Minimizing Upload Latency for Critical Tasks in Cellular-based IoT Networks using Multiple Relays

Shang-Hong Hsu[1], Chi-Han Lin[1], Chih-Yu Wang[2], Wen-Tsuen Chen[1,2,3]
[1]Department of Computer Science, National Tsing-Hua University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[3]Institute of Information Science, Academia Sinica, Taiwan
E-mail:a85547219@gmail.com, finalspaceman@gmail.com, cywang@citi.sinica.edu.tw, chenwt@iis.sinica.edu.tw

***Abstract*—Most existing or developing IoT communication standards are based on the assumption that IoT services only require low data rate transmission and therefore can be supported by limited resources such as narrow-band channels. This assumption rules out those IoT services with burst traffic, critical tasks, and low latency requirements. In this paper, we propose to utilize idle devices in IoT networks to boost the transmission data rate for critical tasks through multiple concurrent transmissions. This approach virtually expands the existing narrow-band IoT protocols to support channel aggregation in order to realize low latency services for critical tasks in IoT networks. We propose task-balance method (TBM) and first-link descending order (FDO) to determine the relay order and data partition in a given relay set. We theoretically prove that the optimal relay configuration that minimizes the uploading latency can be derived in polynomial time. We then show that relay selection problem is NP-hard and propose a greedy algorithm to approximate the optimal solution within a $1/2$ performance lower bound. The simulation results shows that the proposed approach can reduce the latency of critical tasks up to 76% comparing with traditional approaches.**

I. Introduction

Internet of Things (IoT) devices are expected to spring up among us to improve the quality of life in the near future. Most IoT devices have wireless communication capability including local access such as Wi-Fi, bluetooth, and/or long-range access such as LTE, LoRA. The long-range access techniques is more appealing for the larger coverage, but the transmission data rate is usually limited in exchange of better energy efficiency and massive device access supports. This limitation partly comes from the maximum bandwidth each device allows to utilize. For instance, a narrow-band IoT device in LTE systems can only access one channel within 180kHz bandwidth at a time, even if multiple channels exist in the system. This leads to a 250kbps downlink bit-rate and 20kbps uplink bit-rate.

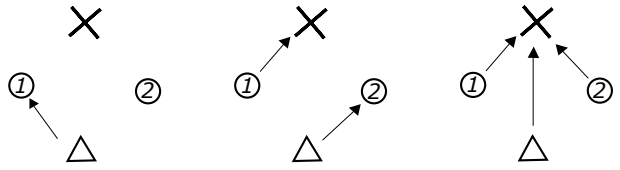
Such a limited per-link data rate may satisfy average requirements of some IoT services but not for those with burst traffic requirements. For instance, a surveillance camera may be monitoring in a lower resolution mode and therefore requires only low data rate. Nevertheless, when an emergency event occurs, it may be required to switch to full resolution mode to collect better images. The captured image also should be uploaded as soon as possible in order to provide real-time information for authorities. The existing narrow-band IoT protocols cannot support this type of applications since such burst uplink traffic will be failed to be delivered in the required latency with the limited per-link data rate.

Device-to-device (D2D) communication refers to exploit hop-by-hop transmissions between multiple users to reduce the base station (BS) intervention. The transmission data rate can be greatly increased due to lower propagation loss. Nevertheless, such a gain comes from the opportunistic proximity between the source and destination, which is not a typical case in cellular-based IoT networks. In celluar-based IoT networks, the final destination of the collected data is mostly a server on the Internet. The devices must deliver the data to a BS first in order to gain the access to Internet. D2D communication is not applicable in such a case.

In many IoT applications, we observe that IoT devices are usually in the idle state or executing non-critical routines. We find that these IoT devices may help boost the transmission rate of certain IoT devices in critical uplink missions by using both limited cellular data rate and D2D communications. The basic idea is to let IoT devices in critical tasks to ask nearby idle devices relaying different portion of the data to the BS through different channels. Through multiple concurrent transmissions, we may boost the transmission data rate and reduce the latency of the critical task. This concept is shown in Fig. 1. At the first step in Fig. 1a, source node transmits a portion of data to device 1. The source then starts transmitting another portion of data to device 2, and device 1 relays the received data to the BS simultaneously, as shown in Fig. 1b. Finally, in Fig. 1c, all devices upload data to the BS concurrently. The advantage of this approach is that the uploading time is overlapped by exploiting neighbor devices' communication capability without breaking the hardware and interface constraints such as low long-range data rate and narrow-band channels. The protocol can be implemented in most existing IoT systems which supports both long-range and short-range access simultaneously, such as narrow-band IoT devices with LTE D2D support, or LoRA devices with Bluetooth or Wi-Fi support.

The idea of multiple concurrent transmission is conceptually

This work is partially supported by the Ministry of Science and Technology of R.O.C. under contract No. MOST105-2221-E-001-003-MY3, MOST103-2218-E-001-002-MY2 and 104-2221-E-001-001-, and by Academia Sinica Thematic Research Program.



(a) Step 1. (b) Step 2. (c) Step 3.

Fig. 1: The concept of boosting. ×: BS. ◯: IoT devices. △: Source node.

similar to carrier aggregation in LTE-Advanced system [1], [2]. Conceptually, these relay devices form a virtual carrier aggregation interface for the critical device in order to utilize multiple channels to deliver the critical task.

