Joint Bandwidth Allocation and Server Scaling for Colocation Edge Computing

Minh N. H. Nguyen and Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University
Email: {minhnhn, cshong}@khu.ac.kr

Abstract: Multi-access Edge Computing (MEC) is recently acknowledged as one of the key pillars for the next revolution of mobile communications area, where the convergence of IT and telecommunications network provides lower latency and more computation capability for cellular base stations (BSs). Nevertheless, a huge capital and operational cost can challenge mobile operators for new BSs and MEC micro-datacenters deployment. Thus, sharing the BS infrastructure, bandwidth, and MEC micro-datacenters for the co-located mobile operators can be an economical solution to provide high performance services with lower expense by exploiting the temporal and spatial difference in traffic loads. Turning this vision to reality, we study a joint bandwidth allocation sharing and shared MEC micro-datacenter scaling in ColoMEC management problem (CMMP) and propose an algorithm based on proximal block coordinate descent technique.

1. Introduction

Nowadays, the emerge of the upcoming mobile communications system, i.e., integrated Multi-access Edge Computing (MEC) deployment in 5G network, which is identified as one of the essential technologies which can support the low latency requirements for future mobile applications and IoT services. The foreseen mobile communications systems are designed to provide efficient and flexible support for edge computing to enable superior performance and quality of experience [1]. Inspired from the MEC sharing, we come up with a new concept of the Colocation Edge Computing (ColoMEC) that allows multiple mobile operators to share not only the same BS tower but also their radio and computation resources colocated at the same edge site.

Based on previous works [2] of sharing backup power for colocation BSs, in this work we develop a new model of sharing bandwidth and MEC micro-datacenter among co-located mobile network operators at ColoEdge sites. Different from recent works, we consider a joint optimal decision of user association, bandwidth sharing, and MEC micro-datacenter scaling for co-located mobile operators at multi-ColoEdge sites. Accordingly, a Central ColoMEC Orchestrator will decide which BSs mobile users can be associated with, the amount of bandwidth allocation for each colocation BSs and the service rate of the MEC micro-datacenter based on the traffic arrival at each ColoEdge site.

2. System Model

In this paper, we first present a Central ColoMEC Orchestrator operating the system of colocation BSs and shared MEC micro-datacenters that provide the communication and computation services for co-located mobile network operators as in Fig. 1. The main function of the Central ColoMEC Orchestrator is jointly controlling the user association, bandwidth allocation and MEC micro-datacenter scaling for the Colocation Edge Computing system. At each ColoEdge site $j$, which belongs to the ColoEdge sites set $\mathcal{S}$, each operator in the set of operators $\mathcal{O}$ provides their services at a colocation BS [2] and a shared MEC micro-datacenter.

In the service coverage, we consider a virtual user $u$ in the users set $\mathcal{U}_i$ of the operator $i$, of which the traffic follows an inhomogeneous Poisson point process with arrival rate. We extend the queueing based analysis for the communication flow-level cost from our previous work [2]. The traffic load of the operator $i$ needs to be processed at the MEC micro-datacenter at the ColoEdge site $j$ is defined as

$$\gamma_j(p_u) := \sum_{u \in \mathcal{U}_i} \gamma(u)p_j(u),$$

where $p_j(u)$ is the associated routing probability of user $u \in \mathcal{U}_i$ to the ColoEdge site $j$. Then for each user $u$, the downlink transmission rate can be served by BS $j$ is denoted by $c_j(u)$ which follows Shannon capacity as in [2].
Then BS load density is defined as \( \beta_g(u, w_i) := \frac{\gamma(u)}{c_g(u, w_i)} \), which defines the fraction of active transmission time required to deliver the traffic load \( \gamma(u) \) from BS \( j \) for user \( u \).

Accordingly, the aggregate arrival traffic load of the shared MEC micro-datacenter at site \( j \) is defined as
\[
\gamma_j := \sum_{i \in \mathcal{U}} \gamma_i(p_j) = \sum_{i \in \mathcal{U}} \gamma(u)p_j(u).
\]

Feasible set of the BS utilization of the operator \( i \) is
\[
\mathcal{F}_i = \left\{ p_i = \{p_{ij}(w_{ij}) = \sum_{w \in \mathcal{U}} \beta_{ij}(u, w_{ij})p_{ij}(u), \right.  \\
0 \leq p_{ij}(w_{ij}) \leq 1 - \epsilon,  \\
\sum_{j \in \mathcal{S}} p_{ij}(u) = 1,  \\
p_{ij}(u) = 0, \forall j \notin \mathcal{S}_i, \forall u \in \mathcal{U}_i,  \\
0 \leq p_{ij}(u) \leq 1, \forall j \in \mathcal{S}, \forall u \in \mathcal{U}_i, \right\}
\]

Thus, we define the flow-level cost as \( \phi(p_j) = \frac{1}{1 - p_j} \).

