Generalized Nash Equilibrium Game for Radio and Computing Resource Allocation in Co-located MEC

Chit Wutyee Zaw*, Nguyen H. Tran†*, Walid Saad‡*, Zhu Han§*, Choong Seon Hong*
*Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, Republic of Korea
†School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
‡Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061
§Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204-4005

***Abstract*—The tower sharing approach has been widely used by Mobile Network Operators (MNOs) to save their Capital Expenditure (CAPEX) by sharing the physical infrastructure hosted by a third party tower provider. In addition, multiple Computing Resource Providers (CRP) are deploying their servers at towers by cooperating with tower providers to grant low latency, real time services to users. Thus, the resource allocation has become a challenging issue where users of different MNOs need to share the computing resources provided by CRPs. In this paper, the joint allocation of uplink, downlink and computing resources is considered to minimize the end to end latency of users where the offloading process is modeled as a network of queues. Since the resource allocation of MNOs and the CRP are coupled with each other, we formulate it as a Generalized Nash Equilibrium Problem (GNEP). We propose a penalty based algorithm to solve the formulated GNEP with an effective initialization approach to improve the performance of the algorithm. Then, we perform the simulation to analyze the performance of the algorithm.**

***Index Terms*—Colocated edge computing, Generalized nash equilibrium, Mobile edge computing, Resource allocation**

I. Introduction

Tower sharing among multiple Mobile Network Operators (MNOs) has been a popular approach since MNOs can reduce their Capital Expenditure by sharing the physical infrastructure deployed by a third party tower provider so as to increase their coverage to deliver the growing demand of users. Wireless Network Virtualization is also a driving force behind this where both radio resources (active infrastructure) and physical devices (passive infrastructure) are shared among multiple Mobile Virtual Network Operators. In this paper, we consider the passive infrastructure approach where MNOs control their own radio resources. Moreover, Computing Resource Providers (CRP) such as IBM and Vapor are cooperating with tower providers to deploy servers at towers so that they can provide real time, low latency applications to users at the edge.

Resource allocation has been a hot topic in Mobile Edge Computing (MEC) recently. An energy efficient radio resource allocation and offloading in multi-cell environment is proposed in [1] where the computing resource allocation at the edge server is not considered. The dynamic task offloading and scheduling IoT services in MEC is proposed in [2] where the radio resource allocation is not considered. Most of the previous works ignore the queueing model to calculate the latency. However, authors in [3] proposed the queue network for an energy efficient resource provision by scaling the CPU of a server while the radio resource allocation is not considered. Another queuing model in MEC is proposed by [4] where authors considered the latency and reliability aware task offloading by controlling the transit power of users and CPU cycles of the MEC server. The wireless transmission and cloud execution are modeled as Poisson processes in [5] where an upper bound on the delay is considered.

The Generalized Nash Equilibrium Problem (GNEP) has become a prominent approach in solving the resource allocation problem since it captures the coupling among players. The formulation of service provisioning in multi-cloud environment using GNEP can be found in [6] and [7]. Authors in [8] proposed a GNEP based algorithm to solve the offloading decision by scheduling users in MEC. GNEP and its solution approaches are discussed [9]. Penalty based algorithms are studied in [10] to solve GNEP efficiently.

The offloading of tasks to MEC servers requires the radio resources of MNOs for transmitting, receiving data, and the computing resources of the CRP for processing the tasks. This strong coupling among providers causes the resource allocation problem challenging. In this paper, we formulate this as GNEP and propose a penalty based algorithm to solve the formulated GNEP. Our contributions are as follows.

- The task offloading of users is modeled as a network of queues. The end to end latency is calculated based on its performance. Due to the queue network, the arrival at a queue is strongly dependent on the departure of tasks at its preceding queue. This makes the resource allocation of MNOs and CRP challenging.
- Since MNOs and CRP are two different entities, we formulate the resource allocation problem as GNEP where they have the conflicting interest in minimizing the latency. We also prove the existence of the Generalized Nash Equilibrium (GNE) of the formulated problem.
- To solve the GNEP, we propose a penalty based resource allocation algorithm by transforming it to Nash Equilibrium Problem (NEP). Since the penalty based algorithms relies on the parameters, we propose an initialization approach to improve the algorithm performance.
- We then perform the simulation to analyze the performance of our proposed algorithm with respect to user loads, initialization approaches and penalty parameters.

978-1-7281-5089-5/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Kyunghee Univ. Downloaded on August 03,2020 at 10:14:40 UTC from IEEE Xplore. Restrictions apply.

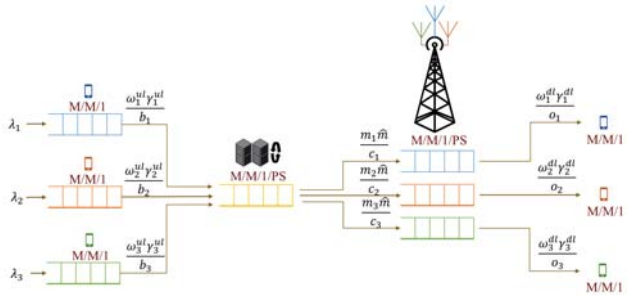

Fig. 1. The network of queues in Co-located Edge Computing

The rest of the paper is organized as follows. We present the system model in Section II. The resource allocation problem is formulated in Section III. This problem is transformed into GNEP in Section IV. A penalty based distributed algorithm is proposed in Section V. We present the simulation results in Section VI and conclude the paper in Section VII.

