On joint cooperative relaying, resource allocation, and scheduling for mobile edge computing networks

Journal:	*IEEE Transactions on Communications*
Manuscript ID	TCOM-TPS-21-0848.R2
Manuscript Type:	Transactions Paper Submissions
Date Submitted by the Author:	n/a
Complete List of Authors:	Biswas, Nilanjan; Universite catholique de Louvain, Electrical Engineering; Birla Institute of Technology Wang, Zijian; Universite catholique de Louvain, Electrical Engineering Vandendorpe, Luc; Universite catholique de Louvain, Electrical Engineering Mirghasemi, Hamed; Universite catholique de Louvain, Electrical Engineering
Keyword:	Mobile communication, Communication, Optimization methods

On joint cooperative relaying, resource allocation, and scheduling for mobile edge computing networks

Nilanjan Biswas; Zijian Wang; Luc Vandendorpe, *IEEE Fellow*; Hamed Mirghasemi

Abstract

In this paper, we consider Internet of Things (IoT) based mobile edge computing (MEC) system. IoT devices are controlled by access points (APs), which do not have access to any licensed spectrum. This puts limitations on IoT devices' offloading capability and hence effectively reduces the number of computed bits. Under such scenario, we consider spectrum sharing, where IoT networks get access to licensed spectrum by helping a primary network (owns licensed spectrum) in cooperative relaying. We aim to maximize sum of the primary network's communication rate and IoT networks' total number of computed bits by jointly optimizing relay selection, licensed spectrum allocation, local computation, and offloading sequence selection. Moreover, we consider important constraints, e.g., the guaranteed rate for the primary network, IoT devices' available energy, and available licensed spectrum resource blocks (RBs). We observe that the optimization problem is combinatoric in nature, which becomes more complicated when IoT devices' scheduling order comes into the formulation. From our analysis, we devise an optimal and computationally efficient algorithm. In the result section, we compare with relevant benchmark systems and show the efficacy of our proposed model. Moreover, we also show how IoT and primary networks benefit from spectrum sharing.

Keywords

Mobile edge computing, partial offloading, spectrum sharing, relaying, bipartite matching

Nilanjan Biswas and Hamed Mirghasemi were with Universite Catholique de Louvain, Louvain la Neuve, Belgium. Zijian Wang and Luc Vandendorpe are with Universite Catholique de Louvain, Louvain la Neuve, Belgium. Emails: nilanjan.biswas058@gmail.com,{zijian.wang,luc.vandendorpe}@uclouvain.be, hamed.mirghasemi@gmail.com.

I. INTRODUCTION

In modern wireless communication systems, we find the emergence of different smart networks, e.g., smart home, smart transportation, smart healthcare, smart agriculture, etc. Internet of Things (IoT) is envisioned as a technique, which may help us in implementing smart networks. According to Cisco [1], we are going to have billions of IoT devices by the year 2022. With the advent of different computation-intensive applications in smart networks, one obvious question arises which is: how computation power and energy-constrained IoT devices would be able to support these applications? IoT devices may take help of cloud computing servers by offloading their computational tasks. However, there are some limitations in offloading to cloud servers. Distant cloud servers may increase latency and energy consumption [2].

To confront this issue, European Telecommunications Standards Institute has proposed a technology called Mobile edge computing (MEC) [3]. Unlike cloud servers, MEC servers are installed at access points (APs) at the edge of networks. IoT devices may offload their computational tasks to MEC servers and substantially get benefited, which has been discussed in [4]. One of the major issues in computational task offloading for IoT networks may be the availability of spectrum. In [5], it has been predicted that by the year 2025, we may expect 2.95 billion smart home sensors. Now, at this point, we face another important question, i.e., how can we support massive IoT networks with our existing spectrum? IoT devices may operate in unlicensed spectrum, e.g., Low Power Wide Area Networks (LPAWN) and ZigBee networks. However, only unlicensed spectrum may not be sufficient to support massive IoT networks. In this context, licensed spectrum sharing for IoT networks has gained significant attention [6].

A. Literature Survey

1) MEC for IoT networks: Typical IoT nodes may be computation power and energy constrained. This has led to adopting MEC for IoT networks [7]. In MEC networks, we find two different offloading modes, i.e., binary and partial [8]. In binary offloading, computational tasks are either offloaded for remote computation at MEC servers or computed locally. Typically, for granular computational tasks, partial offloading may be chosen, where computational tasks may be computed both at MEC servers and locally. Related literature includes [9]–[22], where authors have considered either binary or partial offloading. Further, we can divide these papers into single [11], [13], [15], [20], [21] and multiple device [9], [10], [12], [14], [16]–[19],

[22] offloading scenarios. One common thing is that authors of these papers have considered centralized optimization problem, where they have optimized different metrics, e.g., energy consumption, energy efficiency, total number of computed bits etc.. For multi-device computation at a MEC server, we find two different approaches, i.e., dividing the computation power among different devices [9], [10], [14], [17], [18], [22] and dividing effective computation duration among different devices [16], [17], [19]. Devices are allocated MEC server's processing power one after another for computation power division over time. For the second case, a MEC server acts as a virtual machine, such that, at same time different cores of the MEC server's processing unit are allocated to different devices.

2) Spectrum Sharing for IoT networks: According to [1], global mobile data traffic would rise exponentially and by the year 2022, we may expect it to be 77 Exabytes per month. For regulators, this is a matter of concern as the existing spectrum may not be sufficient to accommodate the data traffic. European Commission, in their report [5], has discussed that with help of spectrum sharing, we may reduce the spectrum requirement to 19 GHz from 76 GHz for 5G communications. There are two different approaches for spectrum sharing [23], i.e., *property rights* and *spectrum commons*. Access to licensed spectrum is received in exchange for appropriate remuneration in *property rights*. Cooperative relaying may be an effective technique for *property rights* [20]–[22], [24], [25]. In *spectrum commons*, spectrum sensing is performed to detect opportunities for spectrum sharing [26]–[28]. For spectrum sensing, licensed spectrum should be continuously monitored to find the spectrum holes. Whereas, for cooperative relaying, licensed spectrum can be used by both licensed and unlicensed networks concurrently. There may be situations (e.g., for cell-edge users), when quality of service (QoS) for wireless users in licensed spectrum may not be maintained instead of full utilization of licensed spectrum [29]. Cooperative relaying may help in improving QoS while reducing spectrum resource requirement. For massive IoT networks, which may have billions of IoT nodes, allocation of exclusive spectrum may not be feasible under the already existing spectrum crunch scenario. For IoT networks, we find [6], [28], [30]–[32], where authors have discussed the efficacy of spectrum sharing for IoT networks in radar and cellular spectrum. One can find standardization effort by the 3rd Generation Partnership Project (3GPP) on spectrum sharing for IoT networks in [33].

This paper is based on a promising marriage of MEC and spectrum sharing for IoT networks.

It is very important to properly design parameters related to licensed spectrum sharing and MEC. In that regard, we find following gaps in literature:

Literature gaps: (a) For IoT networks, in context of MEC, authors of [9]–[19] have jointly designed computation and communication parameters from a single optimization problem. However, none of them have considered spectrum sharing. (b) In context of multiple licensed spectrum RB sharing, we find [27], [28], where authors have scheduled unlicensed users in those RBs. However, we observe that in [27], RB allocation has not been considered in optimization problem. Though authors of [28] have considered spectrum allocation as a binary optimization parameter, they have later relaxed that to be continuous. Moreover, none of [27], [28] has considered MEC. (c) Though authors of [20]–[22] have considered both spectrum sharing and MEC for IoT networks, they have some limitations. [20]–[22] are based on single licensed spectrum and single IoT network. Moreover, authors of [20] have not considered joint design for parameters.

B. *Motivation and Contributions*

Existing literature gap, which we have discussed in last section, motivates us in considering a generalized system model with spectrum sharing and MEC for IoT networks and then going for joint design of spectrum sharing and MEC parameters. We consider a primary network (owns licensed spectrum) and multiple IoT networks, which try to get access to that licensed spectrum. In the primary network, there is a base station (BS), which communicates with multiple primary receivers (PRs) in multiple licensed spectrum RBs (RBs are orthogonal to each other and hence can be visualised as multiple licensed spectrum). In each IoT network, there is an AP, which controls multiple IoT devices. APs may help the primary network in cooperative relaying and as remuneration get access to the licensed spectrum.

Following, we discuss our contributions in this paper in brief, which, to the best of our knowledge, have not been considered earlier in literature:

1) Joint design of relay selection, RB allocation, local computation and offloading: The BS either chooses an AP as a relay to communicate with a PR or directly communicates in an RB. We optimally design relay selection parameters, RB allocation, along with different computation (i.e., local computation and offloading) parameters. We observe that the problem is non-trivial due to its combinatoric nature, which arises due to multiple licensed RBs' allocation. Unlike [28], we find optimal allocation algorithm for these RBs, which is computationally efficient too.

2) Optimal scheduling for offloading and computation: We consider a simplified computation model at MEC servers, where instead of dividing a MEC server's computation power among IoT devices, full computation power is allocated to IoT devices over orthogonal time durations. We find that authors of [16], [17], [19] have considered such computation model in their work. However, authors of those papers have assumed infinite computation power at MEC servers. Unlike a cloud server, a MEC server is highly constrained by computation resources. Hence infinite computation power assumption at MEC servers may not be realistic. In this paper, unlike [16], [17], [19], we consider finite computation power at MEC servers and design optimal order for IoT devices' offloading and computation.

3) Showing efficacy of spectrum sharing: In simulation, we show that both primary and IoT networks get multiple times improvement with help of spectrum sharing in the communication rate and total number of computed bits, respectively.

The rest of the paper is organized as follows. Section II describes our system model and optimization problem. In Section III, we discuss the optimal solution procedure of our optimization problem. The complexity and efficiency of our proposed optimal algorithm are discussed in Section IV. Different useful and interesting observations and their corresponding discussions are given in Section V. We conclude in Section VI.

II. System Model

In Fig. 1, we show the logical diagram of our system model. We consider a primary network and M $(1 \leq m \leq M)$ number of IoT networks. In the primary network, there is a primary base station (BS) and multiple primary receivers (PRs). It has total bandwidth of B, which is divided into N $(1 \leq n \leq N)$ equal bandwidth RBs (denoted by RB_n). The BS communicates with PRs in different RBs. Each IoT network has energy and computation power-constrained IoT devices, which are operated by an AP. A computationally powerful MEC server is installed at an AP. Due to the unavailability of any licensed spectrum for computational task offloading, IoT networks suffer from lower number of computed bits. For the primary network, we consider downlink communication, i.e., from the BS to PRs. Due to poor channel gains between the BS and PRs, received rates at PRs may be low. Under such circumstances, we consider cooperation among the primary network and IoT networks, such that, these networks can satisfy their requirements. The BS may communicate with a PR either by direct or by relay transmission. In relay transmission, the BS chooses an AP as a relay to communicate with a PR in an RB. As remuneration, the AP

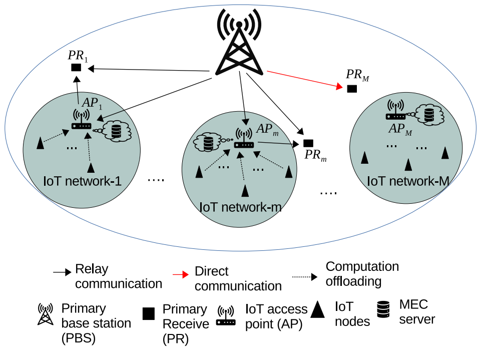


Fig. 1: Logical diagram of system model

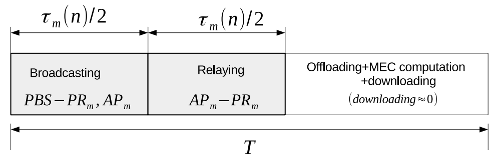


Fig. 2: Different phases based on communication and computation at the BS, PR_m, $\mathrm{IoT}_{k,m}$, and AP_m

gets access to the RB, in which it relays data. During this process, the AP maintains an effective relaying rate above a minimum requirement, which we consider in (6) later.

We denote APs by $\mathrm{AP}_m, m \in \{1, .., M\}$, and IoT devices under AP_m by $\mathrm{IoT}_{k,m}$, $1 \leq k \leq K_m$, where K_m is number of IoT devices under AP_m. Without loss of generality, we assume that there is M number of PRs, which receive either only direct transmission or both direct and relay transmission via APs from the BS. The corresponding PR, to which AP_m may relay data is predefined [1] and for the sake of notation simplicity, we denote it by PR_m. Relay mode activation and RB allocation is characterized by parameter $a_{j,m}(n) \in \{0, 1\}, j = 0, 1$, where $a_{0,m}(n) = 1$ means the BS directly communicates with PR_m in RB_n and $a_{1,m}(n) = 1$ means AP_m helps the BS in relaying to PR_m in RB_n. In Fig. 1, we show relay communications and computation offloading by black coloured solid and dashed arrow lines, respectively, and direct communications by red coloured solid lines. We have shown a scenario, where PR_1 and PR_m receive data directly from the BS and via AP_1 and AP_m, respectively. However, PR_M is receiving only direct communication from the BS, such that, IoT devices under AP_M do not get access to any RB for computational task offloading. The primary network operates in slotted time frames with frame duration of T. For direct transmission to a PR, the BS allocates the PR an RB for the time duration T. In relay transmission, the BS communicates with a PR via associated AP. In Fig. 2, we show corresponding frame structure assuming AP_m relays data to PR_m in RB_n. At the beginning of frame, the BS sends data to AP_m and PR_m for $\tau_m(n)/2$ duration, where $\tau_m(n)$

[1]We assume that the BS has prior information about locations for APs and PRs. Based on that, the BS associates a PR with an AP, which is at closest Euclidian distance to the AP. Such models have been adopted in [34], [35]. However, PRs association with APs may be considered as another design variables, which is beyond the scope of this paper.

is the total relaying time duration for PR_m in RB_n. After receiving the BS's data, AP_m relays it to PR_m for $\tau_m(n)/2$ duration. Once the relaying process is over, the BS gives access of RB_n to AP_m, in which AP_m schedules $\text{IoT}_{k,m}, \forall k$, for offloading. $\text{IoT}_{k,m}, \forall m, k$, may perform local computation from the very beginning of a frame without hampering the offloading procedure because of having separate hardware blocks for communication and computation. For a quick finding, in Table I, we list different important notations (including which are used later) and their descriptions, which we use in this paper. We consider following assumptions in our system

TABLE I: Different notations and their descriptions

Symbol	Description	Symbol	Description
N	Number of RBs	M	Number of AP, PR pairs
$B_n = B/N$	Bandwidth of $\text{RB}_n, \forall n$	R^{IoT}	Total number of computed bits for IoT networks
K_m	Number of IoTs under AP_m	$\text{IoT}_{k,m}$	k^{th} IoT under AP_m
$a_{j,m}(n) \in 0, 1, j = 0, 1$	Indicator of AP_m relays data to PR_m in RB_n (then $a_{1,m}(n) = 1$),or direct transmission to PR_m in RB_n (then $a_{0,m}(n) = 1$)	$\text{IoT}_{[k],m}(n)$	k^{th} ordered IoT scheduled for offloading in RB_n under AP_m
$\tau_m(n)$	Relaying time duration of AP_m in RB_n	p_{BS}	BS transmission power
p_{AP}	Transmission power for $\text{AP}_m, \forall m$,	$d_{\text{I}_k,\text{P}_m}$	Distance between $\text{IoT}_{k,m}$ and PR_m
$h_{\text{A}_m,\text{P}_m}(n)$	Channel gain between AP_m and PR_m in RB_n	d_{B,A_m}	Distance between BS and AP_m
$h_{\text{I}_k,\text{A}_m}(n)$	Channel gain between $\text{IoT}_{k,m}$ and AP_m in RB_n	d_{B,P_m}	Distance between BS and PR_m
$h_{\text{B},\text{A}_m}(n)$	Channel gain between BS and AP_m in RB_n	$d_{\text{A}_m,\text{P}_m}$	Distance between AP_m and PR_m
$h_{\text{B},\text{P}_m}(n)$	Channel gain between BS and PR_m in RB_n	κ	Path loss exponent
$t_{k,m}^{\text{Loc}}$	Local computation time duration for $\text{IoT}_{k,m}$	$r_{j,m}^{\text{Pri}}(n)$	Received rate at PR_m in RB_n
$r_{k,m}^{\text{Loc}}$	Number of locally computed bits at $\text{IoT}_{k,m}$	$p_{k,m}^{\text{Off}}(n)$	Offloading power for $\text{IoT}_{k,m}$ in RB_n
$t_{k,m}^{\text{Off}}(n)$	Offloading duration for $\text{IoT}_{k,m}$ in RB_n	$r_{k,m}^{\text{Off}}(n)$	Number of offloaded bits for $\text{IoT}_{k,m}$ in RB_n

model: (a) Time frame duration is less than coherence time, hence, different channel gains remain fixed during our analysis period, (b) The BS is the central controller and it knows locations of different wireless devices and instantaneous channel gain values, (c) IoT devices and PRs either transmit or receive for having single transceiver chain.

