A Computation-Communication Tradeoff Study for Mobile Edge Computing Networks

Kuikui Li, Meixia Tao, Zhiyong Chen
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, P. R. China
Email: {kuikuili, mxtao, zhiyongchen}@sjtu.edu.cn

***Abstract*—In this paper, we exploit computation replication to reduce the communication time for task offloading in mobile edge computing networks, by introducing a tradeoff between computation load, and communication latency defined as a pair of the normalized uploading time (NULT) and normalized downloading time (NDLT). The key idea of replication is to let mobile users offload their tasks to multiple edge nodes for repetitive execution so as to enable the transmission cooperation in computed results downloading, and consequent interference mitigation across users. We develop an achievable communication latency pair at a given computation load, where the NULT is optimal and the NDLT is within a multiplicative gap of 2 to an information theoretic lower bound. We show that in the specific interval, the NDLT can be traded by the computation load in an inversely proportional function.**

I. Introduction

Mobile edge computing (MEC) has emerged as a promising paradigm to assist mobile users in offloading heavy computational tasks to edge servers for execution [1]. The completion of task offloading consists of two delivery phases, where mobile users first *upload* the input data of tasks to MEC servers via the uplink channel, and then *download* the computed results through the downlink channel [2]–[4]. However, wireless network is becoming interference-limited due to the finite bandwidth resource. It forces the data delivery to be the bottleneck of task offloading, especially for many emerging mobile applications requiring large amounts of data transfer over wireless networks, like virtual and augmented reality [5]. The severe interference can significantly increase the total delivery time consumed in these types of task offloading.

Motivated by this issue, our work aims to design a novel task offloading scheme to reduce the communication latency in multi-user multi-server MEC networks. In specific, we consider an MEC network, where a set of N mobile users offload computationally heavy tasks to a set of M computing-enabled edge nodes (ENs). Each task has an input data denoted as W_j and an output data denoted as $\widetilde{W}_j$, $j = 1, 2, \cdots, N$. We define the *computation load* r as the average number of ENs to compute a task, i.e., the degrees of the replication for executing a task. The *communication latency* is defined as the transmission time pair, denoted as (τ^u, τ^d), for uploading the task inputs and downloading the computed outputs. It naturally arises the following fundamental question:

- *Given a computation load r, what is the minimum achievable communication latency (τ^u, τ^d) for completing task offloading in such MEC systems?*

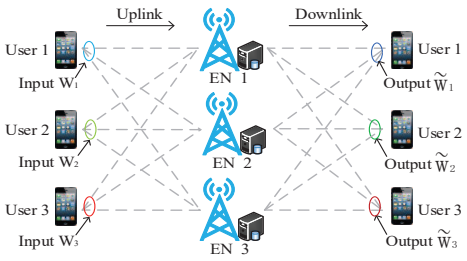

Fig. 1. A three-user three-server MEC network. This figure only shows the case that computation load $r = 3$. The uplink channel is the multicast interference channel whose per-receiver DoF is 1 achieved via TDMA, while the downlink is the MISO broadcast channel with full transmitter cooperation whose per-receiver DoF is 1 achieved by using zero-forcing precoding. Solid circle denotes that the channels inside it carry the same information.

Our work attempts to address the above question by proposing a novel task offloading scheme with computation replication. The main idea of computation replication is to let mobile users offload their computation tasks to multiple ENs for repeated execution so as to enable the multiple ENs to cooperatively transmit the computed results back to users in the downlink. This transmission cooperation can neutralize interferences across users, and hence reduce the downloading time.

We illustrate the advantage of computation replication through an example in 3-user 3-server MEC networks. First, consider that each user offloads its task to all 3 servers for repeated computing. As shown in Fig. 1, in the uploading phase, each user takes turn to multicast its task input to all 3 ENs, consuming 3 time slots in total. In the downloading phase, all the ENs have the same message to transmit, and the downlink channel thus becomes a virtual MISO broadcast channel, and all computed results can be delivered to users within 1 time slot by using zero-forcing precoding [6]. For the baseline scheme where each task is offloaded to an individual EN, both uplink and downlink channels are 3-user interference channels [7]. Both task uploading and downloading can be completed within 2 time slots by using interference alignment. It is seen that by computation replication, the downloading time is reduced from 2 time slots to 1 time slot while the uploading time is increased from 2 time slots to 3 time slots. For those computation tasks whose output data size is much larger than the input data size or the input data size is negligible, computation replication can bring significant benefits in reducing the overall communication latency. This example only presents a particular case $r=3$ for $M=N=3$. A general case for any feasible computation load r in the MEC network with M servers and N users is studied in this work.

Contributions. We consider binary offloading where the computation tasks are not dividable. For a given computation load r, we propose a task assignment scheme where each task is offloaded to r different ENs for repeated computing, and each EN has an even assignment of $\frac{Nr}{M}$ tasks. By utilizing the duplicated computation results on multiple ENs, transmission cooperation can be exploited in the data downloading phase to mitigate multi-user interferences via interference neutralization, and hence improve the transmission rate of the downlink channel. The main distinction in the communication latency (τ^u, τ^d) analysis lies at the degree of freedom (DoF) analysis of circular cooperative interference-multicast channels. We obtain the optimal per-receiver DoF for the uplink channel, and an order-optimal achievable per-receiver DoF for the downlink channel. Based on these DoF regions, we develop an order-optimal achievable communication latency pair at computation load r, where the NULT is optimal compared to a lower bound obtained by using genie-aided arguments, and the NDLT is within a multiplicative gap of 2 to a lower bound obtained based on the optimal DoF of MISO broadcast channels.

Moreover, we reveal that the NDLT is an *inversely proportional function* in the interval $\frac{MN}{M+N} \leq r \leq M$, which presents the *computation-communication tradeoff*. We further reveal that the decrease of NDLT is also at the expense of the increase in the NULT, which thus forms another NULT-NDLT tradeoff.

Related works. The idea of computation replication has been utilized in distributed computer systems, like MapReduce and Spark, to enable coded multicast opportunities for data shuffling across servers [8], or alleviate the random server straggling to shorten the response time [9], [10]. These frameworks are different from the task uploading and downloading procedures between users and ENs in wireless MEC systems. Our work is an attempt to exploit this idea in MEC systems to enable data-level transmission cooperation in downloading the computed results to mitigate the interference across users in wireless networks. This is motivated by the interference neutralization technique enabled by transmitter cooperation in interference networks [6], [11].

Note that transmitter cooperation has been utilized in our previous work [12] that considers partial offloading where each task can be divided arbitrarily, and the uplink and downlink channels are cooperative X-multicast channels. In this paper, we consider binary offloading, and the uplink and downlink channels become circular cooperative interference-multicast channels. Moreover, [12] aims to minimize the total communication and computation time, and hence cannot give a formulation on the relationship between computation load and communication latency in MEC systems.

Notations: $[a:b]$ denotes the set $\{a+1, a+2, \ldots, b\}$, $[K]$ denotes the set $\{1, 2, \ldots, K\}$.

II. Problem Formulation

A. MEC Network Model

We consider an MEC network consisting of M single-antenna ENs and N single-antenna users, as shown in Fig. 1 with $M = N = 3$. Each EN is equipped with a computing server and they all communicate with all users via a shared wireless channel. Denote by $\mathcal{M} = \{1, 2, \ldots, M\}$ the set of ENs and $\mathcal{N} = \{1, 2, \ldots, N\}$ the set of users. The communication link between each EN and each user experiences both channel fading and an additive white Gaussian noise. Let $h_{ij}(g_{ji})$ denote the uplink (downlink) channel fading from user $j \in \mathcal{N}$ (EN $i \in \mathcal{M}$) to EN $i \in \mathcal{M}$ (user $j \in \mathcal{N}$). It is assumed to be independent and identically distributed (i.i.d.) as some continuous distribution.

The network is time-slotted. At each time slot, each user generates an independent computation task to be offloaded to the ENs for execution. The computation task on each user j, for $j \in \mathcal{N}$, is characterized by the input data to be computed, denoted as W_j, with size $|W_j| = L$ bits, the computed output data, denoted as $\widetilde{W}_j$, with size $|\widetilde{W}_j| = \widetilde{L}$ bits[1]. We consider the binary computation task offloading model [2], which requires a task to be executed as a whole. It is suitable for simple tasks that are tightly integrated in structure and are not separable.