Our idea is also similar with cooperative networks. There exists a rich set of literature about cooperative networks involving full-duplex and half-duplex. Some focus on improving resource utilization efficiency and link reliability achieved on full-duplex communication way [3]. Unfortunately, most IoT devices use half-duplex mode to transmit data due to the limitation of hardware. Therefore, we focus on half-duplex relay-based communication in cooperative networks. The performance enhancement brought by the single relay model has been extensively studied [4], [5]. In addition to single relay, multiple relay selection problem is also addressed in cooperative networks [6]–[8]. In these works, multiple relays are organized as a virtual MIMO system and forward source's data to the destination. However, the relays receive the whole source's data by broadcasting, which is different with our work in which each relay only receives and handles parts of the data. In addition, most works focus on improving the reliability of transmissions by exploiting multiple channel diversity. In contrast, our goal is to minimize the transmission latency of critical tasks.

In this paper, we propose a novel semi-sequential relay communication approach to reduce the latency of burst uplink traffic for critical tasks. Our goal is to minimize the upload latency through concurrent transmissions with multiple devices in order to utilize more bandwidth of the system. The challenges of this approach are two folds: 1) selecting the optimal relay node set, and 2) determine optimal relay order and data partition. We first propose task-balance method (TBM) and first-link descending order (FDO) to determine the relay order and data partition in a given relay set. The proposed algorithms sort the helpful idle devices according to their D2D signal quality and calculate the optimal data partition which should be delivered by the relay nodes. We theoretically prove that the algorithm provides the optimal relay configuration that minimizes the uploading latency and runs in polynomial time. We then show that the UE relay selection problem is NP-hard and propose a greedy algorithm which can approximate the optimal solution within a $1/2$ performance lower bound in polynomial time. Our major contributions are as follows:

1) We propose a novel semi-sequential relay communication approach to boost the uplink data rate of IoT devices with limited per-link data rate. The simulation results showed that the proposed approach significantly reduces the latency of burst uplink traffic up to 76%.
2) We prove that there exists an optimal relay order and data partition strategy for a given relay node set to minimize the overall transmission time. In addition, the optimal solution can be found in polynomial time.
3) We show that the multiple relay selection problem is NP-hard, and then an approximation algorithm with 1/2 performance lower bound is proposed

In the rest of this paper, A formal system model and the problem definition are provided in Section II. Then, given a relay order, the optimal relay set and data partition strategy is discussed in Section III. Based on the results from Section III, we then show the hardness of the multiple relay selection problem and then propose a greedy algorithm in Section IV. The simulation results are illustrated in Section V. Finally, we draw our conclusions in Section VI.

II. System Model

We consider a cellular system that consists of one base station (BS), one source device, and N devices. The source requests an upload transmission service and attempts to send D-bit data to the BS. All devices including the source are capable of using D2D communications to perform direct transmissions among devices. Nevertheless, they all equip with exactly one antenna and cannot perform D2D communication and conventional cellular transmission simultaneously. There are $\mathcal{F}$ channels with equal bandwidth available in the system, and each device can utilize at most one channel at a time. Unlike the devices, a BS usually equips with multiple antennas and can simultaneously receive data from $\mathcal{F}$ channels. Our goal is to minimize the upload transmission time, i.e., the upload latency, for the device with critical tasks in cellular-based IoT networks.

A. Semi-sequential Relay Approach

We exploit the idle devices, D2D communications, and all available channels to minimize the transmission time. We assume that the data can be divided into multiple parts, each with different amounts. The idle devices may help relay the data concurrently through transmitting on different channels. We utilize these characteristics and propose a semi-sequential relay approach to transform the single-link transmission from one device to multiple-link concurrent transmissions from multiple relay devices. The key of reducing the total transmission time is to maximize the overlap of the these concurrent transmissions.

An example is illustrated in Fig. 2a, in which the source, denoted by s, is assumed to upload 100M-bit data to the BS. We use this example to illustrate the effect of relay order and partition on the transmission time. In the traditional single-link transmission approach, the source spends 100Mb/6Mbps = 16.67s to finish its transmission. Nevertheless, if source s respectively sends 35, 20, and 40 Mbits to devices 1, 2, and 4 for relay, and finally sends 5 Mbits to the BS by itself, the total upload time could be then reduced by relaying these partitioned data to the BS via devices 1, 2, and 4. Intuitively, the total upload time equals the longest duration among the

concurrent transmissions. In this example, the transmission $s \rightarrow 4 \rightarrow b$ is $\frac{35}{27} + \frac{20}{24} + \frac{40}{28} + \frac{40}{18}$ = 5.78s. Nevertheless, if we adjust the relay order to be devices 4, 1, and 2 with the same data allocation, the longest transmission time occurs on path $s \rightarrow 2 \rightarrow b$, and the total upload time therefore becomes $\frac{40}{28} + \frac{35}{27} + \frac{20}{24} + \frac{20}{24}$ = 4.40s. Clearly, both relay order and data partition may influence the total upload time.

B. Problem Formulation

We formally formulate the semi-sequential relay approach as a relay selection and ordering problem. Assume that source s attempts to send D-bit data to BS b and has N idle neighbors (i.e., relay candidates). Let $\mathcal{F}$ denote the number of channels in the current LTE-system and $\mathbb{M}_s$ denote the set of relay candidates of source s, where $|\mathbb{M}_s| = N$. Let $\mathbb{L}_s = \{l_{s,i} | i \in \mathbb{M}_s\}$ and $\mathbb{L}_b = \{l_{j,b} | j \in \mathbb{M}_s\}$ denote the set of links from source s to relay candidate i and the set of links from relay candidate j to BS b, respectively. Since each link has individual link bandwidth, we further define $\mathbb{C}_s = \{C_{s,i} | l_{s,i} \in \mathbb{L}_s\}$ and $\mathbb{C}_b = \{C_{j,b} | l_{j,b} \in \mathbb{L}_b\}$ as the transmission bit-rate on $l_{s,i}$ and $l_{j,b}$, respectively.

