Let’s Share The Resource When We’re Co-located: Colocation Edge Computing

Minh N. H. Nguyen*, Chit Wutyee Zaw*, Kitae Kim*, Nguyen H. Tran†, Choong Seon Hong*

*Department of Computer Science and Engineering, Kyung Hee University
†School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia

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 networks provides the low latency and computation capability for cellular base stations (BSs). As a result, this is a great opportunity for mobile network operators by deploying new services and applications at BSs. Nevertheless, huge capital and operational cost can challenge mobile operators for the deployment of new BSs and MEC micro-datacenters. Colocation Edge Computing (ColoMEC) is a new concept where multiple operators share not only the same BS tower but also their radio and computation resources colocated at the edge sites. In order to reduce the operational cost of a ColoMEC system, the limited bandwidth at over-utilized colocated BSs can be extended by sharing the bandwidth among BSs, while shared MEC micro-datacenters can be scaled based on the arrival traffic loads. Thus, sharing the BS infrastructure, bandwidth, and MEC micro-datacenters among the co-located mobile operators can be an economical solution to provide high-performance services with low expenses by exploiting the temporal and spatial difference in traffic loads. Turning this vision into reality, we study a joint bandwidth allocation sharing and MEC micro-datacenter scaling in ColoMEC management problem (ColoMEC – MP). To solve ColoMEC – MP problem, we propose an algorithm based on proximal block coordinate descent technique by iteratively solving the decoupled convex subproblems (i.e., user association, bandwidth allocation, and MEC micro-datacenter scaling) with additional proximal terms. To improve the convergence of the proposed algorithm, we propose a greedy initialization for the user association which is based on the link capacity at each user. Our simulation demonstrates the superiority of the algorithm in terms of the operational cost compared with fixed service rate of shared MEC micro-datacenters strategies.

Index Terms—Colocation Edge Computing, User Association, Bandwidth Allocation, Server Scaling Management.

I. INTRODUCTION

Nowadays, the emergence of the upcoming mobile communications system, i.e., integrated Multi-access Edge Computing (MEC) deployment in 5G network, is considered as an evolution of cloud computing that brings the computation capability down to the network edge [1]. According to the recent white paper of ETSI [1], 3GPP 5G system specifications readily enable a collaborative operation between MEC system and a 5G system for the next generation mobile base stations. Furthermore, edge computing 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 an efficient and flexible support for edge computing to enable the superior performance and quality of experience [1]. Therefore, edge computing recently receives lots of attention in wireless communications research [2]–[9]. Furthermore, the concept of MEC sharing is initially studied in MEC slicing with regards to virtual multi-tenant service providers sharing [5], [6], [10]. Recently, Vapor IO company has successfully deployed multiple micro-datacenters at the base of cell towers in Kinetic Edge city to possibly serve hundreds of millions of end users and IoT devices within milliseconds [11].

A. Edge Colocation Computing

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. By leveraging the new ColoMEC system, mobile network operators can not only deploy new BSs and MEC micro-datacenters in a lower capital cost, but also lower operational cost based on the temporal and spatial difference in traffic loads. On one hand, the sharing of bandwidth allocation among BSs allows more bandwidth to be allocated to the high-utilized colocated BSs to reduce the transmission delay as well as the BS energy consumption. On the other hand, the scaling of shared MEC micro-datacenters can balance between the processing delay and energy consumption at colocation edge sites by optimally scaling the service rate of servers. Therefore, we pose an under-investigated ColoMEC management problem for the shared communication and computation resources regarding the operational cost of co-located mobile operators.

We illustrate the ColoMEC system in Figure[1] in which the co-located mobile operators provide communication and computation services at ColoEdge sites with colocation BSs and shared MEC micro-datacenters. Similar to [2], our considered operational cost of co-located mobile operators does not only comprise the processing and transmission delay performance, but also the energy cost of BSs and MEC micro-datacenters. Since the more energy MEC micro-datacenters and colocation BSs consume, the better delay performance of user tasks processing and transmission is. As a consequence, there is a fundamental trade-off between user delay performance and energy cost of the mobile operators. In addition, the installation cost of renewable energy (RE) becomes less and less expensive. In recent sustainable mobile networks [12]–[14], the authors examine hybrid power models, in which...
BSs can consume energy from both electric grid and the RE procurement from solar power, wind turbines. Thus, we also consider the renewable energy (RE) and electric grid as the power sources for the colocation BSs and shared MEC micro-datacenters in ColoMEC system as in [12]. Accordingly, the considered energy cost is the monetary cost to purchase the required energy usage from the utility grid after using the available energy from RE procurement.

### B. Problems and Challenges

We observe that the delay performance and energy consumption of the co-located mobile operators rely on intrinsic decision variables such as the shared MEC micro-datacenter scaling, the allocated bandwidth for colocation BSs, and user association among ColoEdge sites. In a considered region, the mobile devices can be associated with one of the ColoEdge sites and receive different transmission rates with respect to the distance between device location and BSs. Accordingly, these devices can be associated with the nearest BS but this leads to the over-utilization of some BSs or MEC micro-datacenters and under-utilization in the others. Thus, the user association can affect the transmission delay, the quality of service (QoS), and BS energy consumption of the mobile operators [15], [16]. In addition to the user association decision, we also consider the bandwidth sharing among BSs and operators as in [17]. The sharing of limited bandwidth should be carried out based on the traffic arrival pattern of users. Accordingly, the bandwidth sharing scheme should provide more bandwidth for the over-utilized BSs to reduce the overall delay by faster user traffic delivery. The joint decision of user association and bandwidth allocation substantially constitutes the communication cost of colocation BSs which includes the transmission delay and BS energy consumption.

On the other hand, the user association decision of each operator induces the arrival traffic loads that need to be processed at shared MEC micro-datacenters at ColoEdge sites. Since the statistical multiplexing of arrival loads produces a better delay performance with a lower operational cost, the traffic loads from all operators are aggregated and jointly processed at the shared MEC micro-datacenters. The processing delay and energy consumption of MEC micro-datacenters are controlled by the server scaling mechanism such as the service rate scaling to serve the aggregated arrival traffic loads. The server scaling techniques can be operated in practice based on CPU scaling of virtual machines as in [18]. Therefore, three joint intrinsic decisions of the mobile operators are coupling in the trade-offs between overall delay and energy cost minimization objectives. Thus, our research challenge is: \textit{how to optimally control the joint MEC micro-datacenter server scaling, bandwidth sharing, and user association decisions regarding the overall delay performance and energy cost of the ColoMEC system?}

### C. Contributions

In order to provide an efficient approach for the co-located mobile operators, in this work, we develop a ColoMEC management optimization problem design and analysis, which can be summarized as follows:

- **In Section [II]** we first study a system model of ColoEdge sites with colocation BSs and shared MEC micro-datacenters. Then, we introduce network operational cost based on delay performance and energy cost. Finally, we pose the joint user association, bandwidth sharing and MEC micro-datacenter scaling optimization in the ColoMEC – MP problem, which is in form of a multi-convex problem.

- **In Section [III]** we develop a centralized solution approach based on an iterative algorithm, namely PCMM algorithm for solving the ColoMEC – MP problem. Accordingly, the PCMM algorithm provides suboptimal solutions by iteratively solving three decoupled subproblems with additional proximal terms such as user association problem, bandwidth sharing problem and shared MEC micro-datacenter server scaling problem. We also provide a greedy based initialization for the PCMM algorithm to boost the convergence and reduce the number of iterations.

- **In Section [IV]** we further provide extensive numerical results to compare the performance PCMM algorithm with two fixed MEC micro-datacenter scaling according to the mobile operator cost. Finally, we present the conclusions in Section [V].

### D. Related Works

To highlight the differences of Colocation Edge Computing, Cooperation Edge Computing and Edge Slicing, we present a table of comparison in Table [I]. The designs of the three
models are different from each other where in the cooperative
model, the geographically distributed BSs collaborate together
for load balancing, energy efficiency and achieving the mini-
mum latency at the same time regardless the sharing of the
infrastructure, energy, computing, radio resources among the
co-located mobile operators [4], [19], [20], [24]. On the other
hand, the slicing approach focuses on the customization and
flexibility of the requested services by creating the logical
slices which are the combination of physical resources tailored
to meet the specific requirements of services [21]–[23], [25]–
[27]. In contrast to these approaches, our proposed model, the
colocation of base stations and micro-datacenters, allows the
operators to share their radio and computing resources for the
higher utilization of the resources. Since the physical infras-
tructure and energy are shared among the multiple operators,
the operators can save the Capital and Operational Expenditure
(CAPEX, OPEX) in deploying and maintaining multiple base
stations which are required to be installed at different locations
to meet the growing demands of users.