B. Task Offloading Procedure

Before the task offloading procedure begins, the system needs to decide which EN or which set of ENs should each task be assigned to for execution. We denote by $\mathcal{T}_i$ the set of tasks assigned to EN i, $i \in \mathcal{M}$. Every task must be computed, so we have $\bigcup_{i\in\mathcal{M}} \mathcal{T}_i = \{W_1, \cdots, W_N\}$.

Definition 1. *For a given task assignment scheme $\{\mathcal{T}_i\}$, the computation load r, $1 \leq r \leq M$, is defined as the total number of tasks computed at all the M ENs, normalized by the number of tasks N, i.e.,*

$$r \triangleq \frac{\sum_{i\in\mathcal{M}} |\mathcal{T}_i|}{N} \tag{1}$$

Similar to [8], the computation load r can be interpreted as the average number of ENs to compute each task and hence is a measure of computation repetition.

Given a feasible task assignment strategy $\{\mathcal{T}_i\}$, the overall offloading procedure contains two delivery phases, an input data uploading phase and an output data downloading phase.

1) Uploading phase: Each user j employs an encoding function to map its task inputs W_j and channel coefficients $\mathbf{H} \triangleq [h_{ij}]_{i\in\mathcal{M}, j\in\mathcal{N}}$ to a length-T^u codeword $\mathbf{X}_j \triangleq (X_j(t))_{t=1}^{T^u}$, where $X_j(t) \in \mathbb{C}$ is the transmitted symbol at time $t \in [T^u]$. Each codeword has an average power constraint of P_u, i.e., $\frac{1}{T^u}||\mathbf{X}_j|| \leq P_u$. Then, the received signal $Y_i(t) \in \mathbb{C}$ of each EN i at time $t \in [T^u]$ is given by

$$Y_i(t) = \sum_{j=1}^{N} h_{ij}(t) X_j(t) + Z_i(t), \qquad \forall i \in \mathcal{M}, \tag{2}$$

where $Z_i(t) \sim \mathcal{CN}(0, 1)$ is the noise at EN i. Each EN i uses a decoding function to map received signals $(Y_i(t))_{t=1}^{T^u}$ and channel coefficients $\mathbf{H}$ to the estimate $\{\hat{W}_j : W_j \in \mathcal{T}_i\}$ of its assigned task inputs. The error probability is given by

$$P_e^u = \max_{i\in\mathcal{M}} \mathbb{P}\left(\bigcup_{W_j\in\mathcal{T}_i} \left\{ \hat{W}_j \neq W_j \right\} \right). \tag{3}$$

[1]Note that the extension to the general case where each task has distinct input (or output) data size causes the uploading (or downloading) times for different tasks unaligned, which is intractable for further analysis.

2) Downloading phase: After receiving the assigned task input data and executing them at the server, each EN i obtains the output data of its assigned tasks, $\left\{\widetilde{W}_j : W_j \in \mathcal{T}_i\right\}$, and begins to transmit these computed results back to users. The computed results downloading is similar to the task uploading operation. Briefly, each EN encodes the task outputs $\left\{\widetilde{W}_j : W_j \in \mathcal{T}_i\right\}$ and channel coefficients $\mathbf{G} \triangleq [g_{ji}]_{j\in\mathcal{N}, i\in\mathcal{M}}$ into a codeword of block length T^d over the downlink interference channel, with an average power constraint of P_d. Each user j decodes its desired task output data $\widetilde{W}_j$ from its received signals and $\mathbf{G}$, and obtain the estimate $\widehat{\widetilde{W}}_j$. The error probability is given by

$$P_e^d = \max_{j\in\mathcal{N}} \mathbb{P}\left(\widehat{\widetilde{W}}_j \neq \widetilde{W}_j\right). \tag{4}$$

A task offloading policy, denoted as $\left(\{W_{j,\Phi}\}, \left(L, \widetilde{L}\right), r\right)$, with computation load r consists of a sequence of task assignment schemes $\{W_{j,\Phi}\}$, task input uploading schemes with time T^u, and task output downloading schemes with time T^d, indexed by the task input and output data size pair $\left(L, \widetilde{L}\right)$. It is said to be feasible when the error probabilities P_e^u and P_e^d approach to zero when $L \to \infty$ and $\widetilde{L} \to \infty$.

C. Performance Metric

We characterize the performance of the considered MEC network by the computation load r as well as the asymptotic communication times for task input uploading and output downloading.

Definition 2. *The normalized uploading time (NULT) and normalized downloading time (NDLT) for a given feasible task offloading policy with computation load r are defined, respectively, as*

$$\tau^u(r) \triangleq \lim_{P_u\to\infty} \lim_{L\to\infty} \frac{\mathbb{E}_{\mathbf{H}}[T^u]}{L/\log P_u}, \tag{5}$$

$$\tau^d(r) \triangleq \lim_{P_d\to\infty} \lim_{\widetilde{L}\to\infty} \frac{\mathbb{E}_{\mathbf{H}}[T^d]}{\widetilde{L}/\log P_d}. \tag{6}$$

The NULT (or NDLT) definition follows the NDT defined in [13]. Note that $L/\log P_d$ (or $\widetilde{L}/\log P_d$) is the reference time to transmit the task input (or output) data of L (or $\widetilde{L}$) bits for one task in a Gaussian point-to-point baseline system in the high SNR regime. Thus, an NULT (or NDLT) of $\tau^u(r)$ (or $\tau^d(r)$) indicates that the time required to upload (or download) the tasks of all users is $\tau^u(r)$ (or $\tau^d(r)$) times of this reference time period.

Definition 3. *A communication latency pair $(\tau^u(r), \tau^d(r))$ at a computation load r is said to be achievable if there exists a feasible task offloading policy $\left(\{W_{j,\Phi}\}, \left(L, \widetilde{L}\right), r\right)$. The optimal communication latency region is the closure of the set of all achievable communication latency pairs $\left\{(\tau^u(r), \tau^d(r)\right\}$ at all possible computation load r's, i.e.,*

$$\mathscr{T}(r) \triangleq \text{closure}\left\{\left(\tau^u(r), \tau^d(r)\right) : \forall \left(\tau^u(r), \tau^d(r)\right) \text{ is achievable}, \forall r \in [1, M]\right\}. \tag{7}$$

Our goal is to characterize the optimal communication latency pair at any given feasible computation load r.

III. Main Results

Theorem 1. *(Inner bound). An achievable communication latency pair $\left(\tau_a^u(r), \tau_a^d(r)\right)$ at an integer computation load $r \in [M]$, for binary task offloading in the MEC network with M ENs and N users, is given by*

$$\tau_a^u(r) = \min\left\{1 + \frac{Nr}{M}, N\right\}, \tag{8}$$

$$\tau_a^d(r) = \min\left\{1 + \frac{N}{M}, \frac{N}{r}\right\}, \tag{9}$$

when $\frac{N}{M} \in \mathbb{Z}^+$. If $\frac{N}{M}$ is not an integer, we can always find δ_1 and δ_2 such that $\frac{N+\delta_1}{M-\delta_2}$ is the closest integer to $\frac{N}{M}$ and the above results still hold by replacing N with $N+\delta_1$ and M with $M-\delta_2$. An inner bound of the optimal communication latency region, denoted as $\mathscr{T}_{in}$, is given by the union of all achievable latency pairs $(\tau^u(r), \tau^d(r))$ at all integer computation load r's satisfying $\tau^u(r) \geq \tau_a^u(r)$ and $\tau^d(r) \geq \tau_a^u(r)$ for $\forall r \in [M]$.

We prove the achievability of Theorem 1 in Section IV.