To calculate the total upload time t, the transmission complete time of each P_i is needed, which is related to the relay order of device i and the amount of data relayed through P_i. Let's assume that the source s selects n idle neighbors including itself as its relay set, denoted by $\mathbb{M}'_s \subseteq \mathbb{M}_s \cup \{s\}$. For an $\mathbb{M}'_s$, source s also determines their order of transmission, i.e., relay order, denoted by $\mathcal{M}_s = (m_1, \cdots, m_p, \cdots, m_n, s)$, where $m_p \in \mathbb{M}'_s$. The source s then determines the amount of data each device is responsible to relay. Specifically, it determines an assignment of data allocation of relay order $\mathcal{M}_s$, denoted by $\mathcal{D}_s = (D_1, \cdots, D_p, \cdots, D_n, D_s), \sum_{p=1}^{n} D_p + D_s = D$, where the source s sends D_p-bit data to device m_p sequentially following the relay order $\mathcal{M}_s$. Then, source s directly sends the rest D_s bits to BS b. Intuitively, the total upload time T is the longest transmission complete time among P_{m_p}, i.e., $T = \max(t_{m_p,b} + D_p/C_{m_p,b})$, where $t_{i,b}$ represents the start time that the data are sent on the link from device i to BS b. Thus, the objective of this semi-sequential relay system becomes a min-max problem as follows:

$$\min_{\mathbb{M}'_s, \mathcal{M}_s, \mathcal{D}_s} \max_{m_p \in \mathbb{M}'_s} (t_{m_p,b} + \frac{D_p}{C_{m_p,b}}), \tag{1}$$

where

$$t_{m_p,b} = t_{s,m_p} + \frac{D_p}{C_{s,m_p}}, \tag{2}$$

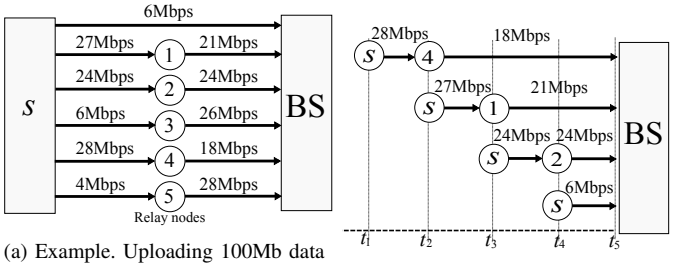
$$t_{s,m_{p+1}} = t_{m_p,b}, \quad 1 \leq p \leq n, \tag{3}$$

$$t_{s,m_1} = 0, \tag{4}$$

subject to

$$D_p \in (0, D], \tag{5}$$

$$\sum_{p=1}^{n} D_p + D_s = D. \tag{6}$$



(a) Example. Uploading 100Mb data from UE to BS.

(b) Relay order.

Fig. 2: The example of semi-sequential relay approach.

$$|\mathbb{M}'_s| \leq \mathcal{F} \tag{7}$$

The upload time of a relay consists of three components: the waiting latency before source s transmits data to the relay (i.e., t_{s,m_p} in Eq. (2)), the propagation delay from source s to device m_p (i.e., $D_p/C_{s,m_p}$ in Eq. (2)), and the propagation delay from relay m_p to BS b (i.e., $D_p/C_{m_p,b}$ in Eq. (1)). Eq. (2) shows the relation of start time between $\mathbb{L}_s$ and $\mathbb{L}_b$ on P_{m_p}. Eq. (3) shows the relation of start time between l_{s,m_p} and $l_{s,m_{p+1}}$, and guarantees that source s transmits data to at most one device at any time during the transmission. In Eq. (4), We show the boundary condition of t_{s,m_p}, i.e., t_{s,m_1}. Since source s requires to send D-bit data to BS b, we formulate the constraints of data size in Eq. (5) and (6). Finally, Eq. (7) shows the maximum number of relay devices can be used.

In Eq. (1), the optimal solution of this problem is to find the optimal relay set $\mathbb{M}'^*_s$, optimal transmission order $\mathcal{M}^*_s$ of $\mathbb{M}'^*_s$, and optimal data allocation $\mathcal{D}^*_s$.

III. Optimal Relay Order and Data Partition

We first consider the case that the number of relay candidates is less than the number of total available channels, i.e., $N < \mathcal{F}$. Under this setting, any relay set $\mathbb{M}'_s$ is feasible as Eq. (7) always holds. We now show that the optimal relay selection, relay order, and data allocation can be derived in polynomial time under this setting.

Intuitively, for any idle neighbor of source s, if the transmission data rate from source s to the neighbor is less than that from s to BS b, the neighbor's relay will never be helpful to reduce the overall transmission time. Thus, the source s should always avoid selecting such neighbors as candidate relays, that is, a device i will be added into the set of relay candidates only if $C_{s,i} > C_{s,b}$ For convenience, for a device m_p in relay order $\mathcal{M}_s$, C_{s,m_p} and $C_{m_p,b}$ are respectively abbreviated to F_p and S_p to represent the data rate of the first and second link. Specially, $C_{s,b}$ is abbreviated by S_s.