Possible application: We can say our system model is very much realistic from facts: (a) cooperative relaying is used in Long Term Evolutions (LTE) networks [36] and (b) deployment of IoT networks in LTE band is found for Narrow-Band IoT (NB-IoT) networks [33], [37]. We can consider IoT access points of our system model as relays for LTE base stations and in return IoT networks get access to LTE spectrum.

A. Primary network rate

The BS may communicate with PR_m either by direct transmission or by relay transmission via AP_m. We consider decode and forward relaying at APs. If the BS communicates with PR_m

in RB_n, then following [38], we can write received rate at PR_m as:

$$r_{j,m}^{\text{Pri}}(n)=\begin{cases}\frac{1}{2}\tau_m(n)B_n\log_2\{1+\min(\gamma_{\mathrm{B,A}_m}(n),\gamma_{\mathrm{B,P}_m+\mathrm{A}_m,\mathrm{P}_m}(n))\}, \mathrm{AP}_m \text{ relays(j=1)},\\ TB_n\log_2\left\{1+\frac{p_{\text{BS}}}{\sigma^2}\left|h_{\mathrm{B,P}_m}(n)\right|^2 d_{\mathrm{B,P}_m}{}^{-\kappa}\right\}, \text{direct transmission(j=0)},\end{cases}. \quad (1)$$

where $B_n = B/N$ is the bandwidth of an RB, $\gamma_{\mathrm{B,A}_m}(n)=\frac{p_{\text{BS}}}{\sigma^2}\left|h_{\mathrm{B,A}_m}(n)\right|^2 d_{\mathrm{B,A}_m}^{-\kappa}$, $\gamma_{\mathrm{B,P}_m+\mathrm{A}_m,\mathrm{P}_m}(n)=\frac{p_{\text{BS}}}{\sigma^2}\left|h_{\mathrm{B,P}_m}(n)\right|^2 d_{\mathrm{B,P}_m}^{-\kappa}+\frac{p_{\text{AP}}}{\sigma^2}\left|h_{\mathrm{A}_m,\mathrm{P}_m}(n)\right|^2 d_{\mathrm{A}_m,\mathrm{P}_m}^{-\kappa}$, and, p_{BS} and p_{AP} are the BS's and AP_m's transmission powers, respectively, σ^2 denotes total noise power at the receiver (we consider identical noise power at all receivers), $h_{\mathrm{B,P}_m}(n)$, $h_{\mathrm{B,A}_m}(n)$, and $h_{\mathrm{A}_m,\mathrm{P}_m}(n)$ are channel gains between $\mathrm{BS}-\mathrm{PR}_m$, $\mathrm{BS}-\mathrm{AP}_m$, and $\mathrm{AP}_m-\mathrm{PR}_m$, respectively, in RB_n. We consider block fading, such that, channel gains remain same throughout a time frame in RB_n. However, channel gains in different RBs may be different. Distances between different entities are denoted by $d_{x,y}$, where $x\in\{\mathrm{B},\mathrm{A}_1,...,\mathrm{A}_m\}, y\in\{\mathrm{A}_1,...,\mathrm{A}_m,\mathrm{P}_1,...,\mathrm{P}_m\}$ (where B, A_{1-m}, and P_{1-m} stand for BS, APs, and PRs), and κ is path loss exponent. So, the primary network's effective rate is expressed as:

$$R^{\text{Pri}}=\sum_{j=0}^{1}\sum_{m=1}^{M}\sum_{n=1}^{N}a_{j,m}(n)r_{j,m}^{\text{Pri}}(n). \quad (2)$$

B. *Total number of computed bits for IoT networks*

In IoT networks, total number of computed bits are based on computation at MEC servers (installed at APs) and local computation at IoT devices. Following we discuss about them.

1) Computation at $\mathrm{AP}_m, \forall m$: AP_m gets access to an RB if it relays data to PR_m in that RB. After that, AP_m schedules $\mathrm{IoT}_{k,m}, k=1,..,K_m$, in orthogonal time duration for partial offloading of their computational tasks. Let us consider AP_m relays data to PR_m in RB_n. We denote offloading time duration for $\mathrm{IoT}_{k,m}$ by $t_{k,m}^{\text{Off}}(n)$, such that, number of offloaded bits become:

$$t_{k,m}^{\text{Off}}(n)B_n\log_2\left(1+\frac{p_{k,m}^{\text{Off}}(n)}{\sigma^2}|h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}^{-\kappa}\right), \quad (3)$$

where $p_{k,m}^{\text{Off}}(n)$, $h_{\mathrm{I}_k,\mathrm{A}_m}(n)$, and $d_{\mathrm{I}_k,\mathrm{P}_m}$ are $\mathrm{IoT}_{k,m}$'s offloading power, channel gain, and distance, respectively, to AP_m. It is to be noted that offloading power for $\mathrm{IoT}_{k,m}$ depends on offloading time duration, which depends on the RB. That is why offloading power has also RB index. Total number of computed bits at the MEC server at AP_m becomes:

$$r_m^{\text{MEC}}=\sum_{k=1}^{K_m}\sum_{n=1}^{N}a_{1,m}(n)t_{k,m}^{\text{Off}}(n)B_n\log_2\left(1+\frac{p_{k,m}^{\text{Off}}(n)}{\sigma^2}|h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}^{-\kappa}\right). \quad (4)$$

2) Local Computation at $\mathrm{IoT}_{k,m}, \forall k, m$: We characterize the computation power of $\mathrm{IoT}_{k,m}$ by a positive parameter tuple $\langle C^{\mathrm{Loc}}_{k,m}, f^{\mathrm{Loc}}_{k,m} \rangle$, where $C^{\mathrm{Loc}}_{k,m}$ is the number of required CPU cycles to compute a bit at $\mathrm{IoT}_{k,m}$ and $f^{\mathrm{Loc}}_{k,m}$ is $\mathrm{IoT}_{k,m}$'s CPU frequency (CPU cycles/second). Total number of locally computed bits at $\mathrm{IoT}_{k,m}$ becomes:

$$r^{\mathrm{Loc}}_{k,m} = \frac{t^{\mathrm{Loc}}_{k,m} f^{\mathrm{Loc}}_{k,m}}{C^{\mathrm{Loc}}_{k,m}},$$

where $t^{\mathrm{Loc}}_{k,m}$ is the local computation time duration at $\mathrm{IoT}_{k,m}$. Therefore, we can write total number of bits, which are computed locally at IoT devices under AP_m as:

$$r^{\mathrm{Loc}}_{m} = \sum_{k=1}^{K_m} r^{\mathrm{Loc}}_{k,m}.$$

Hence, effective number of computed bits (i.e., at IoT devices and at MEC servers) becomes:

$$R^{\mathrm{IoT}} = \sum_{m=1}^{M} r^{\mathrm{MEC}}_{m} + \sum_{m=1}^{M} r^{\mathrm{Loc}}_{m}. \tag{5}$$

C. Different constraints

Following we discuss different constraints in our system model.

1) Guaranteed rate at $\mathrm{PR}_m, \forall m$, for RB sharing: The primary network allows AP_m to access an RB under guaranteed communication rate (i.e., denoted by Q_m) at PR_m via relaying. For example, if the BS allocates RB_n to PR_m, then AP_m is allowed to access RB_n only when the following constraint is satisfied:

$$a_{1,m}(n) r^{\mathrm{Pri}}_{1,m}(n) \geq a_{1,m}(n) Q_m, \forall m, n. \tag{6}$$

2) Computation completion at $\mathrm{IoT}_{k,m}$ and MEC server at $\mathrm{AP}_m, \forall k, m$: There are two different computation procedure in our system model. One at IoT devices and another one is at MEC servers at APs. Both of these computation procedures should end by time frame duration T. For local computation at $\mathrm{IoT}_{k,m}$, we get:

$$t^{\mathrm{Loc}}_{k,m} \leq T. \tag{7}$$

Now, we discuss about the computation procedure at the MEC server at AP_m. In order to compute $\mathrm{IoT}_{k,m}$'s task, the MEC server at AP_m has to become idle and also should receive the IoT device's computational task, which we have shown in Fig. 3. In Fig. 3, the MEC server at

AP_m computes an IoT device's computational task only after receiving it. AP_m maintains order among IoT devices for partial offloading and computation at the MEC server. We consider same order for both offloading and computation for IoT devices, which has been proved to be optimal in [39] in terms of minimum computation completion time. Without loss of generality, let us consider that AP_m gets access to RB_n for relaying to PR_m in that RB. We denote k-th scheduled IoT device under AP_m in RB_n by $\text{IoT}_{[k],m}(n)$. $\text{IoT}_{[k],m}(n)$ partially offloads computational task for $t^{\text{Off}}_{[k],m}(n)$ duration and MEC server at AP_m computes the task for $t^{\text{MEC}}_{[k],m}(n)$ duration. We denote its computation end time instant at the MEC server at AP_m by $I_{[k],m}(n)$. We can write

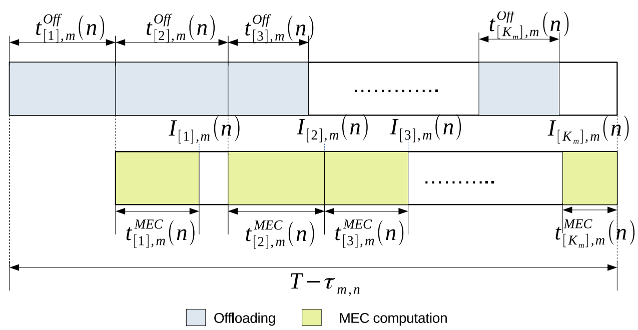


Fig. 3: Offloading and computation for m^{th} IoT network in RB_n

following assuminng $\text{IoT}_{[1],m}(n)$ starts offloading from the time instant $t = 0$:

$$I_{[1],m}(n) = t^{\text{Off}}_{[1],m}(n) + t^{\text{MEC}}_{[1],m}(n); I_{[k],m}(n) = \max\left\{\sum_{i=1}^{[k]} t^{\text{Off}}_{[i],m}(n), I_{[k-1],m}(n)\right\} + t^{\text{MEC}}_{[k],m}(n), k = 2, ., K_m, \tag{8}$$

where $t^{\text{MEC}}_{[k],m}(n) = t^{\text{Off}}_{[k],m}(n) \cdot B_n \cdot \log_2\left(1 + \frac{p^{\text{Off}}_{[k],m}(n)}{\sigma^2}|h_{\text{I}_{[k]},\text{A}_m}(n)|^2 d^{-\kappa}_{\text{I}_{[k]},\text{A}_m}\right) \cdot C^{\text{MEC}}_{[k],m}/f^{\text{MEC}}_m$, $t^{\text{Off}}_{[k],m}(n)$, $p^{\text{Off}}_{[k],m}(n)$, $h_{\text{I}_{[k]},\text{A}_m}(n)$, and $d_{\text{I}_{[k]},\text{A}_m}$ are k^{th} scheduled IoT device's offloading time duration, power, channel gain to AP_m, and distance to AP_m, respectively. f^{MEC}_m is the CPU frequency at the MEC server at AP_m, and $C^{\text{MEC}}_{[k],m}$ is required CPU to compute a bit at the MEC server for $\text{IoT}_{[k],m}(n), \forall n$. The MEC server at the AP_m can start computing $\text{IoT}_{[k],m}(n)$'s computational task only after receiving $\text{IoT}_{[k],m}(n)$'s task and finishing computation of $\text{IoT}_{[k-1],m}(n)$'s computational task. This fact is considered in (8) by considering *maximum* function. It is to be noted that we may get K_m! permutations (! is the symbol of factorial) of IoT devices' scheduling sequence at AP_m. We try to find out optimal sequence from our optimization problem. Therefore, we define following set with IoT devices according to their offloading sequence in RB_n, which is one of our optimization parameters in optimization problem $P1$ in Section III.

$$\mathcal{S}_m(n) = \{\text{IoT}_{[1],m}(n), ..., \text{IoT}_{[K_m],m}(n)\}. \tag{9}$$

Following constraint takes care of timely computation completion at the MEC server at AP$_m$:

$$\tau_m(n) + I_{[k],m}(n) \leq T, \quad \forall m, k, n. \tag{10}$$

In (10), we show that relaying and IoT$_{[k],m}(n)$'s remote computation should be completed by end of the frame duration, i.e., T, which has been shown in Fig. 3.

3) Limitation of energy at IoT$_{k,m}, \forall k, m$: We consider IoT$_{k,m}$ has finite energy at its battery, which it uses judiciously for local computation and offloading based on AP$_m$'s accessibility to an RB. Following we write energy consumption at IoT$_{k,m}$, where the first term is for local computation [12] and the other for offloading:

$$e_{k,m} = w_{k,m} t^{\text{Loc}}_{k,m} (f^{\text{Loc}}_{k,m})^3 + \sum_{n=1}^{N} a_{1,m}(n) t^{\text{Off}}_{k,m}(n) p^{\text{Off}}_{k,m}(n), \tag{11}$$

where $w_{k,m}$ is coefficient depending on chip architecture. Energy constraint at IoT$_{k,m}$ becomes:

$$e_{k,m} \leq E_{k,m}. \tag{12}$$

D. Optimization problem

In our system model, there are two different networks with two different interests. For the primary network, the BS will try to maximize the communication rate, whereas, in IoT networks, total number of computed bits is of interest while maintaining latency constraint in (7) and (10). For applications, e.g., image, video processing, number of computed bits may define quality of service. Therefore, optimization of number of computed bits is important, which has been considered in [12], [19]. We consider a single weighted objective function while considering both networks' interests, and form following optimization problem:

$$P1: \max_{\alpha, \mathbf{A}, \mathbf{P}, \mathbf{S}, \mathbf{T^r}, \mathbf{T^L}, \mathbf{T^O}} \alpha R^{\text{Pri}} + (1-\alpha) R^{\text{IoT}} \tag{13a}$$

$$\text{s.t.:} \quad (C1) \; (6), (7), (10), (12),$$

$$(C2) \; \sum_{j=0}^{1} \sum_{m=1}^{M} a_{j,m}(n) \leq 1, \forall n, \tag{13b}$$

$$(C3) \; \sum_{j=0}^{1} \sum_{n=1}^{N} a_{j,m}(n) \leq 1, \forall m, \tag{13c}$$

where $\alpha \in [0,1]$ is corresponding weight on the primary network's rate, $\mathbf{T}^{\mathrm{r}} = [\tau_m(n)], \mathbf{A} = [a_{j,m}(n)]$, $\mathbf{T}^{\mathrm{L}} = [t_{k,m}^{\mathrm{Loc}}], \mathbf{S} = [\mathcal{S}_m(n)], \mathbf{T}^{\mathrm{O}} = [t_{k,m}^{\mathrm{Off}}(n)]$, and $\mathbf{P} = [p_{k,m}^{\mathrm{Off}}(n)]$. The BS communicates with a PR using an RB either by direct transmission or via an AP as relay, which is considered in (13b). (13c) means that PR_m and $\mathrm{AP}_m, \forall m$, is allocated an RB at maximum. It can be observed that optimization problem $P1$ is combinatorial in nature. We may find the optimal solution by exhaustive search, whose complexity increases substantially for large values of M and N.

In next section, we propose an optimal and computationally efficient approach to solve $P1$, which we zest as follows:(a) We map optimization problem $P1$ with bipartite graph problem for both $N \geq M$ (Section III.A) and $N < M$ (Section III.B), (b) For $N \geq M$, we get optimization problem $P2$, which further can be divided into $P3$ and $P4$ based on whether IoT AP relays data and hence gets access to RB or not, (c) Bipartite matching is done in optimization problem $P5$ based on received values from $P3$ and $P4$, (d) Optimization problem $P3$ is non-convex; without losing optimality, we write $P3$ in form $P3.1$. Still, we observe optimization problem $P3.1$ is non-trivial due to presence of IoT nodes' scheduling order. Following Proposition 3.2 and 3.3, we simplify it and rewrite $P3.1$ in form of $P3.2$, which is convex, (e) We solve $P3.2$ optimally by solving its corresponding dual problem (Chapter 5 of [40]). Dual variables are found from (34) and using those, primal variables are found from $P3.LP$ [15].

III. Optimal Graph-Based Solution procedures

In this section, we discuss about the solution procedure for the optimization problem $P1$. We map it with bipartite graph problem to get an optimal and efficient solution. Based on number of RBs and AP, PR pairs, we may get two different cases, i.e., $N \geq M$ and $N < M$. We analyse these cases and find optimal solutions for both. Before going into further detail, in following (14) and (15), we rewrite (13a) and (11), which help us in our subsequent analysis.

$$\begin{aligned}
&\alpha R^{\mathrm{Pri}} + (1-\alpha) R^{\mathrm{IoT}} \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N} a_{1,m}(n) \left[\alpha r_{1,m}^{\mathrm{Pri}}(n) + (1-\alpha)\{ \sum_{k=1}^{K_m} t_{k,m}^{\mathrm{Off}}(n) B_n \log_2 \left(1 + \frac{p_{k,m}^{\mathrm{Off}}(n) |h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}^{-\kappa}}{\sigma^2} \right) + r_m^{\mathrm{Loc}} \} \right] + \\
&\sum_{m=1}^{M}\sum_{n=1}^{N} a_{0,m}(n) \left[\alpha r_{0,m}^{\mathrm{Pri}}(n) + (1-\alpha) r_m^{\mathrm{Loc}} \right] + (1-\alpha) \sum_{m=1}^{M} r_m^{\mathrm{Loc}} \left[1 - \sum_{n=1}^{N} (a_{0,m}(n) + a_{1,m}(n)) \right],
\end{aligned} \tag{14}$$

$$e_{k,m} = \sum_{n=1}^{N}\left[\sum_{j=0}^{1} a_{j,m}(n) w_{k,m} t^{\text{Loc}}_{k,m} (f^{\text{Loc}}_{k,m})^3 + a_{1,m}(n) t^{\text{Off}}_{k,m}(n) p^{\text{Off}}_{k,m}(n)\right] + (1 - \sum_{n=1}^{N}\sum_{j=0}^{1} a_{j,m}(n)) w_{k,m} t^{\text{Loc}}_{k,m}(f^{\text{Loc}}_{k,m})^3. \tag{15}$$

A. Analysis for $N \geq M$

For $N \geq M$, we have provision to allocate RBs to all PRs. Based on this fact, we present following proposition.

Proposition 3.1: For $N \geq M$, each PR receives an RB either for direct or relay transmission, such that, (13b) holds at equality, i.e., $\sum_{j=0}^{1}\sum_{n=1}^{N} a_{j,m}(n) = 1, \forall m$.

Proof: We prove this from the objective function of $P1$ as has been given in (14), where the last line is for the scenario when a PR does not receive direct/relay transmission. From (14), it can be observed that first two terms, which are multiplied with $a_{1,m}(n)$ and $a_{0,m}(n)$, respectively, are larger than the last term. It means that we can always increase the value for the objective function of $P1$ by allocating an RB to a PR either for direct transmission or for relay transmission. From this fact, we can conclude that as for $N \geq M$ we have provision of allocating RB to a PR either for direct or for relay transmission, i.e., $\sum_{j=0}^{1}\sum_{n=1}^{N} a_{j,m}(n) = 1, \forall m$. ∎

Under Proposition 3.1, last terms of (14) and (15) become zero, so, we can rewrite $P1$ as:

$$P2: \max_{\alpha,\mathbf{A},\mathbf{P},\mathbf{S},\mathbf{T^r},\mathbf{T^L},\mathbf{T^O}} \sum_{m=1}^{M}\sum_{n=1}^{N} a_{0,m}(n)[\alpha r^{\text{Pri}}_{0,m}(n) + (1-\alpha) r^{\text{Loc}}_{m}]$$
$$+ \sum_{m=1}^{M}\sum_{n=1}^{N} a_{1,m}(n)[\alpha r^{\text{Pri}}_{1,m}(n) + (1-\alpha)\{\sum_{k=1}^{K_m} t^{\text{Off}}_{k,m}(n) B_n \log_2\left(1 + \frac{p^{\text{Off}}_{k,m}(n)}{\sigma^2}|h_{\text{I}_k,\text{A}_m}(n)|^2 {d_{\text{I}_k,\text{P}_m}}^{-\kappa}\right) + r^{\text{Loc}}_{m}\}] \tag{16a}$$

s.t.:$(C1)$ (6), (7), (10), (13b),

$$(C2) \quad \sum_{n=1}^{N}[\sum_{j=0}^{1} a_{j,m}(n) w_{k,m} t^{\text{Loc}}_{k,m}(f^{\text{Loc}}_{k,m})^3 + a_{1,m}(n) t^{\text{Off}}_{k,m}(n) p^{\text{Off}}_{k,m}(n)] \leq E_{k,m}, \tag{16b}$$

$$(C3) \quad \sum_{j=0}^{1}\sum_{n=1}^{N} a_{j,m}(n) = 1, \forall m. \tag{16c}$$

As the optimization problem $P2$ is combinatorial in nature, it is not straightforward to solve it. Therefore, we try to find out simple ways to solve it. In that direction, we observe that one reason behind non-convexity of (16a) is multiplicative terms (i.e., associated with α). Therefore, we do our further analysis for given value of $\alpha = \bar{\alpha}$, which also helps in dividing $P2$ into two optimization problems, which we discuss below. For a given pair of m and n, we have at most one non-zero element (i.e., $a_{j,m}(n)$) due to constraints as given in (13b) and (16c). From this fact and for $\alpha = \bar{\alpha}$, we may divide $P2$ into following two optimization problems, i.e., $P3$ and $P4$, for $a_{1,m}(n) = 1$ and $a_{0,m}(n) = 1$, respectively.

$$P3: \max_{\mathbf{p_m(n)},,\mathcal{S}_m(n),\tau_m(n),\mathbf{t}^{\mathrm{L}},\mathbf{t}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n})} \bar{\alpha} r^{\text{Pri}}_{1,m}(n) + (1-\bar{\alpha})[\sum_{k=1}^{K_m} t^{\text{Off}}_{k,m}(n) B_n \log_2 \left(1 + \frac{p^{\text{Off}}_{k,m}(n)}{\sigma^2} |h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d^{-\kappa}_{\mathrm{I}_k,\mathrm{P}_m}\right) + r^{\text{Loc}}_m] \tag{17a}$$

$$\text{s.t.:}(C1)\ \ r^{\text{Pri}}_{1,m}(n) \geq Q_m, \tag{17b}$$

$$(C2)\ \ t^{\text{Loc}}_{k,m} \leq T, \forall k, \tag{17c}$$

$$(C3)\ \ \tau_m(n) + I_{[k],m}(n) \leq T, \forall k, \tag{17d}$$

$$(C4)\ \ w_{k,m} t^{\text{Loc}}_{k,m} (f^{\text{Loc}}_{k,m})^3 + t^{\text{Off}}_{k,m}(n) p^{\text{Off}}_{k,m}(n) \leq E_{k,m}, \forall k, \tag{17e}$$

where $\mathbf{t}^{\mathrm{L}} = [t^{\mathrm{L}}_{1,m}, .., t^{\mathrm{L}}_{K_m,m}]$, $\mathbf{t}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}) = [t^{\mathrm{O}}_{1,m}(n), .., t^{\mathrm{O}}_{K_m,m}(n)]$, $\mathbf{p_m}(\mathbf{n}) = [p_{1,m}(n), .., p_{K_m,m}(n)]$, and $\mathcal{S}_m(n)$ has been defined in (9). Optimal solution procedures for $P3$ is given in Section III-A1.

$$P4: \max_{\mathbf{T}^{\mathrm{L}}} \ \bar{\alpha} r^{\text{Pri}}_{0,m}(n) + (1-\bar{\alpha}) r^{\text{Loc}}_m \tag{18a}$$

$$\text{s.t.:}\ \ (C1) t^{\text{Loc}}_{k,m} \leq T, \forall k, (C2) w_{k,m} t^{\text{Loc}}_{k,m} (f^{\text{Loc}}_{k,m})^3 \leq E_{k,m}, \forall k. \tag{18b}$$

We discuss about solution steps for optimization problem $P4$ in Section III-A2.

At this point, we define following term, which is used in subsequent equations:

$$j^* = \arg \max_{j \in \{0,1\}} R_{j,m}(n), \tag{19}$$

where $R_{1,m}(n)$ and $R_{0,m}(n)$ are corresponding solutions for the above mentioned optimization problems $P3$ and $P4$, respectively.

Consequently, for given value of $\alpha = \bar{\alpha}$, we can rewrite optimization problem $P2$ in form of following optimization problem, i.e., $P5$, without loss of optimality:

$$P5: \ \max_{\mathbf{A}^*} \sum_{m=1}^{M} \sum_{n=1}^{N} a_{j^*,m}(n) R_{j^*,m}(n) \tag{20a}$$

$$\text{s.t.:}(C1) \sum_{m=1}^{M} a_{j^*,m}(n) \leq 1, \forall n; (C2) \sum_{n=1}^{N} a_{j^*,m}(n) = 1, \forall m, \tag{20b}$$

where $\mathbf{A}^* = [a_{j^*,m}(n)]$, j^* is found from (19) for given pair of m and n. From the nature of optimization problem $P5$, it can be observed that we can map $P5$ with a bipartite graph. Later, in Section III-A3, we show that the bipartite graph can be written in form of balanced bipartite graph, such that, we can apply efficient Hungarian algorithm to solve it.

In Fig. 4, we present a pictorial overview about relations between optimization problems $P2$, $P3$, $P4$, and $P5$ ($P3 - P5$ will be discussed in following Section III-A1, III-A2, and III-A3).

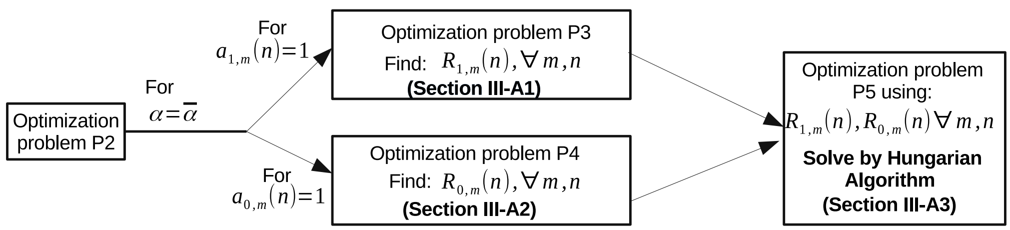


Fig. 4: Relation between optimization problems $P3$, $P4$, and $P5$

1) Solution procedure for $P3$: We observe that $P3$ is a mixed integer non-linear optimization problem, where the set for IoTs' scheduling order, i.e., $\mathcal{S}_m(n)$, is discrete, and rest parameters are continuous. We may consider different offloading orders for IoTs and solve optimization problem different times to get the optimal result, which may be computationally complex as we can order $\text{IoT}_{1,m,n}$ - $\text{IoT}_{K_m,m,n}$ in $K_m!$ ways. Therefore, we try to find an efficient way to solve optimization problem. We observe that for a given scheduling order for IoTs, $P3$ is not convex. One reason behind that is non-convexity of the term as given in (3), which has been used in $P3$. We resolve this by introducing $e^{\text{Off}}_{k,m}(n) = t^{\text{Off}}_{k,m}(n)p^{\text{Off}}_{k,m}(n), \forall k$, such that, we can rewrite number of offloaded bits by $\text{IoT}_{k,m,n}$ as:

$$t^{\text{Off}}_{k,m}(n)B_n \log_2\left(1 + \frac{e^{\text{Off}}_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{t^{\text{Off}}_{k,m}(n)\sigma^2}\right). \tag{21}$$

It is shown in [40] that as $f(x) = \log_2(1+x)$ is concave, it's perspective function, i.e., $f(x,y) = y\log_2(1+x/y)$, is also concave. Following that fact, we can say that the term in (21) is concave. However, still optimization problem $P3$ is non-convex due to the constraint as given in (17d). From the expression of $I_{[k],m}(n)$ (in (8)), it can be observed that there is a term $t^{\text{MEC}}_{[k],m}(n)$, which depends on (21) and is concave according to our previous discussion. Therefore, we are getting concave function less than a constant in constraint as given in (17d), which may make the set over optimization variables discontinuous and hence non-convex. Following [11], [13], [15], [41], we resolve this problem by replacing the term in (21) with an auxiliary optimization parameter , i.e., $r^{\text{Off}}_{k,m}(n)$. As we consider the auxiliary optimization parameter, we need to consider following constraint in (22). It is to be noted that constraint in (22) holds equality in optimal case, as we can increase the value for objective function in (23a) by increasing $r^{\text{Off}}_{k,m}(n)$.:

$$r^{\text{Off}}_{k,m}(n) \leq t^{\text{Off}}_{k,m}(n) B_n \log_2 \left(1 + \frac{e^{\text{Off}}_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{t^{\text{Off}}_{k,m}(n)\sigma^2}\right). \tag{22}$$

We write modified optimization problem as:

$$P3.1: \max_{\mathcal{S}_m(n),\tau_m(n),\mathbf{t}^{\text{L}},\mathbf{r}^{\text{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{t}^{\text{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{e}^{\text{O}}_{\mathbf{m}}(\mathbf{n})} \alpha r^{\text{Pri}}_{1,m}(n) + (1-\alpha)[\sum_{k=1}^{K_m} r^{\text{Off}}_{k,m}(n) + r^{\text{Loc}}_m] \tag{23a}$$

$$\text{s.t.: } (C1)(17\text{b}) - (17\text{d}), (22); (C2)\;\; w_{k,m} t^{\text{Loc}}_{k,m} (f^{\text{Loc}}_{k,m})^3 + e^{\text{Off}}_{k,m}(n) \leq E_{k,m}. \tag{23b}$$

where $\mathbf{r}^{\text{O}}_{\mathbf{m}}(\mathbf{n}) = [r^{\text{Off}}_{1,m}(n), .., r^{\text{Off}}_{k,m}(n)]$, $\mathbf{e}^{\text{O}}_{\mathbf{m}}(\mathbf{n}) = [e^{\text{Off}}_{1,m}(n), .., e^{\text{Off}}_{k,m}(n)]$. Now, we try to find out the order for IoTs for their offloading and computation at the MEC server, i.e., $\mathcal{S}_m(n)$. In that direction, we present following proposition.

Proposition 3.2: For $f^{\text{Loc}}_{k,m} = f^{\text{L}}_m, w_{k,m} = w^{\text{L}}_m, C^{\text{Loc}}_{k,m} = C^{\text{IoT}}_m$, and $C^{\text{MEC}}_{[k],m} = C^{\text{M}}_m$, $\forall k$, IoT$_{p,m}$ should be scheduled before IoT$_{q,m}$ under following condition:

$$E_{p,m}|h^n_{p,m}|^2 d^{-\kappa}_{p,m} > E_{q,m}|h^n_{q,m}|^2 d^{-\kappa}_{q,m}.$$

Proof: Proof is given in Appendix A. ■

Now, we try to simplify the constraint as given in (17d) and present following proposition:

Proposition 3.3: In optimal case, we can consider following two equations in place of (17d):

$$t^{\text{Off}}_{[k],m}(n) \geq t^{\text{MEC}}_{[k-1],m}(n), k = 2, .., K_m, \tag{24a}$$

$$\tau_m(n) + \sum_{k=1}^{p} t^{\text{Off}(n)}_{[k],m} + t^{\text{MEC}}_{[p],m}(n) \leq T, p = 1, .., K_m. \tag{24b}$$

Proof: Proof is given in Appendix B. ■

Following Proposition 3.2 and 3.3, we can rewrite optimization problem $P3.1$ as follows:

$$P3.2: \max_{\tau_m(n),\mathbf{T}^{\text{L}},\mathbf{r}^{\text{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{t}^{\text{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{e}^{\text{O}}_{\mathbf{m}}(\mathbf{n})} \alpha r^{\text{Pri}}_{1,m}(n) + (1-\alpha)[\sum_{k=1}^{K_m} r^{\text{Off}}_{k,m}(n) + r^{\text{Loc}}_m] \tag{25a}$$

$$\text{s.t.: } (17\text{b}), (17\text{c}), (21), (23\text{b}), (24\text{a}), (24\text{b}). \tag{25b}$$

$$L(\tau_m(n),\mathbf{t}^{\mathrm{L}},\mathbf{r}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{t}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{e}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta},\boldsymbol{\beta}) = \alpha r^{\mathrm{Pri}}_{1,m}(n) + (1-\alpha)\sum_{k=1}^{K_m}[r^{\mathrm{Off}}_{k,m}(n) + r^{\mathrm{Loc}}_{k,m}] - \sum_{k=1}^{K_m}\lambda_k(t^{\mathrm{Loc}}_{k,m} - T) - \sum_{k=1}^{K_m}\eta_k\left\{r^{\mathrm{Off}}_{k,m}(n) - t^{\mathrm{Off}}_{k,m}(n)B_n\log_2\left(1+\frac{e^{\mathrm{Off}}_{k,m}(n)|h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}{}^{-\kappa}}{t^{\mathrm{Off}}_{k,m}(n)\sigma^2}\right)\right\} - \sum_{k=1}^{K_m}\mu_k\left\{w_{k,m}t^{\mathrm{Loc}}_{k,m}(f^{\mathrm{Loc}}_{k,m})^3 + e^{\mathrm{Off}}_{k,m}(n) - E_{k,m}\right\} - \sum_{k=2}^{K_m}\zeta_k(t^{\mathrm{MEC}}_{k-1,m}(n) - t^{\mathrm{Off}}_{k,m}(n)) - \sum_{k=1}^{K_m}\beta_k\left\{\tau_m(n) + \sum_{p=1}^{k}t^{\mathrm{Off}}_{p,m}(n) + t^{\mathrm{MEC}}_{k,m}(n) - T\right\}. \tag{26}$$

We observe that optimization problem $P3.2$ is convex. Hence, we can solve the dual problem of $P3.2$ to get the optimal result as has been discussed in Chapter 5 of [40]. In order to make it simple in terms of notation, without loss of generality, we consider that $\mathrm{IoT}_{k,m}$ is scheduled at k^{th} order for offloading and computing at AP_m. Corresponding Lagrangian for $P3.2$ is given in (26), where $\boldsymbol{\eta} = [\eta_1,..,\eta_{K_m}]$, $\boldsymbol{\lambda} = [\lambda_1,..,\lambda_{K_m}]$, $\boldsymbol{\mu} = [\mu_1,..,\mu_{K_m}]$, $\boldsymbol{\zeta} = [\zeta_2,..,\zeta_{K_m}]$, and $\boldsymbol{\beta} = [\beta_1,..,\beta_{K_m}]$, are Lagrange multipliers for constraints as given in (17c), (21),(23b),(24a),(24b). We can write the dual function $g(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta},\boldsymbol{\beta})$ for $P3.2$ as follows:

$$\max_{\tau_m(n),\mathbf{t}^{\mathrm{L}},\mathbf{r}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{t}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{e}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n})} L(\tau_m(n),\mathbf{t}^{\mathrm{L}},\mathbf{r}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{t}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{e}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta},\boldsymbol{\beta}); \text{s.t.: (17b)}.$$

We find optimal values by solving following dual problem:

$$\max_{\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta},\boldsymbol{\beta}} g(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\beta}), \text{s.t.: } \boldsymbol{\eta}\succcurlyeq 0, \boldsymbol{\lambda}\succcurlyeq 0, \boldsymbol{\mu}\succcurlyeq 0, \boldsymbol{\zeta}\succcurlyeq 0, \boldsymbol{\beta}\succcurlyeq 0, \tag{27}$$

where $\boldsymbol{x}\succcurlyeq 0$ means all elements of vector $\boldsymbol{x}$ is greater equal to zero. For given values of $\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta}$, and $\boldsymbol{\beta}$, We can decompose optimization problem in (27) into following sub-problems, where $\tau^{\mathrm{lb}}_m(n) = Q_m/[\frac{B_n}{2}\log_2\{1+\min(\gamma_{\mathrm{B},\mathrm{A}_m}(n),\gamma_{\mathrm{B},\mathrm{P}_m+\mathrm{A}_m,\mathrm{P}_m}(n))\}]$ is evaluated from (17b):

$$\max_{\tau_m(n)} \quad \alpha r^{\mathrm{Pri}}_{1,m}(n) - \sum_{k=1}^{K_m}\beta_k\tau_m(n); \text{s.t.: } \tau_m(n)\geq\tau^{\mathrm{lb}}_m(n) \tag{28a}$$

$$\max_{\mathbf{t}^{\mathrm{L}}}(1-\alpha)\sum_{k=1}^{K_m}r^{\mathrm{Loc}}_{k,m} - \sum_{k=1}^{K_m}\lambda_k t^{\mathrm{Loc}}_{k,m} - \sum_{k=1}^{K_m}\mu_k w_{k,m}t^{\mathrm{Loc}}_{k,m}(f^{\mathrm{Loc}}_{k,m})^3; \text{s.t.: } 0\leq t^{\mathrm{Loc}}_{k,m}\leq T, \tag{28b}$$

$$\max_{\mathbf{t}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{e}^{\mathrm{O}}_{\mathbf{m}}(\mathbf{n})}\sum_{k=1}^{K_m}\eta_k t^{\mathrm{Off}}_{k,m}(n)B_n\log_2\left(1+\frac{e^{\mathrm{Off}}_{k,m}(n)|h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}{}^{-\kappa}}{t^{\mathrm{Off}}_{k,m}(n)\sigma^2}\right) - \sum_{k=1}^{K_m}\mu_k e^{\mathrm{Off}}_{k,m}(n) + \sum_{k=1}^{K_m}\zeta_k t^{\mathrm{Off}}_{k,m}(n) - \sum_{k=1}^{K_m}\left\{\beta_k\sum_{p=1}^{k}t^{\mathrm{Off}}_{p,m}(n)\right\}; \text{s.t.: } \zeta_1 = 0, \tau^{\mathrm{lb}}_m(n)\leq t^{\mathrm{Off}}_{k,m}(n)\leq T, 0\leq e^{\mathrm{Off}}_{k,m}(n)\leq E_{k,m}, \tag{28c}$$

$$\max_{\mathbf{r_m^O(n)}} \quad (1-\alpha)\sum_{k=1}^{K_m} r_{k,m}^{\text{Off}}(n) - \sum_{k=1}^{K_m} \eta_k r_{k,m}^{\text{Off}}(n) - \sum_{k=2}^{K_m} \zeta_k t_{k-1,m}^{\text{MEC}}(n) - \sum_{k=1}^{K_m} \beta_k t_{k,m}^{\text{MEC}}(n),$$
$$\text{s.t.:} r_{k,m}^{\text{Off}}(n) \le TB_n \log_2\left(1 + \frac{E_{k,m}|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{T\sigma^2}\right), \tag{28d}$$

Next, we find optimal solutions for $\tau_m(n), \mathbf{t}^{\text{L}}, \mathbf{r_m^O(n)}, \mathbf{t_m^O(n)}$, and $\mathbf{e_m^O(n)}$ from Karush-Kuhn-Tucker's conditions, which we present in following propositions.

Proposition 3.4: Optimal solution for $t_{k,m}^{\text{Off}}(n)$ and $e_{k,m}^{\text{Off}}(n)$ from (28c) are:

$$t_{k,m}^{\text{Off}}(n) = \begin{cases} T - \tau_m^{\text{lb}}(n), & d_{k,m} < 0 \\ \left[0, T - \tau_m^{\text{lb}}(n)\right], & d_{k,m} = 0 \\ 0, & d_{k,m} > 0 \end{cases} \tag{29}$$

$$e_{k,m}^{\text{Off}}(n) = \rho_{k,m}(n) t_{k,m}^{\text{Off}}(n), \tag{30}$$

where, $\rho_{k,m}(n) = \left[\frac{\eta_k B_n}{\mu_k \ln 2} - \frac{\sigma^2}{|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}\right]^+$, $[x]^+ = \max\{0, x\}$, $d_{k,m} = \sum_{i=k}^{K_m} \beta_i - \zeta_k + \frac{\eta_k B_n \rho_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{\ln(2)(\sigma^2 + \rho_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa})} - \eta_k B_n \log_2\left(1 + \frac{\rho_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{\sigma^2}\right)$.

Proof: Proof is given in Appendix C. ∎

We follow same procedures like Proposition 3.4 for proving following Proposition 3.5, 3.6, and 3.7, which we have excluded due to brevity.

Proposition 3.5: Optimal solution for $\tau_m(n)$ from (28a) is:

$$\tau_m(n) = \begin{cases} T, & f < 0 \\ \left[\tau_m^{\text{lb}}(n), T\right], & f = 0 \\ \tau_m^{\text{lb}}(n), & f > 0 \end{cases} \tag{31}$$

where $f = \sum_{i=1}^{K_m} \beta_i - (1-\alpha)\frac{B_n}{2}\log_2\{1 + \min(\gamma_{\text{B},\text{A}_m}(n), \gamma_{\text{B},\text{P}_m+\text{A}_m,\text{P}_m}(n))\}$.

Proposition 3.6: Optimal solution for $t_{k,m}^{\text{Loc}}$ from (28b) is:

$$t_{k,m}^{\text{Loc}} = \begin{cases} T, & u_k < 0 \\ [0, T], & u_k = 0 \\ 0, & u_k > 0 \end{cases} \tag{32}$$

where, $u_k = \lambda_k + \mu_k w_{k,m}(f_{k,m}^{\text{Loc}})^3 - (1-\alpha) f_{k,m}^{\text{Loc}}/C_{k,m}^{\text{Loc}}$.

Proposition 3.7: We consider $\zeta_{K_m+1}=0$ and find optimal solution for $r^{\text{Off}}_{k,m}(n)$ from (28d) as:

$$r^{\text{Off}}_{k,m}(n)=\begin{cases} t^{\text{Off}}_{k,m}(n)B_n\log_2\left(1+\frac{e^{\text{Off}}_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{t^{\text{Off}}_{k,m}(n)\sigma^2}\right), v_k<0 \\ \left[0, t^{\text{Off}}_{k,m}(n)B_n\log_2\left(1+\frac{e^{\text{Off}}_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{t^{\text{Off}}_{k,m}(n)\sigma^2}\right)\right], v_k=0 \\ 0,\;\; v_k>0, \end{cases} \tag{33}$$

where $v_k=\beta_k\frac{C^{\text{MEC}}_{k,m}}{f^{\text{MEC}}_m}+\eta_k-(1-\alpha)-\zeta_{k+1}\frac{C^{\text{MEC}}_{k,m}}{f^{\text{MEC}}_m}$.

