u^*(s) \) and \( t_d^*(s) \) are given below.

Problem 3 (Uploading Duration Minimization for \( s \in S \))

\[
t_u^*(s) \triangleq \min_{r_{u,0}, \rho_{u,t_u}} t_u \\
\text{s.t.} \quad \rho_{u,k,n} \geq 0, \quad k \in K, \quad n \in N, \quad (5), (7), (9).
\]

Problem 4 (Downloading Duration Minimization for \( s \in S \))

\[
t_d^*(s) \triangleq \min_{r_{d,0}, \rho_{d,t_d}} t_d \\
\text{s.t.} \quad \rho_{d,k,n} \geq 0, \quad k \in K, \quad n \in N, \quad (6), (8), (10).
\]

\(^2\)The optimization w.r.t. the subcarrier and transmission rate allocation is equivalent to but more tractable than that w.r.t. the subcarrier and transmission power allocation.
Let \((\rho_u^*(s), r_u^*(s), t_u^*(s))\) and \((\rho_d^*(s), r_d^*(s), t_d^*(s))\) denote the optimal solutions of Problem 3 and Problem 4, respectively. Let \(\lambda_k^*(s), k \in \mathcal{K}, \mu_k^*(s), \alpha_k^*(s)\) and \(\beta_k^*(s), k \in \mathcal{K}\) denote the optimal Lagrange multipliers w.r.t. the constraints in (7), (9), (8) and (10), respectively. Define the solutions of Problem 3 and Problem 4, respectively. Let \(s^0\) be the solution of Problem 1. First, (9) and (10) can be equivalently converted to:

\[
\text{minimize} \quad \sum_{k \in \mathcal{K}} \lambda_k^0 \left( \mu_{h0} B \left( \frac{\rho_{h0} B}{\lambda_{h0}^0 \ln 2} \right) \right)^{1/2} - \sum_{k \in \mathcal{K}} \lambda_k^0 \left( \mu_{h0} B \left( \frac{\rho_{h0} B}{\lambda_{h0}^0 \ln 2} \right) \right)^{1/2} \ln \left( \frac{\mu_{h0} B \left( \frac{\rho_{h0} B}{\lambda_{h0}^0 \ln 2} \right)^{1/2}}{\lambda_{h0}^0} \right),
\]

where \((x)^\dagger \triangleq \max\{0, x\}\) and \((x)^\ddagger \triangleq \max\{1, x\}\). By exploiting the problem structures and using the KKT conditions, we have:

**Theorem 1** Suppose that for all \(s\in \mathcal{S}\) and \(n \in \mathcal{N}\), there exist unique \(k_n^0\) and \(k_n^\dagger\) such that \(X(h_{k_n^0, n}, \lambda_{k_n^0}^*(s), \mu_{k_n^0}^*(s)) = \min_{j \in \mathcal{J}} X(h_{j, n}, \lambda_{j}^*(s), \mu_{j}^*(s))\) and \(X(h_{k_n^\dagger, n}, \alpha_{k_n^\dagger}^*(s), \beta_{k_n^\dagger}^*(s)) = \min_{j \in \mathcal{J}} X(h_{j, n}, \alpha_{j}^*(s), \beta_{j}^*(s))\). Then, Problem 1 and Problem 2 are equivalent.

Based on Theorem 1, we can solve Problem 2 instead using Algorithm 1.

**Algorithm 1**: Algorithm for Obtaining An Optimal Solution

**Output** \((s^*, \rho^*, r^*, t^*, U^*)\).

1. Set \(x = 0\), and sort \(s\in \mathcal{S}\) by \(U(s)\) from largest to smallest, i.e., \(U(s(1)) \geq \ldots \geq U(s(\mathcal{L})).\)
2. **repeat**
3. Set \(x = x + 1\), and solve Problem 3 and Problem 4 for \(s = s(x)\), using standard convex optimization techniques.
4. **until** \(t_u^*(s(x)) + t_d^*(s(x)) + \sum_{k \in \mathcal{K}} t_o(k, s(x)) \leq T\)
5. Set \(U^* = U(s(x)), s^* = s(x), \rho^* = \rho^*(s(x)), r^* = r^*(s(x))\) and \(t^* = t^*(s(x))\).