In the following, we will show three key properties of the optimal solution, which are 1) the optimal relay set is the set of all relay candidates of source s, 2) the optimal relay order is the order of devices sorted by $C_{s,i}$ in descending order, and 3) the optimal data allocation method is to let the transmissions of all devices (i.e., $m_1, \cdots, m_n$,s) finish simultaneously.

To minimize the transmission time, we first propose task-balance method (TBM). We then provide three theorems as follows: Theorem 1 proves that, given a relay order, our proposed TBM can determine its optimal data allocation, which

makes all transmissions finish simultaneously to minimize the upload time of this relay order. Theorem 2 further points out the optimal relay order of a relay set in TBM is first-link descending order (FDO). Finally, Theorem 3 describes the optimality of FDO combining with TBM among all possible relay strategies in the proposed semi-sequential relay approach.

We first introduce a data allocation method, which lets all transmissions of relay candidates finish simultaneously. The high level concept is to calculate a data allocation ratio between a relay and its next relay in a relay order, which can be achieved by making the transmission time of a relay's second link equal the total transmission time of its next relay. For example, assume that there is a relay order $\mathcal{M}_s = (m_1, m_2, s)$. According to our scenario, relay m_2's data receiving should be after relay m_1's data receiving, and source s finally uploads data after relay m_2's data receiving. Therefore, if a data allocation $\mathcal{D}_s = (D_1, D_2, D_s)$, which makes the transmissions of m_1, m_2, and s complete simultaneously (i.e., $t_{m_1,b} + \frac{D_1}{S_1} = t_{m_2,b} + \frac{D_2}{S_2}$ and $t_{m_2,b} + \frac{D_2}{S_s} = t_{s,b} + \frac{D_s}{S_s}$), exists, we can derive the equation, $\frac{D_1}{S_1} = \frac{D_2}{F_2} + \frac{D_2}{S_2}$ and $\frac{D_2}{S_2} = \frac{D_s}{S_s}$ from Eqs. (2) and (3). Finally, we have $\frac{D_1}{D_2} = \frac{S_1(F_2+S_2)}{F_2 S_2}$ and $\frac{D_2}{D_s} = \frac{S_2}{S_s}$. In general case, the allocation ratio can be formally formulated as $D_1{:}\cdots{:}D_p{:}\cdots{:}D_n{:}D_s = r_1{:}\cdots{:}r_p{:}\cdots{:}r_n{:}r_s$, where

$$r_p = S_p \prod_{j=2}^{p} F_j \prod_{k=p+1}^{n} (F_k + S_k), \text{ and } r_s = S_s \prod_{j=2}^{n} F_j. \quad (8)$$

By applying this allocation method to a relay order $\mathcal{M}_s$, the total transmission time, denoted by $T(\mathcal{M}_s)$, equals the complete time of each relay's transmission (i.e., $T(\mathcal{M}_s) = t_{m_1,b} + \frac{D_1}{S_1} = \frac{D_1}{F_1} + \frac{D_1}{S_1}$). However, the total transmission time can be further reduced by removing a relay m_p from $\mathcal{M}_s$ and re-allocating m_p's data to other devices if

$$T(\mathcal{M}_s) > T(\mathcal{M}_s \backslash \{m_p\}),\ 1 \le p \le n. \quad (9)$$

Due to a limited number of pages, we skip the derivation here. After our derivation based on Eqs. (2)-(4) and (9), the relay m_p should be removed if $X_p \ge 1$, where

$$X_p = \frac{F_{p+1} \sum_{x=p+1}^{n} S_x \prod_{j=p+1}^{x} F_j \prod_{z=x+1}^{n} (F_z + S_z)}{F_p \prod_{x=p+1}^{n} (F_x + S_x)}. \quad (10)$$

From Eq. (10), we observe that: 1) whether m_p should be removed only depends on the data rates of all subsequent devices, i.e., all m_q in $\mathcal{M}_s$ where $p < q \le n$, and 2) the relay m_p should be kept in the relay order $\mathcal{M}_s$ as long as $F_p > F_q, p < q \le n$.

This allocation method and above observations motivate the proposed task-balance method (TBM), as shown in Algorithm 1. Given a relay order $\mathcal{M}_s$, TBM first removes some devices m_p that $X_p \ge 1$, and obtains a new relay order $\mathcal{M}'_s$. TBM then calculates the proportion of data allocation of $\mathcal{M}'_s$. Note that TBM may alter the relay order by removing some of the relays. We define $\mathcal{M}'_s$ is sequentially consistent with $\mathcal{M}_s$ if for every $m_p, m_q \in \mathcal{M}_s, p < q$ and corresponding $m_{p'}$ and $m_{q'}$ in $\mathcal{M}'_s$, we have $p' < q'$ (i.e., the order of $m_{p'}$ is still before $m_{q'}$ in $\mathcal{M}'_s$).

Algorithm 1 Task-balance method (TBM)

Input:
The relay order $\mathcal{M}_s$
Output:
The proportion of data allocation $\mathcal{R}_s$
The new relay order $\mathcal{M}'_s$

```
M'_s ← M_s
for all m_i ∈ M'_s, i from n to 1 do
    Calculate X_p using Eq. (10).
    if X_p ≥ 1 then
        Remove m_p from M'_s
    end if
end for
for all M'_s do
    Calculate r_i using Eq. (8).
end for
for all r_i do
    R_i = r_i / Σ_i r_i
end for
R_s ← (R_1, R_2, ...)
return R_s, M'_s
```

Theorem 1. *Given a relay order $\mathcal{M}_s$, the optimal relay order that is sequentially consistent with $\mathcal{M}_s$ and its optimal data allocation can be obtained by TBM.*

Theorem 1 can be proved by the following two steps: 1) Given a relay order $\mathcal{M}_s$ that $X_p < 1, \forall m_p \in \mathcal{M}_s$, the optimal data allocation can be obtained by TBM. 2) Given any relay order, TBM can find a relay order $\mathcal{M}'_s$ that $X'_p < 1, \forall m'_p \in \mathcal{M}'_s$ and is sequentially consistent with $\mathcal{M}_s$ so that $\mathcal{M}'s$ transmission time is lower than all other relay orders that are sequentially consistent with $\mathcal{M}_s$. The detailed proof is omitted here due to page limitation.

Next, we show that the optimal relay order is the descending order of the data rate of each relay's first link when TBM is performed.

Theorem 2. *The optimal relay order, denoted by $\mathcal{M}^*_s$, is first-link descending order (FDO), or the descending order of the data rate of each device's first link. That is, $\mathcal{M}^*_s = (m^*_1, ..., m^*_p, ..., m^*_n, s)$, where $F^*_p \ge F^*_{p+1}, \forall p \in \{1, ..., n-1\}$.*

Proof. Assume that a non-descending relay order $\mathcal{M}_s = (m_1, ..., m_p, ..., m_n, s)$ exists, and its total transmission time $T(\mathcal{M}_s)$ cannot be reduced by interchanging the order of any two relays. Since $(F_1, ..., F_p, ..., F_n)$ is non-descending, we respectively discuss the following two cases:

Case 1: $F_1 < F_2$

Let $\mathcal{M}'_s = (m_2, m_1, ..., m_n, s)$ be the relay order that relays m_1 and m_2 are interchanged. After allocating data based on Eq. (8), the data allocation ratio is $r'_2{:}r'_1{:}...{:}r'_n{:}r'_s$. Since $T(\mathcal{M}_s)$ is the shortest according to our supposition, we have $T(\mathcal{M}_s) < T(\mathcal{M}'_s)$, which implies

$$\begin{aligned} T(\mathcal{M}_s) &= \frac{D \cdot r_1}{r_1 + r_2 + ... + r_n + r_s} \left(\frac{1}{F_1} + \frac{1}{S_1}\right) \\ &< \frac{D \cdot r'_2}{r'_2 + r'_1 + ... + r'_n + r'_s} \left(\frac{1}{F_2} + \frac{1}{S_2}\right) = T(\mathcal{M}'_s) \\ &\Rightarrow F_2 < F_1, \end{aligned}$$

which is a contradiction. Note that we only calculate the transmission time of the first relay in $\mathcal{M}_s$ and $\mathcal{M}'_s$ since all transmissions are finished simultaneously.

Case 2: $F_p < F_{p+1}, p \in \{2, ..., n-1\}$

Let $\mathcal{M}'_s = (m_1, ..., m_{p+1}, m_p, ..., m_n, s)$ be the relay order that relays m_p and m_{p+1} are interchanged. After allocating data, the data allocation ratio is r'_1:...:r'_{p+1}:r'_p:...:r'_n:r'_s. Since $T(\mathcal{M}_s)$ is the shortest according to our supposition, we have $T(\mathcal{M}_s) < T(\mathcal{M}'_s)$, which implies

$$\begin{aligned}
&\frac{D \cdot r_1}{r_1 + ... + r_p + r_{p+1} + ... + r_n + r_s}(\frac{1}{F_1} + \frac{1}{S_1}) \\
&< \frac{D \cdot r'_1}{r'_1 + ... + r'_{p+1} + r'_p + ... + r'_n + r'_s}(\frac{1}{F_1} + \frac{1}{S_1}) \\
&\Rightarrow F_p > F_{p+1},
\end{aligned}$$

which is also a contradiction.

In summary, based on Case 1 and Case 2, the total transmission time can always be reduced by interchanging the order of two relays m_p and m_{p+1} in $\mathcal{M}_s$ if $F_p < F_{p+1}, \exists p \in \{1, ..., n-1\}$. Similar to bubble sort [9], after several iterations, $T(\mathcal{M}_s)$ is minimized only if $F_p \geq F_{p+1}, \forall p \in \{1, ..., n-1\}$. We defined this order as the first-link descending order (FDO). □

Theorem 3. *If the number of relay candidates is less than the number of channels, i.e., $N < \mathcal{F}$, the upload time from source s to BS b can be minimized by arranging all relay candidates of s according to FDO and then performing TBM.*

Theorem 3 is a direct result from Theorem 1 and Theorem 2 and we omitted the proof here. We therefore can obtain the optimal relay order by FDO and data allocation by TBM. The corresponding upload time is given by Eq. (11). One thing worth noting that the line 1 to 3 of Algorithm 1 is unnecessary when the relay order has been sorted based on FDO since $X_p < 1$ always holds for all devices $m_p \in \mathcal{M}_s$.

We use an example to illustrate the whole procedure of optimizing the upload transmission. In Fig. 2a, devices 1, 2, and 4 are selected as source s's relay candidates due to their higher data rate of first link compared to the data rate from s to BS b. According to FDO, the optimal relay order $\mathcal{M}^*_s$ is (4,1,2, s), which can be found in $\mathcal{O}(N \log N)$ time by sorting. One thing worth noting is that the total number of possible relay orders is $\sum_{m=1}^{N} \binom{N}{m} m!$, which requires tremendous computation compared to FDO. TBM is then performed to determine the data allocation of $\mathcal{M}^*_s$, namely $\mathcal{D}^*_s$. Based on Eq. (8), the allocation ratio of three devices is $18 \cdot (27+21) \cdot (24+24) : 27 \cdot 21 \cdot (24+24) : 27 \cdot 24 \cdot 24 : 27 \cdot 24 \cdot 6 = 32 : 21 : 12 : 3$. Finally, if the total data size $D = 100$ Mbits, devices 1, 2, and 4 respectively forward $\frac{100 \cdot 32}{32+21+12+3}$, $\frac{100 \cdot 21}{32+21+12+3}$, and $\frac{100 \cdot 12}{32+21+12+3}$ Mbits to BS b. Subsequently, source s transmits $\frac{100 \cdot 3}{32+21+12+3}$ to b, as shown in Fig. 2b. By Eq. (11), the total upload time equals $T_{\text{TBM}}(\mathcal{M}^*_s) = 100 \cdot \frac{32}{68} \cdot (\frac{1}{28} + \frac{1}{18}) = 4.29$ s. The upload time of our proposed semi-sequential relay approach is only $\frac{4.29}{(100/6)} = 25.74\%$ to the one under conventional single-link transmission.

IV. Multiple Relay Selection

In this section, we analyze the multiple relay selection problem when $N \geq \mathcal{F}$. $N \geq \mathcal{F}$ comes from the fact that the number of channels or resource blocks is limited in wireless systems. Since some resource may be occupied by other existing transmissions, the resource allocated to source s is limited and, therefore, the number of allowed concurrent transmissions is also limited. As we mentioned in Section III, we assume there are only $\mathcal{F}$ channels, or the system only allows $\mathcal{F}$ concurrent transmissions. Therefore, we can only select up to $\mathcal{F}$ relay nodes including source node.

Recall that, if $N < \mathcal{F}$, given a relay set $\mathbb{M}_s$, the optimal relay order $\mathcal{M}_s$ can be obtained by applying FDO to $\mathbb{M}_s$. When $N \geq \mathcal{F}$, an intuitive scheme is to choose the first $\mathcal{F}$ relays in $\mathcal{M}_s$ according to FDO, which may lead to a sub-optimal solution. For example, in Fig. 2a, if $\mathcal{F} = 2$, (4,1) should be the optimal solution derived by FDO and its transmission time equals $100 \cdot \frac{32}{59}(\frac{1}{28} + \frac{1}{18}) = 4.95$ s. However, the transmission time of relay set (1,2) equals $100 \cdot \frac{7}{12}(\frac{1}{27} + \frac{1}{21}) = 4.93$ s < 4.95 s. Therefore, FDO is not the optimal choice to determine the relay set.

We may show that the relay selection problem is a quadratic constraint optimization problem, which has been shown to be NP-hard [10], [11]. Therefore, we seek approximation algorithms with performance bounds.

A. Greedy Algorithm

We propose a greedy algorithm to obtain the relay set when $N \geq \mathcal{F}$. In each round, from the set of relay candidates, we select a device that can reduce the most transmission time. Let $G(x, \mathcal{M}_s)$ denote the reduced transmission time if device x joins $\mathcal{M}_s$, which can be formulated by

$$G(x, \mathcal{M}_s) = T^{opt}(\mathcal{M}_s) - T^{opt}(\mathcal{M}_s \cup \{x\}). \quad (12)$$

Here, $\mathcal{M}_s \cup \{x\}$ represents that device x is inserted into $\mathcal{M}_s$ before source s. For example, if $\mathcal{M}_s = (m_1, m_2, s)$, $\mathcal{M}_s \cup \{x\} = (m_1, m_2, x, s)$.

Algorithm 2 shows the comprehensive procedure including $N < \mathcal{F}$ and $N \geq \mathcal{F}$. It terminates when the number of selected relays is $\min(\mathcal{F}, |\mathbb{M}_s|)$ or when all relay candidates cannot reduce the time (i.e., $G(x, \mathcal{M}_s) \leq 0$). In summary, when $N \geq \mathcal{F}$, the time complexity is $\mathcal{O}(\mathcal{F}^3)$ due to line 8 of Algorithm 2. when $N < \mathcal{F}$, the time complexity is $\mathcal{O}(N^2)$ due to line 17, i.e., Algorithm 1.

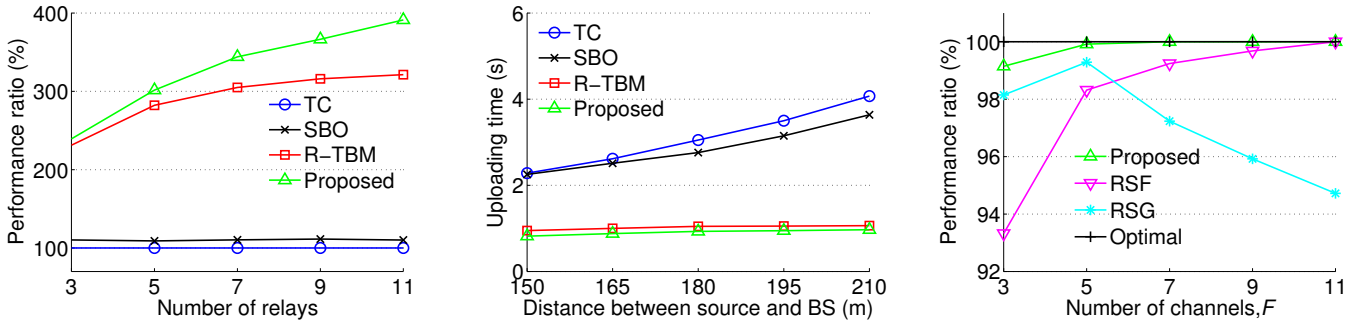
Theorem 4. *The proposed greedy algorithm guarantees a $\frac{1}{2}$-approximation to the optimal solution for the proposed multiple relay selection problem.*

We can prove this by verifying that our objective function (i.e., Eq. (12)) satisfies the properties of matroid and nondecreasing submodular set function [12], [13]. The detailed proof is omitted here due to page limitation.

V. Simulation Results

We use NS3 to simulate the LTE-based IoT network. Our scenario contains two BSs, namely BS 1 and BS 2, at a

$$T^{opt}(\mathcal{R}) = \frac{D\prod_{i=1}^{n-1}(F_i+S_i)}{F_1(S_1\prod_{i=2}^{n-1}(F_i+S_i)+\sum_{i=2}^{n-1}S_i\prod_{j=1}^{i}F_j\prod_{k=i+1}^{n-1}(F_k+S_k)+(S_s+S_n)\prod_{i=1}^{n}F_i)} \tag{11}$$



(a) Effect of relay nodes (b) Effect of distance (c) Performance gap

Fig. 3: Simulation results

Algorithm 2 Proposed Method for Semi-Sequential Relay

Input:
The set of relay candidates: $\mathbb{M}_s$
Number of available channels: $\mathcal{F}$
The data size of source s: D
Output:
The relay order: $\mathcal{M}_s$
The data allocation of $\mathcal{M}_s$: $\mathcal{D}_s$

$N \leftarrow |\mathbb{M}_s|$
$\mathcal{M}_s \leftarrow (s)$
if $N < \mathcal{F}$ **then**
  $\mathcal{M}_s \leftarrow \mathcal{M}_s \cup \mathbb{M}_s$
  Re-arranging $\mathcal{M}_s$ by FDO
else
  for $i = 1$ *to* $\mathcal{F}$ **do**
    Find $x = \arg_{x\in\mathbb{M}_s}\max(G(x,\mathcal{M}_s))$ by Eq. (12)
    if $G(x,\mathcal{M}_s) \leq 0$ **then**
      break;
    end if
    $\mathbb{M}_s \leftarrow \mathbb{M}_s \backslash \{x\}$
    $\mathcal{M}_s \leftarrow \mathcal{M}_s \cup \{x\}$
    Re-arranging $\mathcal{M}_s$ by FDO
  end for
end if
$(R_1, ..., R_i, ...) \leftarrow$TBM($\mathcal{M}_s$)
$\mathcal{D}_s \leftarrow D \cdot (R_1, ..., R_i, ...)$
return $\mathcal{M}_s$, $\mathcal{D}_s$

distance of 500 m. We first randomly scatter 30 devices around BS 1 within a radius of 250 m, each of which connects to BS 1. These devices are treated as typical LTE users and have occupied certain resources in the LTE network. We then randomly scatter 50 devices around BS 2 within a radius of 250 m. These are treated as inter-cell interference sources. We deploy the IoT network within BS 1 by additionally scattering one critical node (treated as the source device) and several normal nodes (treated as relay candidates) near BS 1, in the cell, all of which are connected to BS 1. In addition, each relay candidate connects to the source node through LTE D2D connections. The remaining simulation parameters are listed in Table I.

Our proposed method, denoted by *Proposed* in the simulation, is based on Algorithm 2 combining TBM and FDO in Section III and greedy relay selection in Section IV. For comparison, we use 5 other naive methods in the simulations: 1) *Traditional Communication (TC):* the traditional approach that the source directly transmits data to BS 1, 2) *Single-Best Relay only (SBO):* a cooperative approach that selects only one best relay node to relay. For the scenario that the number of channels is more than the number of relay candidates, i.e., $N < \mathcal{F}$, we further simulate 3) *Random ordering with TBM (R-TBM):* a random relay order which is applied with TBM. This helps us justify the optimality of proposed FDO method. For the case that the number of channels is less than or equal to the number of relay candidates ($\mathcal{F} \leq N$), we further simulate two different relay node selection and ordering methods: 4) *Gain-based Relay Selection and Ordering (RS-G):* the relay set is determined by Eq. (12) and its order is in descending order, and 5) *First-Link-based Relay Selection and Ordering (RS-F):* determine the relay order based on FDO and choose the first $\mathcal{F}$ relays to be the relay set.

TABLE I: Simulation Configuration

Parameter	Value
Carrier Frequency	2 GHz [14]
Uploading File Size	10 Mbits
eNodeB Tx power	46 dBm [15]
D2D node Tx power	23 dBm [15]
Path loss (cell link)	128.1+37.6log R [15]
Path loss (D2D link)	46.8+16.9log R [14]
Shadowing loss	$N(0, 8$ dB$)$ [15]
Available MCSs	29 possible MCSs [16]
TTI	1 ms [15]
Granularity of CSI feedback	100 TTIs [15]
Granularity of scheduling	1 TTI [15]
♯ of Runs	50

A. Number of Relay Nodes

In the first simulation, we show how the number of relay candidates influence the performance gain. The distance between source node and BS is 180 m and we randomly scatter a given number (3 - 11) of relay candidates within a radius of 50 m. We simulate with 50 channels. In other words, there is

no constraint on the number of relay nodes. We measure the performance of each method by the average upload time of the critical task.