With optimal primal variables defined in Proposition 3.4 to 3.7 for given $(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta},\boldsymbol{\beta})$, we can evaluate optimal dual variables with help of ellipsoid method. We use sub-gradients for $(\eta_k,\lambda_k,\mu_k,\zeta_k,\beta_k)$ in ellipsoid method as follows:

$$\left\{\begin{array}{l} r^{\text{Off}}_{k,m}(n)-t^{\text{Off}}_{k,m}(n)B_n\log_2\left(1+\frac{e^{\text{Off}}_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{t^{\text{Off}}_{k,m}(n)\sigma^2}\right), t^{\text{Loc}}_{k,m}-T, w_{k,m}t^{\text{Loc}}_{k,m}(f^{\text{Loc}}_{k,m})^3 \\ e^{\text{Off}}_{k,m}(n)-E_{k,m}, t^{\text{MEC}}_{k-1,m}(n)-t^{\text{Off}}_{k,m}(n), \tau_m(n)+\sum_{p=1}^{k}t^{\text{Off}}_{p,m}(n)+t^{\text{MEC}}_{k,m}(n)-T \end{array}\right\}. \tag{34}$$

From ellipsoid method, we receive optimal dual variables, i.e., $(\boldsymbol{\eta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*,\boldsymbol{\zeta}^*,\boldsymbol{\beta}^*)$, from which, we evaluate optimal $\boldsymbol{\rho^*_m(n)}=\left\{\rho^*_{k,m}(n)\right\}_{k=1}^{K_m}$ and then solve following linear programming:

$$P3.LP:\max_{\tau_m(n),\mathbf{t}^{\text{L}},\mathbf{r}^{\text{O}}_{\mathbf{m}}(\mathbf{n}),\mathbf{t}^{\text{O}}_{\mathbf{m}}(\mathbf{n})}\quad \alpha r^{\text{Primary}}_{k,n,l}+(1-\alpha)\left[\sum_{k=1}^{K_m}r^{\text{Off}}_{k,m}(n)+r^{\text{Loc}}_m\right] \tag{35a}$$

s.t.:$(C1)$(17c), (24a), (24b),

$$(C2)r^{\text{Off}}_{k,m}(n)\le t^{\text{Off}}_{k,m}(n)B_n\log_2\left(1+\frac{\rho^*_{k,m}(n)|h_{\text{I}_k,\text{A}_m}(n)|^2 d_{\text{I}_k,\text{P}_m}{}^{-\kappa}}{\sigma^2}\right),\forall k, \tag{35b}$$

$$(C3)\frac{\tau_m(n)B_n}{2}\log_2\{1+\min(\gamma_{\text{B},\text{A}_m}(n),\gamma_{\text{B},\text{P}_m+\text{A}_m,\text{P}_m}(n))\}\ge Q_m, \tag{35c}$$

$$(C4)w_{k,m}t^{\text{Loc}}_{k,m}(f^{\text{Loc}}_{k,m})^3+\rho^*_{k,m}(n)\le E_{k,m}. \tag{35d}$$

In following algorithm, we show different solution steps to solve optimization problem $P3.2$:

Algorithm 1: Optimal algorithm for solving $P3.2$

1 Initialize: $\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta},\boldsymbol{\beta}$;
2 **repeat**
3 Evaluate $\tau_m(n)$ and $r^{\text{Off}}_{k,m}(n), t^{\text{Off}}_{k,m}(n), e^{\text{Off}}_{k,m}(n), t^{\text{Loc}}_{k,m}, \forall k$ from Proposition 3.4 to 3.7;
4 Calculate subgradient from (34), and update $\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\eta}$ using ellipsoid method;
until *Until* $\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\zeta},\boldsymbol{\beta}$ *converge*;
5 **Result**: Optimal values: $\rho^*_{k,m}(n)$, $\tau^*_m(n)$, $r^{\text{Off}*}_{k,m}(n), t^{\text{Off}*}_{k,m}(n), e^{\text{Off}*}_{k,m}(n), t^{\text{Loc}*}_{k,m}, \forall k$.

2) Solution procedure for $P4$ in (18a)-(18b)*:* For $a_{0,m}(n) = 1$, PR_m receives direct transmission from the BS in RB_n, to which AP_m does not get access. It can be observed that, there is no parameter to be optimized for direct transmission for the primary network. Only, we need to find out $\mathrm{IoT}_{k,m}$'s, $\forall k$, local computation durations. From (18a)-(18b), we can conclude that in optimal case, we get $t^{\mathrm{Loc}}_{k,m} = \min(T, \frac{C^{\mathrm{Loc}}_{k,m} E_{k,m}}{w_{k,m}(f^{\mathrm{Loc}}_{k,m})^3})$.

3) Mapping $P5$ in (20a)-(20b) *with a balanced bipartite graph:* In this section, we discuss about the solution procedure of optimization problem $P5$, from where we get optimal mapping between RBs and different AP, PR pairs. We observe that structure of $P5$ matches with maximum weighted bipartite matching (MWBM) problem in the domain of *bipartite graph* [42], which can be solved using computationally efficient Hungarian algorithm for balanced partite graph [2].

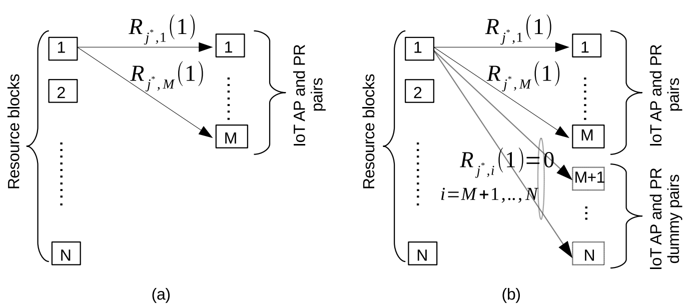


Fig. 5: Bipartites showing RBs and AP, PR pairs mapping for $(N > M)$ (j^* is solved from (19)) (a) unbalanced (b) balanced after adding dummy AP, PR pairs

In optimization problem $P5$, we can construct two partites with RBs and AP, PR pairs, which is shown in Fig. 5(a). $R_{j^*,m,1}, m = 1, .., M$, are corresponding weights for the mapping between RB_1 and AP_m, PR_m pair for j^*, which can be solved from (19). Following this same procedure, we can map other RBs (i.e., $n = 2, .., N$) with AP, PR pairs, which we have not shown in Fig. 5 to keep it simple. For $N > M$, we include dummy AP, PR pairs, i.e., blocks $M+1$ to N on the right side of Fig. 5(b). However, these dummy pairs do not exist in real, such that, they do not have any effect on optimization problem. We consider corresponding weights for dummy pairs as zero. In general, we can write different weights in Fig. 5(b) as given below for $v \in \{1, .., N\}$:

$$R_{j^*,u}(v) \begin{cases} \neq 0, u \in \{1, .., M\}, \\ = 0, u \in \{M+1, .., N\}. \end{cases}$$

[2]In a balanced bipartite graph, both partites have identical number of vertices.

After incorporating dummy nodes, we get identical number of RBs and AP, PR pairs, such that, we can map our problem to balanced bipartite graph and solve using Hungarian algorithm [43].

In following Algorithm 2, we show optimal steps for solving $P2$ for $N \geq M$.

Algorithm 2: Optimal algorithm for solving $P2$ for $N \geq M$

```
Divide α ∈ [0, 1] into L grids, i.e., α = {0, 1/L, 2/L, .., 1};
for l=1:L+1 do
    ᾱ = α(l);
    for m=1:M do
        for n=1:N do
            evaluate R_{1,m}(n) and R_{0,m}(n) by solving P3 and P4, respectively;
            Find j* from (19)
    Find value for utility function following Section III-A3
Choose optimal value for the objective function in P2 and hence α, for which we get it;
Result: Optimal solution for P2 for N ≥ M
```

B. Analysis for $N < M$

For $N < M$, all PRs and hence all APs would not get access to RBs, i.e., for given RB_n and PR_m, AP_m pair, we may get $a_{0,m}(n) = a_{1,m}(n) = 0$. Therefore, unlike for $N \geq M$, the last term in (14) will be there for $N < M$, which corresponds to non-allotment of RBs. In other words, we can say that, for given association between RBs and AP, PR pairs (i.e., $\mathbf{A}$), we get: Total utility = utility for alloted RBs + utility for without allotment of RBs. From (14), it can be observed that we always get higher utility value for first two cases (i.e., for allotment of RBs) than the last one (i.e., without allotment of RBs). It means that in optimal case, we should allot all RBs for $N < M$, i.e., $\sum_{m=1}^{M} \sum_{j=0}^{1} a_{j,m}(n) = 1, \forall n$. We can do this allotment by performing same mapping procedure between RBs and AP, PR pairs, which we did for $N \geq M$ in Section III-A. Following, we write RBs and AP, PR pairs mapping optimization problem for $N < M$:

$$P6: \max_{\mathbf{A}^*} \sum_{m=1}^{M} \sum_{n=1}^{N} a_{j^*,m}(n) R_{j^*,m}(n); \text{s.t.:} (C1) \sum_{m=1}^{M} a_{j^*,m}(n) = 1, \forall n; (C2) \sum_{n=1}^{N} a_{j^*,m}(n) \leq 1, \forall m,$$

where all terms have been discussed in Section III-A. We follow Section III-A3 to translate $P6$ in form of balanced bipartite graph problem (by adding dummy RBs). Due to brevity, we have not shown the graphical representation for bipartite matching for $N < M$ in this paper. After translating $P6$ into balanced bipartite graph problem, we can solve it using Hungarian algorithm.

After solving $P6$, we get AP, PR pairs, which get access to RBs. Rest AP, PR pairs do not get access to RBs. Corresponding IoT devices under the AP (whose associated PR does not get access to any RB), perform local computation only. In Algorithm 3, we represent our findings.

Algorithm 3: Optimal algorithm for solving $P2$ for $N < M$

1 Divide $\alpha \in [0,1]$ into L grids, i.e., $\alpha = \{0, 1/L, 2/L, .., 1\}$;
2 **for** *l=1:L+1* **do**
3   Follow step-3 to 7 of Algorithm 2;
4   Follow Section III-A3 for optimal mappings between RBs and AP, PR pairs using Hungarian algorithm (dummy RBs to be added to create balanced bipartite);
5   Find set for PRs' indexes (which are APs' indexes too), i.e., $\mathcal{M} = \{m | a_{j^*,m}(n) = 0\}$, which do not get access to any RB;
6   IoT$_{k,m}, \forall k$, under AP$_m, m \in \mathcal{M}$, perform local computation only, whose durations can be found from step-4 of this algorithm. Evaluate value for utility function;
7 Choose optimal value for the objective function in $P2$ and hence α, for which we get it;
8 **Result:** Optimal solution for $P2$ for $N < M$

Discussion on multiple RB allocation: In our system model, we consider single transceiver chain at IoT devices and PRs. We have considered that PRs and IoT devices access RBs following TDMA, and hence, they may not be able to access multiple RBs due to time synchronization problem. Therefore, allocation of multiple RBs to an AP, PR pair may not be relevant in our system model. However, our system model can be easily extended for multiple transceiver chains at PRs and IoT devices to make suitable for multiple RBs allocation. If we keep provision for allocating up-to N RBs, then the constraint as given in (13c) becomes $\sum_{j=0}^{1}\sum_{n=1}^{N} a_{j,m}(n) \leq N$, for which, we may not apply our so far discussed solution procedure. If we want to allocate up-to N number of RBs to an AP, PR pair, then without loss of optimality, we may consider N number of copies for that pair while solving the optimization problem. Therefore, modified set for AP, PR pairs become $\mathcal{M} = \{1, 1_{d,1}, 1_{d,2}, .., 1_{d,N-1}, ..., M, M_{d,1}, M_{d,2}, .., M_{d,N-1}\}$, where for the pair AP$_M$, PR$_M$, we have considered dummy nodes $M_{d,1}$ to $M_{d,N-1}$ to make N number of copies for the pair. Now, constraints as given in (13b) and (13c), will become $\sum_{j=0}^{1}\sum_{m\in\mathcal{M}} a_{j,m}(n) \leq 1, \forall n$ and $\sum_{j=0}^{1}\sum_{n=1}^{N} a_{j,m}(n) \leq 1, \forall m \in \mathcal{M}$, respectively. After this operation, we get $M.N$ number of AP, PR pairs. Therefore, we need to append $N(M-1)$ number of dummy RBs to make both partites balanced to use Hungarian algorithm, which we have discussed earlier. Due to brevity, we have not included detailed discussion on this. One may refer to [44], [45] in this regard.

IV. Discussion on computational complexity

In this section, we will discuss computational complexity for solving the problems for $N \geq M$ and $N < M$, which have been discussed in Algorithm 2 and 3, respectively. At first, we discuss about the scenario for $N \geq M$, which is solved by Algorithm 2. It can be observed that there are three different steps, i.e., step-6 (solves $P3$ and $P4$), 7 (finds j^*), and 8 (solves $P5$), which mainly contributes behind the complexity for Algorithm 2. Following we discuss about them:

1) In step-6 of Algorithm 2, we find values for $R_{1,m}(n)$ and $R_{0,m}(n)$. Value for $R_{1,m}(n)$ is found from optimization problem $P3$, which is rewritten in form of optimization problem $P3.2$. In Algorithm 1, we show steps for solving $P3.2$. Complexity for step-3 and setp-4 of Algorithm 1 are $O(4K_m+1)$ and $O(5K_m)$, respectively. So effective complexity can be written as $O(5K_m+1)$. After considering the convergence order of ellipsoid method from [46], we can say effective complexity of step-2 to 4 of Algorithm 1 is $O(K_m^3 \log \frac{R}{r})$, where R and r are parameters of ellipsoid algorithm. In step-5 of Algorithm 1, we solve a linear programming. Complexity of solving the linear programming is $O(\frac{K_m^3}{\ln K_m})$ [47]. Now, as step-3 of Algorithm 2 runs for MN times, we can write effective complexity for finding $R_{1,m}(n)$ at step-3 of Algorithm 2 becomes $O(MN(\sum_{m=1}^{M} K_m^4 \log \frac{R}{r} + \frac{K_m^3}{\ln K_m}))$. We solve $R_{0,m}(n)$ following Section III-A2, which would have effective complexity of $O(MN\sum_{m=1}^{M} K_m)$. Therefore we can write the total complexity of step-3 of Algorithm 2 as $O(MN(\sum_{m=1}^{M} K_m^4 \log \frac{R}{r} + \frac{K_m^3}{\ln K_m}) + MN\sum_{m=1}^{M} K_m)$.

2) In step-7 of Algorithm 2, we find whether we should do direct transmission or relay transmission. As there are just two possibilities, the complexity of this step is $O(MN)$.

3) In step-8 of Algorithm 2, we solve one assignment problem using Hungarian algorithm. Following [43] we can find the complexity of step-8 of Algorithm 2 as $O(\max(N,M)^3)$.

Now, it can be observed that step-6 to 8 of $P2$ are run for L number of times based on different values for α. Therefore, effectively, we can write complexity for $P2$ is $O[L\{MN(\sum_{m=1}^{M} K_m^4 \log \frac{R}{r} + \frac{K_m^3}{\ln K_m}) + MN\sum_{m=1}^{M} K_m + MN + \max(N,M)^3)\}]$, which is polynomial in time and far efficient compared to exhaustive search. It can be observed that except step-5 and 6 of Algorithm 3, all other steps of Algorithm 3 match with Algorithm 2, which do not increase complexity for Algorithm 3 separately. Hence, we conclude that for $N < M$ also complexity remains same.

V. Results and discussions

In this section, we show our proposed model's efficacy from simulation results. We show different parameters' values in Table II, which we have considered. As IoT devices may be deployed in LTE spectrum [33], we have considered the spectrum bandwidth according to LTE spectrum bandwidth. Among other parameters, path loss exponent $\kappa = 3$, distances between

TABLE II: Different parameters' values

Parameters	Values	Parameters	Values
Time frame duration (T)	1 ms.	Coefficient related to hardware architecture (w_k)	10^{-28}
Total system bandwidth (B)	3 MHz.	Receiver noise power spectral density (σ^2)	-174 dBm/Hz
IoT$_{k,m}$'s CPU frequency ($f^{\text{Loc}}_{k,m}$), $\forall k, m$	10^{10} Hz.	BS transmission power (p_{BS})	43 dBm
Required CPU cycles/bit ($C^{\text{Loc}}_{k,m}$),$\forall k, m$	10^6 CPU cycles/bit	AP transmission power (p_{AP})	30 dBm

the BS to PRs, the BS to APs, APs to PRs, and APs to IoT devices are distributed uniformly in between 200 to 300 meters, 70 to 80 meters, 50 to 60 meters, and 20 to 30 meters, respectively. We assume that different channels in our system model follows complex Normal distribution, such that, we can say that $|h_{\text{A}_m-\text{P}_m}(n)|^2, |h_{\text{I}_k-\text{A}_m}(n)|^2, |h_{\text{B}-\text{A}_m}(n)|^2$, and $|h_{\text{B}-\text{P}_m}(n)|^2$ follow exponential distribution. We assume unity mean exponential distribution for all channel gains. We consider $\alpha = 0.5$ for Fig. 7, 8, 9, and 10.

We show a comparative result in Section V-A. Except that, we also compare with following two baselines to show our proposed model's efficacy:

1) *No spectrum sharing (NSS):* To show efficacy of spectrum sharing, we analyse the framework for NSS. Under NSS, the BS directly communicates with PRs and IoT devices perform local computing only for unavailability of spectrum for offloading. Corresponding optimization problem for the primary network under NSS becomes:

$$NSS : \max_{\mathbf{B}} \sum_{m=1}^{M}\sum_{n=1}^{N} b_{m,n} r^{\text{Pri}}_{0,m}(n); \text{s.t.:}(C1) \sum_{n=1}^{N} b_{m,n} \leq 1, \forall m, (C2) \sum_{m=1}^{M} b_{m,n} \leq 1, \forall n,$$

where $\mathbf{B} = [b_{m,n}]$, $b_{m,n} \in \{0, 1\}$, is the indicator of RB$_n$'s allocation to PR$_m$ (i.e., $b_{m,n} = 1$). We can map the above optimization problem with balanced bipartite graph and solve using Hungarian algorithm. We define the primary network's gain from spectrum sharing as below, where the numerator and the denominator are found by solving $P2$ and NSS, respectively:

$$\frac{\text{Primary network's total rate with spectrum sharing}}{\text{Primary network's total rate without spectrum sharing}} \tag{37}$$

Similarly, for IoT networks, gain from spectrum sharing is defined as:

$$\frac{\text{Total number of computed bits with spectrum sharing}}{\text{Total number of computed bits without spectrum sharing}} \tag{38}$$

where the numerator and the denominator can be found from $P2$ and $P4$, respectively.

2) *Non-joint (NJ) optimization:* In NJ optimization, relaying and computation parameters are

found from two different optimization problems. Related literature includes [20], where authors have considered NJ optimization problem in their system model context. Following, in $NJ.P1$ and $NJ.P2$, we discuss two optimization problems for primary network's rate and IoT networks' total computed number of bits maximization, while finding relaying time duration and computation parameters, respectively. From $NJ.P1$, we find out if APs help in relaying or not, i.e., $\mathbf{A}$, which is used in subsequent optimization problem $NJ.P2$. We solve $NJ.P1$ and $NJ.P2$ for fixed $\alpha = \bar{\alpha}$ and find optimal value for α by search operation. We divide total duration equally for relaying and offloading, which is given in constraint (i.e., $(C2)$ in $NJ.P1$ and $NJ.P2$).

$$NJ.P1 : \max_{\mathbf{A},\mathbf{T}^{\mathrm{r}}} \sum_{m=1}^{M}\sum_{n=1}^{N} \bar{\alpha} R^{\mathrm{Pri}}; \text{s.t.:}(C1)(6),(16\text{b}),(16\text{c}); (C2)\tau_m(n) \leq T/2, \forall m, n.$$

We denote solutions from $NJ.P1$ by $\mathbf{A} = \bar{\mathbf{A}}$ and $\mathbf{T}^{\mathrm{r}} = \bar{\mathbf{T}}^{\mathrm{r}}$, which we use to solve following:

$$NJ.P2 : \max_{\mathbf{P},\mathbf{R},\mathbf{S},\mathbf{T}^{\mathrm{L}}} \sum_{m=1}^{M}\sum_{n=1}^{N} (1-\bar{\alpha}) R^{\mathrm{IoT}}; \text{s.t.:}(C1)(7),(12); (C2) I_{[k],m}(n) \leq T/2, \forall k, m, n.$$

A. Optimality and efficacy check for our proposed scheduling order

In our system model, unlike [16], [17], [19], we have considered finite computation power at MEC servers. Therefore, offloading and computation scheduling order design is very important for our system model, which has not been considered in [16], [17], [19]. In this section, we show optimality of our proposed method to find scheduling order in Proposition 3.2 comparing with exhaustive search (i.e., best one from all possible combinations of scheduling order). Moreover, we also compare with the scenario when there is no fixed scheduling order, i.e., randomly choosing any order to show efficacy of our proposed method. We have considered one AP and PR pair (i.e., $M = 1$), ten IoT devices under the AP (i.e., $K_m = 10$), $f_m^{\mathrm{AP}} = 10^{16}$ Hz, and Q_m =1000 bits, to generate Fig. 6. Stored energy at IoT nodes follow uniform distribution in between 0.1 to 2 Joules. We have compared in term of weighted sum of relaying rate and total number of computed bits in Fig. 6(a), and total number of computed bits in Fig. 6(b). It can be observed that in both cases, our proposed scheduling order exactly matches with the best one from exhaustive search, i.e., search among all $K_m!$ permutations, which we have discussed earlier. Hence, we can conclude that our proposed scheduling order in Proposition 3.2 is optimal. Moreover, in both Fig. 6(a) and (b), we can observe that compared to random scheduling order, we get better result, which proves efficacy of our proposed method. As we increase α, gap between received values from our proposed scheduling order and random scheduling order reduces. For higher values of α, the primary network's rate gets more importance, which increases relaying time duration, and hence reduces offloading time duration. As offloading time reduces, effective computation mainly depends on local computation, which is the reason behind reduced value for total number of computed bits in Fig. 6(b) for varying α. This is also the reason behind lower gap in values received for optimal and random scheduling for higher values of α in Fig. 6(a).

B. Varying number of IoT devices and RBs

In Fig. 7(a) and (b), we analyse the effect of varying number of IoT devices under APs and RBs on primary network's optimal rate and IoT networks' total number of computed bits, respectively. We consider identical number of IoT devices under each AP, i.e., $K_m = K, \forall m$. In Fig. 7(a) and (b), we have considered $Q_m = 1000$ bits, $f_m^{\text{AP}} = 10^{16}$ Hz, $\forall m$, and uniformly distributed energies (i.e., between 1 to 2 Joules) at IoT devices. In Fig. 7(a), we have considered $N = M = 5$. It is observed that while increasing K_m, primary network's rate reduces, whereas, IoT networks' total number of computed bits increases. For our setup in Fig. 7(a), we observe that offloading rate for IoT networks is higher than primary network's communication rate. Therefore, time duration for offloading is tried to increase while reducing relaying time duration. From these reasons, for increasing K_m, we may explain the rising and falling patterns in IoT networks' computed bits and primary network's communication rate, respectively. However, we observe saturation for higher values of K_m. Due to time constraint, we may not increase/decrease offloading time/relaying time durations, which may explain the saturation phenomenon. In Fig. 7(b), we have considered

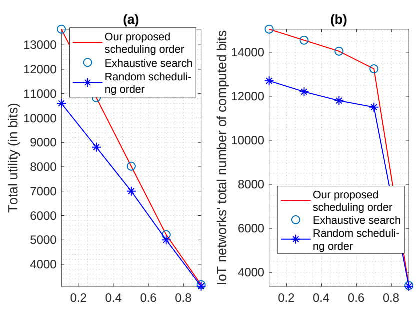


Fig. 6: Comparison with exhaustive search and random scheduling order in term of (a) Total utility (b) Number of computed bits

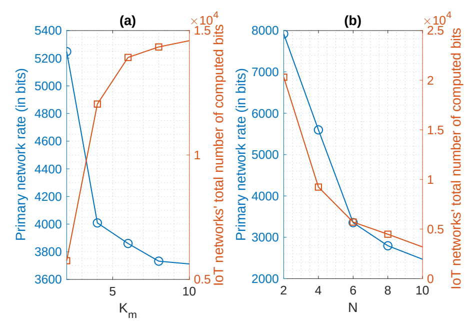


Fig. 7: Results for varying (a) number of IoT devices under $\text{AP}_m, \forall m$ (i.e., K_m) (b) number of RBs (i.e., N)

$M = 5, K_m = 10, \forall m$. We observe that as we increase number of RBs (i.e., N), both primary network's rate and IoT networks' total number of computed bits reduce. Effective bandwidth for an RB reduces as we increase N. Now as there are constraints on both relaying and offloading time durations, for lower bandwidth, both parameters decrease. Moreover, for $N > M$, $N - M$ number of RBs are not allocated, which may be another reason behind the decreasing pattern.

C. Varying stored energy at IoT nodes

In Fig. 8, we show how the primary network's rate and the IoT network's total computed number of bits change with stored energy at IoT nodes. We consider identical number of IoT devices under each AP, i.e., $K_m = K, \forall m$, $Q_m = 1000$ bits, $f_m^{\text{AP}} = 10^{16}$ Hz, $\forall m$, and $N = M = 5$. We analyse for two different cases, i.e., stored energy at IoT nodes follow uniform distribution in between 0.5 to 1 Joule and 2.5 to 3 Joule. We observe that with increasing number of IoT nodes under an AP, IoT network's total number of computed bit increases, whereas, the primary network's rate decreases. This can be explained following same analogy, which we gave for Fig. 7. IoT nodes will be able to offload more with more stored energy, which increases offloading time and hence reduces relaying time. Therefore, IoT network's total computed bit number increases and primary network's rate reduces.

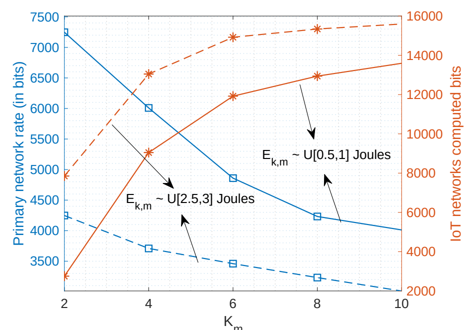


Fig. 8: Result for varying $E_{k,m}$ and K_m ($U[a, b]$:uniformly distributed between a and b)

D. Comparison with No Spectrum Sharing (NSS)

We compare with NSS model in Fig. 9, which has been discussed earlier in this section. We have considered $N = 4, M = 8$, $K_m = K = 10, \forall m$, uniformly distributed stored energies (i.e., in between 1 to 2 Joules) at IoT devices, and $Q_m = 1000$ bits, $\forall m$. In Fig. 9(a), we plot the primary network's total rate and IoT networks' total number of computed bits for varying MEC servers computation power, i.e., $f_m^{\text{AP}} = f^{\text{AP}}, \forall m$. It can be observed that with spectrum sharing, we get higher values for both metrics compared to NSS. For spectrum sharing, IoT networks' total number of computed bits increases and primary network's total rate decreases for varying f^{AP}. MEC servers can compute more computational tasks with higher computation power, and hence, offloading time duration increases, which reduces relaying time duration. This may explain those rising and falling patterns. Relaying time duration cannot be lowered below a level due to required rate at PRs, which also put constraint on offloading durations, which are the reasons behind saturation in values for both metrics in higher values of f^{AP}. For NSS, the primary network performs direct transmission and IoT network performs local computation only. Hence, both metrics do not depend on f^{AP}, which can be observed in Fig. 9(a). In Fig. 9(b), we emphasize on the benefit of spectrum sharing in our system model by comparing it with NSS. Following (37) and (38), we have plotted Fig. 9(b). It can be observed that the primary network's gain decreases, whereas, IoT network's gain increases with f^{AP}. This can be explained from our previous discussion. From Fig. 9(b), it can be observed that for our set-up, the primary network and IoT networks can gain almost eight times and six times, respectively, from spectrum sharing.

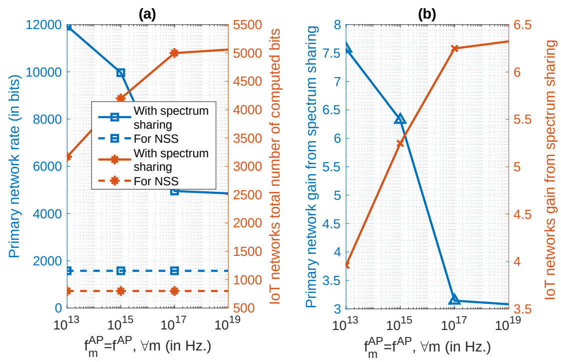


Fig. 9: (a) Comparison with NSS in term of primary network's rate and IoT network's number of computed bits (b) Spectrum sharing gain (from (37) and (38))

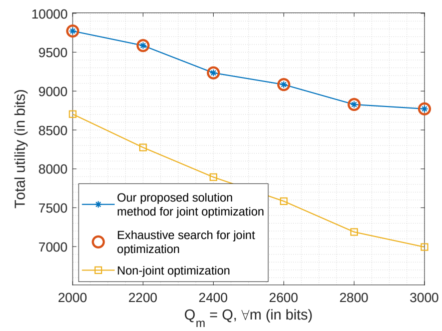


Fig. 10: Comparison with non-joint optimization (given in beginning of Section V) and exhaustive search in term of total utility

E. Comparison with Non-Joint (NJ) optimization and exhaustive search

In Fig. 10, we analyse advantage for joint optimization problem (i.e., joint design of relaying and computation parameters in $P1$) over non-Joint (NJ) optimization problem, which we have discussed in beginning of this section and has been considered in [20]. We also validate our proposed solution method for solving the joint optimization problem by comparing with exhaustive search, where we consider all possible combinations of RB allocation and communication method selection (i.e., relaying or direct communication) to a PR. We consider best possible result from all combinations. We consider $N = M = 4$, $f_m^{\mathrm{AP}} = 10^{16}$ Hz, $\forall m$, $K_m = K = 6, \forall m$, and uniformly distributed energies (i.e., between 1 to 2 Joules) at IoT devices. We show the comparison result for varying $Q_m = Q, \forall m$. It can be observed that joint optimization outperforms NJ one in term of total utility as has been defined in (13a). APs may not be able to satisfy higher required rate at PRs, under which direct communication may take place to PRs and APs do not get access to RBs. Therefore, both primary network's rate and IoT networks' total computed bits reduce with increasing Q_m. This particular phenomenon occurs with higher chance as we go on increasing Q_m, which is the reason behind increasing gap while varying Q_m in Fig. 10. Our proposed solution method provides excatly same result like exhaustive search, which proves our proposed solution method's optimality.

VI. Conclusion

Spectrum management for massive IoT networks is an important aspect. In this paper, we consider MEC for IoT networks in the context of spectrum sharing, where IoT networks get access to licensed spectrum after helping a primary network by cooperative relaying. We have formulated an optimization problem, where both relaying and MEC parameters have been optimized jointly. The optimization problem is observed to be non-trivial due to the presence of binary variables, which makes the problem combinatoric. From our analysis, we could map the optimization problem with a bipartite matching problem in Graph theory, which we solved using optimal and efficient Hungarian algorithm. Moreover, we have also proposed optimal scheduling order for IoT devices, which is very important in our system model. In the result section, we have shown that both primary and IoT networks benefit from the cooperation, which we show in Fig. 9(b). Distributed design may be a relevant future work for our model.

Acknowledgement

This work was supported by F.R.S.-FNRS under the EOS program (EOS project 30452698) and by INNOVIRIS under the COPINE-IOT project.

APPENDIX A: Proof of Proposition 3.2

We prove this proposition based on following two facts: (a) Under identical values for computation parameters', i.e., $f_{k,m}^{\text{Loc}} = f_m^{\text{L}}$, $w_{k,m} = w_m$, $C_{k,m}^{\text{Loc}} = C_m^{\text{L}}$, IoTs locally compute identical number of bits and consume identical energy to compute a bit, i.e.,: $\mathcal{E}_{k,m}^{\text{L}} = \left(w_m t_{k,m}^{\text{Loc}} (f_m^{\text{L}})^3\right) / \left(\frac{t_{k,m}^{\text{Loc}} f_m^{\text{L}}}{C_m^{\text{L}}}\right) = w_m (f_m^{\text{L}})^2 C_m^{\text{L}}, k = 1, 2,$ (b) For identical requirement of CPU to compute a bit at the AP, i.e., $C_{[k],m}^{\text{MEC}} = C_m^{\text{M}}$, the AP takes identical time for all users to compute identical amount of task.

As IoTs behave identically in terms of local computation and the AP treats all IoTs in identical way while computing their tasks, the only difference remains in offloading at IoTs. It is intuitive that among $\text{IoT}_{p,m}$ and $\text{IoT}_{q,m}$, the AP will give more preference to the IoT, which can offload more data for identical offloading duration. Therefore, we can judge an IoT's offloading capability by it's stored energy, channel gain and distance to the AP. In other words, from (21), we can say that IoT_p should be scheduled before IoT_q for $E_{p,m}|h_{p,m}^n|^2 d_{p,m}^{-\kappa} > E_{q,m}|h_{q,m}^n|^2 d_{q,m}^{-\kappa}$.

APPENDIX B: Proof of Proposition 3.3

From (8), we can get following two cases: Case1: $\sum_{i=1}^{k} t_{[i],m}^{\text{Off}}(n) < I_{[k-1],m}(n)$; Case2: $\sum_{i=1}^{k} t_{[i],m}^{\text{Off}}(n) \geq I_{[k-1],m}(n)$. At first, let us analyse Case1 for $k = 2$, which can be generalised later:

$$t_{[1],m}^{\text{Off}}(n) + t_{[2],m}^{\text{Off}}(n) < I_{[1],m}(n) \implies t_{[2],m}^{\text{Off}}(n) < t_{[1],m}^{\text{MEC}}(n). \tag{39}$$

Now, from (23a), it can be observed that objective function increases as we increase offloading rate. Also from (21), we observe that offloading rate increases with offloading duration without violating any constraint in $P3.1$. Therefore, the relation in (39) is not necessarily true. Hence, we conclude that in optimal case, we get $t_{[2],m}^{\text{Off}}(n) \geq t_{[1],m}^{\text{MEC}}(n)$. We can easily generalize this and get the relation as given in (24a). We use the relation as given in (24a), then, from (8) we get:

$$I_{[k],m}(n) = \sum_{i=1}^{k} t_{[k],m}^{\text{Off}}(n) + t_{[k],m}^{\text{MEC}}(n). \tag{40}$$

From (40) and (17d), we get (24b). So, in optimal case, we replace (17d) by (24a) and (24b).

APPENDIX C: Proof of Proposition 3.4

We prove this with help of KKT stationarity conditions. We get the relation as given in (30) from the relation received from partial derivative of Lagrangian with respect to $e_{k,m}^{\text{Off}}(n)$, i.e., $\frac{\partial L}{\partial e_{k,m}^{\text{Off}}(n)} = \eta_k \log_2(e) B_n |h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}{}^{-\kappa} / \{\sigma^2 (1 + \frac{e_{k,m}^{\text{Off}}(n) |h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}{}^{-\kappa}}{t_{k,m}^{\text{Off}}(n)\sigma^2})\} - \mu_k = 0$. Similarly, while deriving (29), we consider partial derivative of Lagrangian with respect to $t_{k,m}^{\text{Off}}(n)$, i.e., $\frac{\partial L}{\partial t_{k,m}^{\text{Off}}(n)} = -\sum_{i=k}^{K_m} \beta_i - \eta_k \log_2(e) B_n \left[\frac{\frac{e_{k,m}^{\text{Off}}(n) |h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}{}^{-\kappa}}{t_{k,m}^{\text{Off}}(n)\sigma^2}}{1 + \frac{e_{k,m}^{\text{Off}}(n) |h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}{}^{-\kappa}}{t_{k,m}^{\text{Off}}(n)\sigma^2}} \right] + \zeta_{k+1} + \eta_k B_n \log_2 \left(1 + \frac{e_{k,m}^{\text{Off}}(n) |h_{\mathrm{I}_k,\mathrm{A}_m}(n)|^2 d_{\mathrm{I}_k,\mathrm{P}_m}{}^{-\kappa}}{t_{k,m}^{\text{Off}}(n)\sigma^2}\right)$.

REFERENCES

[1] C. V. N. Index, "Cisco visual networking index: global mobile data traffic forecast update, 2017–2022," *Cisco*, 2019.

[2] D. Sabella, A. Vaillant, P. Kuure, U. Rauschenbach, and F. Giust, "Mobile-edge computing architecture: The role of mec in the internet of things," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 84–91, 2016.

[3] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—a key technology towards 5g," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[4] X. Sun and N. Ansari, "Edgeiot: Mobile edge computing for the internet of things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, 2016.

[5] E. Union, "Identification and quantification of key socio-economic data to support strategic planning for the introduction of 5g in europe — smart 2014/0008, 2016."

[6] L. Zhang, Y.-C. Liang, and M. Xiao, "Spectrum sharing for internet of things: A survey," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 132–139, 2018.

[7] P. Corcoran and S. K. Datta, "Mobile-edge computing and the internet of things for consumers: Extending cloud computing and services to the edge of the network," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 73–74, 2016.

[8] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[9] Z. Gao, W. Hao, and S. Yang, "Joint offloading and resource allocation for multi-user multi-edge collaborative computing system," *IEEE Transactions on Vehicular Technology*, 2021.

[10] Z. Li, N. Zhu, D. Wu, H. Wang, and R. Wang, "Energy-efficient mobile edge computing under delay constraints," *IEEE Transactions on Green Communications and Networking*, 2021.

[11] X. Li, R. Fan, H. Hu, N. Zhang, X. Chen, and A. Meng, "Energy-efficient resource allocation for mobile edge computing with multiple relays," *IEEE Internet of Things Journal*, 2021.

[12] S. Bi and Y. J. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 4177–4190, 2018.

[13] X. Hu, K.-K. Wong, and K. Yang, "Wireless powered cooperation-assisted mobile edge computing," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2375–2388, 2018.

[14] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268–4282, 2016.

[15] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4188–4200, 2018.

[16] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1784–1797, 2017.

[17] H. Sun, F. Zhou, and R. Q. Hu, "Joint offloading and computation energy efficiency maximization in a mobile edge computing system," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3052–3056, 2019.

[18] M. Qin, N. Cheng, Z. Jing, T. Yang, W. Xu, Q. Yang, and R. R. Rao, "Service-oriented energy-latency tradeoff for iot task partial offloading in mec-enhanced multi-rat networks," *IEEE Internet of Things Journal*, 2020.

[19] F. Zhou and R. Q. Hu, "Computation efficiency maximization in wireless-powered mobile edge computing networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3170–3184, 2020.

[20] B. Liu, J. Bai, Y. Ma, J. Wang, and G. Lu, "Energy-efficient resource allocation for wireless powered cognitive mobile edge computing," in *Proceedings of IEEE International Conference on Communications (ICC) Workshops*, 2019.

[21] N. Biswas, H. Mirghasemi, and L. Vandendorpe, "Joint optimization of relaying rate and energy consumption for cooperative mobile edge computing," in *Proceedings of WiOpt*, 2020.

[22] ——, "Sharing is caring: A mobile edge computing perspective," in *Proceedings of International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 1298–1303.

[23] J. M. Peha, "Approaches to spectrum sharing," *IEEE Communications magazine*, vol. 43, no. 2, pp. 10–12, 2005.

[24] O. Simeone, I. Stanojev, S. Savazzi, Y. Bar-Ness, U. Spagnolini, and R. Pickholtz, "Spectrum leasing to cooperating secondary ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 203–213, 2008.

[25] S. S. Kalamkar, J. P. Jeyaraj, A. Banerjee, and K. Rajawat, "Resource allocation and fairness in wireless powered cooperative cognitive radio networks," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3246–3261, 2016.

[26] N. Biswas, G. Das, and P. Ray, "Buffer-aware user selection and resource allocation for an opportunistic cognitive radio network: A cross-layer approach," *IEEE/ACM Transactions on Networking*, 2022.

[27] K. Kulkarni and A. Banerjee, “Multi-channel sensing and resource allocation in energy constrained cognitive radio networks,” *Physical Communication*, vol. 23, pp. 12–19, 2017.

[28] W. Sun, Q. Song, J. Zhao, L. Guo, and A. Jamalipour, “Adaptive resource allocation in swipt-enabled cognitive iot networks,” *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 535–545, 2021.

[29] A. Khan, M. H. Rehmani, and A. Rachedi, “Cognitive-radio-based internet of things: Applications, architectures, spectrum related functionalities, and future research directions,” *IEEE wireless communications*, vol. 24, no. 3, pp. 17–25, 2017.

[30] T. Van der Vorst, J. Veldman, and J. van Rees, “The wireless internet of things: Spectrum utilisation and monitoring,” *Radiocommunications Agency Netherlands, Utrecht, Tech. Rep*, vol. 1618, p. v2, 2016.

[31] W. Lu, P. Si, G. Huang, H. Han, L. Qian, N. Zhao, and Y. Gong, “Swipt cooperative spectrum sharing for 6g-enabled cognitive iot network,” *IEEE Internet of Things Journal*, 2020.

[32] W. Lu, S. Hu, X. Liu, C. He, and Y. Gong, “Incentive mechanism based cooperative spectrum sharing for ofdm cognitive iot network,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 662–672, 2019.

[33] A. Hoglund, X. Lin, O. Liberg, A. Behravan, E. A. Yavuz, M. Van Der Zee, Y. Sui, T. Tirronen, A. Ratilainen, and D. Eriksson, “Overview of 3gpp release 14 enhanced nb-iot,” *IEEE Network*, vol. 31, no. 6, pp. 16–22, 2017.

[34] H. K. Mendis and F. Y. Li, “Achieving ultra reliable communication in 5g networks: A dependability perspective availability analysis in the space domain,” *IEEE Communications Letters*, vol. 21, no. 9, pp. 2057–2060, 2017.

[35] A. Roy, P. Chaporkar, A. Karandikar, and P. Jha, “Online radio access technology selection algorithms in a 5g multi-rat network,” *IEEE Transactions on Mobile Computing*, 2021.

[36] Y. Yang, H. Hu, J. Xu, and G. Mao, “Relay technologies for wimax and lte-advanced mobile systems,” *IEEE Communications Magazine*, vol. 47, no. 10, pp. 100–105, 2009.

[37] Nokia. ”LTE-M: Optimizing lte for the internet of things white paper”,white paper, tech. rep., 2015; accessed 17th april 2021. [Online]. Available: https://lafibre.info/images/4g/201508-nokia-lte-m-white-paper.pdf

[38] L. Vandendorpe, R. T. Durán, J. Louveaux, and A. Zaidi, “Power allocation for ofdm transmission with df relaying,” in *Proceedings of IEEE International Conference on Communications (ICC)*, 2008.

[39] S. M. Johnson, “Optimal two-and three-stage production schedules with setup times included,” *Naval research logistics quarterly*, vol. 1, no. 1, pp. 61–68, 1954.

[40] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[41] H. Lim and T. Hwang, “Energy-efficient computing for wireless powered mobile edge computing systems,” in *Proceedings of Vehicular Technology Conference (VTC2019-Fall)*. IEEE, 2019, pp. 1–5.

[42] D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Upper Saddle River, 2001, vol. 2.

[43] C. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

[44] D. Bertsekas, *Network optimization: continuous and discrete models*. Athena Scientific, 1998.

[45] A. Mukhopadhyay and G. Das, “Low complexity fair scheduling in lte/lte-a uplink involving multiple traffic classes,” *IEEE Systems Journal*, 2020.

[46] S. Bubeck, “Convex optimization: Algorithms and complexity,” *arXiv preprint:1405.4980*, 2014.

[47] K. M. Anstreicher, “Linear programming in O($[n^3$ / $\ln n]$ L) operations,” *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 803–812, 1999.

Response to reviewers' comments

Title of the manuscript: On joint cooperative relaying, resource allocation, and scheduling for mobile edge computing networks

Manuscript ID: TCOM-TPS-21-0848.R1

We sincerely express our gratitude to the editor for considering this manuscript for review. We thank the anonymous reviewers for their insightful comments that helped in re-organizing and enriching the modified manuscript. We have shown the modifications in the modified manuscript by using blue colour.

Editor's comments:

Comment 1:

The math expression part is hard to follow and needs to improve

Answer: We apologize for the difficulty. In our modified manuscript, we have made changes to improve readability related to the mathematical expressions. We have discussed about those changes in response to both reviewers' last comments.

Comment 2:

There is lack of validations on the math derivations

Answer: The editor's comment is well accepted. In our modified manuscript, we have incorporated some changes, which would help in understanding our mathematical derivations. We have followed previous works while deriving some of our mathematical expressions, e.g., Equation 1 and 11. In the modified manuscript, we have referred to those. Moreover, we have also updated Figure 10 (which was Figure 9 before) while adding a result for exhaustive search to validate our proposed optimal solution method. In exhaustive search, we have considered all possible combinations of RB allocation to a PR and mode of communication (i.e., relaying and direct communication). Then we have chosen the best possible result from all combinations. Except these, we have followed some more steps to increase the readability, which we have discussed in response to the second reviewer's last comment.

Comment 3:

The novelty of the problem formulation should be well justified, compared to many existing works on multiple licensed spectrum resource block with multiple licensed users

Answer: We thank the editor for this comment. In our work, we have considered two major aspects for IoT networks, i.e., spectrum sharing and mobile edge computing. We have explored potential marriage between those two aspects, which we feel is very much important for IoT networks and have been consistently omitted in previous literature. We agree that in literature, there are works on multiple licensed spectrum sharing. However, to the best of our knowledge, none of them have considered spectrum sharing and mobile edge computing jointly. Hence, their system models are very different from ours. In our system model, we consider multiple licensed resource block with multiple licensed users. Except this, we consider multiple relays and multiple IoT networks to make our system model a generalized one. We form an optimization problem considering joint design of both spectrum sharing and mobile edge computing parameters, which include relay selection, resource block allocation, IoT nodes' scheduling. Though the optimization problem is combinatoric and hence non-trivial, we develop an optimal and efficient algorithm, which solves the problem in polynomial time.

In response to second reviewer's first comment, we have categorized the available literature, where our work has been shown in red box. In our modified manuscript, we could not include the hierarchy due to space constraint. However, we have modified 'Literature Survey' in Section I.A by including relevant papers, i.e., [7], [25], [27], and [28]. Moreover, we have also included 'Literature gap' at the end of 'Literature survey' section in the modified manuscript and then modified the section 'Motivations and contributions'.

Comment 4:

The optimization problem (P1) should be justified, and channel gain should be clarified

Answer: We apologize for not being clear about the channel gains in our earlier manuscript. In our modified manuscript, we have discussed the channel gain parameters, which we have added in response to first reviewer's second comment.