Recently, there is an increasing interest in mobile edge
computing studies for future mobile applications. Many of
them focus on the task offloading and resource allocation for
BSs and MEC servers of a single mobile operator such as
[7]–[9], [28]–[32]. One of the key challenges in MEC is the
joint allocation of communication and computing resources
where the recent works such as [29]–[32] address this problem
in which a single operator controls the joint allocation of
both communication and computing resources of its users.
The computation offloading in fog radio access networks is
proposed in [28] where users can choose the computation
offloading modes, which are local, fog or cloud computation
offloading, in order to minimize the weighted latency and
energy consumption. Different from a single edge site envi-
nronment, we consider a multi-site environment with a central
ColoMEC Orchestrator to manage all of BSs and MEC micro-
datacenters. In the related model for the ultra-dense context
[5], the authors focus on energy-aware mobility management.
And in [4], the authors propose the load balancing scheme
for a network of MEC-enabled BSs while our formulation is
dedicated to the user association and scaling management of
MEC micro-datacenters. To the best of our knowledge, the
most relevant MEC models with regards to multiple service
providers are MEC slicing [5], [6] where the authors formulate
stochastic games for the long-term utility performance for the
users task offloading from multi-tenants service providers.
However, in [5], the authors only consider channel assignment
and the energy allocation for users is also considered in [6].

Based on previous works of backup power sharing for
colocation BSs [15], [16], we develop a new model of sharing
bandwidth and MEC micro-datacenter among co-located
mobile network operators at ColoEdge sites in this work.
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 Or-
chestrator will decide BSs that 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. Our ColoMEC
management problem and solution approach are explained in
detail later in the next sections.

II. COLOMEC MANAGEMENT PROBLEM
A. System Model

We first present a Central ColoMEC Orchestrator operating
a system of colocation BSs and shared MEC micro-datacenters
that provides the communication and computation services for
co-located mobile network operators as shown in Figure[1]

The main function of the Central ColoMEC Orchestrator is jointly
control 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 set
of ColoEdge sites \( S \), each operator in the set of operators \( O \)
provides their services at a colocation BS [16] and a shared
MEC micro-datacenter. Although sharing BS infrastructure
and MEC micro-datacenter introduces an economical solution
for mobile network operators, the existing BSs deployment

<table>
<thead>
<tr>
<th>TABLE I: Comparison of the systems</th>
</tr>
</thead>
</table>
| **Design Scenario** | BSs and MEC micro-datacenters of dif-
ferent operators are deployed together
at the same location and share the in-
frastucture, energy, computing, radio
resources among the operators. | Geographically distributed base stations
collaborate together (no sharing of re-
sources) [4], [19], [20]. | Physical resources such as radio, com-
puting and storage are packed together
into logical slices which are specific to
requirements of services [21]–[23]. |
| **Objective** | Cost Effective (CAPEX, OPEX), Re-
source Utilization | Load Balancing, Energy Efficiency | Customizability, Flexibility, Service-
oriented [5], [6], [25]–[27] |

| TABLE II: The summary table of important notations |
|-----------------|-----------------|
| **Var** | **Definition** |
| \( t \) | Index denoting mobile operator |
| \( j \) | Index denoting ColoEdge site |
| \( \lambda(u) \) | Arrival rate at user \( u \) |
| \( \gamma(u) \) | Traffic load from user \( u \) |
| \( e_{ij}(u) \) | Transmission rate of user \( u \) of the operator \( i \) from BS \( j \) |
| \( \phi_{ij}(u) \) | BS load for user \( u \) of operator \( i \) at BS \( j \) |
| \( \rho_{ij} \) | Routing probability associated with user \( u \) of operator \( i \) at site \( j \) |
| \( w_{ij} \) | BS Load utilization of operator \( i \) at site \( j \) |
| \( f_{ij} \) | Bandwidth allocation for the operator \( i \) at site \( j \) |
| \( \phi_{(s)}(\rho_{i}) \) | Flow-level cost of the operator \( i \) |
| \( \psi_{i}(\rho_{i}) \) | Power consumption of operator \( i \) |
| \( \phi_{(d)}(f_{ij}, \gamma_{j}) \) | Processing delay of MEC micro-datacenter at site \( j \) |
| \( P(f_{ij}, \gamma_{j}) \) | Power consumption of MEC micro-datacenter at site \( j \) |
| \( \Sigma(f, w, \gamma) \) | Energy consumption of the ColoMEC system |
| \( p_{grid} \) | Renewable Energy procurement |
equation (1). According to M/G/1 processor sharing queueing

deinitions.

B. Computation delay at shared MEC micro-datacenters

Communication cost and energy cost for each ColoEdge site in next subsections.

\[ \gamma_j := \sum_{u \in \mathcal{O}_j} \gamma(u)p_{ij}(u) \]  

Based on the traffic loads model, we formulate the communication cost and energy cost for each ColoEdge site in next subsections.

B. Computation delay at shared MEC micro-datacenters

The computation time of MEC micro-datacenter at site \( j \) is the processing time of the arrival traffic load \( \gamma_j \) as in the equation (1). According to M/G/1 processor sharing queuing model [33], the average response time of the MEC micro-datacenter at each site \( j \) with the traffic load \( \gamma_j \) (bits/s) is as

\[ \phi_p(f_j, \gamma_j) = \frac{1}{\mu_j \cdot f_j - \gamma_j}, \]

where \( \mu_j \) is the maximum service rate (processed bits/s) of MEC micro-datacenter which can be different based on the MEC micro-datacenter processing power, \( f_j \) is the service rate scaling factor of the MEC micro-datacenter at site \( j \), of which the traffic load should be greater than the arrival traffic load \( \gamma_j \). Thus, the service rate scaling factor is bounded (i.e., \( f_j \subset \left[ \frac{1}{\mu_j}, 1 \right] \)).

Therefore, the aggregated processing delay of shared MEC micro-datacenters at all ColoEdge sites is defined as

\[ \phi_p(f, \gamma) := \sum_{j \in \mathcal{S}} \phi_p(f_j, \gamma_j) = \sum_{j \in \mathcal{S}} \frac{1}{\mu_j \cdot f_j - \gamma_j}. \] (2)

C. Communication flow-level cost of colocation BSs

For each user \( u \) of the operator \( i \) served by BS \( j \), the downlink transmission rate (bits/s) follows Shannon capacity as

\[ c_{ij}(u, w_{ij}) = w_{ij} \log_2 \left( 1 + \frac{P_{ij}g_{ij}(u)}{\sigma^2 + I_{ij}(u)} \right), \]

where \( w_{ij} \) is the fraction of bandwidth allocated to the operator \( i \) at the BS \( j \), \( P_{ij} \) denotes the transmission power of operator \( i \) at BS \( j \) and \( g_{ij}(u) \) denotes the channel gain from the BS \( j \) to user \( u \) including path loss, shadowing, and other factors. In addition, \( \sigma^2 \) denotes the noise power and \( I_{ij}(u) \) denotes the average interference received by user \( u \). Various radio propagation models can also be used to model the path loss in dB and accounted for shadow fading effect.

**BS load** of the operator \( i \) at the BS \( j \) is defined by the transmission time required to deliver the traffic load \( \gamma(u) \) in (bits/s) from BS \( j \) to user \( u \) as follows

\[ \beta_{ij}(u, w_{ij}) := \frac{\gamma(u)}{c_{ij}(u, w_{ij})}. \]

The summary of the processed traffic load model at the ColoEdge site \( j \) is illustrated in Figure 2.