II. System Model

A single tower model is considered where a set of MNOs, $\mathcal{J} = \{1, \cdots, J\}$, and a CRP are co-located together. The tasks generating at user u follows a homogeneous Poisson process with rate λ_u where input file and required CPU cycles of the tasks follow the exponential distribution with means b_u and c_u respectively. The results of the tasks computed by the MEC servers are also assumed to follow the exponential distribution with mean o_u. The latency has three parts which are the time required to upload the input data to MEC servers, process the task on MEC servers and send the result back to the user. Each stage is modeled as a queue which is shown in Fig. 1 in which the service rates are defined as the number of tasks departing the serving stations per second. These are often defined as the departure rates as well. Thus, the arrival rate of a queue is same as the departure rate of the preceding queue.

A. Communication Model

We assume that MNOs own different frequency bands. For each MNO, FDD LTE frequency band allocation is considered for uplink and downlink transmissions where these transmissions operate on two different frequencies. The bandwidth resources of these uplink and downlink frequency bands are then orthogonally allocated to users.

1) Uplink Transmission: The average rate of uplink transmission of user $u \in \mathcal{U}_j$ is

$$R_u^{\text{ul}} = \omega_u^{\text{ul}} \hat{\omega}_j^{\text{ul}} \log_2 \left(1 + \frac{p_u g_u}{n_0}\right) \coloneqq \omega_u^{\text{ul}} \gamma_u^{\text{ul}}, \tag{1}$$

where ω_u^{ul} is the fraction of the uplink bandwidth allocated to user u, $\hat{\omega}_j^{\text{ul}}$ is the total uplink bandwidth owned by MNO j, p_u is the constant transmit power of user u, g_u is the uplink channel gain, and n_0 is the additive white Gaussian noise. The utilization of user u is $\upsilon_u = \frac{\lambda_u b_u}{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}$. The time required to upload the input file with the average file size b_u is

$$t_u^{\text{ul}} = \frac{b_u}{\omega_u^{\text{ul}} \gamma_u^{\text{ul}} [1 - \upsilon_u]}. \tag{2}$$

Each user u is considered as a M/M/1 queue with task arrival rate, λ_u, and service rate, $\frac{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}{b_u}$, which is the number of tasks departing from the uplink transmission queue per second.

2) Downlink Transmission: The expected workload which is the result of user u from MEC servers follows an exponential distribution with the mean, o_u. The result arrives at the BS j with the rate $\sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u}$ which is the service rate of MEC servers for the task of user u where $\hat{m}$ is the total CPU capacity of the MEC servers. The average rate of downlink transmission of user $u \in \mathcal{U}_j$ is

$$R_u^{\text{dl}} = \omega_u^{\text{dl}} \hat{\omega}_j^{\text{dl}} \log_2 \left(1 + \frac{p_j g_u}{n_0}\right) \coloneqq \omega_u^{\text{dl}} \gamma_u^{\text{dl}}, \tag{3}$$

where p_j is the constant transmit power of the BS of MNO j for $u \in \mathcal{U}_j$ and ω_u^{dl} is the fraction of the downlink bandwidth allocated to user u. The utilization of MNO j is $\rho_j = \frac{\sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u}}{\sum_{u \in \mathcal{U}_j} \frac{\omega_u^{\text{dl}} \gamma_u^{\text{dl}}}{o_u}}$. The time required to send the result back to user u is

$$t_u^{\text{dl}} = \frac{o_u}{\omega_u^{\text{dl}} \gamma_u^{\text{dl}} [1 - \rho_j]}, \tag{4}$$

where each MNO j which can be considered as BS j is modeled as a multiclass M/M/1 processor sharing queue with service rate $\sum_{u \in \mathcal{U}_j} \frac{\omega_u^{\text{dl}} \gamma_u^{\text{dl}}}{o_u}$ as in [11].

B. Computing Model

The tasks of users which need to be computed at the MEC servers arrive with the rate $\sum_{j=1}^{J} \sum_{u \in \mathcal{U}_j} \frac{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}{b_u}$. The utilization of the MEC server is $\psi = \frac{\sum_{j=1}^{J} \sum_{u \in \mathcal{U}_j} \frac{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}{b_u}}{\sum_{j=1}^{J} \sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u}}$ where m_u is the fraction of CPU cycles allocated to user u. The time required to compute the task of user u is

$$t_u^{\text{p}} = \frac{c_u}{m_u \hat{m} [1 - \psi]}, \tag{5}$$

where the MEC server is modeled as multiclass M/M/1 processor sharing queue with service rate $\sum_{j=1}^{J} \sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u}$ as in [11].

Notations: The vector representations of the decision variables which are used throughout the paper are defined as $\mathbf{W}_j^{ul} \coloneqq [\omega_u^{ul}]_{u \in \mathcal{U}_j}^T \in \mathbb{R}_+^{|\mathcal{U}_j|}$, $\mathbf{W}^{ul} \coloneqq [\mathbf{W}_1^{ul}, ..., \mathbf{W}_N^{ul}]^T \in \mathbb{R}_+^{|\mathcal{U}|}$, $\mathbf{W}_j^{dl} \coloneqq [\omega_u^{dl}]_{u \in \mathcal{U}_j}^T \in \mathbb{R}_+^{|\mathcal{U}_j|}$, $\mathbf{W}^{dl} \coloneqq [\mathbf{W}_1^{dl}, ..., \mathbf{W}_N^{dl}]^T \in \mathbb{R}_+^{|\mathcal{U}|}$, $\mathbf{m}_j \coloneqq [m_u]_{u \in \mathcal{U}_j}^T \in \mathbb{R}_+^{|\mathcal{U}_j|}$, $\mathbf{m} \coloneqq [m_u]_{u \in \mathcal{U}}^T \in \mathbb{R}_+^{|\mathcal{U}|}$.

III. Problem Formulation

A. Objective Function

The objectives of both MNOs and CRP are to provide services with the minimum latency. Since the latency directly influences the energy consumption of the providers which means that if the time required for the transmissions and computation is long, the providers need to run their devices for the longer period. Thus, providing the services with the low latency can also benefit the providers in terms of the energy consumption. The objective function is defined as follows:

$$\Theta = \sum_{j=1}^{J} \sum_{u \in \mathcal{U}_j} t_u^{\text{ul}} + t_u^{p} + t_u^{\text{dl}}. \tag{6}$$

Authorized licensed use limited to: Kyunghee Univ. Downloaded on August 03,2020 at 10:14:40 UTC from IEEE Xplore. Restrictions apply.

B. Constraints

1) Bandwidth Resource Constraints: For MNO j, the bandwidth resources allocation must be less than or equal to 1.

$$\sum_{u\in\mathcal{U}_j} \omega_u^{\text{ul}} \leq 1, \tag{7}$$

$$\sum_{u\in\mathcal{U}_j} \omega_u^{\text{dl}} \leq 1. \tag{8}$$

2) Computing Resource Constraint: The fraction of CPU resources allocated to users must be less than or equal to 1.

$$\sum_{j=1}^{J}\sum_{u\in\mathcal{U}_j} m_u \leq 1. \tag{9}$$

3) Queue Utilization Constraints of Users, BSs and MEC Servers: For queues to be stable, their utilization must be less than 1. In this paper, we consider that the serving stations are allowed to be busy at most $(1-\epsilon)\times 100\%$ of the time.

$$\upsilon_u \leq 1-\epsilon, \qquad \forall u\in\mathcal{U}_j, \forall j\in\mathcal{J}, \tag{10}$$

$$\rho_j \leq 1-\epsilon, \qquad \forall j\in\mathcal{J}, \tag{11}$$

$$\psi \leq 1-\epsilon. \tag{12}$$

C. Optimization Problem of CRP

The problem of CRP can be formulated as

$$\begin{aligned} \underset{\mathbf{m}}{\text{minimize}} \quad & \Theta_{\text{CRP}}(\mathbf{m},\mathbf{W}^{\text{ul}},\mathbf{W}^{\text{dl}}) \\ \text{subject to} \quad & (9),(11) \text{ and } (12), \end{aligned} \tag{13}$$

where $\Theta_{\text{CRP}}(\mathbf{m},\mathbf{W}^{\text{ul}},\mathbf{W}^{\text{dl}}) = \Theta$.

D. Optimization Problem of MNOs

The objective of MNO j is to minimize the latency according to the resource allocation of other MNOs and CRP. The constraint in (12) is rewritten as follows:

$$\frac{\sum_{u\in\mathcal{U}_j}\frac{\omega_u^{\text{ul}}\gamma_u^{\text{ul}}}{b_u}+\alpha_{-j}^{\text{ul}}}{\sum_{u\in\mathcal{U}_j}\frac{m_u\hat{m}}{c_u}+\alpha_{-j}^{\text{m}}} \leq 1-\epsilon, \tag{14}$$

where $\alpha_{-j}^{\text{ul}} = \sum_{\substack{i=1\\ i\neq j}}^{J}\sum_{u\in\mathcal{U}_i}\frac{\omega_u^{\text{ul}}\gamma_u^{\text{ul}}}{b_u}$, $\alpha_{-j}^{\text{m}} = \sum_{\substack{i=1\\ i\neq j}}^{J}\sum_{u\in\mathcal{U}_i}\frac{m_u\hat{m}}{c_u}$. The problem of MNO j can be formulated as

$$\begin{aligned} \underset{\mathbf{W}_j^{\text{ul}},\mathbf{W}_j^{\text{dl}}}{\text{minimize}} \quad & \Theta_{\text{MNO}_j}(\mathbf{m},\mathbf{W}^{\text{ul}},\mathbf{W}^{\text{dl}}) \\ \text{subject to} \quad & (7),(8),(10),(11) \text{ and } (14), \end{aligned} \tag{15}$$

where $\Theta_{\text{MNO}_j}(\mathbf{m},\mathbf{W}^{\text{ul}},\mathbf{W}^{\text{dl}}) = \sum_{u\in\mathcal{U}_j} t_u^{\text{ul}} + t_u^{\text{p}} + t_u^{\text{dl}}$.

IV. Formulation as GNEP

Since the latency is decreasing with the resource allocation but increasing with the queue utilization, the higher resource allocation at a queue would increase the latency at its following station with the resource constraints. Thus, MNOs and CRP compete each other to minimize the total latency of users while guaranteeing the queue utilization constraints. Thus, we formulate the resource allocation problem as GNEP where the strategy set of a player depends on other players' strategies.