Therefore, the overall flow-level cost of all operators is defined as \( \Phi(p) = \sum_{j \in \mathcal{S}} \phi_j(p_j) = \sum_{j \in \mathcal{S}} \frac{1}{1 - p_j} \).

The delay optimal strategy of the network operator for all BSs is equivalent to minimize the following cost
\[
\Phi(p) = \sum_{j \in \mathcal{S}} \phi_j(p_j) = \sum_{j \in \mathcal{S}} \frac{1}{1 - p_j}.  \tag{1}
\]

### Computational delay:

According to M/G/1 processor sharing queueing model [3], the average response time of the MEC micro-datacenter at each site \( j \) with the traffic load \( \gamma_j \) is as
\[
\phi_j(f_j, \gamma_j) = \frac{1}{u_j \times f_j - \gamma_j},
\]

where \( u_j \) is the service rate of an edge server and \( f_j \in [0, 1] \) is the service rate scaling factor of the MEC micro-datacenter at site \( j \).

The aggregate computational delay of all ColoEdge sites as
\[
\phi_p(f_j, \gamma_j) := \sum_{j \in \mathcal{S}} \phi_j(f_j, \gamma_j) = \sum_{j \in \mathcal{S}} \frac{1}{u_j \times f_j - \gamma_j}.  \tag{2}
\]

### Power and Energy model:

The power consumption of BS \( j \) is followed a linear model as in [2] as
\[
\psi_j(p_j) = (1 - q_j)p_j + q_jQ,
\]

where \( Q \) is the maximum power of BS. Thus, in the considered control period \( \Delta \), then BS energy usage is
\[
\mathcal{E}(p_j) = \psi_j(p_j) \times \Delta.
\]

The linear power model of MEC micro-datacenter at site \( j \) in [4] as
\[
P(f_j, \gamma_j) = (1 - q_j)\frac{\gamma_j}{\mu_j \times f_j}P_{\text{max}}(f_j) + \eta P_{\text{max}}(f_j),
\]

where \( \mu_j \times f_j \) is the service rate of the MEC micro-datacenter, \( \eta \) is the fraction of idle power consumption over overall power consumption, and \( P_{\text{max}}(f_j) \) is the maximum power consumption of MEC micro-datacenters. Specifically, the MEC micro-datacenter utilization will be decreased when we increase the scaling factor \( f_j \). As the power consumption of server is a quadratic function of CPU cycle frequency [4], for simplicity, we assume the maximum power consumption of MEC micro-datacenter \( P_{\text{max}}(f_j) \) is also a quadratic function of the service rate scaling factor \( f_j \) as follows
\[
P_{\text{max}}(f_j) = \omega f_j^2 P_{\text{max}},
\]

where \( \omega \) is the PUE of the MEC micro-datacenter and \( P_{\text{max}} \) is the maximum power of MEC micro-datacenter, respectively. Then, the power model of MEC micro-datacenter is defined as
\[
P(f_j, \gamma_j) = \omega P_{\text{max}}((1 - \eta)\frac{\gamma_j}{\mu_j \times f_j} + \eta f_j^2).  \tag{4}
\]

Thus, in the considered control period \( \Delta \), the MEC micro-datacenter energy usage is
\[
\mathcal{E}(p_j, \gamma_j) = P(f_j, \gamma_j) \times \Delta.
\]

Energy usage from electricity grid follows
\[
\mathcal{E}(f_j, p, w) := \sum_{j \in \mathcal{S}} \left( \sum_{i \in \mathcal{O}_j} \mathcal{E}_p(p_{ij}) + \mathcal{E}(f_j, \gamma_j) \right) - \mathcal{E}_{\text{gen}},  \tag{5}
\]

where \( \mathcal{E}_{\text{gen}} \) is the renewable energy procurement.

### 3. ColoMEC Management Problem

In this paper, our objective of the CMMP problem reflects the trade-off of delay performance and energy cost by unit prices \( \kappa_1, \kappa_2 \) and the electricity price \( p_{\text{grid}} \) in \$. These costs follows the definition from equation (1), (2) and (5). Therefore, we ColoMEC management problem (CMMM) as follows

\[
\min_{f, p, w} \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} \mathcal{E}_p(p_{ij}) + \sum_{j \in \mathcal{S}} \mathcal{E}(f_j, p, w)
\]

s.t.
\[
\sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} w_{ij} = w_{\text{tot}},  \text{(Shared BW among operators)}
\]

\[
\begin{align*}
w_{ij} & \in (0, w_{\text{tot}}), \forall j \in \mathcal{S}, \forall i \in \mathcal{O}_j, \\
\gamma_j & \leq f_j \leq 1, \forall j \in \mathcal{S}, \\
\mu_j & < f_j, \forall j \in \mathcal{S}, \\
\rho_i & \in \mathcal{F}_i, \forall i \in \mathcal{S}.
\end{align*}
\]

In this problem, we control the service rate scaling factor \( f \) of the shared MEC micro-datacenters, user association vector \( p \), and the bandwidth allocation \( w \). At the ColoEdge sites, the operators are first allocated the fraction amount from the total shared bandwidth as in the first constraint. The stationary condition of the analyzed processing queue requires the service rate (i.e., \( \mu_j \times f_j \)) should be greater than the arrival loads \( \gamma_j \).