**Definition 1 (Feasibility):** The set \( \mathcal{F}_i \) of feasible BS loads (or utilization) \( \rho_i \) is the set of BS utilization, i.e., \( \rho_i := \{\rho_{ij}\} \) of the mobile operator \( i \) with the association probability variable, \( p_i(u) \) is defined as

\[ \rho_{ij} := \frac{\gamma_{ij}(u)}{c_{ij}(u, w_{ij})}. \]
where $Q$ is the maximum power of BSs. Our BS energy model follows a linear energy model in which we have the static power and the usage power consumption which is proportional to the BS utilization $\rho_{ij}$ [36]. The $\vartheta$ in the formula defines the fraction of the static power and usage power portion of the measured maximum power consumption of a BS. Accordingly, when the BS is fully utilized (i.e., $\rho_{ij} \to 1$), it consumes the maximum power as $Q$ which includes power consumption of transmit antennas, power amplifiers, and others.

Thus, within the considered control period $\Delta_t$, the energy usage of the operator $i$ at the BS $j$ is

$$E_c(\rho_{ij}) = \psi_{ij}(\rho_{ij}) \times \Delta_t.$$  \hspace{1cm} (6)

When the operator $i$ does not exist at the ColoEdge site $j$, the energy consumption is zero, (i.e., $E_c(\rho_{ij}) = 0$).

**MEC Micro-datacenter Energy:**

We also refer to [37] for the linear power model of shared MEC micro-datacenter at site $j$ for the aggregated traffic from all operators $\gamma_j$ as

$$P(f_j, \gamma_j) = (1 - \eta)\gamma_j \mu_j \times f_j P_{\text{max}, j}(f_j) + \eta P_{\text{max}, j}(f_j),$$

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 $P_{\text{max}, j}(f_j)$ is the maximum power consumption of MEC micro-datacenters. The fraction $\gamma_j / \mu_j \times f_j$ is the utilization of MEC micro-datacenter, which depends on the service rate scaling factor $f_j$. 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 [37], the maximum power consumption of MEC micro-datacenter $P_{\text{max}, j}(f_j)$ is analogously a quadratic function of the service rate scaling factor $f_j$ as follows

$$P_{\text{max}, j}(f_j) = \omega f_j^2 P_{\text{max}},$$

where $\omega$ is the Power usage effectiveness (PUE) of the MEC micro-datacenter and $P_{\text{max}}$ is the maximum power of MEC micro-datacenter, respectively. Intuitively, the MEC micro-datacenter requires more power when it needs high CPU cycle and produces a higher service rate to serve the user tasks. Then, the power model of MEC micro-datacenter is defined as

$$P(f_j, \gamma_j) := (1 - \eta)\frac{\gamma_j}{\mu_j} \omega f_j P_{\text{max}} + \eta \omega f_j^2 P_{\text{max}}$$

$$= \omega P_{\text{max}} ((1 - \eta)\frac{\gamma_j}{\mu_j} f_j + \eta f_j^2).$$

Consequently, in the considered control period $\Delta_t$, the energy consumption of the shared MEC micro-datacenter at ColoEdge site $j$ is as

$$E_p(f_j, \gamma_j) = P(f_j, \gamma_j) \times \Delta_t.$$  \hspace{1cm} (7)

According to the energy consumption for the communication and computation at ColoEdge sites in the equation (6) and (7) respectively, the remaining energy that needs to be purchased from utility grid after using the available renewable energy procurement is defined as

$$\mathcal{F}_i = \{ \rho_i | \rho_{ij}(w_{ij}) = \sum_{u \in \mathcal{U}_i} \beta_{ij}(u, w_{ij}) \rho_{ij}(u),$$

$$0 \leq \rho_{ij}(w_{ij}) \leq 1 - \epsilon,$$

$$\sum_{j \in \mathcal{S}} \rho_{ij}(u) = 1,$$

$$\rho_{ij}(u) = 0, \forall j \notin \mathcal{S}_i, \forall u \in \mathcal{U}_i,$$

$$0 \leq \rho_{ij}(u) \leq 1, \forall j \in \mathcal{S}, \forall u \in \mathcal{U}_i,$$

where $\epsilon$ is an arbitrarily small positive constant. The sum of the routing probability of traffic flow from the user $u$ to all the BSs should be one. However, we force the association probability of the user $u$ to zero if the operator $i$ does not exist at the ColoEdge site $j$. Similar to [16], [35], the feasible set $\mathcal{F}_i$ can be analogously proved to be convex.

**Flow-level cost:** Based on the queuing analysis [35] with the M/GI/1 multi-class processor sharing system, the expected total number of flows of the operator $i$ at the BS $j$ is defined as

$$L_{ij} = \frac{\rho_{ij}}{1 - \rho_{ij}}.$$

The flow-level cost of the operator $i$ at the BS $j$ can be defined as

$$\phi_c(\rho_{ij}) = L_{ij} + 1 = \frac{1}{1 - \rho_{ij}}.$$

The application of the delay optimal strategy for each operator $i$ requires minimizing the following cost

$$\phi_c(\rho_i) = \sum_{j \in \mathcal{S}} \phi_c(\rho_{ij}) = \sum_{j \in \mathcal{S}} \frac{1}{1 - \rho_{ij}}.$$  \hspace{1cm} (4)

According to the Little Law, to minimize the average delay is equivalent to minimizing the expected total number of flows. For the analytical purpose, we use the flow-level costs $\phi_c(\rho_i)$ as an equivalent average flow delay cost, which reflects the QoS of the operator $i$. The further detailed discussion of the colocation BS cost model can be referred to our previous work [16].

Therefore, the overall flow-level cost of all operators is defined as

$$\phi_c(\rho_i) = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \phi_c(\rho_{ij}) = \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{O}_j} \phi_c(\rho_{ij})$$  \hspace{1cm} (5)

Note that the right-hand side of the equation (5) eliminates the delay cost of the operators that does not exist in the ColoEdge sites.

**D. Power and Energy model**

Similar to the delay performance at BSs and shared MEC micro-datacenters, we formulate the energy usage model of ColoEdge sites in this subsection.

**BS Energy:** The BS power consumption of the operator $i$ at site $j$ follows a linear model as in [36] which depends on the BS utilization $\rho_{ij}$ as

$$\psi_{ij}(\rho_{ij}) = (1 - \vartheta)\rho_{ij} Q + \vartheta Q,$$

where $Q$ is the maximum power of BSs. Our BS energy model follows a linear energy model in which we have the static power and the usage power consumption which is proportional to the BS utilization $\rho_{ij}$ [36]. The $\vartheta$ in the formula defines the fraction of the static power and usage power portion of the measured maximum power consumption of a BS. Accordingly, when the BS is fully utilized (i.e., $\rho_{ij} \to 1$), it consumes the maximum power as $Q$ which includes power consumption of transmit antennas, power amplifiers, and others.

Thus, within the considered control period $\Delta_t$, the energy usage of the operator $i$ at the BS $j$ is

$$E_c(\rho_{ij}) = \psi_{ij}(\rho_{ij}) \times \Delta_t.$$  \hspace{1cm} (6)

When the operator $i$ does not exist at the ColoEdge site $j$, the energy consumption is zero, (i.e., $E_c(\rho_{ij}) = 0$).

**MEC Micro-datacenter Energy:**

We also refer to [37] for the linear power model of shared MEC micro-datacenter at site $j$ for the aggregated traffic from all operators $\gamma_j$ as

$$P(f_j, \gamma_j) = (1 - \eta)\gamma_j \mu_j \times f_j P_{\text{max}, j}(f_j) + \eta P_{\text{max}, j}(f_j),$$

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 $P_{\text{max}, j}(f_j)$ is the maximum power consumption of MEC micro-datacenters. The fraction $\gamma_j / \mu_j \times f_j$ is the utilization of MEC micro-datacenter, which depends on the service rate scaling factor $f_j$. 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 [37], the maximum power consumption of MEC micro-datacenter $P_{\text{max}, j}(f_j)$ is analogously a quadratic function of the service rate scaling factor $f_j$ as follows

$$P_{\text{max}, j}(f_j) = \omega f_j^2 P_{\text{max}},$$

where $\omega$ is the Power usage effectiveness (PUE) of the MEC micro-datacenter and $P_{\text{max}}$ is the maximum power of MEC micro-datacenter, respectively. Intuitively, the MEC micro-datacenter requires more power when it needs high CPU cycle and produces a higher service rate to serve the user tasks. Then, the power model of MEC micro-datacenter is defined as

$$P(f_j, \gamma_j) := (1 - \eta)\frac{\gamma_j}{\mu_j} \omega f_j P_{\text{max}} + \eta \omega f_j^2 P_{\text{max}}$$

$$= \omega P_{\text{max}} ((1 - \eta)\frac{\gamma_j}{\mu_j} f_j + \eta f_j^2).$$

Consequently, in the considered control period $\Delta_t$, the energy consumption of the shared MEC micro-datacenter at ColoEdge site $j$ is as

$$E_p(f_j, \gamma_j) = P(f_j, \gamma_j) \times \Delta_t.$$  \hspace{1cm} (7)

According to the energy consumption for the communication and computation at ColoEdge sites in the equation (6) and (7) respectively, the remaining energy that needs to be purchased from utility grid after using the available renewable energy procurement is defined as
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2020.2982679, IEEE Transactions on Vehicular Technology

\[ E(f, p, w) := \sum_{j \in S} \left( \sum_{i \in O_j} E_c(\rho_{ij}) + E_p(f_j, \gamma_j) \right) - E_{\text{gen}}, \]  

where \( E_{\text{gen}} \) is the renewable energy procurement in a certain control period.

### E. Problem Formulation

In the scope of this paper, we consider the overall operational cost that consists of the processing delay, communication flow-level cost, and energy costs of all mobile operators. Correspondingly, the Central ColoMEC Orchestrator needs to control efficiently the joint MEC service rate scaling, user association, and bandwidth allocation decisions to minimize the overall cost regarding the traffic loads of users, dynamic electricity price, and the RE procurement in a certain time period. Therefore, we propose the ColoMEC Management Problem of the mobile operator as follows

**ColoMEC Management Problem (ColoMEC – MP)**

\[
\text{min.} \quad \kappa_1 \sum_{j \in S} \sum_{i \in O_j} \phi_c(\rho_{ij}) + \kappa_2 \sum_{j \in S} \phi_p(f_j, \gamma_j) \\
+ p_{\text{grid}} \times E(f, p, w) \\
\text{s.t.} \quad \sum_{j \in S, i \in O_j} w_{ij} = w_{\text{tot}}, (\text{Shared BW among operators}) \tag{9} \\
\quad w_{ij} \in (0, w_{\text{tot}}), \forall j \in S, \forall i \in O_j, \tag{10} \\
\quad \gamma_j \leq f_j \leq 1, \forall j \in S, \tag{11} \\
\quad \rho_i \in \mathcal{F}_i, \forall i \in \mathcal{O}. \tag{12}
\]

**Remark 1.** BS load \( \rho_i \) is the function of bandwidth allocation \( w_i \) and user association \( p_i \), while the shared MEC micro-datacenter traffic loads \( \gamma_j \) is the function of the user association decisions from all operators at the ColoEdge site \( j \) that is defined in the equation 7.

The objective of the ColoMEC – MP 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 $\text{S}$. These costs follow the definition from equation (2), (5), (8) in the previous subsections. The feasible set of BS load \( \rho_i \) in constraint (12) is defined in the definition of feasibility (3), which is also a convex set. Straightforwardly, the problem is non-convex due to the coupling of our three decision variables in the objective function.

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 fraction of the total shared bandwidth is first allocated to the operators as in the constraint (12). As the aforementioned discussion in our system model, sharing bandwidth scheme allows overall delay reduction of the overloaded BSs by allocating more bandwidth to rapidly transmit the user data. The decisions of bandwidth allocation and user association jointly affect the communication cost and energy consumption of BSs based on the temporal, spatial arrival traffic difference of operators. Consequently, there exists a coupling in bandwidth allocation and user association decisions among operators and BSs. In the shared MEC micro-datacenters operations, the aggregated traffic load \( \gamma_j \) at each ColoEdge site \( j \) is processed within the processing delay \( \phi_p(f_j, \gamma_j) \) based on the scaling factor of the service rate \( f_j \). Accordingly, user association decisions of co-located operators induce the amount of arrival traffic load at shared MEC micro-datacenters which result in another coupling between MEC micro-datacenter scaling and user association decisions of operators.

### III. 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 the multi-convex problem 38. The convexity proofs of subproblems are shown in the next subsection. In this section, we propose a proximal block coordinate descent based algorithm and a greedy-based initialization approach for the ColoMEC – MP problem.

To solve this multi-convex problem, we design an iterative algorithm, namely PCMM, based on Proximal Block Coordinate Descent (PBCD) algorithm 38. The block coordinate descent method follows Gauss-Seidel update technique, which cyclically solves the optimization problem over each block of variables while fixing the remaining blocks at their last updated values. The additional proximal terms guarantee the convergence of the iterative algorithm to local optimal solutions regardless of the strongly convex condition of the objective for each block of variables 38. We transform our subproblems such as User Association problem, Bandwidth Allocation problem and MEC Micro-datacenters Scaling problem to strongly convex problems by adding the proximal terms of decision variables to the objective functions. Accordingly, PCMM first updates the user association of all operators by solving the User Association problem with an additional proximal term for the given allocated bandwidth of colocation BSs and service rate scaling factor of the shared MEC micro-datacenters. After updating the user association decision, the allocated bandwidth of each operator and the service rate scaling of the shared MEC micro-datacenter at ColoEdge sites are updated by the proximal Bandwidth Allocation problem and MEC Micro-datacenters Scaling problem respectively. The whole process is repeated until it achieves the convergence condition which is shown at line 10 in Alg. 7. Specifically, we describe these sub-problems for each iteration as following.

At the iteration \( k + 1 \), given the previous service rate scaling factor \( f^{(k)} \) of all MEC micro-datacenters and the bandwidth allocation \( u^{(k)} \), the ColoMEC – MP problem can be simplified to the UAP problem with an additional proximal term for the user association decisions as follows:

\[ f^{(k+1)} = \text{argmin} \quad \phi_c(\rho_i^{(k)}) + \phi_p(f_j^{(k)}, \gamma_j^{(k)}) + p_{\text{grid}} \times E(f^{(k)}, p^{(k)}, w) \]

\[ s.t. \quad \sum_{j \in S, i \in O_j} w_{ij}^{(k+1)} = w_{\text{tot}} \]

\[ w_{ij}^{(k+1)} \in (0, w_{\text{tot}}), \forall j \in S, \forall i \in O_j \]

\[ \gamma_j^{(k+1)} \leq f_j^{(k)}, \forall j \in S \]

\[ \rho_i \in \mathcal{F}_i, \forall i \in \mathcal{O} \]
User Association Problem (UAP)

\[ \begin{align*}
\text{min. } & \quad \kappa_1 \sum_{j \in S} \sum_{i \in O_j} \phi_c(p_{ij}(w_{ij}^{(k)})) + \kappa_2 \sum_{j \in S} \phi_p(f_j^{(k)}, \gamma_j) \\
+ & \quad p_{grid} \sum_{j \in S} \left( \sum_{i \in O_j} E_c(p_{ij}(w_{ij}^{(k)})) + E_p(f_j^{(k)}, \gamma_j) \right) \\
+ & \quad \frac{\varphi}{2} \sum_{j \in S} \sum_{i \in O_j} ||p_{ij} - p_{ij}^{(k)}||^2
\end{align*} \]

s.t. \( \gamma_j < \mu_j f_j^{(k)} \) \hspace{1cm} \quad \quad \quad (13)
\( \rho_i(w_i^{(k)}) \in F_i, \forall i \in O. \) \hspace{1cm} \quad \quad \quad (14)

Remark 2. BS load \( \rho_i(w_i^{(k)}) \) of the operator \( i \) is a function of the user association \( p_i \).

Lemma 1. The UAP problem is a convex problem.

**Proof:** As in [17] and our previous work [16], the BS load \( \rho_i(w_i^{(k)}) \) is the linear function, the flow-level cost of operator \( i \) at BS \( j \), \( \phi_c(p_{ij}(w_{ij}^{(k)})) \), is the convex function of user association (i.e., \( p_{ij} \)), and \( F_i \) is a convex feasible set. In addition, the proximal term is also a convex function and the arrival traffic load at the shared MEC micro-datacenter \( j \) is a linear function of \( p_{ij} \). Accordingly, \( E_c(p_{ij}(w_{ij}^{(k)})) \) are the linear functions of \( p_{ij} \). Moreover, we prove the processing delay is also a convex function of \( p_{ij} \) by validating its second derivative.

\[ \frac{\partial^2 C(p)}{\partial p_{ij}(u)^2} = \frac{\partial^2 C(p)}{\partial \gamma_j} \frac{\partial \gamma_j}{\partial p_{ij}(u)} = \frac{\kappa_2 \gamma(u)}{(\mu_j \times f_j^{(k)} - \gamma_j)^2} \]
\[ = \frac{\kappa_2 \gamma(u)}{(\mu_j \times f_j^{(k)} - \sum_{i \in S, w \in U_i} \gamma(u)p_{ij}(u))^2 + 2\kappa_2 \gamma^2(u)} \]
\[ = \frac{\kappa_2 \gamma(u)}{(\mu_j \times f_j^{(k)} - \sum_{i \in S, w \in U_i} \gamma(u)p_{ij}(u))^3} \]
\[ = \frac{2\kappa_2 \gamma^2(u)}{(\mu_j \times f_j^{(k)} - \lambda_j)^3} > 0. \] \hspace{1cm} (15)

Thus, the processing delay cost is a convex function of \( p_{ij} \). Finally, the objective function is convex because it is the summation of the convex and linear functions of user association \( p_{ij} \). Therefore, the UAP problem is a convex problem that can be solved by using off-the-shelf solver (i.e., ECOS [39]).

Algorithm 1 Proximal ColoMEC Management (PCMM) algorithm

1. The central ColoMEC Orchestrator updates \( \mathcal{E}_{gen}, p_{grid} \).
2. Initialize \( k = 0, p_0, f_0, w_0 \).
3. **repeat**
   4. Compute \( p^{(k+1)} \) from UAP problem given \( f^{(k)}, w^{(k)}, p^{(k)} \) at the central ColoMEC Orchestrator;
   5. Compute \( w^{(k+1)} \) from BAP problem given \( p^{(k+1)}, w^{(k)} \) at the central ColoMEC Orchestrator;
   6. for ColoEdge site \( j \in S \) do
      7. Compute traffic load \( \tilde{\gamma}_j(p^{(k+1)}) \) in equation (1);
      8. Compute \( f_j^{(k+1)} \) from Lemma 2 given \( \tilde{\gamma}_j(p^{(k+1)}), f_j^{(k)} \);
   **end for**
10. **until** \( |f^{(k+1)} - f^{(k)}| \leq \epsilon_1, ||w^{(k+1)} - w^{(k)}|| \leq \epsilon_2 \).

Bandwidth Allocation Problem (BAP)

\[ \begin{align*}
\text{min. } & \quad \kappa_1 \sum_{j \in S} \sum_{i \in O_j} \phi_c(p_{ij}^{(k+1)}, w_{ij}) + \frac{\varphi}{2} \sum_{j \in S} \sum_{i \in O_j} (w_{ij} - w_{ij}^{(k)})^2 \\
+ & \quad p_{grid} \sum_{j \in S} \sum_{i \in O_j} E_c(p_{ij}(w_{ij}^{(k)})) + E_p(f_j^{(k)}, \gamma_j) \\
\text{s.t. } & \quad \sum_{j \in S, i \in O_j} w_{ij} = w_{tot}, \\
& \quad w_{ij} \in (0, w_{tot}), \forall j \in S, \forall i \in O_j.
\end{align*} \]

Remark 3. BS load \( \rho_j(p_j^{(k+1)}) \) is the function of the bandwidth allocation \( w_j \). As in the \([17]\), given the user association \( p^{(k+1)} \), the BS load function is redefined as \( \tilde{p}_j :\tilde{p}_j(w_j) := \frac{\tilde{\rho}_j}{w_j} \), where \( \tilde{\rho}_j \) is the referenced BS load of operator \( i \) at ColoEdge site \( j \) when \( w_{ij} = 1 \) (i.e., \( \tilde{\rho}_j := \rho_j(1, p^{(k+1)}(1)) \)).

According to the Feasibility definition \([3]\), BS utilization is less than one (i.e., \( \rho_j(w_j) \leq 1 - \epsilon \)), which implies that \( w_{ij} \geq (1 - \epsilon)\tilde{\rho}_j \). Thus, an equivalent problem for the BAP is defined as

\[ \begin{align*}
\text{min. } & \quad \kappa_1 \sum_{j \in S, i \in O_j} w_{ij} \tilde{\rho}_j + \kappa_3 \sum_{j \in S, i \in O_j} \tilde{\rho}_j w_{ij} \\
+ & \quad \frac{\varphi}{2} \sum_{j \in S, i \in O_j} (w_{ij} - w_{ij}^{(k)})^2 \\
\text{s.t. } & \quad \sum_{j \in S, i \in O_j} w_{ij} = w_{tot}, \\
& \quad w_{ij} \in (1 - \epsilon)\tilde{\rho}_j, w_{tot}, \forall j \in S, \forall i \in O_j, \quad (16)
\end{align*} \]

where \( \kappa_3 = p_{grid}Q(1 - \theta)A_{\lambda} \).

It is straightforward to prove the convexity of the problem by validating that the second derivative of the objective \( C(w) \) is non-negative as follows

\[ \frac{\partial^2 C(w)}{\partial w_{ij}^2} = \frac{-\kappa_1 \tilde{\rho}_j}{(w_{ij} - \tilde{\rho}_j)^2} + \frac{\varphi}{w_{ij}^2} \]
\[ \frac{\partial^2 C(w)}{\partial w_{ij}^2} = \frac{2\kappa_1 \tilde{\rho}_j}{(w_{ij} - \tilde{\rho}_j)^2} + \frac{2\kappa_3 \tilde{\rho}_j}{w_{ij}^2} + \varphi. \]

Consequently, the second derivative of \( C(w) \) is non-negative because of the constraint \([16]\).

Thus, the objective function of
the BAP problem is a convex function and BAP is a convex problem.

Similarly, given the user association decision \( p^{(k+1)} \), the ColoMEC-MP problem can be simplified to the MMSP problem as follows

**MEC Micro-datacenter Scaling problem (MMSP)**

\[
\begin{align*}
\min_f & \quad \kappa_2 \sum_{j \in S} \phi_p(f_j, \tilde{\gamma}_j) + \rho_{grid} \sum_{j \in S} \mathcal{E}_p(f_j, \tilde{\gamma}_j) \\
& \quad + \frac{\eta}{2} \sum_{j \in S} (f_j - f_j^{(k)})^2 \\
\text{s.t.} & \quad \frac{\tilde{\gamma}_j}{\mu_j} < f_j \leq 1, \quad \forall j \in S,
\end{align*}
\]

where \( \tilde{\gamma}_j := \gamma_j(p^{(k+1)}) \).

After substituting and removing the constants in the objective function, we obtain the following equivalent problem

\[
\begin{align*}
\min_f & \quad \kappa_2 \sum_{j \in S} \frac{1}{\mu_j} f_j - \tilde{\gamma}_j + \kappa_4 \left( (1 - \eta) \frac{\tilde{\gamma}_j}{\mu_j} f_j + \eta f_j^2 \right) \\
& \quad + \frac{\eta}{2} \sum_{j \in S} (f_j - f_j^{(k)})^2 \\
\text{s.t.} & \quad \frac{\tilde{\gamma}_j}{\mu_j} < f_j \leq 1, \quad \forall j \in S,
\end{align*}
\]

where \( \kappa_4 = \rho_{grid} \omega \rho_{max} \Delta_t \).

Moreover, we observe that the MMSP problem can be solved for each ColoEdge site due to the independent characteristic of decision variables in the problem.

**Independent MEC Micro-datacenter Scaling problem (iMMSP)**

\[
\begin{align*}
\min_{f_j} & \quad \frac{\kappa_2}{\mu_j} f_j - \tilde{\gamma}_j + \kappa_4 \left( (1 - \eta) \frac{\tilde{\gamma}_j}{\mu_j} f_j + \eta f_j^2 \right) \\
& \quad + \frac{\eta}{2} (f_j - f_j^{(k)})^2 \\
\text{s.t.} & \quad \frac{\tilde{\gamma}_j}{\mu_j} < f_j \leq 1.
\end{align*}
\]  \( \tag{17} \)

**Lemma 2.** The iMMSP problem is a strongly convex function. According to Karush-Kahn-Tucker (KKT) [40], the iMMSP problem is feasible and has an unique optimal solution if there is an unique solution in \( \left( \frac{\tilde{\gamma}_j}{\mu_j}, 1 \right) \) of the following equation

\[
A + (2\eta + \varphi) f = \frac{B}{(\mu \times f - \tilde{\gamma})^2}
\]  \( \tag{18} \)

where \( A = \kappa_4 (1 - \eta) \frac{\tilde{\gamma}_j}{\mu_j} - \varphi f^{(k)} \) and \( B = \kappa_2 \mu \).

**Proof:** In this proof, we skip index \( j \) of a specific ColoEdge site. We consider the first and second derivative of the objective of the iMMSP problem, denoted by \( C(f) \)

\[
\frac{dC(f)}{df} = \frac{-\mu \kappa_2}{(\mu \times f - \tilde{\gamma}^2) + \kappa_4 \left( (1 - \eta) \frac{\tilde{\gamma}_j}{\mu_j} + 2\eta f \right) + \varphi (f - f^{(k)})}
\]

\[
\Rightarrow \frac{d^2C(f)}{df^2} = \frac{2\mu^2 \kappa_2}{(\mu \times f - \tilde{\gamma})^3} + 2\eta \kappa_4 + \varphi.
\]

According to the constraint (17), \( \frac{d^2C(f)}{df^2} > 0 \). Thus, the objective function is a strongly convex function and iMMSP is a strongly convex problem.

**Algorithm 2 Greedy Initialization Strategy (GIS)**

1: The central ColoMEC Orchestrator updates \( \mathcal{E}_{gen}, \rho_{grid} \)
2: Initialize \( p = 0 \);
3: for each operator \( i \in O \) do
4: \hspace{1em} for \( u \in \mathcal{U}_i \) do
5: \hspace{2em} Find the nearest BS \( j \) of user \( u \) based on the link capacity;
6: \hspace{2em} Set \( p_{ij}(u) = 1 \);
7: \hspace{1em} end for
8: end for
9: Compute \( w \) from BAP problem given \( p \) without adding proximal term (i.e., \( \varphi = 0 \)) at the central ColoMEC Orchestrator;
10: for ColoEdge site \( j \in S \) do
11: \hspace{1em} Compute traffic load \( \gamma_j(p) \) in equation (1);
12: \hspace{1em} Compute \( f_j \) from Lemma 2 given \( \gamma_j(p) \) with \( \varphi = 0 \);
13: end for

**Central ColoMEC Orchestrator**

**Mechanical micro-datacenter Scaling at ColoEdge site 1**

**Algorithm module deployment.**

\[ f \]

Using KKT condition, we have the following condition

\[
\frac{dC(f)}{df} = \frac{-\mu \kappa_2}{(\mu \times f - \tilde{\gamma}^2) + \kappa_4 \left( (1 - \eta) \frac{\tilde{\gamma}_j}{\mu_j} + 2\eta f \right) + \varphi (f - f^{(k)}) = 0}
\]

\[
\Rightarrow \kappa_4 (1 - \eta) \frac{\tilde{\gamma}_j}{\mu_j} - \varphi f^{(k)} + (2\eta + \varphi) f = \frac{-\mu \kappa_2}{(\mu \times f - \tilde{\gamma})^2}.
\]  \( \tag{19} \)

It is straightforward that the left-hand side of the equation (19) is an increasing function due to \( 2\eta + \varphi > 0 \). When \( f \in \left( \frac{\tilde{\gamma}_j}{\mu_j}, 1 \right) \), the right-hand side of the equation (19) is a decreasing function since \( \frac{-\mu \kappa_2}{(\mu \times f - \tilde{\gamma})^2} > 0 \). Therefore, the problem is feasible if there exist a unique solution in \( \left( \frac{\tilde{\gamma}_j}{\mu_j}, 1 \right) \) of the equation (19).

The convergence analysis note for the PCMM algorithm is shown in Appendix A. The complexity of a single iteration is \( O(|\mathcal{U}| \times |S| \times |O|) \) based on dimensions of the block variables in each subproblem. Accordingly, total execution time of algorithm is \( T \times O(|\mathcal{U}| \times |S| \times |O|) \) where \( T \) is the total number of iterations.

**Greedy Initialization Strategy:** In the PCMM algorithm, the number of iterations depends on the initialization of the decision variables (i.e., \( p^{(0)}, f^{(0)}, w^{(0)} \)). Therefore, we propose a greedy strategy based on the transmission rate of each user as an approximated distributed approach for UAP problem.
for the initialization stage. Specifically, users are allowed to associate with the nearest ColoEdge site for the association step. By leveraging the benefit of bandwidth sharing among the BSs and operators by solving BAP problem, over-utilized BSs can receive more fraction of the total bandwidth to reduce the flow-level cost. In addition, this greedy strategy enables users to perform association with the colocation BSs individually. Note that, the GIS only requires one iteration since the fixed user association decision can be decoupled among the joint decisions. Different from GIS, the PCMM algorithm requires multiple communications rounds until the convergence condition is satisfied.

According to the theoretical global convergence analysis in Appendix A, one of the essential assumptions requires that the initial point should be close to the critical points. Based on the observation of the converged solution of the user association, most of the interior virtual users still associate with the closest base station. Thus, the GIS initialization strategy provides an initial solution that is close to the converged solution and it can boost the convergence of the PCMM algorithm. The implementation can be performed in two ways for the more practical approach. First, the PCMM algorithm can be stopped when the infinitesimal overall cost improvement is obtained and, second, only GIS step can be run.

The PCMM algorithm can be deployed at the Central ColoMEC Orchestrator to control the bandwidth sharing and user association decision, while the shared MEC micro-datacenter scaling modules can be conducted at the ColoEdge sites as shown in Figure 3. Thus, this deployment scheme allows the shared MEC micro-datacenter to be deployed and managed by a third party MEC operator (e.g., Vapor IO [41]). For another setting, where MEC micro-datacenters are deployed by the mobile operators, the shared MEC micro-datacenter scaling modules can also belong to the Central ColoMEC Orchestrator.

IV. PERFORMANCE EVALUATION

<table>
<thead>
<tr>
<th>Types</th>
<th>Hours</th>
<th>Prices ($/Wh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-Peak</td>
<td>[14:00, 20:00]</td>
<td>0.12196</td>
</tr>
<tr>
<td>Mid-Peak</td>
<td>[6:00, 14:00], [20:00, 22:00]</td>
<td>0.07134</td>
</tr>
<tr>
<td>Off-Peak</td>
<td>[22:00, 6:00]</td>
<td>0.02182</td>
</tr>
</tbody>
</table>

TABLE III: Time-Off-Use Tariff of a weekday in summer [41].

A. Simulation Settings

We consider a system of five ColoEdge sites in a $1 \times 1$ km$^2$ region as in Figure 1. In this region, the mobile users of each operator are divided into 100 virtual groups based on the location. The traffic flows of users are generated according to the Poisson Point Process. Users can be associated with all BSs in the considered region which can affect the BS utilization. In order to control the traffic load, we assume that each data request has the download size which follows the log normal distribution with mean value $1/\mu(u) = 1$. As an example of the heterogeneous service demands, the traffic loads of three operators in the considered region are generated and represented by the heat map (normalized to one) as shown in Figure 4a. First, the traffic is randomly generated for operator 1 and the traffic load of operator 2 is densely generated around the center of ColoEdge site 2 while the traffic is produced densely in corner regions for operator 3. Figure 4 shows the traffic pattern of each operator at 5 p.m., the input parameters for one-day ColoMEC operation such as electricity price, renewable energy procurement, and normalized traffic loads. Accordingly, one-day ColoMEC operation includes the decision making for 96 time slots where each duration is 15 minutes. The traffic load of users is lowest when people are sleeping (12 a.m. - 6 a.m.) but it will be high for the rest of the day. The electricity price is shown in Table II. For the RE, the highest energy generation occurs during the day time (up to 600 Wh) since the main popular source of RE is from the procured solar energy. However, we consider the energy generation during nights which comes from the other RE sources (e.g., wind energy).

According to the communication model of urban macro-cells, the simulation parameters follow the instructions of the WiMAX evaluation methodology document [42]. Specifically, we consider that the specifications of all colocation BSs are similar to each other such as BS height and MT height which are $32 \text{ m}$ and $1.5 \text{ m}$ respectively. In the simulation, we consider that there is no inter-operator interference and Gaussian-like noise inter-cell interference is static with lognormal shadow fading where standard deviation is $8 \text{ dB}$, the maximum BSs
transmission power is 40 Watts and COST 231 path loss model as similar settings in [16].

Furthermore, the dynamic electricity price in one day follows the Table III and the monetary unit cost of delay performance is $\kappa_1 = 5\ \$/$\text{unit}, $\kappa_2 = 50\ \$/$\text{unit}$. The maximum BS power $Q$ is 865 Watts [35], the portion of static power and active power of BSs and MEC micro-datacenters are the same which is $\vartheta = \eta = 0.2$. The maximum power of MEC micro-datacenter $P_{\text{max}}$ is 1000 Watts, and PUE $\omega$ is 1.5 [42]. The maximum service rate of shared MEC micro-datacenters $\mu$ is 150 processed Mbits/s. The parameter $\rho$ is set to 0.005 and the convergence condition thresholds $\epsilon_1, \epsilon_2$ are $2 \times 10^{-9}$ and $10^{-5}$ respectively.

B. Simulation Results

1) Performance of PCMM algorithm at 5 p.m.: The performance of PCMM is analyzed at the typical time slot corresponding to 5 p.m. when both the electric price and traffic load are high. In Figure 5 we first illustrate the unique optimal solutions of the iMMSP problems of ColoEdge sites using KKT condition as in Lemma 2. The dotted lines represent the left-hand sides of the equation (19) which is a decreasing function for each ColoEdge site. The solid straight lines are linear functions which are the right-hand sides of the equation (19). If the problems are feasible, there are unique solutions which are presented by the black dots for ColoEdge sites in Figure 5.

For the convergence of the PCMM algorithm, we show the service rate scaling factor and total cost convergence along with the number of iterations in Figure 6. The convergence of the total cost is compared between two different approaches: GIS and Random initialization. Using the GIS initialization, the algorithm converges to a suboptimal solution after 219 iterations while the Random initialization takes almost a double number of iterations to converge as shown in Figure 6a (i.e., the random initialization requires 403 iterations on average of 50 realizations in Figure 6b). With different random initial points, the PCMM algorithm can converge to different stationary solutions at different number of iterations. The box plot illustrates the statistical distribution and shows the high variance of the required number of iterations for the random initialization. Therefore, GIS strategy can not only improve the convergence performance but also avoids unpredictable behaviors unlike the random initialization. Furthermore, the improvement of the total cost from the initial point of GIS to the stationary solution is insignificant. Accordingly, the initial solution of GIS strategy or intermediate solutions can be applied if the operational time budget is limited. On the other hand, random initialization does not provide the low cost from the initialization and intermediate solutions during the first 50 iterations.

Besides, the values of the scaling factors $f$ of ColoEdge sites vary with the traffic load and association of each site as shown in Figure 7. The scaling factor $f$ of site 2 is the

![Fig. 5: KKT Unique solutions.](image1)

![Fig. 6: GIS and Random initialization strategy comparison on convergence performance.](image2)

![Fig. 7: Service rate scaling factor convergence with GIS initialization.](image3)
highest since that site is received more user traffic loads according to the traffic arrival pattern in Figure 4a. In addition, the convergence of bandwidth allocation of three operators is shown in Figure 8. The bandwidth allocation converges to a suboptimal solution after 218 iterations. In fact, the allocation of bandwidth depends on the traffic loads and the association of users at the edge site. When the arrival traffic load at the ColoEdge site is high, the larger amount of bandwidth is needed to be allocated to that site to reduce the communications delay. The three operators have a different allocation of bandwidth according to the association decision in Figure 10 and the traffic pattern in Figure 4a at 5 p.m. Specifically, from these two operators, operator 3 has larger site coverage in Figure 10a and the highest associated traffic loads at site 3, thus, the more bandwidth is allocated to this BS which is taken from BS2 as shown in Figure 8c.

In addition, we compare the bandwidth allocation at ColoEdge site 2 of three different strategies which are PCMM, 50% and 100% activation of the shared MEC micro-datacenter in Figure 9. The bandwidth allocation is strongly connected with the user association coverage in the following results. With the two activation strategies of the MEC micro-datacenter, the coverage of the BS2 for all operators which is shown in Figure 10 is smaller, thus, less user traffic is served at this ColoEdge site, then less bandwidth is required. For the 50% MEC micro-datacenter activation strategy, the site coverage is significantly reduced due to the limitation of the arrival traffic load by the constraint (11); thus, it has the lowest bandwidth allocation for the BS2. Besides, PCMM and the full activation strategy produce the same allocation due to their user association as shown in Figure 10b and 10c for operator 2 and 3, respectively. For operator 1, the full activation strategy gives the lower bandwidth allocation than PCMM which is also affected by their lower user association coverage as in Figure 10a.

Figure 11 shows the comparison of flow-level cost, processing delay, energy cost and total cost of three strategies. In Figure 11a, the flow level cost, which is the communication
cost of the 50% activation approach, is the highest while the other two approaches have a similar result due to the limitation of the arrival traffic load where user requests cannot be associated with their nearest BSs. However, PCMM and the full activation strategy have comparable flow-level cost since they have the similar user association and bandwidth allocation. When MEC micro-datacenters are operated in the full mode, the processing delay is the lowest but the energy cost is the highest as shown in Figure 11b and 11c. The 50% activation strategy produces the highest delay because of the low service rate of the MEC micro-datacenter, while PCMM provides the lower processing delay than the 50% activation approach since the scaling factor of the majority of edge sites are higher than 0.5 as shown in Figure 7. Although 50% acquires lower the energy consumption of MEC micro-datacenters, it also consumes more energy of colocation BSs compared to PCMM; thus, the energy cost of the two approaches is similar to each other. Consequently, PCMM has the lowest total cost compared to 50% and 100% MEC micro-datacenters activation strategies.

2) Performance of PCMM algorithm for one-day ColoMEC operation: The performance of PCMM algorithm is analyzed for the whole day including 96 time slots with various electricity price, traffic demands, and renewable energy procurement. The comparison of energy usage among the three strategies are presented in Figure 12. The energy usage of the full MEC micro-datacenters activation strategy is the highest while PCMM consumes slightly more energy compared to the 50% activation as in Figure 11c. The energy usage increases with respect to the traffic load of users according to the demand of users as in Figure 4b. The top figure in the energy usage in Figure 12 shows that there is a drop around 6 a.m due to the fact that energy generation of the RE sources gets higher at that time as in Figure 4b. We can see a sudden high in energy usage around 5 p.m. because the energy generation from RE sources decrease at that time. For PCMM, there is a small increase around 10 p.m. since the energy price is decreased at that time. The energy cost has a similar trend as the energy price which is shown in Figure 4b. The only difference is a small rise after 5 p.m. because of the low generation of renewable energy.

Finally, Figure 13 shows the comparison of flow-level cost, processing delay and 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 Figure 13a. Specifically, when the overall traffic loads are low before 6 a.m. as in Figure 4b, three strategies obtain similar flow-level communications cost. After 6 a.m., the flow-level cost of 50% MEC activation strategy is considerably higher compared to the proposed PCMM and full MEC activation strategy. Besides, 50% MEC activation strategy produces the highest processing delay because of the lowest service rate of MEC micro-datacenters as shown in Figure 13b. However, for PCMM, since the objective is to minimize the delay and energy cost, the energy cost dominate its processing delay in the MMSP problem; thus, it forces to reduce the service rate when the energy prices increase. Accordingly, the processing delay has a similar trend as the energy price as shown in Figure 4b. On the other hand, when MEC micro-datacenters are operated in the full mode, the processing delay is not affected by the traffic loads or the energy cost and it achieves the lowest value. In overall, the total cost of the PCMM algorithm has the lowest value in time slots when the traffic loads are high compared to fixed service rate of shared MEC micro-datacenters strategies as shown in Figure 12. Accordingly, our proposed algorithm has the lowest flow-level communications
cost, low processing delay and energy consumption throughout the whole day operation.