In Fig. 3a, we let the performance of *Traditional Communication* be the baseline and show the performance gain of other methods. We observe that all methods involving multiple relay (*Proposed* and *R-TBM*) provides a significant performance boost, that is, a significant reduction in the transmission time. The gain increase as more and more relay candidates are available to help transmission despite the fact that the improvement is diminishing as the number of nodes increases. We also observe that Proposed method performs better than *R-TBM* thanks to the optimal relay order derived by FDO.

B. Distance

Fig. 3b illustrates the upload time of proposed methods versus the distance between the source with BS. We scatter 7 relay candidates within a radius of 50 meters of the source node to help transmit its data. The transmission time of *Traditional Communication* and *SBO* significantly increases as the distance increases. *Proposed* and *R-TBM*, on the other hand, maintain a stable performance regardless of the distance. It should be noted that with the increase of the distance between the source and BS, the improvement provided by the proposed method becomes more significant. It eventually achieves 4.2 times performance improvement at 210 meters.

C. Relay Selection

Finally, we exam the performance of proposed relay selection process. We distribute 11 relay candidates but limit the the number of available channels to $\mathcal{F}$. Therefore, the source node can request at most $\mathcal{F}$ devices to transmit the data. We first use the Brute force method to derive the optimal relay set which minimize the transmission time and let it be the baseline. We compare it with the three kind of different selection-ordering method: *Proposed*, *RS-G*, and *RS-F*. The results are shown in Fig. 3c. The performance of *RS-G* is better than that of *RS-F* when there are less relay devices available, but worse when the number of available channels increases. This is due to the differences in relay selection process. RS-G can correctly identify those relay nodes who potentially contribute most to the reduction in transmission time at early stage, but the performance loss in suboptimal relay order eventually becomes significant when the number of channels increases. On the other hand, *RS-F* provides the optimal relay order in any given relay set. When the relay set is close to the complete relay candidate set, the solution is closed to the optimal solution provided by the Brute algorithm. The performance of *Proposed* method is the best of the three and is approximate to optimal. This is because it selects appropriate relay nodes and sorts them in the correct order.

VI. Conclusions

In this paper, we design a novel semi-sequential relay communication approach for cellular-based IoT networks to reduce the latency of critical tasks through concurrent transmissions with multiple idle IoT devices. We formulate the problem as a relay selection and ordering problem. We theoretically prove that the proposed algorithms, TBM and FDO, provide the optimal relay configuration that minimizes the uploading latency when the relay set is given. The optimal solution for the relay selection problem, which is NP-hard in general, can be approximated by the proposed greedy algorithm with a $1/2$ performance lower bound in polynomial time. The performance enhancement is evaluated through simulations. We observed that the proposed approach can reduce the transmission time of critical tasks up to 76%.

References

[1] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger, and F. Xu, "Overview of 3GPP LTE-Advanced carrier aggregation for 4G wireless communications," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 122–130, February 2012.

[2] S. Wang and J. S. Thompson, "Signal processing implementation of virtual carrier for supporting M2M systems based on LTE," in *Proceedings of IEEE VTC-Spring*, May 2015, pp. 1–5.

[3] D. Korpi, J. Tamminen, M. Turunen, T. Huusari, Y. S. Choi, L. Anttila, S. Talwar, and M. Valkama, "Full-duplex mobile device: pushing the limits," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 80–87, September 2016.

[4] Y.-W. P. Hong, W.-J. Huang, and C.-C. J. Kuo, *Cooperative communications and networking: technologies and system design*. Springer Science & Business Media, 2010.

[5] B. Talha and M. Ptzold, "Channel models for mobile-to-mobile cooperative communication systems: A state of the art review," *IEEE Vehicular Technology Magazine*, vol. 6, no. 2, pp. 33–43, June 2011.

[6] Z. Yi and I. m. Kim, "Diversity order analysis of the decode-and-forward cooperative networks with relay selection," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1792–1799, May 2008.

[7] S. Narayanan, M. D. Renzo, F. Graziosi, and H. Haas, "Distributed spatial modulation: A cooperative diversity protocol for half-duplex relay-aided wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 2947–2964, May 2016.

[8] R. Mesleh, S. S. Ikki, E.-H. M. Aggoune, and A. Mansour, "Performance analysis of space shift keying (SSK) modulation with multiple cooperative relays," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 201, 2012.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. The MIT Press, 2009.

[10] P. M. Pardalos and G. Schnitger, "Checking local optimality in constrained quadratic programming is NP-hard," *Operations Research Letters*, vol. 7, pp. 33–35, November 1978.

[11] K. G. Murty and S. N. Kabadi, "Some NP-Complete problems in quadratic and nonlinear programming," *Mathematical programming*, vol. 39, no. 1, pp. 117–129, March 1987.

[12] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[13] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, *An analysis of approximations for maximizing submodular set functions—II*, M. L. Balinski and A. J. Hoffman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1978.

[14] M. Condoluci, L. Militano, A. Orsino, J. Alonso-Zarate, and G. Araniti, "LTE-direct vs. WiFi-direct for machine-type communications over LTE-A systems," in *Proceedings of IEEE PIMRC*, 2015, pp. 2298–2302.

[15] H. S. Liao, P. Y. Chen, and W. T. Chen, "An efficient downlink radio resource allocation with carrier aggregation in LTE-Advanced networks," *IEEE Transactions on Mobile Computing*, vol. 13, no. 10, pp. 2229–2239, October 2014.

[16] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2 (Release 12)*, December 2014.