Theorem 2. *(Outer bound). The optimal communication latency pair $\left(\tau^{u^*}(r), \tau^{d^*}(r)\right)$ at any feasible computation load $r \in \left\{r : \sum_{i=1}^M a_i = Nr, a_i \in [0:N], \forall i \in \mathcal{M}\right\}$[2], for binary task offloading in the MEC network with M ENs and N users, is lower bounded by*

$$\tau^{u^*}(r) \geq \min\left\{1 + \frac{Nr}{M}, N\right\}, \tag{10}$$

$$\tau^{d^*}(r) \geq \frac{N}{\min\{M, N\}}. \tag{11}$$

An outer bound of the optimal communication latency region, denoted as $\mathscr{T}_{out}$, is given by the union of all latency pairs $(\tau^u(r), \tau^d(r))$ at all feasible computation load r's satisfying

$$\tau^u(r) \geq \left\{\frac{Nr}{M} + 1, N\right\}, \ \tau^d(r) \geq \frac{N}{\min\{M, N\}} \tag{12}$$

for $\forall r \in \left\{r : \sum_{i=1}^M a_i = Nr, a_i \in [0:N], \forall i \in \mathcal{M}\right\}$.

The full proof of Theorem 2 is given in Section V. Here, we give the main idea of converse proof. First, we use genie-aided arguments to derive a lower bound on the NULT of any given feasible task assignment with computation load r. Then, we optimize the lower bound over all feasible task assignment schemes to obtain the minimum NULT for a given computation load r. For NDLT, since the downlink channel capacity in this problem cannot exceed that of a virtual MISO broadcast channel with N single-antenna receivers and a M-antenna transmitter, so we have the lower bound of NDLT in (11). By comparing the achievable NULT and NDLT to their lower bounds, we prove the optimality of NULT and the gap of NDLT, which are given below.

Corollary 1. *(Gap of NULT, NDLT). At a computation load $r \in [M]$, the achievable NULT in (8) is **optimal**, and the*

[2]Note that different from the achievable inner bound in Theorem 1, the outer bound in converse proof given in Section V holds for any feasible r for binary offloading, and this feasible region contains the integer set $\{r : r \in [M]\}$.

achievable NDLT in (9) is within a multiplicative gap of 2 to its minimum.

The proof of Corollary 1 is also given in Section V.

Fig. 2 shows the inner bound $\mathscr{T}_{in}$ and outer bound $\mathscr{T}_{out}$ of the optimal communication latency region in the MEC network with $M=N=10$.

Now, we demonstrate how the computation load r affects the achievable communication latency $(\tau_a^u(r), \tau_a^d(r))$. By discussing the $\min$ function terms in Eq. (8) and Eq. (9), we have the monotonicity of the achievable computation-communication function $(\tau_a^u(r), \tau_a^d(r))$:

- The NULT $\tau_a^u(r)$ increases strictly with the computation load r for $1 \le r \le M - \frac{M}{N}$, and then keeps a constant N for $M - \frac{M}{N} \le r \le M$.
- The NDLT $\tau_a^d(r)$ keeps a constant $1+\frac{N}{M}$ for $1 \le r \le \frac{MN}{M+N}$, and then is **inversely proportional** to the computation load r for $\frac{MN}{M+N} \le r \le M$.

Remark 1. *The achievable computation-communication function $(\tau_a^u(r), \tau_a^d(r))$ has two corner points $\left(1+\frac{N^2}{M+N}, 1+\frac{N}{M}\right)$ and $\left(N, \frac{N^2}{MN-M}\right)$, corresponding to $r=\frac{MN}{M+N}$ and $r=M-\frac{M}{N}$, respectively. They are explained as follows:*

- *For input data uploading, before r increases to $M-\frac{M}{N}$, the NULT is increasing since more traffic is introduced in the uplink. When r grows to more than $M-\frac{M}{N}$, there is no need to increase the NULT since all tasks can be uploaded by using TDMA within N time slots.*
- *For output data downloading, before r increases to $\frac{MN}{M+N}$, the increasing transmission cooperation gain brought by computation replication cannot exceed the fixed interference alignment gain in terms of the NDLT reduction, and hence only interference alignment is utilized and the NDLT keeps fixed. When r grows to more than $\frac{MN}{M+N}$, the transmission cooperation gain is larger than the fixed interference alignment gain, transmitter cooperation is thus utilized in data downloading and the NDLT begins to decrease with r.*

It can be easily proved that $M-\frac{M}{N} \ge \frac{MN}{M+N}$ due to $M, N \ge 2$. Hence, we have the following remark to reveal the tradeoff between computation load and communication latency, and illustrate the interaction between the NULT and NDLT.

Remark 2. *It is concluded that there exists a **computation-communication tradeoff** in the specific interval of the computation load, namely, the achievable NDLT $\tau_a^d(r)$ can be decreased in an **inversely proportional** way by increasing the computation load in the interval $\frac{MN}{M+N} \le r \le M$. In addition, in this tradeoff, the NDLT is decreased also at the expense of the increase in NULT, which thus forms another NULT and NDLT tradeoff. Hence, computation replication can significantly shorten the communication latency for offloading the applications whose computed results delivery time dominates the entire offloading time, such as AR applications of which the output data size is much larger than the input data size.*

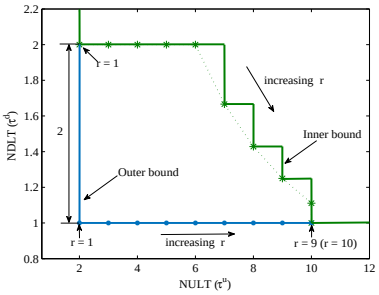

Fig. 2. The inner bound $\mathscr{T}_{in}$ and outer bound $\mathscr{T}_{out}$ of the optimal communication latency $\mathscr{T}$ in the MEC network with $M=N=10$. The achievable NULT is optimal while the gap of NDLT is within 2.

It is further seen from Fig. 2 that the achievable inner bound is composed of three sections corresponding to three different intervals of the computation load, and the lower envelop of the inner bound at $5 \le r \le 9$ presents the tradeoff between NULT and NDLT, which is in an inversely proportional form.

IV. Achievable Task Offloading Scheme

1) Task uploading: Consider that the system parameters M and N satisfy $\frac{N}{M} \in \mathbb{Z}^+$ such that $Nr=Mn$ holds for $\forall r \in [M]$, where r is the integer computation load and n is an integer in $[N]$. By doing so, in the proposed task assignment method, we let each task be executed at exactly r different ENs and let each EN execute n distinct tasks with even load. Note that if $\frac{N}{M}$ is not an integer, we can inject δ_1 (≥ 0) tasks, or let δ_2 (≥ 0) ENs idle and use the remaining ENs for task offloading, such that $\frac{N+\delta_1}{M-\delta_2}$ is the integer closest to $\frac{N}{M}$, denoted as n_1. In this way, we still have $(N+\delta_1)r=(M-\delta_2)rn_1$ for $\forall r \in [M-\delta_2]$, and can use the new $N+\delta_1$ and $M-\delta_2$ to replace N and M to obtain the corresponding analytical results.

To ensure even task assignment on each EN, we perform circular assignment. Specifically, the set of tasks assigned to EN $i \in \mathcal{M}$ is given by

$$\mathcal{T}_i = \left\{W_{j+1} : j \in [(i-1)n : (in-1)] \,(\mathrm{mod}\, N)\right\}. \quad (13)$$

An example of the task uploading for $M=N=4$ and $r=3$ is shown in Fig. 3.

Given the above task assignment in (13), the uplink channel formed by uploading the N tasks to their corresponding ENs is referred to as *the circular interference-multicast channel with multicast group size r*. This channel is different from the X-multicast channel with multicast group size r defined in [14], [15], where any subset of r receivers can form a multicast group, resulting in $\binom{M}{r}$ multicast groups, and each transmitter needs to communicate with all the $\binom{M}{r}$ multicast groups. In our considered circular interference-multicast channel, there are only N multicast groups which are performed circularly by the M receivers and each transmitter only needs to communicate with one multicast group. The optimal per-receiver DoF of this uplink channel is given as follows.

Lemma 1. *The optimal per-receiver DoF of the circular interference-multicast channel with N transmitters and M receivers satisfying $\frac{N}{M} \in \mathbb{Z}^+$ and multicast group size r is*

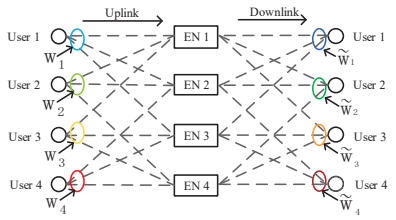

Fig. 3. Illustration of even task assignment for $M = N = 4$ and $r = 3$. The tasks *circularly* assigned to 4 ENs are $\{W_1, W_2, W_3\}$, $\{W_4, W_1, W_2\}$, $\{W_3, W_4, W_1\}$, and $\{W_2, W_3, W_4\}$. The uplink channel is the circular interference-multicast channel whose per-receiver DoF is $\frac{3}{4}$, while the downlink is the circular cooperative interference channel whose per-receiver DoF is $\frac{3}{4}$. Solid circle denotes that the channels inside it carry the same information.

given by

$$DoF_r^u = \max\left\{\frac{Nr}{Nr+M}, \frac{r}{M}\right\},\ r \in [M]. \tag{14}$$

Proof: First, we use partial interference alignment scheme proposed in [7] to achieve a DoF of $\frac{Nr}{Nr+M}$ for each receiver. Then, we compare it to the DoF of $\frac{r}{M}$ achieved by TDMA. The detailed achievable scheme and optimality proof are given in Appendix A. ■

The traffic load for each EN to receive its assigned tasks is $\frac{Nr}{M}L$ bits. Based on [13, Remark 3] and [14, Remark 1], the NULT for each EN at computation load r can be calculated as

$$\tau_a^u(r) = \frac{\frac{Nr}{M}}{DoF_r^u} = \min\left\{\frac{Nr}{M} + 1, N\right\},\ r \in [M]. \tag{15}$$

2) Results downloading: After computing all offloaded tasks, ENs begin to transmit the computed results back to users via downlink channels. Recall that each task is computed at r different ENs and each EN i has the computed results of n different tasks $\mathcal{T}_i$, as given in (13). Each user j wants the computed results $\widetilde{W}_j$ for $\forall j \in \mathcal{N}$, which is owned in r different ENs. Multiple ENs with the same computed results can exploit transmission cooperation to neutralize interferences across users [6], [11]. The computation results downloading for $M = N = 4$ and $r = 3$ is shown in Fig. 3. We refer to the downlink channel formed by downloading the N tasks as *the circular cooperative interference channel with transmitter cooperation group size* r. This channel is different from the cooperative X channel with transmitter cooperation group size r defined in [14], [16], where any subset of r transmitters can form a cooperation group, resulting in $\binom{M}{r}$ groups in total, and each transmitter cooperation group has messages to send to all receivers. In our considered downlink channel, there are only N cooperation groups which are performed circularly by the M transmitters and each group only needs to communicate with one receiver. An achievable per-receiver DoF of this downlink channel is given as below.

Lemma 2. *An achievable per-receiver DoF of the circular cooperative interference channel with M transmitters and N receivers satisfying $\frac{N}{M} \in \mathbb{Z}^+$ and transmitter cooperation group size r is given by*

$$DoF_r^d = \max\left\{\frac{M}{N+M}, \frac{r}{N}\right\},\quad r \in [M], \tag{16}$$

and it is within a multiplicative gap of 2 to the optimal DoF.

Proof: We first use interference neutralization to achieve a DoF of $\frac{r}{N}$ for each user, and then compare it with the per-receiver DoF of $\frac{M}{N+M}$ achieved by only using interference alignment. Please refer to Appendix B for the full proof. ■

The traffic load for each user to download its task output data is L bits. Hence, by [13, Remark 3] and [14, Remark 1], the NDLT for each user at computation load r is given by

$$\tau_a^d(r) = \frac{1}{DoF_r^d} = \min\left\{\frac{N}{M}+1, \frac{N}{r}\right\},\quad r \in [M]. \tag{17}$$

Based on Eq. (15) and Eq. (17), we thus have the achievable communication latency pair $(\tau_a^u(r), \tau_a^d(r))$ at an integer computation load $r \in [M]$ for binary offloading. An inner bound of the optimal communication latency region can be given by the union of all achievable latency pairs $(\tau^u(r), \tau^d(r))$ at all integer computation load r's satisfying

$$\tau^u(r) \geq \tau_a^u(r),\ \tau^d(r) \geq \tau_a^u(r),\ \forall r \in [M]. \tag{18}$$

V. Converse Proof of Outer Bound

A. Lower bound of NULT

We prove the lower bound of NULT at any feasible computation load $r \in \left\{r : \sum_{i=1}^M a_i = Nr, a_i \in [0:N], \forall i \in \mathcal{M}\right\}$, i.e., $\tau^{u^*}(r) \geq \min\left\{\frac{Nr}{M}+1, N\right\}$. First, we use genie-aided arguments to derive a lower bound on the NULT of any given feasible task assignment with a computation load r. Then, we construct an optimization problem and solve its optimal solution to acquire the lower bound on the minimum NULT of all feasible task assignment schemes.

Given a computation load r. Consider an arbitrary task assignment scheme where the number of tasks assigned to each EN i is denoted as a_i, $\forall i \in \mathcal{M}$, and satisfies

$$\sum_{i=1}^M a_i = Nr, \tag{19}$$

$$a_i \in [0:N],\quad i \in \mathcal{M}. \tag{20}$$

Note that we only need consider $a_i > 0$ case since $a_i = 0$ means no task is assigned to EN i and we can remove EN i from the EN set $\mathcal{M}$, which will not change the results. Consider the following three disjoint subsets of task input data (or messages):

$$\mathcal{W}_r = \{W_{j,\mathcal{S}_j} : j \in \mathcal{N}, i \in \mathcal{S}_j\}, \tag{21}$$

$$\mathcal{W}_t = \{W_{j,\mathcal{S}_j} : j = t_o, i \notin \mathcal{S}_j\}, \tag{22}$$

$$\overline{\mathcal{W}} = \{W_{j,\mathcal{S}_j} : j \neq t_o \text{ and } i \notin \mathcal{S}_j\}, \tag{23}$$

where $W_{j,\mathcal{S}_j}$ denotes the input message of the task from user j assigned to all ENs in subset $\mathcal{S}_j$, and t_o denotes one of the users that do not offload their tasks to EN i, i.e., $\mathcal{W}_r \cap \mathcal{W}_t = \varnothing$. It is seen that the set $\mathcal{W}_r$ indicates the messages that EN i need decode, i.e., $|\mathcal{W}_r| = a_i$; The set $\mathcal{W}_t$ is a nonempty set with cardinality $|\mathcal{W}_t| = 1$ when EN i is not assigned all N tasks, i.e., $a_i < N$, since user t_o exists in this case; Otherwise, we have $\mathcal{W}_t = \varnothing$ for $a_i = N$. We will show that set $\mathcal{W}_r \cup \mathcal{W}_t$ has the maximum number of messages that can be decoded by EN i.

Let a genie provide the messages $\overline{\mathcal{W}}$ to all ENs, and additionally provide messages $\mathcal{W}_r$ to ENs in $\mathcal{M}/\{i\}$. The

received signal of EN i can be represented as

$$\hat{\mathbf{y}}_i = \sum_{j=1,\neq t_o}^{M} \mathbf{H}_{ij}\mathbf{x}_j + \mathbf{H}_{it_o}\mathbf{x}_{t_o} + \hat{\mathbf{z}}_i, \tag{24}$$

where $\mathbf{H}_{ij}$, $\mathbf{x}_j$, $\mathbf{z}_i$ are diagonal matrices representing the channel coefficients between user j and EN i, signal transmitted by user j, noise received at EN i, over the block length T^u, respectively. Note that we reduce the noise at EN i from $\mathbf{z}_i$ to $\hat{\mathbf{z}}_i$ by a fixed amount such that its received signal $\mathbf{y}_i$ can be replaced by $\hat{\mathbf{y}}_i$. The ENs in $\mathcal{M}/\{i\}$ have messages $\overline{\mathcal{W}}\cup\mathcal{W}_r$, which do not include the message of user t_o. Using these genie-aided information, each EN $k \in \mathcal{M}/\{i\}$ can compute the transmitted signals $\{\mathbf{x}_j : j \neq t_o\}$ and subtract them from the received signal. Thus, the received signal of EN $k \neq i$ can be rewritten as

$$\bar{\mathbf{y}}_k = \mathbf{y}_k - \sum_{j=1,\neq t_o}^{N} \mathbf{H}_{kj}\mathbf{x}_j = \mathbf{H}_{kt_o}\mathbf{x}_{t_o} + \mathbf{z}_k. \tag{25}$$

Since the message $\mathcal{W}_t$ is intended for some ENs in $\mathcal{M}/\{i\}$, denoted as $\mathcal{R}_t$, the ENs in $\mathcal{R}_t$ can decode it. By Fano's inequality and (25), we have

$$H(\mathcal{W}_t|\mathbf{y}_k, \overline{\mathcal{W}}, \mathcal{W}_r) \leq T^u\epsilon, \quad k \in \mathcal{R}_t. \tag{26}$$

Consider EN i, it can decode messages $\mathcal{W}_r$ intended for it. By Fano's inequality, we have

$$H(\mathcal{W}_r|\hat{\mathbf{y}}_i, \overline{\mathcal{W}}) \leq |\mathcal{W}_r|T^u\epsilon. \tag{27}$$

Using genie-aided messages $\overline{\mathcal{W}}$ and decoded messages $\mathcal{W}_r$, EN i can compute the transmitted signals $\{\mathbf{x}_j : j \neq t_o\}$, and subtract them from the received signal. We thus have

$$\bar{\mathbf{y}}_i = \hat{\mathbf{y}}_i - \sum_{j=1,\neq t_o}^{N} \mathbf{H}_{ij}\mathbf{x}_j = \mathbf{H}_{it_o}\mathbf{x}_{t_o} + \hat{\mathbf{z}}_i. \tag{28}$$

By reducing noise and multiplying the constructed signal $\bar{\mathbf{y}}_i$ at EN i by $\mathbf{H}_{kt_o}\mathbf{H}_{it_o}^{-1}$, we have

$$\bar{\mathbf{y}}_i^k = \mathbf{H}_{kt_o}\mathbf{H}_{it_o}^{-1}\bar{\mathbf{y}}_i = \mathbf{H}_{kt_o}\mathbf{x}_{t_o} + \hat{\mathbf{z}}_k^{'}, \tag{29}$$

where $\hat{\mathbf{z}}_k^{'}$ represents the reduced noise. It is seen that $\bar{\mathbf{y}}_i^k$ is a degraded version of $\bar{\mathbf{y}}_k$ at EN k in $\mathcal{R}_t$, so EN i must be able to decode the messages that ENs in $\mathcal{R}_t$ can decode. Thus, we have

$$H(\mathcal{W}_t|\hat{\mathbf{y}}_i, \overline{\mathcal{W}}, \mathcal{W}_r) \leq H(\mathcal{W}_t|\mathbf{y}_k, \overline{\mathcal{W}}, \mathcal{W}_r) \leq T^u\epsilon, \quad i \in \mathcal{R}_t. \tag{30}$$

All the above changes including genie-aided information, receiver cooperation, and noise reducing can only improve capacity. Therefore, we have the following chain of inequalities,

$$(|\mathcal{W}_r| + |\mathcal{W}_t|)L = H(\mathcal{W}_r, \mathcal{W}_t) \overset{(a)}{=} H(\mathcal{W}_r, \mathcal{W}_t|\overline{\mathcal{W}}) \tag{31}$$

$$\overset{(b)}{=} I(\mathcal{W}_r, \mathcal{W}_t : \hat{\mathbf{y}}_i|\overline{\mathcal{W}}) + H(\mathcal{W}_r, \mathcal{W}_t|\hat{\mathbf{y}}_i, \overline{\mathcal{W}}) \tag{32}$$

$$\overset{(c)}{=} I(\mathcal{W}_r, \mathcal{W}_t : \hat{\mathbf{y}}_i|\overline{\mathcal{W}}) + H(\mathcal{W}_r|\hat{\mathbf{y}}_i, \overline{\mathcal{W}}) + H(\mathcal{W}_t|\hat{\mathbf{y}}_i, \mathcal{W}_r, \overline{\mathcal{W}}) \tag{33}$$

$$\overset{(d)}{\leq} I(\mathcal{W}_r, \mathcal{W}_t : \hat{\mathbf{y}}_i|\overline{\mathcal{W}}) + |\mathcal{W}_r|T^u\epsilon + T^u\epsilon \tag{34}$$

$$\overset{(e)}{\leq} I(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{a_i}, \mathbf{x}_{t_o} : \hat{\mathbf{y}}_i|\overline{\mathcal{W}}) + (|\mathcal{W}_r| + 1)T^u\epsilon \tag{35}$$

$$\overset{(f)}{\leq} T^u \log P_u + (|\mathcal{W}_r| + 1)T^u\epsilon, \tag{36}$$

where (a) is due to the independence of messages, (b) and (c) follow from the chain rule, (d) uses Fano's inequalities (27) and (30), (e) is the data processing inequality, and (f) uses the DoF bound of the MAC channel. By dividing on $\frac{L}{\log P_u}$, and taking $P_u \to \infty$ and $\epsilon \to 0$, we have $\tau^u \geq |\mathcal{W}_r| + |\mathcal{W}_t| = \min\{a_i + 1, N\}$.

Thus, for any given feasible task assignment $\mathbf{a} \triangleq [a_i]_{i\in\mathcal{M}}$, the NULT satisfies $\tau^u \geq \min\{a_i+1, N\}$ for $\forall i \in \mathcal{M}$, i.e., the minimum NULT for this given task assignment is lower bounded by

$$\tau^{u^*}(r, \mathbf{a}) \geq \max_{i\in\mathcal{M}} \min\{a_i+1, N\} = \min\left\{\max_{i\in\mathcal{M}} a_i+1, N\right\}. \tag{37}$$

Therefore, the minimum NULT $\tau^{u^*}(r)$ for all feasible task assignment schemes is given by $\tau^{u^*}(r) = \min_{\mathbf{a}} \tau^{u^*}(r, \mathbf{a})$. It can be lower bounded by the optimal solution of the following integer programming problem,

$$\begin{aligned} \text{P1}: \quad & \min_{\mathbf{a}} \ \min\left\{\max_{i\in\mathcal{M}} a_i + 1, N\right\} \\ & s.t. \quad \sum_{i=1}^{M} a_i = Nr \\ & \qquad a_i \in [0:N], \quad i \in \mathcal{M}. \end{aligned} \tag{38}$$

By relaxing the integer constraint $a_i \in [0:N]$ into a real-value constraint $0 \leq a_i \leq N$, the optimal solution is still a lower bound of the minimum NULT $\tau^{u^*}(r)$. Since the objective is equivalent to minimizing the term $\max_{\forall i\in\mathcal{M}} a_i$, the optimal solution can be obtained easily as $a_i^* = \frac{Nr}{M}$, $\forall i \in \mathcal{M}$. Hence, the minimum NULT is lower bounded by

$$\tau^{u^*}(r) \geq \min\left\{\frac{Nr}{M} + 1, N\right\}. \tag{39}$$

The proof of the lower bound of NULT is thus completed. Comparing (39) with (8) in Theorem 1, we see that they are the same. Thus, the proposed equal task assignment and task uploading scheme in Section IV-1 is optimal.

B. Lower bound and gap of NDLT

Let $\mathbf{x_i}$ denote the signal transmitted by each EN i, and $\mathbf{y_j}$ the signal received at each user j, over the block length T^d. Consider the N computed results decoded by N users, we have the following chain of inequalities,

$$N\widetilde{L} = H(\widetilde{W}_1, \cdots, \widetilde{W}_N) \tag{40}$$

$$= I(\widetilde{W}_1, \cdots, \widetilde{W}_N : \mathbf{y}_1, \cdots, \mathbf{y}_N) + H(\widetilde{W}_1, \cdots, \widetilde{W}_N|\mathbf{y}_1, \cdots, \mathbf{y}_N) \tag{41}$$

$$\overset{(g)}{\leq} I(\widetilde{W}_1, \cdots, \widetilde{W}_N : \mathbf{y}_1, \cdots, \mathbf{y}_N) + \sum_{j=1}^{N} H(\widetilde{W}_j|\mathbf{y}_j) \tag{42}$$

$$\overset{(h)}{\leq} I(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M : \mathbf{y}_1, \cdots, \mathbf{y}_N) + NT^d\epsilon \tag{43}$$

$$\overset{(i)}{\leq} \min\{M, N\}T^d \log P_d + NT^d\epsilon, \tag{44}$$

where (g) follows from $H(\widetilde{W}_1, \cdots, \widetilde{W}_N|\mathbf{y}_1, \cdots, \mathbf{y}_N) \leq \sum_{j=1}^{N} H(\widetilde{W}_j|\mathbf{y}_1, \cdots, \mathbf{y}_N) \leq \sum_{j=1}^{N} H(\widetilde{W}_j|\mathbf{y}_j)$, (h) follows from the data processing inequality and Fano's inequality, and (i)

uses the capacity bound of the MISO broadcast channel with a M-antenna transmitter and N single-antenna receivers. By dividing on $\frac{\widetilde{L}}{\log P_d}$, and taking $P_d \to \infty$ and $\epsilon \to 0$, we have

$$\tau^d \geq \frac{N}{\min\{M,N\}}. \tag{45}$$

Hence, the minimum NDLT is lower bounded by $\tau^{d^*} \geq \frac{N}{M}$. It can be easily proved that the multiplicative gap between the achievable NDLT $\tau_a^d(r)$ in Theorem 1 and its lower bound is within 2 for $\frac{N}{M} \in \mathbb{Z}^+$, i.e., $\frac{\tau_a^d(r)}{\tau^{d^*}(r)} \leq \frac{\min\left\{\frac{N}{M}+1, \frac{N}{r}\right\}}{N/M} \leq 2$. We complete the proof of the lower bound and gap of NDLT in binary offloading.

Combining (39) and (45), an outer bound of the optimal communication latency region can be given by the union of all latency pairs $(\tau^u(r), \tau^d(r))$ at all feasible computation load r's satisfying

$$\tau^u(r) \geq \left\{\frac{Nr}{M}+1, N\right\},\ \tau^d(r) \geq \frac{N}{\min\{M,N\}} \tag{46}$$

for $\forall r \in \left\{r : \sum_{i=1}^M a_i = Nr, a_i \in [0:N], \forall i \in \mathcal{M}\right\}$.

APPENDIX

A. Proof of Lemma 1

1) Achievable scheme: We use the partial interference alignment scheme with a $u_s = ns^\Gamma + (s+1)^\Gamma$ symbol extension over the original channel, where $s \in \mathbb{N}$ and $\Gamma = M(N-n)$. Specifically, each transmitter j encodes the task input message W_j into s^Γ independent streams x_j^l, $l \in [s^\Gamma]$, each beamformed along a $u_s \times 1$ column vector $\mathbf{v}_j^l$. So the symbol $\bar{\mathbf{X}}_j$ transmitted at transmitter j can be expressed as

$$\bar{\mathbf{X}}_j = \sum_{l=1}^{s^\Gamma} x_j^l \mathbf{v}_j^l = \bar{\mathbf{V}}_j \mathbf{X}_j, \tag{47}$$

where $\mathbf{X}_j \triangleq (x_j^l)_{l=1}^{s^\Gamma}$ is a $s^\Gamma \times 1$ column vector, and $\bar{\mathbf{V}}_j = [\mathbf{v}_j^l]_{l=1}^{s^\Gamma}$ is a $u_s \times s^\Gamma$ matrix. Then, the received signal at EN i can be written as

$$\bar{\mathbf{Y}}_i = \sum_{j=1}^N \bar{\mathbf{H}}_{ij} \bar{\mathbf{V}}_j \mathbf{X}_j + \bar{\mathbf{Z}}_i, \tag{48}$$

where $\bar{\mathbf{Y}}_i$ and $\bar{\mathbf{Z}}_i$ represent the u_s symbol extension of the received signal Y_i and noise Z_i, respectively. $\bar{\mathbf{H}}_{ij}$ is a $u_s \times u_s$ diagonal matrix representing the u_s symbol extension of the channel, whose l-th diagonal element is $h_{ij}(l)$.

Next, we design the beamforming vectors such that each receiver i can decode the n desired signals $\{\mathbf{X}_{k+1} : k \in [(i-1)n : (in-1)] (\mathrm{mod}\, N)\}$ by zero-forcing the interferences. To align the $N-n$ interference signals at each receiver together in the space with dimension $(s+1)^\Gamma$, beamforming vectors need to satisfy the following conditions:

$$\text{span}(\bar{\mathbf{H}}_{ij}\bar{\mathbf{V}}_j) \prec \text{span}(\mathbf{V}), \forall i \in \mathcal{M}, \tag{49}$$

for $\forall j \in \mathcal{N}, (i,j) \notin \{(i,k+1) : k \in [(i-1)n : (in-1)] (\mathrm{mod} N)\}$, where span($\mathbf{P}$) denotes the space spanned by the column vectors of matrix $\mathbf{P}$, and $\mathbf{V}$ is a $u_s \times (s+1)^\Gamma$ matrix. Now we need to design the column vectors of $\bar{\mathbf{V}}^{[j]}$ and $\mathbf{V}$ to satisfy (49). Let $\mathbf{w}$ be a $u_s \times 1$ column vector $\mathbf{w} = (1, 1, \cdots, 1)^{\mathrm{T}}$. The sets of column vectors of $\bar{\mathbf{V}}_j$ and $\mathbf{V}$, denoted as $\bar{\mathcal{V}}_j$ and $\mathcal{V}$, respectively, are given as below

$$\bar{\mathcal{V}}_j = \underbrace{\left\{\left(\prod_{t\in\mathcal{M}, q\in\mathcal{N}, (t,q)\notin\{(t,k+1):k\in[(t-1)n:(tn-1)](\mathrm{mod}N)\}} \left(\bar{\mathbf{H}}_{tq}\right)^{\alpha_{tq}}\right)\mathbf{w} : \alpha_{tq} \in [0:s-1]\right\}}_{\text{a total of } s^\Gamma \text{ columns}} \tag{50}$$

for $\forall j \in \mathcal{N}$, and

$$\mathcal{V} = \underbrace{\left\{\left(\prod_{t\in\mathcal{M}, q\in\mathcal{N}, (t,q)\notin\{(t,k+1):k\in[(t-1)n:(tn-1)](\mathrm{mod}N)\}} \left(\bar{\mathbf{H}}_{tq}\right)^{\alpha_{tq}}\right)\mathbf{w} : \alpha_{tq} \in [0:s]\right\}}_{\text{a total of } (s+1)^\Gamma \text{ columns}}. \tag{51}$$

We then show that the desired signal streams received at each receiver are linearly independent of each other and interference signal streams such that the desired streams can be decoded by zero-forcing interferences. At any receiver i, the desired signal streams are beamformed along the ns^Γ vectors of $\left[\bar{\mathbf{H}}_{i,i_1}\bar{\mathbf{V}}_{i_1} \ \ \bar{\mathbf{H}}_{i,i_2}\bar{\mathbf{V}}_{i_2} \ \cdots \ \bar{\mathbf{H}}_{i,i_n}\bar{\mathbf{V}}_{i_n}\right]$, where $i_m = (i-1)n + m \,(\mathrm{mod}\, N),\ m \in [n]$. By condition (49), the interference streams at any receiver i from transmitter j are aligned at the column vector space of $\mathbf{V}$ for $j \in [N] / \{k+1 : k \in [(i-1)n : in-1] \,(\mathrm{mod}\, N)\}$. To decode the desired ns^Γ streams successfully, it suffices to show that the $u_s \times u_s$ matrix

$$\mathbf{A}_i = \left[\bar{\mathbf{H}}_{i,i_1}\bar{\mathbf{V}}_{i_1} \ \ \bar{\mathbf{H}}_{i,i_2}\bar{\mathbf{V}}_{i_2} \ \cdots \ \bar{\mathbf{H}}_{i,i_n}\bar{\mathbf{V}}_{i_n} \ \ \mathbf{V}\right] \tag{52}$$

has a full rank of u_s almost surely for $\forall i \in \mathcal{M}$. By the beamforming vectors in (50) and (51), we can observe that the u_s elements in the l-th row of $\mathbf{A}_i$ have the following forms

$$\underbrace{\left\{h_{i,i_m}(l) \prod_{t\in\mathcal{M}, q\in\mathcal{N}, (t,q)\notin\{(t,k+1):k\in[(t-1)n:(tn-1)](\mathrm{mod}N)\}} (h_{tq}(l))^{\alpha_{tq}} : \alpha_{tq} \in [0:s-1], m \in [n]\right\}}_{\text{a total of } ns^\Gamma \text{ elements}} \bigcup$$

$$\underbrace{\left\{\prod_{t\in\mathcal{M}, q\in\mathcal{N}, (t,q)\notin\{(t,k+1):k\in[(t-1)n:(tn-1)](\mathrm{mod}N)\}} (h_{tq}(l))^{\beta_{tq}} : \beta_{tq} \in [0:s]\right\}}_{\text{a total of } (s+1)^\Gamma \text{ elements}}, \tag{53}$$

where $\{h_{ij}(l)\}$ are drawn independently from a continuous probability distribution. We observe from (53) that all the elements of $\mathbf{A_i}$ meet the two conditions of [17, Lemma 1]. Hence, the matrix $\mathbf{A_i}$ is a full-rank matrix for $\forall i \in \mathcal{M}$. Taking s to infinity, the DoF for each receiver achieved by above scheme is given by $\lim_{s\to+\infty} \frac{ns^\Gamma}{ns^\Gamma + (s+1)^\Gamma} = \frac{n}{n+1} = \frac{Nr}{Nr+M}$.

Further, consider the basic scheme that N transmitters deliver their messages to the assigned receivers in the time division manner, which achieves a DoF of $\frac{n}{N} = \frac{r}{M}$ for each receiver. Therefore, the per-receiver DoF of the considered interference-multicast channel with multicast group size r is given by $DoF_r^u = \max\left\{\frac{Nr}{Nr+M}, \frac{r}{M}\right\}$.

2) Converse Proof of Lemma 1: Since in Section V-A, we have proved that the proposed task assignment and uploading scheme in Section IV-1 is information-theoretically optimal,

the per-receiver DoF of the considered uplink channel in (14) of Lemma 1 must also be optimal.

B. Proof of Lemma 2

1) Achievability: We show the achievable schemes in two cases, $r=1$ and $r\geq 2$, respectively.

$i)$ $r=1$: By (13), the task output messages at each transmitter i can be represented as $\left\{\widetilde{W}_j : j\in[(i-1)n+1:in]\right\}$. Let $\Gamma = M(N-1)$ and consider a $u_s = s^\Gamma + n(s+1)^\Gamma$ symbol extension. Each message $\widetilde{W}_j$ are encoded into s^Γ independent streams x_j^l, $l\in[s^\Gamma]$, each beamformed along a $u_s\times 1$ column vector $\mathbf{v}_j^l$, for $\forall j\in\mathcal{N}$. Then, the signal transmitted at transmitter i can be expressed as

$$\bar{\mathbf{X}}_i = \sum_{j=(i-1)n+1}^{in}\sum_{l=1}^{s^\Gamma} x_j^l\mathbf{v}_j^l = \sum_{j=(i-1)n+1}^{in}\bar{\mathbf{V}}_j\mathbf{X}_j, \tag{54}$$

where $\mathbf{X}_j = (x_j^l)_{l=1}^{s^\Gamma}$ is a $s^\Gamma\times 1$ column vector and $\bar{\mathbf{V}}_j = [\mathbf{v}_j^l]_{l=1}^{s^\Gamma}$ is a $u_s\times s^\Gamma$ matrix. The signal received at receiver j can be expressed as

$$\bar{\mathbf{Y}}_j = \sum_{i=1}^{M}\bar{\mathbf{G}}_{ji}\sum_{k=(i-1)n+1}^{in}\bar{\mathbf{V}}_k\mathbf{X}_k + \bar{\mathbf{Z}}_j, \tag{55}$$

where $\bar{\mathbf{G}}_{ji}$ is a $u_s\times u_s$ diagonal matrix representing the u_s symbol extension of the channel, $\bar{\mathbf{Y}}_j$ and $\bar{\mathbf{Z}}_j$ represent the u_s symbol extension of the received signal Y_j and noise Z_j, respectively.

Next, we align the interferences at each receiver j such that the total dimension of the spaces spanned by the interference vectors is $n(s+1)^\Gamma$. Then, the desired s^Γ streams corresponding to the desired signal $\mathbf{X}_j$ can be decoded by zero-forcing the interferences from an $u_s = s^\Gamma + n(s+1)^\Gamma$-dimensional received signal vector. We ensure this by designing the beamforming vectors $\left\{\bar{\mathbf{V}}_j\right\}$ as follows, where the message $\widetilde{W}_j$ desired by receiver j is at transmitter $\lceil\frac{j}{n}\rceil\in\mathcal{M}$ by (13),

$$\left.\begin{aligned}
\text{span}(\bar{\mathbf{G}}_{ji}\bar{\mathbf{V}}_{(i-1)n+1}) &\subset \text{span}(\mathbf{U}_1)\\
\text{span}(\bar{\mathbf{G}}_{ji}\bar{\mathbf{V}}_{(i-1)n+2}) &\subset \text{span}(\mathbf{U}_2)\\
&\vdots\\
\text{span}(\bar{\mathbf{G}}_{ji}\bar{\mathbf{V}}_{(i-1)n+k}) &\subset \text{span}(\mathbf{U}_k)\\
&\vdots\\
\text{span}(\bar{\mathbf{G}}_{ji}\bar{\mathbf{V}}_{in}) &\subset \text{span}(\mathbf{U}_n)
\end{aligned}\right\}\begin{array}{l}\forall j\in\mathcal{N},\forall i\in\mathcal{M},\\ j\neq(i-1)n+k \text{ for } \forall k\in[n],\end{array}$$

where $\mathbf{U}_k$ is a $u_s\times(s+1)^\Gamma$ matrix, $\forall k\in[n]$. Next, we design $\{\mathbf{V}_j\}$ and $\{\mathbf{U}_k\}$ to satisfy above conditions. First, we generate n $u_s\times 1$ column vectors $\mathbf{w}_k = (w_k^l)_{l=1}^{u_s}$, $k\in[n]$. All elements of these n vectors are chosen i.i.d from some continuous distribution whose support lies between a finite minimum value and a finite maximum value. Then, the sets of column vectors of $\bar{\mathbf{V}}_j$ and $\mathbf{U}_k$ are denoted as $\bar{\mathcal{V}}_j$ and $\mathcal{U}_k$, respectively, and are given as follows,

$$\bar{\mathcal{V}}_{(i-1)n+k} = \left\{\left(\prod_{q\in\mathcal{N},t\in\mathcal{M},(q,t)\neq((t-1)n+k,t)}\left(\bar{\mathbf{G}}_{qt}\right)^{\alpha_{qt}}\right)\mathbf{w}_k : \alpha_{qt}\in[0:s-1]\right\} \tag{56}$$

for $\forall i\in\mathcal{M},\forall k\in[n]$, and

$$\mathcal{U}_k = \left\{\left(\prod_{t\in\mathcal{M},q\in\mathcal{N},(q,t)\neq((t-1)n+k,t)}\left(\bar{\mathbf{G}}_{qt}\right)^{\alpha_{qt}}\right)\mathbf{w}_k : \alpha_{qt}\in[0:s]\right\} \tag{57}$$

for $\forall k\in[n]$.

In the following, we show that the desired signal streams are linearly independent with the interference signal streams, and hence can be decoded by zero-forcing the interference. Consider the signal vectors received at any receiver $j=(i-1)n+k$, $i\in\mathcal{M}$, $k\in[n]$. By (56), the desired signal streams are beamformed along the s^Γ vectors of $\bar{\mathbf{G}}_{(i-1)n+k,i}\bar{\mathbf{V}}_{(i-1)n+k}$, while the interference vectors are aligned at the column vector spaces of $\mathbf{U}_{k'},\forall k'\in[n]$. To decode the desired streams successfully, it suffices to show that the $u_s\times u_s$ matrix

$$\mathbf{\Lambda}_j = \mathbf{\Lambda}_{(i-1)n+k} = \left[\bar{\mathbf{G}}_{(i-1)n+k,i}\bar{\mathbf{V}}_{(i-1)n+k}\ \ \mathbf{U}_1\ \ \mathbf{U}_2\ \cdots\ \mathbf{U}_n\right] \tag{58}$$

is a full-rank matrix almost surely for $\forall j\in\mathcal{N}$ or $\forall i\in\mathcal{M}$ and $\forall k\in[n]$. It is seen that the l-th row elements of $\mathbf{\Lambda_j}$ have the following forms,

$$\left\{h_{(i-1)n+k,i}(l)\prod_{q\in\mathcal{N},t\in\mathcal{M},(q,t)\neq((t-1)n+k,t)}(g_{qt}(l))^{\alpha_{qt}}w_k(l) : \alpha_{qt}\in[0:s-1]\right\}\bigcup$$
$$\left\{\prod_{q\in\mathcal{N},t\in\mathcal{M},(q,t)\neq((t-1)n+k',t)}(g_{qt}(l))^{\beta_{qt}}w_{k'}(l) : \beta_{qt}\in[0:s],k'\in[n]\right\}. \tag{59}$$

By (59), we have:

1) The product term in the l-th row of $\mathbf{U}_k$ contains $w_k(l)$ with exponent 1, but do not contain $w_{k'}(l),\forall k'\neq k$. Thus, all the monomial elements in the l-th row of $[\mathbf{U}_1\ \mathbf{U}_2\ \cdots\ \mathbf{U}_n]$ are unique.

2) The equations corresponding to $\mathbf{G}_{(i-1)n+k,i}$ are not contained in the interference alignment relations of (56) for $\mathbf{U}_k$, so the monomial elements in the l-th row of $\mathbf{U}_k$ do not contain $h_{(i-1)n+k,i}$, $\forall i\in\mathcal{M}$. It means that all the monomial terms in $\bar{\mathbf{G}}_{(i-1)n+k,i}\bar{\mathbf{V}}_{(i-1)n+k}$ are different from those in $\mathbf{U}_k$. They are also different from the monomial terms in $\mathbf{U}_{k'}$, $\forall k'\neq k$, due to $w_k(l)$.

Therefore, we can conclude that these u_s vectors in $\mathbf{\Lambda_j}$ are independent, and hence $\mathbf{\Lambda_j}$ is a full-rank matrix. Taking s to infinity, the scheme achieves a per-receiver DoF of $\lim_{s\to+\infty}\frac{s^\Gamma}{s^\Gamma+n(s+1)^\Gamma} = \frac{1}{1+n} = \frac{M}{N+M}$. Comparing it with the DoF of $\frac{1}{N}$ achieved by TDMA, the per-receiver DoF of the considered downlink channel for $r=1$ is given by $DoF_1^d = \max\left\{\frac{M}{N+M},\frac{1}{N}\right\}$.

$ii)$ $r\geq 2$: We first consider interference neutralization enabled by transmitter cooperation. Encode the task output message $\widetilde{W}_j$ into r independent streams x_j^p, $p\in[r]$. For better illus-

tration, each stream x_j^p is given an index $(p-1)N+j$. There are a total of Nr (or Mn) different streams corresponding to all N messages. Based on the index order, these Nr different streams can be divided into N groups, each group with r different streams, where the k-th group is given by

$$\mathcal{Q}_k = \left\{x_j^p \colon (p-1)N+j \in [(k-1)r+1 \colon kr]\right\},\ k \in [N]. \tag{60}$$

Since each message exists at r different transmitters, each stream is also owned by r different transmitters. Each group of streams is downloaded in the time division manner. The downlink channel formed by transmitting each group of r streams can be treated as a MISO broadcast channel with perfect transmitter cooperation, whose sum DoF is r [6], [18] achieved by using interference neutralization. Thus, a DoF of $\frac{r}{N}$ is obtained for each receiver in the $r \geq 2$ case.

Then, we apply asymptotic interference alignment scheme in the $r = 1$ case. Since $\frac{N}{M} = n_1$, the messages at each transmitter i are $\left\{\widetilde{W}_k : k \in [(i-1)n_1+1 : in_1]\right\}$ for $r = 1$ and $\left\{\widetilde{W}_{k+1} : k \in [(i-1)rn_1 : (irn_1-1)] (\mathrm{mod}\, N)\right\}$ for $r \geq 2$. In $r \geq 2$ case, we can let each transmitter only transmit the n_1 messages among the total rn_1 messages, and different transmitters transmit non-overlapped messages. By doing so, we construct a downlink channel with the same information flow as $r = 1$ case. Utilizing the alignment scheme in $r = 1$ case, we thus obtain a per-receiver DoF of $\frac{M}{N+M}$ for $r \geq 2$ case. Comparing above two schemes, a Dof of $\max\left\{\frac{M}{N+M}, \frac{r}{N}\right\}$ is obtained.

Summarizing the per-receiver DoF for $r = 1$ and $r \geq 2$ cases, we thus prove Lemma 2.

2) Converse proof of Lemma 2: For the upper bound, we assume that each transmitter has known all N task output messages so that the full transmitter cooperation among M transmitters is enabled. This assumption can only improve the channel capacity. We thus construct a virtual MISO broadcast channel with N single-antenna receivers and a M-antenna transmitter, whose optimal sum DoF is $\min\{M, N\}$ [19]. The capacity of the considered cooperative interference channel cannot exceed that of this virtual MISO broadcast channel. Hence, the optimal per-receiver DoF $DoF_r^{d^*}$ is upper bounded by $\frac{\min\{M,N\}}{N}$. The gap between this upper bound and the achievable lower bound in (16) satisfies $\frac{DoF_r^{d^*}}{DoF_r^d} \leq 2$.

REFERENCES

[1] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.

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

[3] S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "Communication-aware computing for edge processing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2885–2889.

[4] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1451–1455.

[5] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Commun. Letters*, vol. 6, no. 3, pp. 398–401, Jun. 2017.

[6] V. S. Annapureddy, A. E. Gamal, and V. V. Veeravalli, "Degrees of freedom of interference channels with CoMP transmission and reception," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 5740–5760, Sep. 2012.

[7] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the K-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[8] S. Li, M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr, "A fundamental tradeoff between computation and communication in distributed computing," *IEEE Trans. Inf. Theory*, vol. PP, no. 99, pp. 1–1, 2017.

[9] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Inf. Theory*, vol. PP, no. 99, pp. 1–1, 2017.

[10] D. Wang, G. Joshi, and G. Wornell, "Efficient task replication for fast response times in parallel computation," in *Proc. ACM SIGMETRICS*, 2014.

[11] I. Wang and D. N. C. Tse, "Interference mitigation through limited transmitter cooperation," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 2941–2965, May 2011.

[12] K. Li, M. Tao, and Z. Chen, "Exploiting computation replication in multi-user multi-server mobile edge computing networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018.

[13] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.

[14] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[15] J. Hachem, U. Niesen, and S. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, pp. 1–1, 2018.

[16] A. M. Girgis, Ö. Erçetin, M. Nafie, and T. A. ElBatt, "Degrees of freedom of interference networks with transmitter-side caches," *CoRR*, vol. abs/1712.05957, 2017. [Online]. Available: http://arxiv.org/abs/1712.05957

[17] V. R. Cadambe, S. A. Jafar, "Interference alignment and the degrees of freedom of wireless X networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 3893–3908, Sep. 2009.

[18] T. Gou and S. A. Jafar, "Degrees of freedom of the k user $mtimesn$ mimo interference channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6040–6057, Dec 2010.

[19] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.