Reviewer-1

General comment: The authors responded to/ addressed the reviewers' comments which improved the quality of the paper.

We thank the reviewer for taking time to review our paper.

Comment 1:

In the optimization problem (P1), considering the local computing time and offloading time as decision variables is weird. In general, latency is a performance indicator.

Answer: We thank the reviewer for this comment. We agree that latency is a performance indicator for wireless networks with delay sensitive applications. In our work, we have also targeted delay sensitive applications and therefore, we have considered that tasks will be computed within a stringent time limit (within a time frame duration). This evaluates our proposed algorithm in the most stringent condition. We may relax the delay limit if the application permits. Even in that case, our algorithm would be applicable.

It is to be noted that the number of computed bits would depend on how efficiently we can allocate the available time for offloading and processing the tasks. If we can increase the number of computed bits, the quality of service enhances (e.g., image, video processing). Since total available time duration is constant, allocating time for offloading and computation is vital for increasing the number of computed bits. Therefore, allocation of time indirectly influences the quality of service. Also note that the local computation might take place throughout the whole-time interval T (see Figure 2 of manuscript and descriptions therein). Considering energy constraints, we have made a balance between local processing and offloading time (as wireless transmission due to offloading consumes energy).

We have optimized both division of time and energy consumption to maximize the number of computed bits. In literature, we find following works, which have also considered maximization of total number of computed bits:

- Bi, Suzhi, and Ying Jun Zhang. "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading." IEEE Transactions on Wireless Communications, vol. 17, no. 6, pp: 4177-4190, 2018 (Reference [12] in our manuscript)

- F. Zhou and R. Q. Hu, “Computation efficiency maximization in wireless-powered mobile edge computing networks,” IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 3170–3184, 2020 (Reference [19] in our manuscript)

In [11] and [19], authors have maximized total number of computed bits under available time and energy constraints. Authors have considered time and energy as available resources, which have been evaluated while maximizing total number of computed bits. In both [11] and [19], authors have considered single frequency resource block, which mobile edge computing networks own.

Unlike [11] and [19], in our work, we have considered cooperative relaying-based spectrum sharing, which has gained much attention due to its potential in avoiding unavoidable spectrum gridlock situation due to ever increasing number of wireless devices. We have considered multiple licensed frequency resource blocks, which are shared with mobile edge computing networks under guaranteed benefit for the primary network (owns those licensed multiple frequency resource blocks). We formulate a cross-layer optimization problem, which considers both network layer scheduling and physical layer resource allocation, e.g., transmission time, energy, and frequency.

In the modified manuscript, we have included following discussion at bottom of page no-11 in this regard:

‘For the primary network, the BS will try to maximize the communication rate, whereas, in IoT networks, total number of computed bits is of interest while maintaining latency constraint in (7) and (10). For applications, e.g., image, video processing, number of computed bits may define quality of service. Therefore, optimization of number of computed bits is important, which has been considered in [12], [19].’

Comment 2:

There is no explanation about the channel gain h. The authors need to discuss the channel model.

Answer: We thank the reviewer for pointing out this shortcoming. We apologize for not being clear about it in our previously submitted manuscript. We consider that channel follows complex Normal distribution. Hence, modulus of channel gain follows Rayleigh distribution, and square of that follows exponential distribution. In the modified manuscript, we have added following discussion below Table II in page no-24:

‘We assume that different channels in our system model follows complex Normal distribution, such that, we can say that $|h_{A_m-P_m}(n)|^2, |h_{I_k-A_m}(n)|^2, |h_{B-A_m}(n)|^2$, and $|h_{A_m-P_m}(n)|^2$follow exponential distribution. We assume unity mean exponential distribution for all channel gains.’

Comment 3:
Mathematical representations need further improvement.

Answer: We thank the reviewer for this comment. In our modified manuscript, we have made changes to improve readability related to mathematical expressions. Please refer to the response to the second reviewer’s last comment, where we have discussed about those changes.

Reviewer-2

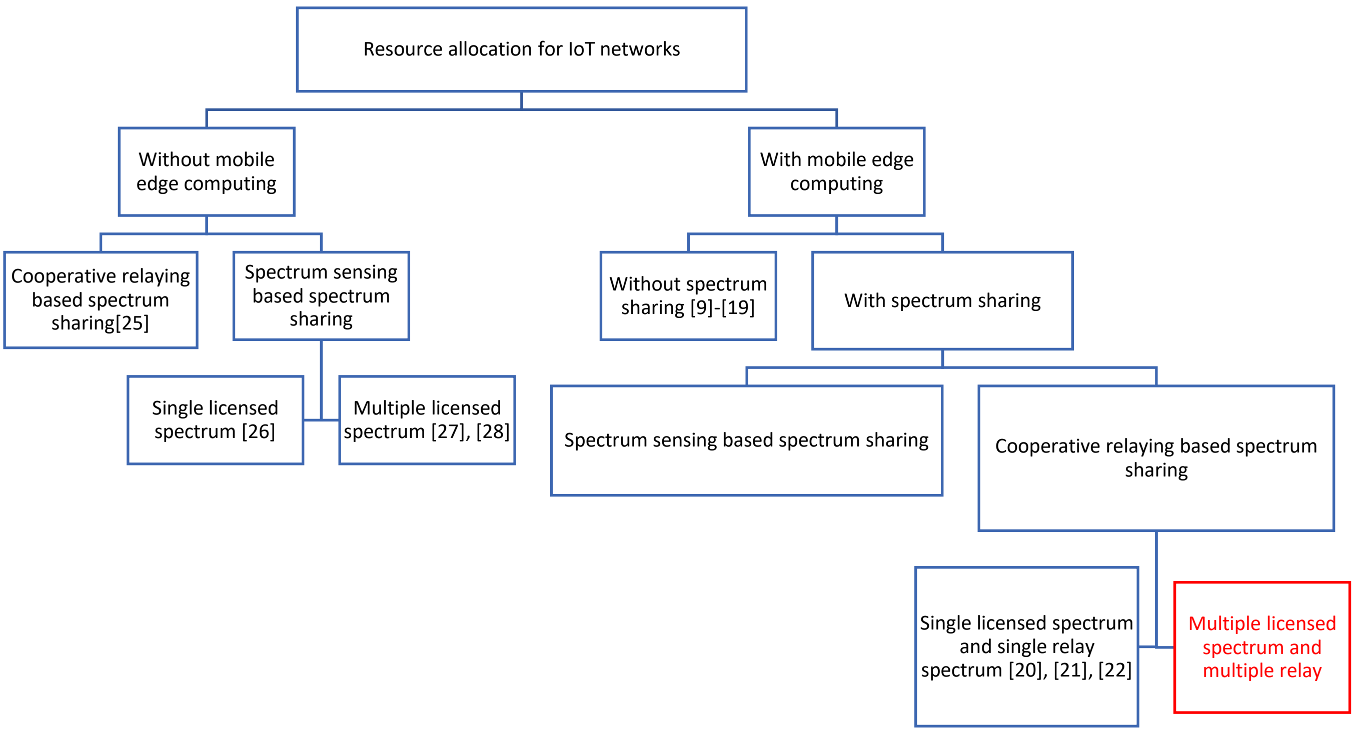
Comment 1:

Spectrum sharing is not a novel issue. There are so many existing works focusing on it. If we look at the submission without edge computing, it is a discussion about "Multiple licensed spectrum resource block, Multiple licensed users".

Answer: We thank the reviewer for this comment. We agree that spectrum sharing has been well studied in recent past. Moreover, it has received much attention for 5G and beyond 5G communications due to its potential in avoiding spectrum scarcity problem. Following the reviewer’s comment, we did another round of literature survey and have added references [27] and [28], where multiple licensed spectrum has been considered, in which unlicensed users have been scheduled.

Spectrum sharing has gained much attention for IoT networks too as massive IoT networks may not be served with existing limited spectrum. Another important aspect for IoT networks is mobile edge computing, which may help typical computation power and energy constrained IoT nodes in computing tasks. Reference [7] (of our modified manuscript) may be referred to in this regard. In this manuscript, we have explored the promising marriage of mobile edge computing and spectrum sharing for IoT networks for a generalized system model. According to our understanding, we feel that it is very important to jointly design parameters related to spectrum sharing and MEC. IoT nodes may not have access to licensed spectrum for offloading their computation tasks to MEC servers, which can be solved with help of spectrum sharing.

Based on our literature survey, we can divide resource allocation for IoT networks, which we have shown in form of hierarchy in next page and have shown our work in red coloured box. In a broad sense, based on the availability of licensed spectrum, we can divide the literature on resource allocation for mobile edge computing based IoT networks into two fields, i.e., without spectrum sharing and with spectrum sharing. With spectrum sharing can be further divided into two, i.e., spectrum sensing based and cooperative relaying based. For spectrum sensing, unlicensed network needs to perform spectrum sensing and monitor licensed network's activity in licensed spectrum in a continuous way. Moreover, as the sensing outcome (i.e., licensed spectrum is being used or not by licensed network) is probabilistic, it may be difficult to maintain the licensed networks' quality of service due to interference, which may happen when the licensed network is falsely detected as idle. These problems do not arise for cooperative relaying-based spectrum sharing. The unlicensed network helps the licensed network by cooperatively relaying its (licensed network's) data and in return gets access to licensed spectrum. For spectrum sensing, it is important to regularly monitor the licensed network's activity. Otherwise, unlicensed network may interfere and degrade licensed network's quality of service. In our work, we consider cooperative relaying-based spectrum sharing for IoT networks.



Following are some important literature gap, which we have observed and motivated us in picking this problem:

- **Allocation of multiple licensed spectrum resource block:** We observe that multiple licensed spectrum resource block allocation can be done using binary variables, which would make the optimization problem combinatoric. Authors of [27] have scheduled unlicensed users in the licensed spectrum resource block. However, they have not considered any variable related to that. Though authors of [28] have considered multiple licensed resource blocks, they have relaxed the binary parameters to be continuous

- **Scheduling of IoT nodes at MEC servers at access points:** We observe that scheduling of IoT nodes is very important for multiple IoT nodes. In this context, both offloading and computation (at MEC server) scheduling arises. We observe this scheduling has been considered in references [16], [17], and [19] of our manuscript under assumption of infinite computation power at MEC servers, which may not hold in realistic scenario. Unlike cloud server, MEC server would be computation power constrained

- **Joint design of spectrum sharing and MEC for generalized scenario:** As we consider both spectrum sharing and MEC, we need to design associated parameters properly. We observe that authors of references [9]-[19] have considered MEC only. Though [20]-[22] have considered both spectrum sharing and MEC, they did their analysis for single licensed spectrum resource block and single relay. Moreover, joint design has not been considered in [20]

From the previously submitted manuscript, our contributions could not be understood clearly. We apologize for that. In the modified manuscript, we have done following changes to put our problem in appropriate place, such that, our contributions become clearer:

- Though we could not add the above shown hierarchy, we have mentioned about those papers (given in boxes of different fields inside the hierarchy) in right context
- We have added ‘Literature gap’ below ‘Literature survey’ (end of Section I.A)
- We have rewritten ‘Literature survey’ and ‘Motivation and contributions’ (Section I.A and I.B)

Comment 2:

The authors do not justify in which edge computing scenario this proposed algorithm will be perfectly suitable.

Answer: In our manuscript, we have considered mobile edge computing for IoT networks. Spectrum sharing for IoT networks is a very important topic and it has gained attention by standardization body 3GPP also. IoT networks can share licensed spectrum to carry on communications with mobile edge computing servers. In our modified manuscript, we have mentioned about application of our proposed system model below Table I in page no-7:

'Possible application: We can say our system model is very much realistic from facts: (a) cooperative relaying is used in Long Term Evolutions (LTE) networks [36] and (b) deployment of IoT networks in LTE band is found for Narrow-Band IoT (NB-IoT) networks [33], [37]. We can consider IoT access points of our system model as relays for LTE base stations and in return IoT networks share LTE spectrum.'

Comment 3:

(a) The mathematics is too complex, which makes it hard to read and believe.
(b) There are too many parameters, yet there are few experiments to support the conclusion

Answer:

(a) We apologize for the difficulty. In the manuscript, we have given number of mathematical equations to describe the optimization problem initially and then for discussing its corresponding solution procedures.

In our modified manuscript, we have considered the reviewer's suggestion and have made following changes to increase the readability of the modified manuscript:

- We have shortened symbols for some parameters, which we have given in blue colours
- We have made corresponding changes in Figures too, where we have used different parameters
- We have followed previous literature to get some mathematical expressions in our manuscript, e.g., Equation (1) and (11) can be derived following references [38] and [12], respectively. In our modified manuscript, we have referred to corresponding sources while describing those equations

- A convex optimization problem can be solved by solving its corresponding dual problem as has been given in Chapter 5 of the reference [40] of our manuscript. We have used this fact while solving the convex optimization problem P3.2 in page no-17 and later. In the modified manuscript, we have mentioned this below Equation 26
- We have removed equation numbers from some equations, which are not referred later or are used as intermediate steps to a final equation, for example number of locally computed bits above Equation 5 of modified manuscript
- We have added short discussion about our solution methodology above Section III in page no: 12. We hope this discussion would give an overview about our solution approach

We have validated one of our important mathematical findings in the result section. In Figure 6 of our manuscript, we have shown that the received result from our proposed algorithm (with polynomial time complexity order) exactly matches with exhaustive search (which is optimal but with non-polynomial computation order). In the modified manuscript, we have changed Figure 10 (which was Figure 9 before) by adding a result for exhaustive search. It is observed that the result for exhaustive search exactly matches with the result received from our proposed algorithm.

(b) The reviewer's comment is well accepted. In the modified manuscript, we have added another important result, i.e., Figure 8, where we have shown how the primary network's rate and the IoT network's total number of computed bits change with stored energy at IoT nodes.

To,
The Editor,
IEEE Transactions on Communications.

03rd June, 2022,

Subject: Cover letter

Dear Editor,

We really appreciate the effort that you and all the reviewers have put forward towards providing insightful and valuable suggestions. With the help of the reviewers' suggestions, we have carefully revised our original manuscript while making relevant changes. Following, we discuss some of them:

- Rewritten 'Literature survey' and 'Motivation and contributions' in 'Introduction' part. Moreover, added 'Literature gap' at end of 'Literature survey' to make our contributions more clear
- Changed mathematical notations and added discussions to make mathematical expressions more readable
- Added a discussion at the end of Section II, where we have talked about our approach for solving optimization problem P1 in zest
- Added justification behind our optimization problem and its relevance
- Added discussion on channel gains
- Added and modified results (Figure 8 and 10, respectively) to validate our mathematical findings
- Added more literature and have discussed about differences with them

We have addressed each individual comment by the reviewers. We believe that the modified manuscript has been significantly improved in terms of both its technical quality and clarity. We thank you and the reviewers for their sincere and valuable efforts.

In our modified manuscript, we have highlighted all modifications associated to reviewers' comments using blue color. Detailed response to every reviewer's comment indicating corresponding changes in the modified manuscript is also attached along-with the submission. We sincerely appreciate the helpful comments and suggestions. We believe that these have greatly strengthened the paper. Thank you for facilitating this submission.

Thanking you,

Your's sincerely,
Nilanjan Biswas, Zijian Wang, Luc Vandendorpe, and Hamed Mirghasemi