V. CONCLUSION

In this paper, we study a new sharing model of colocation BSs infrastructure, bandwidth, and MEC micro-datacenter to reduce the capital and operational expenses for co-located mobile network operators. We then analyze 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. The PCMM algorithm deployment requires user association and bandwidth allocation operation at Central ColoMEC Orchestrator and shared MEC micro-datacenters at ColoEdge sites to provide a more flexible MEC deployment from third party operator (e.g., Vapor IO). In addition, we also enhance the convergence speed by a greedy based initialization algorithm according to the link capacity in order to reduce the number of updates between ColoEdge sites and central ColoMEC Orchestrator.

In the future work, we advocate extending our current ColoMEC model to adopt a practical colocation micro-datacenter concept as in [11], [14] instead of shared micro-datacenters. To overcome the signaling challenge for the global information collection, one of the promising solutions is the decentralized optimization technique (e.g., multi-convex ADMM [45]) where the operator can individually solve the subproblems based on its local information and exchange the required information with the Central ColoMEC Orchestrator for the aggregation. Moreover, the ColoMEC system can be analyzed in a game-strategic design among co-located mobile operators, which can provide a more powerful and flexible operation mechanism.

REFERENCES


A. Convergence Analysis

The proof mainly is based on the work in [38] which shows the global convergence condition and the asymptotic convergence rate by using the assumption of the Kurdyka-Lojasiewicz property.

The general block multi-convex function is in form of

\[
\min_{x \in \mathcal{X}} F(x_1, \ldots, x_s) := f(x_1, \ldots, x_s) + \sum_{i=1}^{s} r_i(x_i)
\]

where variable \( x \) is decomposed into \( s \) blocks \( x_1, \ldots, x_s \), \( f \) is assumed to be a differentiable and block multi-convex function, \( r_i \) is the convex function for each block, and a set \( \mathcal{X} \) is a block multi-convex set. In specific, the function \( f \) is the convex function and the set \( \mathcal{X} \) is a convex set of each block \( x_i \) while other blocks are fixed. Note that the joint constraint among blocks can be modeled in \( \mathcal{X} \), and individual constraints for each block can be modeled as the indicator function \( r_i \) for each block \( x_i \). The convergence of the proximal block coordinate descent analysis requires the following assumptions

**Assumption 1.** \( F \) is continuous in \( \text{dom}(F) \) and \( \inf_{x \in \text{dom}(F)} F(x) < \infty \).

**Assumption 2.** In the analysis [38], the parameter \( L_i^{k-1} \) (i.e., \( L_i^{k-1} \equiv \rho \) in our algorithm) is chosen according to \( 0 < l_i < L_i^{k-1} < L_i \leq \infty, i = 1, \ldots, s \).

Next, we show the Kurdyka-Lojasiewicz property of a function and Kurdyka-Lojasiewicz (KL) inequality.

**Assumption 3.** A function \( \psi(x) \) satisfies the Kurdyka-Lojasiewicz (KL) property at point \( \bar{x} \in \text{dom}(\partial \psi) \) if there exists \( \theta \in (0; 1) \) such that

\[
\frac{|\psi(x) - \psi(\bar{x})|}{\text{dist}(0, \partial \psi(x))} > 0
\]

is bounded around \( \bar{x} \). In other words, in a certain neighborhood \( \mathcal{U} \) of \( \bar{x} \), there exists a function \( \phi(s) = cs^{\theta} \) for some \( c > 0 \) and \( \theta \in (0; 1) \) such that the KL inequality holds

\[
\phi\left(\frac{|\psi(x) - \psi(\bar{x})|}{\text{dist}(0, \partial \psi(x))}\right) \geq 1, \forall x \in \mathcal{U} \cap \text{dom}(\partial \psi),
\]

where \( \psi(x) \neq \psi(\bar{x}) \), \( \text{dom}(\partial \psi) := \{x | \partial \psi(x) \neq \emptyset\} \) and

\[
\text{dist}(0, \partial \psi(x)) := \min \{\|y\| | y \in \partial \psi(x)\}.
\]

In fact, this KL inequality condition can be satisfied at the critical point \( \bar{x} \) for any real analytic functions \( \psi \). In the particular case of the locally strongly convex functions \( \psi(x) \), in a neighborhood \( D \) with constant \( \mu \), for any \( x,y \in D, \forall \gamma(x) \in \partial \psi(x) \), we have

\[
\psi(y) - \psi(x) \geq \langle \gamma(x), y - x \rangle + \frac{\mu}{2} ||x - y||^2 \geq -\frac{1}{\mu} ||\gamma(x)||^2.
\]

Then, \( \mu(\psi(x) - \psi(y)) \leq \text{dist}(0, \partial \psi(x))^2 \) and \( \psi \) satisfies KL inequality with \( \phi(s) = \frac{1}{\mu} s\sqrt{\gamma} \), (i.e., \( \theta = \frac{1}{2} \)), and \( \mathcal{U} = D \cap \{x | \psi(x) \geq \psi(\bar{x})\} \).

**Assumption 4.** The initial point \( x_0 \) is sufficiently close to the critical point \( \bar{x} \), and the value of the function \( F(x_k) \) for each iteration \( k \) is always greater than the value function at \( \bar{x} \) (i.e., \( F(x_k) > F(\bar{x}) \), \( \forall k \geq 0 \)).

Therefore, with regard to the Assumption 1 - 4, the sequence \( \{x^k\} \) of the algorithm converges globally to the closest critical point \( \bar{x} \) (or stationary point) as Theorem 2.8 in [38] by proving the bounded of \( \sum_{k=0}^{\infty} ||x^k - x^{k+1}|| \). According to Theorem 2.9 in [38], the asymptotic convergence rate depends on the parameter \( \theta \) as follows

1) If \( \theta = 0 \), \{\( x^k \}\} converges to \( \bar{x} \) within a finite iterations,
2) If \( \theta \in (0, \frac{1}{2}) \), \( ||x^k - \bar{x}|| \leq C\tau^k, \forall k \geq k_0 \), with \( k_0 > 0 \), \( C > 0 \), \( \tau \in [0, 1) \),
3) If \( \theta \in (\frac{1}{2}, 1] \), \( ||x^k - \bar{x}|| \leq C\tau^{-1+\theta}/(2\theta-1), \forall k \geq k_0 \), with \( k_0 > 0 \), \( C > 0 \).

Note that in the strongly convex case with \( \theta = 1/2 \). When \( \theta = 2/3 \), we obtain the sublinear convergence rate \( O(\frac{1}{k}) \).

B. Estimation error in traffic prediction

Figure 14 shows the convergence performance the given arrival traffic estimation error. In this experiment, the estimation error is added into the real traffic for each virtual user by a normal distribution with zero mean and the variance which is different in percentage from the true values. For example, if the prediction accuracy is 100%, the variance is zero. While the prediction accuracy is 95%, then the variance is different in 5% from the real traffic load. According to different levels of estimation error, the system suffers high values of the total cost compared to the true prediction before obtaining desirable solutions that are close to the stationary ones. Consequently, these estimation errors in the traffic load prediction consequently result in longer running time to obtain the converged solutions. From our observations, with small errors in the traffic load prediction, the initialization points by using the GIS strategy can be strongly affected and far from the solutions in the true prediction and the PCMM algorithm suffers higher costs. In other extreme cases, the scaling factor \( f_j \), at edge site \( j \) can be less than the lower bound \( \frac{1}{\mu} \) due to incorrect traffic predictions. Thus, they violate the constraint (11) of the ColoMEC – MP problem and make the computation cost in the equation (2) become negative values. However, at the converged solutions, the aggregated BS loads are balanced among BSs. Thus, it significantly mitigates the effects of the incorrect prediction in the PCMM algorithm and produces almost the same solutions with the true prediction results.

Fig. 14: Estimation error effect in the convergence of our proposed method.