Let $\mathcal{P}=\{0,1,\cdots,J\}$ be the set of players where CRP is indexed as 0. Let $\mathcal{U}$ be the set of all users where $\mathcal{U}=\cup_{j=1}^{J}\mathcal{U}_j$. Let $\mathbf{x}^0 \coloneqq \mathbf{m}$ and $\mathbf{x}^p \coloneqq [\mathbf{W}_j^{\text{ul}},\mathbf{W}_j^{\text{dl}}]^T$ for $j=1,\cdots,J$ and $p=0,\cdots,J$ where $\mathbf{x}_p$ is a n_p vector which is the strategy of player p.

A. Coupling Constraints

In order to formate as a GNEP, the coupling constraints are rewritten as follows.

1) Queue Utilization of BSs: The constraint in (11) for each MNO j can be written as

$$f_j(\mathbf{W}_j^{\text{dl}},\mathbf{m}_j) \leq 0, \tag{16}$$

where $f_j(\mathbf{W}_j^{\text{dl}},\mathbf{m}_j) = \frac{\sum_{u\in\mathcal{U}_j}\frac{m_u\hat{m}}{c_u}}{\sum_{u\in\mathcal{U}_j}\frac{\omega_u^{\text{dl}}\gamma_u^{\text{dl}}}{o_u}} - 1 + \epsilon$.

This can be written as follows for all MNOs, $j=1,\cdots,J$.

$$f(\mathbf{W}^{\text{dl}},\mathbf{m}) \leq \mathbf{0} \tag{17}$$

where $f(\mathbf{W}^{\text{dl}},\mathbf{m}) = [f_j(\mathbf{W}_j^{\text{dl}},\mathbf{q}_j)]_{j=1,\cdots,J}^T$.

2) Queue Utilization of MEC servers pool: The constraint in (12) can be written as

$$g(\mathbf{W}^{\text{ul}},\mathbf{m}) \leq 0, \tag{18}$$

where $g(\mathbf{W}^{\text{ul}},\mathbf{m}) = \frac{\sum_{j=1}^{J}\sum_{u\in\mathcal{U}_j}\frac{\omega_u^{\text{ul}}\gamma_u^{\text{ul}}}{b_u}}{\sum_{j=1}^{J}\sum_{u\in\mathcal{U}_j}\frac{m_u\hat{m}}{c_u}} - 1 + \epsilon$.

B. Strategy Sets of Players

The strategy set of CRP satisfying its computing resources requirement can be defined as follows:

$$\mathcal{S}_0 = \{\mathbf{m} : \sum_{j=1}^{J}\sum_{u\in\mathcal{U}_j} m_u \leq 1\}.$$

The strategy set of MNO j satisfying bandwidth requirements and the queue stability at users can be defined as

$$\begin{aligned} \mathcal{S}_j = \{(\mathbf{W}_j^{ul},\mathbf{W}_j^{dl}) : & \sum_{u\in\mathcal{U}_j}\omega_u^{ul} \leq 1, \sum_{u\in\mathcal{U}_j}\omega_u^{dl} \leq 1, \\ & \mathbf{W}_j^{\text{ul}} \geq \tilde{\mathbf{W}}_j^{\text{ul}}\}, \end{aligned}$$

where $\tilde{\mathbf{W}}_j^{\text{ul}} = [\frac{\lambda_u b_u}{\gamma_u^{\text{ul}}(1-\epsilon)}]_{u\in\mathcal{U}_j}^T \in \mathbb{R}_{++}^{|\mathcal{U}_j|}$.

C. GNEP Formulation for CRP

The GNEP of CRP is formulated as follows which is the reformation of (13).

$$\begin{aligned} G_0(\mathbf{x}^{-0}): \quad \underset{\mathbf{m}\in\mathcal{S}_0}{\text{minimize}} \quad & \Theta_{\text{CRP}}(\mathbf{m},\mathbf{W}^{\text{ul}},\mathbf{W}^{\text{dl}}) \\ \text{subject to} \quad & (17) \text{ and } (18). \end{aligned} \tag{19}$$

The following lemma specifies the convexity of $G_0(\mathbf{x}^{-0})$.

Lemma 1. *$\mathcal{S}_0$ is a compact, convex set. $\Theta_{CRP}(\mathbf{m},\mathbf{W}^{ul},\mathbf{W}^{dl})$, $f(\mathbf{W}^{dl},\mathbf{m})$ and $g(\mathbf{W}^{ul},\mathbf{m})$ are convex in $\mathbf{m}$.*

Authorized licensed use limited to: Kyunghee Univ. Downloaded on August 03,2020 at 10:14:40 UTC from IEEE Xplore. Restrictions apply.

D. GNEP Formulation for MNOs

The MEC server queue utilization constraint in (14) for MNO j can be rewritten as

$$g_j(\mathbf{W}^{\text{ul}}, \mathbf{m}) \leq 0, \tag{20}$$

where $g_j(\mathbf{W}^{\text{ul}}, \mathbf{m}) = \frac{\sum_{u \in \mathcal{U}_j} \frac{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}{b_u} + \alpha_{-j}^{\text{ul}}}{\sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u} + \alpha_{-j}^{\text{m}}} - 1 + \epsilon$.

The GNEP of MNO j is defined as follows which is also the reformulation of (15).

$$\begin{aligned} G_j(\mathbf{x}^{-j}): \quad & \underset{\mathbf{W}_j^{ul}, \mathbf{W}_j^{dl} \in \mathcal{S}_j}{\text{minimize}} \quad \Theta_{\text{MNO}_j}(\mathbf{m}, \mathbf{W}^{\text{ul}}, \mathbf{W}_j^{\text{dl}}) \\ & \text{subject to} \quad (16) \text{ and } (20) \end{aligned} \tag{21}$$

The convexity of $G_j(\mathbf{x}^{-j}), \forall j$, is described in the following lemma.

Lemma 2. *$\mathcal{S}_j$ is a compact, convex set. $\Theta_{MNO_j}(\mathbf{m}, \mathbf{W}^{ul}, \mathbf{W}_j^{dl}), f_j(\mathbf{W}_j^{dl}, \mathbf{m}_j)$ and $g_j(\mathbf{W}^{ul}, \mathbf{m})$ are convex in $\mathbf{W}_j^{ul}$ and $\mathbf{W}_j^{dl}$.*

Because of the strong dependency among decision variables in the objective functions and coupling constraints, the proof of the convexity is not straightforward. Due to the page limitation, the proof is ignored in this paper.

E. Existence of GNE

The existence of the GNE of the formulated problem is presented in this section.

Theorem 1. *$\mathbf{x}^*$ which is the solution of $G_p(\mathbf{x}^{*,-p}), \forall p$, is a GNE.*

Proof: $G_p(\mathbf{x}^{-p})$ for $p = 0, \cdots, J$ is a convex programming problem with fixed $\mathbf{x}^{-p}$, there is $\mathbf{x}^* = [\mathbf{x}^{*,0}, \cdots, \mathbf{x}^{*,J}]$ where $\mathbf{x}^{*,p}$ is the optimal solution of $G_p(\mathbf{x}^{*,-p})$. ■

V. Penalty based Distributed Algorithm

The formulated GNEP is difficult to solve due to the couplings in the objective function and strategy sets of the players. In this section, the GNEP is reformulated as a NEP where the coupling constraints are penalized in the objective function.

A. Reformation as the Nash Equilibrium Game

The NEP with penalized objective function of CRP is defined as follows:

$$\begin{aligned} \underset{\mathbf{m} \in \mathcal{S}_0}{\text{minimize}} \quad & \Theta_{\text{CRP}}(\mathbf{m}, \mathbf{W}^{\text{ul}}, \mathbf{W}^{\text{dl}}) + \sum_{j=1}^{J} \kappa_0^{\text{BS}_j} f_j(\mathbf{W}_j^{\text{dl}}, \mathbf{m}_j) \\ & + \kappa_0^{\text{mec}} g(\mathbf{W}^{\text{ul}}, \mathbf{m}), \end{aligned} \tag{22}$$

where $\kappa_0^{\text{BS}_j}$ and κ_0^{mec} act as penalty parameters of CRP for the coupling constraints.

The NEP for MNO j with penalized objective function is also formulated as follows:

$$\begin{aligned} \underset{\mathbf{W}_j^{ul}, \mathbf{W}_j^{dl} \in \mathcal{S}_j}{\text{minimize}} \quad & \Theta_{\text{MNO}_j}(\mathbf{m}, \mathbf{W}^{\text{ul}}, \mathbf{W}_j^{\text{dl}}) \\ & + \kappa_j^{\text{BS}_j} f_j(\mathbf{W}_j^{\text{dl}}, \mathbf{m}_j) + \kappa_j^{\text{mec}} g(\mathbf{W}^{\text{ul}}, \mathbf{m}), \end{aligned} \tag{23}$$

Algorithm 1 Penalty based Distributed Algorithm

1: Choose the penalty parameters $\kappa_p^{\text{BS}_j,0}, j = 1, \cdots, N$ and $\kappa_p^{\text{mec},0}$, $p = 0, \cdots, N$
2: $k \leftarrow 0$
3: Choose an initial point for $\mathbf{m}^k, \mathbf{W}^{\text{ul},k}, \mathbf{W}^{\text{dl},k}$ as in (24)
4: CRP solves the problem in (22)
5: each MNO j solves the problem in (23)
6: **if** $f_j(\mathbf{W}_j^{*,\text{dl}}, \mathbf{m}_j^*) \leq 0, \forall j$ and $g(\mathbf{W}^{*,\text{ul}}, \mathbf{m}^*) \leq 0$ **then**
7: $[\mathbf{W}^{*,\text{ul}}, \mathbf{m}^*, \mathbf{W}^{*,\text{dl}}]$ is a GNE.
8: **else**
9: Each player p update their penalty parameters, $\kappa_p^{\text{BS}_j,k+1}, \forall j$ and $\kappa_p^{\text{mec},k+1}$, as follows.
10: $\kappa_p^{\text{BS}_j,k+1} = \begin{cases} \kappa_p^{\text{BS}_j,k} + \Delta_p^{\text{BS}_j,k} & \text{if } f_j(\mathbf{W}_j^{*,\text{dl}}, \mathbf{m}_j^*) > 0 \\ \kappa_p^{\text{BS}_j,k} & \text{if } f_j(\mathbf{W}_j^{*,\text{dl}}, \mathbf{m}_j^*) \leq 0 \end{cases}$
11: $\kappa_p^{\text{mec},k+1} = \begin{cases} \kappa_p^{\text{mec},k} + \Delta_p^{\text{mec},k} & \text{if } g(\mathbf{W}^{*,\text{ul}}, \mathbf{m}^*) > 0 \\ \kappa_p^{\text{mec},k} & \text{if } g(\mathbf{W}^{*,\text{ul}}, \mathbf{m}^*) \leq 0 \end{cases}$
12: $k \leftarrow k + 1$
13: **go to** line number 4
14: **end if**

where κ_j^{BS} and κ_j^{mec} are the penalty parameters of MNO j for the coupling constraints. Note that the penalty parameters of players are different which allows players to have a full control over the game.

The proposed distributed algorithm is implemented at the tower station where each player solves their optimization problem and exchanges the current resource allocation among each other. The algorithm works as follows. First, the initial point is chosen for the resource allocation and penalty parameters. Each player solves their optimization problems at lines 4 and 5. Then, they update their penalty parameters, lines 10 and 11, until the penalized constraints are feasible as stated at line 6.

$$\begin{aligned} \omega_u^{\text{ul},0} &= \frac{\lambda_u b_u}{\gamma_u^{\text{ul}}(1-\epsilon)}, \omega_u^{\text{dl},0} = \frac{o_u \sum_{u \in \mathcal{U}_j} \frac{m_u^0 \hat{m}}{c_u}}{\gamma_u^{\text{dl}} |\mathcal{U}_j| (1-\epsilon)}, \\ m_u^0 &= \frac{c_u \sum_{j=1}^{J} \sum_{u \in \mathcal{U}_j} \frac{\omega_u^{\text{ul},0} \gamma_u^{\text{ul}}}{b_u}}{\hat{m}(1-\epsilon) \sum_{j=1}^{J} |\mathcal{U}_j|}. \end{aligned} \tag{24}$$

Since penalty based algorithms rely on the given parameters, we propose an initialization approach to improve the algorithm performance by considering the stability of queues at the serving stations. First, we consider the worst case scenario for the queues where the serving stations are busy $(1-\epsilon) \times 100\%$ of the time. From this, we assign the lower bound for the resource allocation for each user u as in (24). At each iteration k, the CRP and MNOs solve the convex optimization problems in (22) and (23) respectively where we can derive the optimal solution with KKT conditions. Thus, let its complexity be $\mathcal{O}(\Sigma)$. Each player then needs to perform the penalty updates for the coupling constraints in which the complexity is $\mathcal{O}(J+1)$. Since Σ is much larger than $J+1$, the complexity of the proposed algorithm is $\mathcal{O}(K.|\mathcal{P}|.\Sigma)$ where K is the total number of iterations until convergence.

Authorized licensed use limited to: Kyunghee Univ. Downloaded on August 03,2020 at 10:14:40 UTC from IEEE Xplore. Restrictions apply.

B. Convergence of The Algorithm

The following theorem states the convergence of the proposed algorithm.

Theorem 2. *The penalty based distributed algorithm converges to a GNE if the following two assumptions are satisfied.*

(i) For each $p = 0, \cdots, J$, $\mathcal{S}_p$ is nonempty and compact.
(ii) The e-MFCQ holds at any $\mathbf{x} \in \mathcal{S}$.

The set $\mathcal{S}$ is defined as $\mathcal{S} := \prod_{p=0}^{J} \mathcal{S}_p$. The first assumption is defined in Sections IV-B and IV-C. The extended Mangasarian-Fromovitz constraint qualification (e-MFCQ) holds at any $\mathbf{x} \in \mathcal{S}$ if there exists $\mathbf{v}^{\text{mec}} \in \mathbb{R}^{n_p}$ and $\mathbf{v}^{\text{BS}} = [\mathbf{v}^{\text{BS}_1}, \cdots, \mathbf{v}^{\text{BS}_N}] \in \mathbb{R}^{n_p N}$ such that

$$\nabla_{\mathbf{x}_p} g(\mathbf{x})^T \mathbf{v}^{\text{mec}} < 0, p = 0, \cdots, N \text{ if } g(\mathbf{x}) \geq 0.$$

$$\nabla_{\mathbf{x}_p} f_j(\mathbf{x})^T \mathbf{v}^{\text{BS}_j} < 0, \forall j \in \mathcal{I}_p(\mathbf{x}_p), p = 0, \cdots, N.$$

where $\mathcal{I}_p(\mathbf{x}_p) = \{j \mid f_j(\mathbf{x}_p) \geq 0\}$. This means that the sum of first derivatives with respect to the player's strategy, $\mathbf{x}_p$, is not equal to zero if the coupling constraints are inactive.

Proof: The proof is as follows for player CRP. It is same for other players, MNOs. The first derivative of $f_j(\mathbf{W}_j^{\text{dl}}, \mathbf{m}_j)$ with respect to $m_u, \forall u \in \mathcal{U}_j$ is

$$\frac{\partial f_j}{\partial m_u} = \frac{\hat{m}}{c_u} \frac{1}{\sum_{u \in \mathcal{U}_j} \frac{\omega_u^{\text{dl}} \gamma_u^{\text{dl}}}{o_u}}.$$

If $\frac{\sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u}}{\sum_{u \in \mathcal{U}_i} \frac{\omega_u^{\text{dl}} \gamma_u^{\text{dl}}}{o_u}} - 1 + \epsilon \geq 0$, $\sum_{u \in \mathcal{U}_i} \frac{\omega_u^{\text{dl}} \gamma_u^{\text{dl}}}{o_u} > 0$. Thus, $\frac{\partial f_j}{\partial m_u} \neq 0$. The first derivative of $g(\mathbf{W}^{\text{ul}}, \mathbf{m})$ with respect to m_u is

$$\frac{\partial g}{\partial m_u} = -\frac{\hat{m}}{c_u} \frac{\sum_{j=1}^{N} \sum_{u \in \mathcal{U}_i} \frac{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}{b_u}}{[\sum_{j=1}^{N} \sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u}]^2}.$$

If $\frac{\sum_{j=1}^{N} \sum_{u \in \mathcal{U}_i} \frac{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}{b_u}}{\sum_{j=1}^{N} \sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u}} - 1 + \epsilon \geq 0$, $\sum_{j=1}^{N} \sum_{u \in \mathcal{U}_i} \frac{\omega_u^{\text{ul}} \gamma_u^{\text{ul}}}{b_u} > 0$ and $\sum_{j=1}^{N} \sum_{u \in \mathcal{U}_j} \frac{m_u \hat{m}}{c_u} > 0$. Thus, $\frac{\partial g}{\partial m_u} \neq 0$. ■

Although the algorithm solves the penalized NEP which is the transformation of the formulated GNEP, it converges to a GNE as mentioned above. Thus, there is no gap between the GNE and the equilibrium obtained by the algorithm.

VI. Simulation Results

Three MBSs which are owned by different opeartors are co-located together with a MEC server deployed by a CRP. The simulation is performed using Python. The tasks arrival at users follows a Poisson distribution. The size of input files, required CPU cycles, and output files follows an exponential distribution with the means chosen uniformly from the range given in Table I. The optimization problems are solved using [12]. Since this paper is the first work for the resource allocation in co-located MEC system, there is no previous work to compare with our proposal. However, Fig. 3 shows the comparison of our work with the baseline resource allocation approaches, proportional and uniform allocation.

TABLE I
Simulation Parameters

Parameters	Values
Transmit power of the MBS	43 dbm
Transmit power of users	20 dbm
The uplink bandwidth of MNOs	20 MHz
The downlink bandwidth of MNOs	20 MHz
Power density thermal noise	-174 dBm/Hz
Task Arrival Rate	[0.5, 0.7] tasks/s
CPU speed at MEC server	16 GHz
Input file size	[20, 40] KB
Task size	[0.2, 0.4] GHz
Output file size	[300, 800] KB

Fig. 2 shows the resource allocation to the users of different MNOs with respect to user loads. First, the resource allocation on the balanced and unbalanced user loads is analyzed. The number of MNO3's users is much higher than MNOs 1 and 2 for the latter case. This leads to different MEC resource allocations among MNOs which causes the tighter uplink bandwidth allocation. Although the uplink allocation of MNO 2 is higher in the unbalanced case, it cannot receive the MEC server resources more due to the higher allocation of MNOs 1 and 3. Moreover, the MEC server resources need to be allocated according to the uplink allocation to maintain the queue stability. Since the downlink resources do not have the impact on the uplink and MEC server allocation, they are fully allocated to the users. The impact of under-loaded and overloaded scenarios on the algorithm is also analyzed. The resource allocation of the MEC server is restricted by the uplink allocation in overloaded case. In under-loaded scenario, CRP has the flexible control over the resources since the MEC server's capacity is higher than the requests of users. This causes the game less competitive among MNOs and CRP.

As in (16) and (18), the resource allocation of the MEC server has the huge impact on the penalized constraints. Thus, we analyze the convergence of the algorithm on the MEC server allocation in Fig. 3 and Fig. 4. In Fig. 3, the proposed initialization is compared with the uniform approach, in which all users receive the same amount of resources, and the proportional approach, in which the resource allocation is proportional to the task requirements of users. The convergence of the algorithm with our proposal is much faster than other two approaches because our proposal begins with a feasible point where the penalized constraints are satisfied. The GNEs found by the initialization approaches are different where the MEC server allocation achieved by our proposal is lower than the others, because the higher resource allocation at MEC server results in higher latency due to the limitation of the downlink resources. Thus, the latency obtained by other approaches is 3 times higher than our proposal.

The convergence of the algorithm on different parameters is shown in Fig. 4. The two parameters, penalty parameters, $\kappa_p^{\text{mec,BS}_j}$, and step sizes, $\Delta_p^{\text{mec,BS}_j}$, are fixed alternatively to analyze the algorithm's performance. The results achieved by the algorithm with various parameters are not much different

Authorized licensed use limited to: Kyunghee Univ. Downloaded on August 03,2020 at 10:14:40 UTC from IEEE Xplore. Restrictions apply.

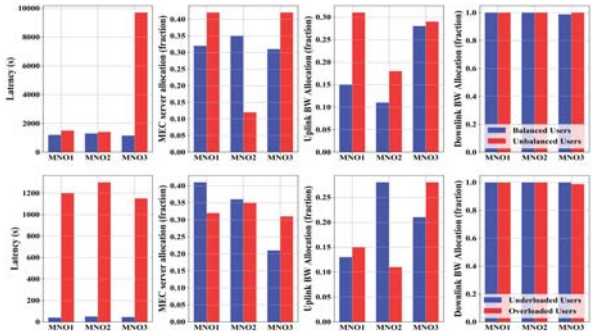

Fig. 2. Resource allocation on user load.

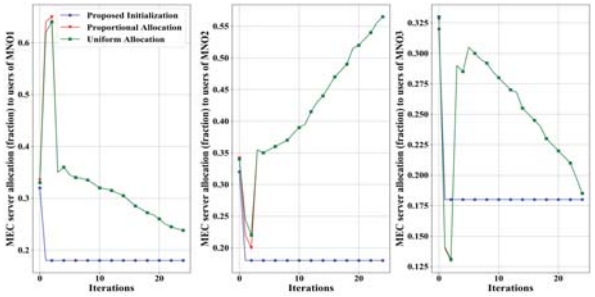

Fig. 3. Algorithm Performance on initialization approach.

due to the proposed initialization. Since the latency is decreasing with the downlink allocation, the MEC server allocation needs to be reduced to achieve the minimum latency. This is also influenced by the uplink allocation when the users are overloaded in the system as we discussed in Fig. 2.

VII. Conclusion

In this paper, the joint radio and computing resources allocation problem is formulated in the co-located edge computing system. The end to end latency is characterized by a network of queues. Due to the strong coupling in the queue network, the computing resource allocation of the CRP and the uplink radio resource allocation of MNOs are strongly dependent on one another. The coupling on the resource allocation decisions of MNOs and CRP is formulated as GNEP. A penalty based distributed algorithm with an effective initialization technique, which converges to an equilibrium faster than the baseline resource allocation approaches, is proposed to solve the GNEP efficiently. However, the penalty parameters have no impact on the performance of the algorithm in which the equilibria achieved with different parameters are similar to each other.

Acknowledgment

This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01287, Evolvable Deep Learning Model Generation Platform for Edge Computing) and Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2015-0-00567, Development of Access Technology Agnostic Next-Generation Networking Technology for Wired-Wireless Converged Networks) and partially supported by US MURI AFOSR MURI 18RT0073, NSF EARS-1839818, CNS1717454, CNS-1731424, and CNS-1702850. Dr. CS Hong is the corresponding author.

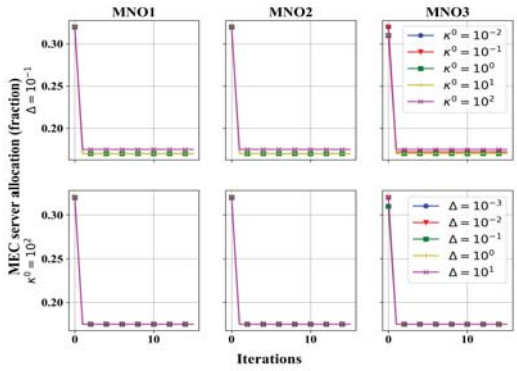

Fig. 4. Algorithm performance on penalty parameters.

References

[1] A. Khalili, S. Zarandi, and M. Rasti, "Joint resource allocation and offloading decision in mobile edge computing," *IEEE Communications Letters*, vol. 23, no. 4, pp. 684–687, April 2019.

[2] H. A. Alameddine, S. Sharafeddine, S. Sebbah, S. Ayoubi, and C. Assi, "Dynamic task offloading and scheduling for low-latency iot services in multi-access edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 3, pp. 668–682, March 2019.

[3] P. Chang and G. Miao, "Resource provision for energy-efficient mobile edge computing systems," in *IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, Dec 2018.

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

[5] T. Zhao, S. Zhou, X. Guo, and Z. Niu, "Tasks scheduling and resource allocation in heterogeneous cloud for delay-bounded mobile edge computing," in *IEEE International Conference on Communications (ICC)*, Paris, May 2017.

[6] P. Liu, X. Mao, F. Hou, and S. Zhang, "Generalized nash equilibrium model of the service provisioning problem in multi-cloud competitions," in *IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Guangzhou, Oct 2018, pp. 1485–1490.

[7] D. Ardagna, M. Ciavotta, and M. Passacantando, "Generalized nash equilibria for the service provisioning problem in multi-cloud systems," *IEEE Transactions on Services Computing*, vol. 10, no. 3, pp. 381–395, May 2017.

[8] D. Nowak, T. Mahn, H. Al-Shatri, A. Schwartz, and A. Klein, "A generalized nash game for mobile edge computation offloading," in *2018 6th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud)*, Bamberg, March 2018.

[9] F. Facchinei and C. Kanzow, "Generalized nash equilibrium problems," *4OR*, vol. 5, no. 3, pp. 173–210, Sep 2007.

[10] M. Fukushima, "Restricted generalized nash equilibria and controlled penalty algorithm," *Computational Management Science*, vol. 8, no. 3, pp. 201–218, Aug 2011.

[11] A. P. Zwart, "Sojourn times in a multiclass processor sharing queue," *Memorandum COSOR*, vol. 9822, 1998.

[12] S. Diamond and S. Boyd, "Cvxpy: A python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, Jan 2016.

Authorized licensed use limited to: Kyunghee Univ. Downloaded on August 03,2020 at 10:14:40 UTC from IEEE Xplore. Restrictions apply.