5 Suboptimal Solution

To further reduce the computational complexity caused by searching for \(s\), in this section, a low-complexity algorithm is proposed to obtain a suboptimal solution of Problem 1. First, (9) and (10) can be equivalently converted to:

\[
\sum_{l \in \mathcal{L}_k} s_{k,l}^2 n_{o,k,l} \sum_{n \in \mathcal{N}} t_{o,k,n} \leq t_o, \quad o \in \mathcal{O}, \quad k \in \mathcal{K}.
\]

The discrete constraints in (1) can be equivalently converted to:

\[
s_{k,l} \in [0, 1], \quad l \in \mathcal{L}_k, \quad k \in \mathcal{K},
\]

\[
s_{k,l}(1 - s_{k,l}) \leq 0, \quad l \in \mathcal{L}_k, \quad k \in \mathcal{K}.
\]

By replacing (16) with a penalty for violating (16) in the objective function of Problem 1 and relaxing the constraints in (3) and (4), we have:

**Problem 5 (Penalized DC Problem of Problem 1)**

\[
\max_{r \geq 0, s, \rho, t} \quad U(s) - \gamma P(s)
\]

**s.t.** (2), (5), (6), (7), (8), (11), (12), (13), (14).

Here, \(\gamma > 0\) represents the penalty parameter, and \(P(s) \triangleq \sum_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}_k} s_{k,l}(1 - s_{k,l})\) denotes the penalty function. By Theorem 1 and [5], we have:

3 As \(X(h_{j, n}, \lambda_j^*(s), \mu_j^*(s))\), \(j \in \mathcal{K}\) are usually different, the condition can be easily satisfied.
Theorem 2 Suppose that the condition in Theorem 1 holds. Then, there exists $\gamma_0 > 0$ such that for all $\gamma > \gamma_0$, Problems 1 and 5 are equivalent.

Based on Theorem 2, we can solve Problem 5 (for a sufficiently large $\gamma > 0$) instead. $P(s)$ is concave, and $U(s)$ and all the constraints are convex. Thus, Problem 5 is a penalized DC problem. A stationary point of Problem 5 can be obtained by using the DC algorithm [5]. The DC algorithm can be conducted multiple times, each with a random initial feasible point of Problem 5, and the stationary point of Problem 5 that achieves the maximum total utility with zero penalty is chosen as a suboptimal solution of Problem 1.

6 Numerical Results
In this section, the proposed solutions are compared with three baseline schemes [3, 4, 1] numerically. Specifically, in Baseline 1, each subcarrier is assigned to the user with the largest channel power [3], and the optimization of the other variables are obtained using DC programming; in Baseline 2, a binary approximation of the optimal solution of the continuous relaxation of Problem 1 [4] is adopted; in Baseline 3, equal power allocation is adopted, and service level selection and subcarrier allocation are obtained in a greedy manner. We set $L_k = 6$, $T = 40 \times 10^{-3}$, $B = 4 \times 10^4$, $N = 64$, $P_k = 1$, $n_0 = 1.66 \times 10^{-16}$, $U_{k,l} = l$. For all $k \in K$, assume $h_k$ follows Rayleigh fading with mean $10^{-6}$, and that $N_{u,k,l}$, $N_{e,k,l}$ and $N_{d,k,l}$ follow the uniform distributions as in [5]. The average performance over 100 realizations of the random variables is evaluated. Fig. 1 shows that the average utility per user (i.e., total utility divided by $K$) of each scheme increases with $F$ and $P_0$, which reflects the tradeoff between user experience and consumptions of both computation resource and communications resource. The two proposed solutions have comparable average utilities and outperform all baseline schemes.

7 Conclusion
This letter established a communication model for offloading scalable tasks in TDD-OFDMA systems and optimized the system utility maximization with respect to the service level selection and resource allocation. The analytical and numerical results provide insights for designing efficient MEC systems.

© IEICE 2021