### 4. Solution Approach

Even though the problem is non-convex, when we fix two of three decision variables, the problem becomes convex, thus, the specific form of this problem is multi-convex problem. The convexity proofs of subproblems are not presented according to the space limitation. We adopt a proximal block coordinate descent based algorithm [5], namely PCMM. The PCMM algorithm will iteratively solves three following subproblems.
by a convex solver (i.e. ECOS [6]) at each iteration \( k \) until the convergence condition is achieved.

**User Association problem**

\[
\begin{align*}
\min_{\rho_i} & \quad \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} \phi_c(p_{ij}(w_{ij}^{(k)})) + \kappa_2 \sum_{j \in \mathcal{S}} \phi_p(f_j^{(k)}, \gamma_j) \\
& + p_{\text{grid}} \sum_{j \in \mathcal{S}} \left( \sum_{i \in \mathcal{O}_j} \phi_c(p_{ij}(w_{ij}^{(k)})) + \phi_p(f_j^{(k)}, \gamma_j) \right) \\
& + \frac{\varphi}{2} \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} \| p_{ij} - p_{ij}^{(k)} \|^2 \\
\text{s.t.} & \quad \gamma_j < h_j f_j^{(k)} \\
& \quad \rho_i(w_{ij}^{(k)}) \in \mathcal{F}_i, \forall i \in \mathcal{O}.
\end{align*}
\]

**Bandwidth Allocation problem**

\[
\begin{align*}
\min_{w_{ij}} & \quad \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} \phi_c(p_{ij}(w_{ij})) + \frac{\varphi}{2} \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} (w_{ij} - w_{ij}^{(k)})^2 \\
& + p_{\text{grid}} \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} \phi_c(p_{ij}(w_{ij}^{(k+1)})) \\
\text{s.t.} & \quad \sum_{j \in \mathcal{S}} w_{ij} = w_{\text{tot}}, \\
& \quad w_{ij} \in (0, w_{\text{tot}}), \forall j \in \mathcal{S}, \forall i \in \mathcal{O}_j.
\end{align*}
\]

**MEC Micro-datacenter Scaling problem**

\[
\begin{align*}
\min_{f_j} & \quad \sum_{j \in \mathcal{S}} \phi_p(f_j, \gamma_j) + p_{\text{grid}} \sum_{j \in \mathcal{S}} \phi_p(f_j, \gamma_j) \\
& + \frac{\varphi}{2} \sum_{j \in \mathcal{S}} (f_j - f_j^{(k)})^2 \\
\text{s.t.} & \quad \gamma_j < h_j f_j, \forall j \in \mathcal{S}.
\end{align*}
\]

5. **Simulation Results**

For an example scenario, we consider a system of five sites. These sites are located in a 1 \( \times \) 1 km\(^2\) region, the generated traffic loads come from 100 virtual groups of users. The monetary unit cost of delay performance is \( 50\$ \). The input parameters for one-day ColoMEC operation such as electricity price, renewable energy procurement, and normalized traffic loads are generated in Fig. 2. The others parameters can be found in [2].

Figure 3 shows the comparison of the total cost of three strategies. The flow-level cost of three strategies which is determined by the arrival traffic load of the users as shown in Fig. 2. Specifically, when the overall traffic loads are low, three strategies obtain similar low total cost. After 6 a.m., the traffic load increasing as in Fig. 2, the 50% MEC activation strategy requires high delay cost for the processing and the full MEC activation strategy needs huge energy cost to process traffic. In overall, the total cost of the PCMM algorithm has the lowest value in high traffic load time slots compared to fixed service rate of shared MEC micro-datacenters strategies.

6. **Conclusion**

In this paper, we study a new sharing model of colocated BS infrastructure, bandwidth, and MEC micro-datacenter to reduce the capital and operational expenses for co-located mobile network operators. We then analyze for a fundamental and under-explored ColoMEC management problem with regards to the joint user association among ColoEdge sites, shared bandwidth allocation and MEC micro-datacenter service rate scaling. As a suboptimal solution of the ColoMEC management problem, we propose the PCMM algorithm based on Proximal Block Coordinate Descent algorithm for solving our multi-convex problem.

![Fig. 2: The input parameters for one-day operation.](image)

**Fig. 3: Total Cost.**

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2017R1A2A2A05000995). *Dr. CS Hong is the corresponding author*

References:


