IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 June 2023, accepted 25 July 2023, date of publication 2 August 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301468

RESEARCH ARTICLE

A Rate Splitting Multiple Access Interface for Clustered Wireless Federated Learning

NICLAS FÜHRLING, (Graduate Student Member, IEEE),
KENGO ANDO, (Graduate Student Member, IEEE),
HYEON SEOK ROU, (Graduate Student Member, IEEE),
AND GIUSEPPE THADEU FREITAS DE ABREU, (Senior Member, IEEE)
School of Computer Science and Engineering, Constructor University, 28759 Bremen, Germany

Corresponding author: Niclas Führling (nfuehrling@constructor.university)

ABSTRACT Consider a wireless federated learning (WFL) system where the edge devices (EDs) performing local training are closely located in a cluster, such that in addition to their private model updates, a locally common (a.k.a. consensus) model can be computed or selected at each cycle of the learning process. For such a clustered WFL (CWFL) paradigm, we design an uplink (UL) radio access scheme based on the rate splitting multiple access (RSMA) architecture, which is shown not only to significantly reduce the latency of WFL in comparison to systems employing time-domain multiple access (TDMA) and non-orthogonal multiple access (NOMA), but also to be more energy efficient, reaching a lower latency with less power than the latter alternatives. To that end, we exploit the cluster-wide consensus model as the common message used to construct the common component of the RSMA scheme and build an optimization problem aimed at minimizing the latency of each CWFL round by means of optimally allocating the corresponding computation times and uplink transmission durations of each ED, taking into account constraints such as energy consumption, data rates and computational capabilities of each ED. By adequately setting a system parameter referred to as the rate-splitting factor, the formulation also applies to systems employing conventional TDMA or NOMA methods, such that the proposed technique can be seen as a generalization of those approaches, in the context of CWFL schemes. Simulation results on the proposed RSMA-interfaced UL scheme for CWFL are given, both with and without the incorporation of optimal NOMA decoding at the base station (BS), the first of which is found to yield only a mild improvement over the latter, indicating that the proposed approach is in fact the key factor in the overall latency reduction achieved.

INDEX TERMS Wireless federated learning, rate splitting multiple access, clustered networks, delay minimization, convex optimization.

I. INTRODUCTION

As 5G systems evolve and the fundamentals of future 6G systems develop, so do potential applications which include a diverse use of Internet-of-Things (IoT) devices in areas such as healthcare [1], self-driving vehicles [2] and Industry 4.0 [3], where it has recently sparkled great interest in the so-called Ambient IoT concept [4]. Given the fact that many of such applications require learning processes over massive amounts of data, often of a sensitive nature, WFL [5] has continuously gained popularity in the last years as an enabling technology that combines the advantages of machine learning with the scalability of wireless systems and the privacy preservation feature of the federated approach.

Indeed, conventional machine learning (ML) schemes relying on distributed data sources are such that learned models are trained with data collected by various EDs and transmitted to a central processing unit (CPU), typically under a request, prediction and feedback protocol [6]. A clear drawback of such an approach is the lack of privacy, since the data of EDs

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

is fully exposed, at least to the CPU. Federated learning (FL) schemes, first proposed in [7], resolve this issue by training local versions of the learned model privately at individual EDs, which are then transmitted to the CPU for aggregation into the global model. Since in the FL approach the information exchanged between the EDs and the CPU consists of ''machine data'' (*e.g.*, neural network parameters), privacy is less of a concern.

Besides improving privacy, it has been in fact shown [1], [2], [3] that FL also tends to build smarter models with less latency and lower power consumption compared to centralized architectures.

Wireless federated learning [8] is a variation of FL in which the EDs are connected to the CPU via wireless links, with advantages in flexibility and ease of implementation, since EDs no longer need be physically tethered to the CPU, making the approach ideal for IoT applications. But WFL itself is not without its own challenges, two of which are the convergence time of the model given heterogeneous computational capabilities at EDs [5], and the need to optimize transmit powers due to the variable link quality between EDs and the BS to which the CPU is connected [9].

In order to address these issues, work on wireless interfaces designed specifically to WFL schemes has started to appear in the wireless literature [10], [11]. To cite a few recent contributions in this area, a PHY-layer scheme utilizing intelligent reflecting surfaces (RIS) for the minimization of latency of WFL systems was proposed in [12]. Working at the interface of MAC and PHY layers, a NOMA scheme for WFL was proposed in [13], where a compute-then-transmit (CT) transmission model was integrated to the access protocol in order to quantify and optimize the impact of dynamic and fixed decoding orders onto the convergence and energy consumption of the WFL system. Finally, the challenges in minimizing the latency of WFL systems was addressed from a network layer perspective in [14], where a 3-layer hierarchical architecture making use of edge servers was proposed.

One aspect of the WFL approach that has not been well exploited, however, is that in many cases EDs are sufficiently close to each other that they can be treated as a cluster [15], [16], [17], giving way to the concept of CWFL. Indeed, advantages of clusterization in FL schemes have already been thoroughly studied and demonstrated from a machine learning viewpoint [18], [19], [20], but approaches that also take advantage clusterization in the design of wireless interfaces for WFL in particular have not been sufficiently investigated. An inspiring example is the work done in [21], where the usage of TDMA-based cell free massive MIMO (CF-mMIMO) for CWFL was considered, and the performance of the scheme as a function of the clusterization of EDs was optimized.

Inspired by the aforementioned line of work, we consider in this paper a RSMA UL interface [22], [23], [24] to connect the EDs to the CPU in CWFL systems. Originally designed for downlink (DL) transmission, RSMA is considered a promising access scheme for sixth-generation (6G) systems, due to its various demonstrated advantages compared to conventional schemes such as space division multiple access (SDMA) and NOMA, which include higher energy efficiency due to lower overhead requirements [25], robustness to imperfect channel state information (CSI) at both transmitter and receiver [26], and fairness in terms of minimum-rate-to-total-throughput [27].

In its original DL setting, RSMA works in such a manner that messages to be sent from a multi-antenna BS to a group of users are split into a codeword common to all messages and thus broadcasted to all users, and private codewords transmitted to each user via beamforming. Each user then combines the common and private messages it received in order to extract its original intended message.

Details on the construction of common and private messages for transmission, as well as their combination at the receivers are beyond the scope of this article and can be found in the abundant RSMA literature, and a description of an implementation on universal software-defined radio platforms (USRPs) can also be found in [28]. In the context of this article, it is sufficient to observe that in order for RSMA to be employed, some information common to all users must be at hand, which is indeed the case in a CWFL system where EDs are locally clustered and thus can exchange short messages among them in a device-to-device (D2D) fashion.

Under such a paradigm, we then propose a mechanism to optimize – while taking into account local constraints – the time and power allocated to each ED at both the computation and the uplink transmission processes, so as to minimize the learning convergence latency of CWFL schemes with RSMA uplink interfaces. The proposed method includes a variation where the BS decodes the messages from the EDs following a NOMA multi-user interference suppression architecture, such that in a sense, the new method generalizes NOMA-based WFL schemes, in the context of CWFL systems.

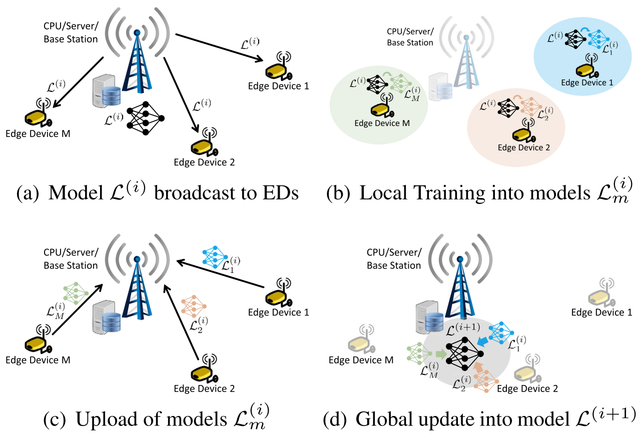
Since the problem formulated as briefly described above is not convex, we further contribute with a convexized relaxation of the latter, obtained via fractional programming (FP) [29], [30], which can then be solved efficiently. The results reveal a strong potential of RSMA as a mechanism to accelerate CWFL.

II. PRELIMINARIES

A. SYSTEM MODEL OF CONVENTIONAL WFL

Consider a multi-user (MU) single-input single-output (SISO) system, where a number M of sparsely-located single-antenna EDs indexed by $m \in \mathcal{M} = \{1, \ldots, M\}$ perform WFL under the service of a single-antenna BS, also referred to as the server or the CPU.

Conventional WFL schemes [2], [5] such as this, operate as illustrated in Figure 1, where the main steps of an i-th cycle of such a system can be briefly described as follows:



(a) Model $\mathcal{L}^{(i)}$ broadcast to EDs (b) Local Training into models $\mathcal{L}_m^{(i)}$

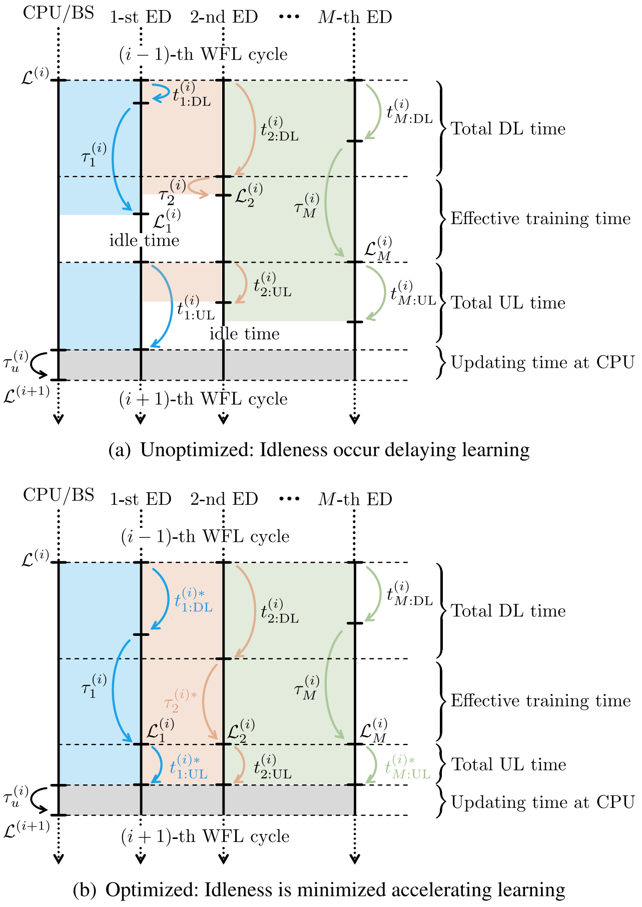
(c) Upload of models $\mathcal{L}_m^{(i)}$ (d) Global update into model $\mathcal{L}^{(i+1)}$

FIGURE 1. Illustration of a conventional WFL system.

- First, a CPU connected to a BS broadcasts,[1] to all M EDs in its surroundings, the information corresponding to the latest update $\mathcal{L}^{(i)}$ of a given learning model $\mathcal{L}$.
- Each of the various EDs then proceeds to train the received model using private data, such that after a delay $\tau_m^{(i)}$, a generic m-th device obtains a locally updated model $\mathcal{L}_m^{(i)}$.
- Next, the EDs upload their locally trained models back to the BS through a multi-access UL wireless interface, such as for instance the TDMA-based approach proposed in [21], or the NOMA-based approach of [13], to cite two specific state-of-the-art (SotA) methods.
- Finally, the uploaded locally-trained models $\mathcal{L}_m^{(i)}$ are combined at the CPU into the updated model $\mathcal{L}^{(i+1)}$, which is then transmitted to the EDs in the next cycle.

Notice that the DL transmission time $t_{m:\mathrm{DL}}^{(i)}$ required to covey the model $\mathcal{L}^{(i)}$ to each m-th ED, depends on the condition (rate) of the channel h_m between the BS and the m-th ED, and therefore can be substantially different from one ED to another. Consequently, if not optimized, the DL transmission times $t_{m:\mathrm{DL}}^{(i)}$ are all distinct, which in turn implies that each ED starts its local training of the model $\mathcal{L}^{(i)}$ at different times. In addition, since the EDs have different computational capabilities, again if not optimized, the training periods (delays) $\tau_m^{(i)}$ are also distinct, such that each ED will be ready to upload its locally trained model $\mathcal{L}_m^{(i)}$ at different times. Given that most efficient in-band wireless multi-access UL schemes are synchronous, the latter results in undesirable delays in the over all WFL process.

To illustrate the problem, we depict in Figure 2 a full cycle of both an unoptimized and an optimized WFL processes. For the sake of this particular illustration, it is assumed that the starting time of wireless transmission must be the same for all devices[2]. In the case of an unoptimized scheme,



(a) Unoptimized: Idleness occur delaying learning

(b) Optimized: Idleness is minimized accelerating learning

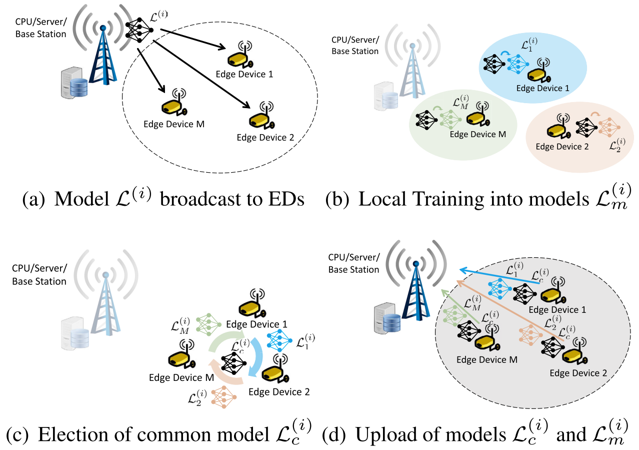
FIGURE 2. Illustration of one cycle of WFL processes with unoptimized (a) and optimized (b) air interfaces and local training delays.

as depicted in Figure 2(a), EDs which completed the receiving and local updating of the model earlier, either due to a fast DL channel (*e.g.* first ED, in blue) or due to a high computation capability (*e.g.* second ED, in orange) must wait until the slowest ED (*e.g.* M-th ED, in green) is ready. Similarly, also during the UL, the CPU must wait until all locally updated models $\mathcal{L}_m^{(i)}$ are received in order to process them into the next globally update model $\mathcal{L}^{(i+1)}$ can be computed.

In contrast to the latter, in an optimized CWFL scheme, as depicted in Figure 2(b), the DL transmission time and the local model training times are optimized so as to synchronize the EDs for UL transmission, and the transmit powers used by each ED are adjusted according to an RSMA interface such that the UL transmission is optimized. Thanks to the optimization of the transmit powers (with corresponding impact onto the transmission times $t_{m:\mathrm{DL}}$ and $t_{m:\mathrm{UL}}$) and processing clock rates (with corresponding impact onto the processing delays τ_m), undesired idleness at each cycle of the FL process can be minimized, resulting in faster performance. Excellent examples of techniques to achieve the latter are the NOMA-based approach of [13] and the TDMA-based CF-mMIMO was considered in [21].

[1] We consider in the article that the links between BS and EDs are SISO. An extension to the multiple-input multiple-output (MIMO) case will be addressed in a follow-up work.

[2] This assumption is made here merely for illustrative purposes, but in practice this would for example correspond to a transmission scheme employing a rateless code [31].



(a) Model $\mathcal{L}^{(i)}$ broadcast to EDs (b) Local Training into models $\mathcal{L}_m^{(i)}$

(c) Election of common model $\mathcal{L}_c^{(i)}$ (d) Upload of models $\mathcal{L}_c^{(i)}$ and $\mathcal{L}_m^{(i)}$

FIGURE 3. Illustration of the proposed CWFL system, where: a) the CPU broadcasts the model $\mathcal{L}^{(i)}$ to a cluster of EDs; which b) is trained privately at each ED into $\mathcal{L}_m^{(i)}$; amongst which c) one is selected a locally common model $\mathcal{L}_c^{(i)}$; all of which d) uploaded to the server via an RSMA interface.

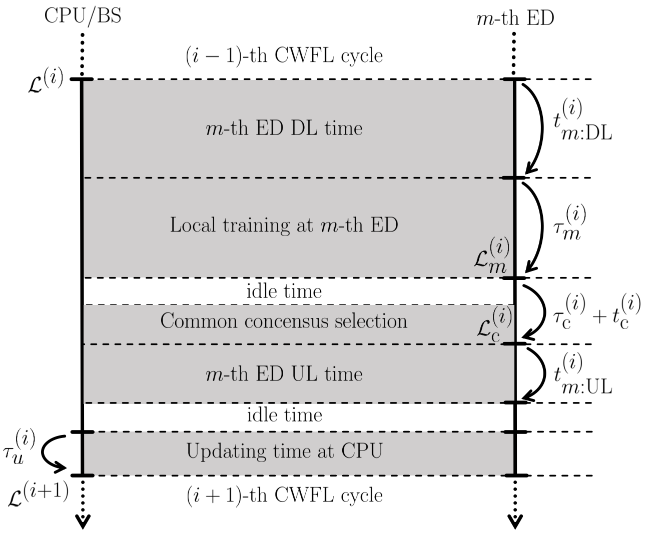


FIGURE 4. One cycle of the proposed CWFL process, from the perspective of the m-th ED, with corresponding times and delays.

B. SYSTEM MODEL OF CLUSTERED WFL

Unlike the model of Section II-A consider that the MU SISO system is as illustrated in Fig. 3, such that the WFL process is performed over a cluster of co-located M single-antenna EDs, under the service of the single-antenna BS/CPU.

As further illustrated in Fig. 4, each i-th cycle of the clustered WFL process starts with the DL broadcasting of the FL model $\mathcal{L}^{(i)}$ from the CPU to the EDs, which then independently train the model with their own private data, obtaining the locally-updated models $\mathcal{L}_m^{(i)}$, as usual.

A first distinction between a CWFL system and a conventional WFL scheme is, however, that after all other EDs finish their own local training, a locally common model $\mathcal{L}_\mathrm{c}^{(i)}$ is elected among all local models $\mathcal{L}_{m=1,\cdots,M}^{(i)}$ of the cluster, relying on the fact that D2D communication can be performed among EDs fast and efficiently.

Another important distinction between the CWFL system here proposed and other CWFL methods is, furthermore, that during the upload step each m-th ED transmits information both on the elected locally common model $\mathcal{L}_\mathrm{c}^{(i)}$ and its own private model $\mathcal{L}_m^{(i)}$, in the forms of the common and private components of an RSMA signal, respectively. Thanks to this approach, the CWFL method not only benefits from advantages of clusterization within the leaning process itself [18], [19], [20], but also from the higher spectral efficiency of RSMA schemes compared to other multi-access methods [22], [23], [24].

With respect to the latter statement, it must be emphasized that since in general different independent users simultaneously accessing a common BS usually do not share common information, is in fact originally envisioned not for UL but for DL transmissions. In the case of CWFL, however, such common information can be readily obtained in the form of the locally common, as briefly explained above. Details on the election mechanism and the RSMA wireless interface required to that end are described in the sequel, and are integral parts of the original contribution of the article.

After the UL transmission of all common and private models to the BS, the CPU finally updates the global model and a new cycle can start. Again, here it is worth emphasizing that in order to optimize the system, the idle times depicted in Figure 4 must be eliminated, which in turn requires the optimization of the transmission rates of the DL and UL transmissions to/from each ED, as well as the local training delays (*i.e.* processing clocks) of each ED. The mechanism to achieve these goals are also part of our contribution, and are described in detail in the next section.

III. PROPOSED RSMA-INTERFACED CWFL SCHEME

A. RSMA TRANSMISSION MODEL

In view of the CWFL model described above, and referring to the DL-RSMA signal design described in [22], [23], and [24], where the transmit signal consists of a message component common to all EDs, and message components private to each ED,[3] the UL transmit signal x_m corresponding to the m-th ED can be described as

$$x_m = \overbrace{p_{\mathrm{c},m}s_\mathrm{c}}^{\substack{\text{precoded common message corresponding to}\\ \text{information on the locally common model}}} + \underbrace{p_{\mathrm{p},m}s_{p,m}}_{\substack{\text{precoded private message corresponding to}\\ \text{local estimate of the model}}} \in \mathbb{C}, \qquad (1)$$

where $p_{\mathrm{c},m} \in \mathbb{C}$ and $p_{\mathrm{p},m} \in \mathbb{C}$ are the complex precoders employed by the m-th ED to its common and private messages components, respectively, while $s_\mathrm{c} \in \mathcal{S}_\mathrm{c}$ and $s_{p,m} \in \mathcal{S}_\mathrm{p}$ are the corresponding common and private RSMA transmit

[3] Notice that the encoding of the private messages $s_{\mathrm{p},m}$ is carried out in knowledge of (that is, excluding) the common message s_c, such that there is no redundancy in the pairs $(s_\mathrm{c}, s_{\mathrm{p},m})$.

symbols, selected from capacity-achieving complex-valued constellations of unit average power.

In light of equation (1), the transmit signal vector collecting the signals transmitted simultaneously by all EDs in the cluster can be compactly described as

$$\mathbf{x} \triangleq \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix} = \mathbf{p}_\mathrm{c} s_\mathrm{c} + \mathbf{p}_\mathrm{p} \odot \mathbf{s}_\mathrm{p} \in \mathbb{C}^{M\times 1}, \tag{2}$$

where the vectors $\mathbf{p}_\mathrm{c} \triangleq [p_{\mathrm{c},1}, \ldots, p_{\mathrm{c},M}]^\mathsf{T} \in \mathbb{C}^{M\times 1}$ and $\mathbf{p}_\mathrm{p} \triangleq [p_{\mathrm{p},1}, \ldots, p_{\mathrm{p},M}]^\mathsf{T} \in \mathbb{C}^{M\times 1}$ carry the scalar[4] complex precoders applied by each ED to the common and private message components, respectively, while $\mathbf{s}_\mathrm{p} \triangleq [s_{\mathrm{p},1}, \ldots, s_{\mathrm{p},M}]^\mathsf{T} \in \mathbb{C}^{M\times 1}$ is the vector of private symbols.

It follows that the UL signal received at the BS is given by

$$y = \mathbf{h}^\mathsf{H}\mathbf{x} + n \in \mathbb{C}, \tag{3}$$

where $\mathbf{h} \in \mathbb{C}^{M\times 1}$ is the channel vector with elements $h_m \triangleq g_m d_m^{-2}$, with $g_m \sim \mathcal{CN}(0, 1)$ denoting the small scale fading component and d_m denoting the distance (in meters) of the path between the m-th ED and the BS, respectively; and $n \sim \mathcal{CN}(0, N_0)$ is the narrowband additive white Gaussian noise (AWGN) at the BS, with power spectral density (*i.e.*, variance per Hz) given by N_0.

In light of equations (2) and (3), the signal-to-interference-and-noise ratio (SINR) corresponding to the common message at the BS is given by

$$\gamma_\mathrm{c} = \frac{|\mathbf{h}^\mathsf{H}\mathbf{p}_\mathrm{c}|^2}{\sum_{m\in\mathcal{M}} |h_m p_{\mathrm{p},m}|^2 + BN_0}, \tag{4}$$

where we emphasize that the simultaneous transmission of the same common message s_c by all EDs resembles a classic distributed beamforming scheme [32].

Assuming that due the gain of the distributed beamforming over the common message is sufficiently high for the common message component s_c to be perfectly recovered and removed from the received signal via interference cancellation (IC) techniques [22], the SINR corresponding to the private message from the m-th ED can then be modeled as

$$\gamma_{\mathrm{p},m} = \frac{|h_m p_{\mathrm{p},m}|^2}{\sum_{i\in\mathcal{M}, i\neq m} |h_i p_{\mathrm{p},i}|^2 + BN_0}. \tag{5}$$

Notice that similarly to the above, NOMA-based successive interference cancellation (SIC) can also be incorporated into the proposed UL-RSMA system, in contrast to usual DL-RSMA where the EDs can only estimate the common message. To that end, taking into account that the total number of possible decoding orders scales combinatorially with M, and that it has been established that the optimal decoding order in single-cell NOMA is independent of power allocation [33], we consider the standard "greedy" decoding order based on the channel gains, where the ED with largest channel gain has the highest decoding priority. In addition, it is assumed for convenience and without loss of generality that the ED indices are ordered with respect to the decoding order, such that ED 1 is decoded first, and ED M is decoded last.

In light of the above, the SINR of the private message of the m-th ED is given by

$$\gamma_{p,m}^{\mathrm{NOMA}} = \frac{|h_m p_{\mathrm{p},m}|^2}{\sum_{i=m+1}^{M} |h_i p_{\mathrm{p},i}|^2 + BN_0}, \tag{6}$$

where the difference from equation (5) is that the interference term in the denominator contains only the interference due to the remaining messages according to the decoding order.

With basis on the latter SINR expressions, the corresponding rates for the common and private messages are respectively given by

$$R_\mathrm{c} = \log_2(1 + \gamma_\mathrm{c}), \tag{7}$$

$$R_{\mathrm{p},m} = \log_2(1 + \gamma_{\mathrm{p},m}), \tag{8}$$

with the total rate R of the RSMA system given by

$$R = R_\mathrm{c} + \sum_{m\in\mathcal{M}} R_{\mathrm{p},m}. \tag{9}$$

B. TIME CONSUMPTION MODELS

Referring to Fig. 4, and following related literature [10], [11], [12], [13], in this subsection we shall quantify the time consumed during each step of one cycle of the proposed RSMA-interfaced CWFL scheme. To that end, we first highlight the slight notational and nomenclature distinction generally employed hereafter, between time costs related to the communication of information among devices, referred to as transmission times and denoted by the letter t; and time costs associated with computing tasks, referred to as delays and denoted by the letter τ.

1) DOWNLINK TRANSMISSION TIME

Adhering to the aforementioned notation and nomenclature, as per Fig. 4, the first time cost associated with an i-th CWFL cycle is the DL transmission time $t_{m:\mathrm{DL}}^{(i)}$ required by the BS to convey the model $\mathcal{L}^{(i)}$ to the m-th ED in the cluster.

Assuming the transmit power of the BS to be sufficiently high such that the DL channels have rates far superior to those of UL channels [11], it can be considered that $t_{m:\mathrm{DL}}^{(i)}$ reduces fundamentally to the propagation delay from the BS to the ED, which can therefore be described as

$$t_{m:\mathrm{DL}}^{(i)} \approx \frac{d_m}{c} \approx \epsilon_m \ll t_{m:\mathrm{UL}}^{(i)}, \; \forall\, m \in \mathcal{M}, \tag{10}$$

where c is the speed of light and ϵ_m is the distinct propagation delay for user m, assumed to be much smaller[5] that the UL transmission time.

[4] The MIMO case incorporating transmit (TX) and receive (RX) beamforming will be considered in a follow-up work.

[5] While much smaller than the UL times, DL times cannot be considered to be synchronous without violating the water-filling principle, such that the minimization of round delays is a fundamentally coupled optimization problem, as described latter in Section IV.

2) LOCAL MODEL UPDATING TIME

Next, consider the computation delay $\tau_m^{(i)}$ associated with the m-th ED training the model $\mathcal{L}^{(i)}$ with local information in order to obtain the local model $\mathcal{L}_m^{(i)}$. Following [13], and assuming that the local training at the m-th ED is based on a dataset of size D_m bits, processed with processor of cycle frequency f_m bits per cycle, and requiring altogether k_m cycles to converge, we have

$$\tau_m^{(i)} = \frac{k_m D_m}{f_m}, \tag{11}$$

where we may emphasize two assumptions employed by the model, namely, that the computational performance of the m-th ED is assumed constant across multiple cycles, such that the superscript (i) can be omitted; and that devices have heterogeneous computational capabilities.[6]

Notice that under the assumption that each ED has a maximum cycle frequency $f_m \leq f_m^{\max}$, a lower bound $\tau_m^{\min}$ of the computational delay may be obtained as

$$\tau_m^{\min} \triangleq \max_{m \in \mathcal{M}} \left(\frac{k_m D_m}{f_m^{\max}} \right), \tag{12}$$

from which we readily obtain

$$\tau_m^{(i)} \geq \tau_m^{\min}, \ \forall\, m. \tag{13}$$

3) LOCALLY COMMON MODEL ELECTION TIME

In principle, the locally common model $\mathcal{L}_c^{(i)}$ could be obtained via any number of distributed consensus algorithms [34], including blockchain-based approaches recently proposed for the FL problem itself [35]. It is well-known, however, that the time and delay associated with obtaining distributed ledger-type local consensus can be very large [36]. We therefore discard such alternatives and consider instead the much simpler and fast alternative of electing one of the local private models $\mathcal{L}_m^{(i)}$ as the locally common model $\mathcal{L}_\text{c}^{(i)}$, taking into account that: a) in the context of WFL, the optimal integration of the local models $\{\mathcal{L}_m^{(i)}\}$ is best performed by the CPU; and b) electing one of the local models $\mathcal{L}_m^{(i)}$ as $\mathcal{L}_\text{c}^{(i)}$ eliminates costs in computing $\mathcal{L}_\text{c}^{(i)}$, such that we have

$$\tau_\text{c}^{(i)} = 0. \tag{14}$$

This leaves us with the task of designing a mechanism to elect[7] the locally common model $\mathcal{L}_\text{c}^{(i)}$ among all available local private models $\mathcal{L}_m^{(i)}$. One alternative to that end is to simply employ a round-robin (RR) scheduling mechanism, whereby the private model of each of the M EDs is selected sequentially, which can be described mathematically as setting $\mathcal{L}_{1+\text{mod}(i-1,M)}^{(i)} \rightarrow \mathcal{L}_\text{c}^{(i)}$.

In that case, suffice it that at each i-th FL cycle, the $(1 + \text{mod}(i-1, M))$-th ED locally broadcasts its private model to the cluster, such that under the assumption that EDs are sufficiently closely located for the D2D channels to have much higher capacity than the UL channels, implies that

$$t_\text{c}^{(i)} \approx 0. \tag{15}$$

Possibly improved variations of the above mechanism can obviously obtained by optimizing the RR scheduling at the CPU itself, based for instance on the contributions of each of the local models in determining the central model at previous WFL cycles. That may have the potential advantage of yielding faster convergence of the WFL process as a whole, if for instance the RR scheduling is optimized to prioritize local models based on their extrinsic information that contribute to consensus carried out at the CPU.

Finally, an alternative to the RR-based election of $\mathcal{L}_\text{c}^{(i)}$ is a game-theoretical selection mechanism inspired by the Dutch auction (DA) protocol [40]. In a Dutch auction, the asked price starts high and is slowly lowered until taken by a bidder.

It is well known that under properly designed "pricing", DAs achieve extremely fast closure with minimal signaling [41], and that the mechanism can be implemented for the efficient and fully-distributed selection of nodes in ad-hoc networks [42] by taking for "price" any parameter independently obtainable at the devices, such as CSI, received signal strength indicator (RSSI), etc.

In an unoptimized WFL scheme, where the DL transmission times $t_{m:\text{DL}}^{(i)}$ from the BS to each ED, as well as the processing delays $\tau_m^{(i)}$ required for the local model update are not synchronized, a simple "price" to consider is the quantity $t_{m:\text{DL}}^{(i)} + \tau_m^{(i)}$. In other words, in this case the DA amounts to an "early bird catches the worm" variation of the classic first-come first-served (FCFS) protocol [43], in which the ED with the earliest ready-to-transmit time is the winner of the auction and therefore simply broadcasts its own updated model to the other EDs in the cluster, as the local common model.

In turn, in an optimized CWFL system, where the quantity $t_{m:\text{DL}}^{(i)} + \tau_m^{(i)}$ is identical to all EDs, a more sensible "price" to base the DA on is the Hamming distances between the bit encodings of the central model $\mathcal{L}^{(i)}$ and those of the local private models $\mathcal{L}_m^{(i)}$. In this case, it is shown in Appendix A that the closure time of the DA is exponentially distributed, with a decaying parameter directly proportional to the number of bidders and inversely proportional to the similarity among local models, which in turn implies that the average time of closure of the DA-based election mechanism decreases with the number of EDs in the cluster as well as with the similarity among private models.[8]

All in all, based on the discussion above, we conclude that both $\tau_\text{c}^{(i)}$ and $t_\text{c}^{(i)}$ can be neglected compared to $\tau_m^{(i)}$ and $t_m^{(i)}$, which is considered next.

[6] Notice that heterogeneous computational capabilities may arise even among devices with identical hardware, due to differences in state variables such as stored energy and memory occupancy.

[7] The problem relates closely to the problem of opportunistic relay selection, for which a rich literature also exists [37], [38], [39] that demonstrate that fast mechanisms with little feedback information can be designed.

[8] The details of the cluster formation and operation is beyond the scope of the article, and can be found in abundant literature, including [16], [17], [18], [19], [20].

4) UPLINK TRANSMISSION TIME

Next, let us turn our attention to the time consumed by the cluster of EDs to transmit their local model updates to the CPU. On that matter, let us start by emphasizing that the signal transmission model described by equations (2) and (3) imply that the UL signals' periods and bandwidths of all EDs must be identical. In other word, we must have

$$t^{(i)}_{m:\text{UL}} = t^{(i)}, \forall\, m \in \mathcal{M}. \tag{16}$$

Denoting therefore the bandwidth of the uplink channels by B, and given the rates described by equations (7) and (8), the total information that is conveyed by the cluster of EDs to the CPU during UL transmission is given by

$$\begin{aligned} Z^{\text{tot}} &= t^{(i)}B\Big(R_{\text{c}} + \sum_{m\in\mathcal{M}} R_{\text{p},m}\Big) \\ &= \underbrace{t^{(i)}B\log_2(1+\gamma_{\text{c}})}_{\text{information conveyed via } s_{\text{c}}} + \sum_{m\in\mathcal{M}} \overbrace{t^{(i)}B\log_2(1+\gamma_{\text{p},m})}^{\text{information conveyed via } s_{\text{p},m}}. \end{aligned} \tag{17}$$

Under the assumptions that a model update requires a minimum amount of information denoted by Z^{min}, such that $Z^{\text{tot}} \geq Z^{\text{min}}$, and that the portions of information conveyed by the common and private messages are complementary, equation (17) suggests, under the synchronicity of the transmission system implied by equation (2), the following bounds on the transmission time during the i-th UL step

$$t^{(i)} \geq \frac{\rho Z^{\text{tot}}}{B\log_2(1+\gamma_{\text{c}})} \tag{18}$$

$$t^{(i)} \geq \frac{(1-\rho)Z^{\text{tot}}}{MB\log_2(1+\min\{\gamma_{\text{p},m}\})}, \tag{19}$$

where ρ is a rate-splitting factor that describes the portions of information in Z^{tot} corresponding to the common and private components of the RSMA messages, respectively, such that

$$0 \leq \rho \leq 1. \tag{20}$$

5) CENTRAL MODEL UPDATING DELAY

Finally, we address the delay $\tau^{(i)}_{\text{u}}$ associated with updating the central model, which in fact can be ignored, under the assumption that the CPU has much higher processing power than the EDs in the cluster, such that we here set

$$\tau^{(i)}_{\text{u}} \approx 0. \tag{21}$$

C. ENERGY CONSUMPTION MODELS

Next, let us address the energy consumption associated with the proposed RSMA-interfaced CWFL system. In alignment with the discussions in the preceding subsection, and under the assumption that the BS has an unlimited energy source, suffice it to that end consider only the costs most critical to the EDs, namely, the energy required privately by each to update the DL central model $\mathcal{L}^{(i)}$ into its locally trained model $\mathcal{L}^{(i)}_m$, and the subsequent UL transmissions of both the elected locally common model $\mathcal{L}^{(i)}_{\text{c}}$ and the private models $\mathcal{L}^{(i)}_m$ from the cluster to the BS.

1) ENERGY CONSUMPTION OF LOCAL MODEL UPDATING

Following related work [10], [13], the energy required for local training of the model at the m-th ED can be modeled as

$$E^{\text{comp}}_m = \zeta_m k_m D_m f^2_m = \zeta_m \frac{k^3_m D^3_m}{\tau^2_m}, \tag{22}$$

where ζ_m is a energy-efficiency constant that captures hardware capabilities of the m-th ED.

2) ENERGY CONSUMPTION OF UL TRANSMISSIONS

In turn, in light of equations (1) and (16), the energy required by the m-th ED to transmit its RSMA signal is given by

$$E^{\text{trans}}_m = (P_{\text{c},m} + P_{\text{p},m})\, t^{(i)}, \tag{23}$$

where we emphasize once again that the signals transmitted by all EDs must have the same duration $t^{(i)}$, and the quantities $P_{\text{c},m} \triangleq |p_{\text{c},m}|^2$ and $P_{\text{p},m} \triangleq |p_{\text{p},m}|^2$ are the powers of the precoders applied by the m-th ED to the common and private messages, respectively.

From equations (22) and (23), the total energy consumed by the m-th ED at the i-th CWFL cycle is given by

$$E^{\text{comp}}_m + E^{\text{trans}}_m = \zeta_m \frac{k^3_m D^3_m}{\tau^2_m} + (P_{\text{c},m} + P_{\text{p},m})t^{(i)} \leq E^{\text{max}}_m, \tag{24}$$

where E^{max}_m denotes an energy constraint of the m-th ED.

IV. ROUND DELAY MINIMIZATION OF CWFL

A. FORMULATION OF RSMA PRECODER PROBLEM

In view of the system model described above, we elaborate in this section an optimization problem aimed at minimizing the total WFL round delay. For the sake of simplicity, and without significant loss of insight, we shall consider in this version of the article the simplified case in which the number of samples in the training data set, the cycle frequency of local processors, the hardware-dependent energy-efficiency constant, and the stored energy at each ED in the cluster are equal for all EDs, that is $D_m = D$, $f_m = f$, and $\zeta_m = \zeta$. In addition, given that the optimization problem is meant to be solved centrally at the CPU, in knowledge of the aforementioned parameters and the CSI, which can be communicated by the EDs or estimated during uplink sessions, we shall also without loss of generality hereafter drop the superscript $(\cdot)^{(i)}$ indicating cycle counter.

With these remarks made, an extension of the problem proposed in [13] to the RSMA scheme introduced here is

$$\min_{\tau,t,\mathbf{p}_{\text{c}},\mathbf{p}_{\text{p}}} \;\tau + t, \tag{25a}$$

$$s.t. C_1 : \tau \geq \tau^{\text{min}}, \tag{25b}$$

$$C_2 : tB\log_2(1+\gamma_{\text{c}}) \geq \rho Z^{\text{tot}}, \tag{25c}$$

$$C_3 : tMB\log_2(1+\min\{\gamma_{\text{p},m}\}) \geq (1-\rho)Z^{\text{tot}}, \tag{25d}$$

$$C_4: 0 \le \rho \le 1, \tag{25e}$$

$$C_5: \zeta \frac{k_m^3 D^3}{\tau^2} + (P_{\mathrm{c},m} + P_{\mathrm{p},m})\, t \le E_m^{\max}, \ \forall m \in \mathcal{M}, \tag{25f}$$

$$C_6: t \ge 0, \tag{25g}$$

$$C_7: P_{\mathrm{c},m} \ge 0, \ \forall m \in \mathcal{M}, \tag{25h}$$

$$C_8: P_{\mathrm{p},m} \ge 0, \ \forall m \in \mathcal{M}, \tag{25i}$$

where the objective function (25a) reflects the fact that the local model updating delay τ given in equation (11) and the uplink transmission time t in equation (17) are the most significant and optimizable time costs of the CWFL cycle, as illustrated in Fig. 4 and described in Subsection (III-B); the constraints C_1:(25b) through C_5:(25f) follow directly from equations (13), (18)-(20), and (24), respectively; and the remaining constraints are due to obvious physical conditions.

We highlight that as a consequence of the proposed incorporation of RSMA for the UL transmission of the local models $\mathcal{L}_{\mathrm{c}}^{(i)}$ and $\mathcal{L}_m^{(i)}$ to the BS, the optimization problem formulated in (25) is a pre-coding design problem, unlike the power control problem presented in [13]. It follows, therefore, that constraints (25h) and (25i) are redundant, as those conditions are satisfied absolutely under the definitions $P_{\mathrm{c},m} \triangleq |p_{\mathrm{c},m}|^2$ and $P_{\mathrm{p},m} \triangleq |p_{\mathrm{p},m}|^2$, such that problem (25) simplifies to

$$\min_{\tau, t, \mathbf{p}_{\mathrm{c}}, \mathbf{p}_{\mathrm{p}}} \ \tau + t, \tag{26a}$$

$$s.t. \quad C_1: \tau \ge \tau^{\min}, \tag{26b}$$

$$C_2: tB \log_2(1 + \gamma_{\mathrm{c}}) \ge \rho Z^{\mathrm{tot}}, \tag{26c}$$

$$C_3: tMB \log_2(1 + \min\{\gamma_{\mathrm{p},m}\}) \ge (1-\rho) Z^{\mathrm{tot}}, \tag{26d}$$

$$C_4: 0 \le \rho \le 1, \tag{26e}$$

$$C_5: \zeta \frac{k_m^3 D^3}{\tau^2} + (P_{\mathrm{c},m} + P_{\mathrm{p},m})\, t \le E_m^{\max}, \forall m \in \mathcal{M}, \tag{26f}$$

$$C_6: t \ge 0. \tag{26g}$$

B. RELAXATION OF RSMA PRECODER VIA FRACTIONAL PROGRAMMING

Notice that the optimization problem (26) is not convex due to constraints C_2 and C_3, with an additional challenge posed by the coupling of the optimization variables t, $P_{\mathrm{c},m}$, and $P_{\mathrm{p},m}$ in C_5. To mitigate this challenge, we introduce in the sequel a fractional programming-based variation of the problem, enabled by the quadratic transform described in [29]. In particular, C_2 and C_3 are convexized by applying the quadratic transform unto γ_{c} and $\gamma_{\mathrm{p},m}$, which yields

$$\gamma_{\mathrm{c}}^{\mathrm{qt}} = 2\Re\{v_{\mathrm{c}}^* \mathbf{h}^{\mathrm{H}} \mathbf{p}_{\mathrm{c}}\} - v_{\mathrm{c}}^* \Big(\sum_{m \in \mathcal{M}} |h_m p_{\mathrm{p},m}|^2 + BN_0 \Big) v_{\mathrm{c}}, \tag{27a}$$

Algorithm 1 Delay Minimization for CT-RSMA

```
Set $i = 1$
Initialize auxiliary variables $\{v_{\mathrm{c}}^{(0)}, v_{\mathrm{p},m}^{(0)}\}$, $\forall m \in \mathcal{M}$
repeat
    repeat
        Solve problem (29) on $\mathbf{p}_{\mathrm{p}}$, $\mathbf{p}_{\mathrm{c}}$, $\tau$, with fixed $t$
        Update $\{v_{\mathrm{c}}^{(i)}, v_{\mathrm{p},m}^{(i)}\}$ using (28a) and (28b)
        Set $i = i + 1$
    until inner loop convergence criterion is satisfied
    Solve problem (29) on $t$ and $\tau$, with fixed $\mathbf{p}_{\mathrm{p}}$, $\mathbf{p}_{\mathrm{c}}$
until outer loop convergence criterion is satisfied
```

$$\gamma_{\mathrm{p},m}^{\mathrm{qt}} = 2\Re\{v_{\mathrm{p},m}^* h_m p_{\mathrm{p},m}\} - v_{\mathrm{p},m}^* \Big(\sum_{i=m+1}^{M} |h_i p_{\mathrm{p},i}|^2 + BN_0 \Big) v_{\mathrm{p},m}, \tag{27b}$$

where v_{c} and $v_{\mathrm{p},m}$ are auxiliary resulting from the quadratic transformation of each fractional term, which are updated at each iteration according to the closed-form expressions

$$v_{\mathrm{c}}^{\star} = \frac{\mathbf{h}^{\mathrm{H}} \mathbf{p}_{\mathrm{c}}}{\sum_{m \in \mathcal{M}} |h_m p_{\mathrm{p},m}|^2 + BN_0}, \tag{28a}$$

$$v_{\mathrm{p},m}^{\star} = \frac{h_m p_{\mathrm{p},m}}{\sum_{i=m+1}^{M} |h_i p_{\mathrm{p},i}|^2 + BN_0}. \tag{28b}$$

Substituting the pair of equations (27) into the SINR terms of constraints (26c) and (26d), and replacing the min operator in inequality (26d) by the enforcement of that constraint for all m yields the convexized reformulated problem

$$\min_{\tau, t, \mathbf{p}_{\mathrm{c}}, \mathbf{p}_{\mathrm{p}}} \ \tau + t, \tag{29a}$$

$$s.t. C_1: \tau \ge \tau^{\min}, \tag{29b}$$

$$C_2: tB \log_2(1 + \gamma_{\mathrm{c}}^{\mathrm{qt}}) \ge \rho Z^{\mathrm{tot}}, \tag{29c}$$

$$C_3: tMB \log_2(1 + \gamma_{\mathrm{p},m}^{\mathrm{qt}}) \ge (1-\rho) Z^{\mathrm{tot}}, \ \forall m \in \mathcal{M}, \tag{29d}$$

$$C_4: 0 \le \rho \le 1, \tag{29e}$$

$$C_5: \zeta \frac{k_m^3 D^3}{\tau^2} + \Big(|p_{\mathrm{c},m}|^2 + |p_{\mathrm{p},m}|^2 \Big) t \le E_m^{\max}, \ \forall m \in \mathcal{M}, \tag{29f}$$

$$C_6: t \ge 0. \tag{29g}$$

The direct solution of problem (29) over t, τ, $\mathbf{p}_{\mathrm{c}}$, and $\mathbf{p}_{\mathrm{p}}$ simultaneously is, however, quite challenging, due to the coupling of these optimization variables within C_5. This latter difficulty can, nevertheless, be circumvented by applying a nested multi-staging approach [44], [45], as follows.

After the initialization of the auxiliary variables $v_{\mathrm{c}}^{\star}$ and $v_{\mathrm{p},m}^{\star}$ with feasible values, the problem is first solved in an inner loop for a fixed value of t, over the remaining variables $\mathbf{p}_{\mathrm{p}}$, $\mathbf{p}_{\mathrm{c}}$ and τ. The inner loop is repeated until convergence of $\mathbf{p}_{\mathrm{p}}$, $\mathbf{p}_{\mathrm{c}}$ and τ, or until a desired maximum number of iterations, and is required because C_2 and C_3 are convexized via FP, which is fundamentally an iterative quadratic majorization method that must be solved till local convergence before the problem can be solved on the remaining variables. That is, in fact, the next stage of the algorithm, which in the outer loop solves the problem t and τ, with all with the variables $\mathbf{p}_\mathrm{p}$ and $\mathbf{p}_\mathrm{c}$ fixed.

The auxiliary variables v_c and $v_{\mathrm{p},m}$ are updated at each step of the loop, which is terminated upon convergence or after a fixed amount of iterations. The procedure described above is summarized as a pseudo-code in Algorithm 1.

C. A NOTE ON THE STATE-OF-THE-ART

Before comparing the proposed method against the SotA, in particular the NOMA-based WFL scheme recently presented in [13], let us first describe the relationship between these two methods. To that end, we compare directly the optimization problem in (29) with that described by [13, Eq.(23)].

First, notice that by setting the rate-splitting factor ρ in equation (29) to zero, not only the constraints C_2 and C_4 are void, but also the system reduces to one in which all information is transmitted as private components of the message, resulting in a NOMA transmission model similar to that assumed in [13]. In addition, still as a consequence of setting $\rho = 0$, it can be seen that constraint C_3 of problem (29), namely, the limitation on the total information conveyed by ED m as given in inequality (29c), is also equivalent to [13, Const. C_2 of Eq. (23)]. To elaborate, these constraints differ only in the fact that the optimization variable in [13, Const. C_2 of Eq.(23)] is the vector of consumed energy $\mathbf{E} = [E_1, \cdots, E_M]$, whereas here the latter is mapped onto a vector of transmit powers $\mathbf{p}_\mathrm{p}$ (in case of $\rho = 0$). Since E and $\mathbf{p}_\mathrm{p}$ relate to each other directly by multiplication of the corresponding transmission time t, the two constraints are fully equivalent. Notice also that the energy constraint formulated as [13, Const. C_1 of Eq.(23)] requires the additional constraint $E_n \geq 0$, whereas here a corresponding constraint is unnecessary, since C_5 is formulated directly with basis on the power of a precoder (*i.e.*, $|p_{\mathrm{p},m}|^2$ for the case of $\rho = 0$), which is inherently non-negative.

With these remarks made, it is clear that the proposed problem formulation described by equation (29) is a generalization of that of [13, Eq. (23)], such that in the proposed method the information can be split into common and private transmit components at an arbitrary ratio as governed by ρ, and which incorporates the latter as a special case the proposed method reduces to by setting $\rho = 0$.

We further emphasize that even in that reduced case, however, two key differences between the two approaches exist. The first is on how the problems are solved. In particular, the non-convexity of [13, Const. C_2 of Eq.(23)] is handled in [13] by rewriting the optimization problem into [13, Eq.(26)], while here we instead employ fractional programming technique described in [29] and [30]. And the second is that here the problem addressed concerns precoding, that is, the optimum adjustment of both the power and the phase of transmitted signals, while [13] focuses only on power allocation.

TABLE 1. Simulation parameters.

Parameter	Value	Parameter	Value
$f_m^{\max}$	1.5GHz	D_m	1 Mbit
B	1 MHz	N_0	-174 dBm/Hz
ζ	10^{-27}	k_m	$\sim \mathcal{U}(10, 40)$
M	10 EDs	d_m	$\sim \mathcal{U}(0, 500m)$

In view of all the above, in the comparisons to follow all results obtained with $\rho = 0$ can be considered as slightly improved variations of [13].

V. PERFORMANCE EVALUATION

In this section, the effectiveness of the proposed algorithm is evaluated via computer simulations and comparisons against the most relevant SotA technique, namely the NOMA-based WFL scheme recently proposed in [13]. For the sake of simplicity, and to allow direct comparison, all simulations are performed under set-ups in which nodes transmit files with identical size, although the number of clock cycles k_m required for local processing are distinct, taken randomly from a uniform distribution. Simulation parameters are as shown in Table 1.

The algorithm is implemented in MATLAB, with the minimization problem solved using the CVX optimization package. Our first results are shown in Figure 5, which compares the latency achieved by systems with $M = 10$ and $M = 20$ EDs under different values of rate-splitting factor ρ. It can be seen that in general, the RSMA approach ($\rho > 0$) outperforms[9] the NOMA alternative ($\rho = 0$), with the optimum value of ρ – *i.e.*, the rate splitting factor that leads to lowest latencies – increasing with M.

Next, Figure 6 compares the performance of the NOMA-based SotA scheme ($\rho = 0$) with that of the proposed RSMA-interfaced method as function of the number of EDs in the system, with the optimal rate splitting factor used at in each case. The results show that when the number of EDs M in the system increases and the optimal value of ρ is employed accordingly, the gain in latency reduction over the NOMA alternative becomes even larger.

To offer another perspective on the significance of the advantage of RSMA over NOMA as the most suitable transmission model for wireless federated learning, we further compare in Figures 7(a) and 7(b) the performance of CWFL system with different cluster sizes M and different splitting ratios ρ, as a function of the data size Z. The results not only corroborate those of Figure 5, but also reveal that the relative gain achieved by optimizing ρ is more significant as the cluster size M and the data size Z increase.

Obviously, the results shown in Figures 5 to 7(b) are theoretical, since the focus of this work was solely on the optimization of the transmission parameters associated with the integration wireless transmissions with FL approach in

[9] With the system parameters of Fig. 5, the proposed method is shown to achieve a latency gain of 4.2% - 6.3% over the SotA method [13].

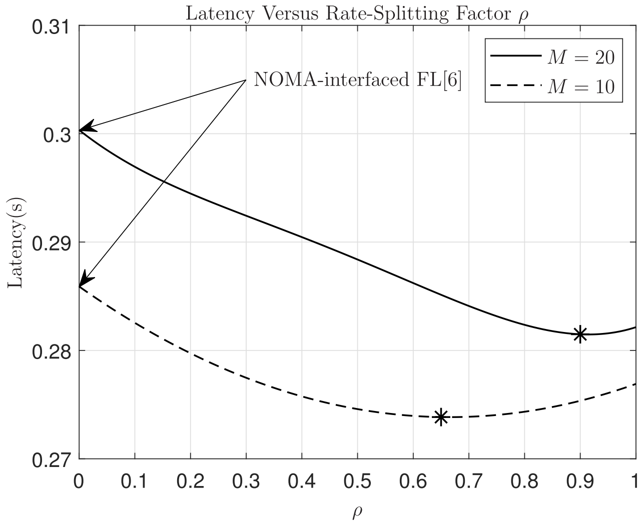


FIGURE 5. Latency as a function of ρ, with $E_m^{\max} = 2$J and $Z = 0.8$ Mbits, where $\rho = 0$ corresponds to the NOMA-interfaced WFL SotA method of.

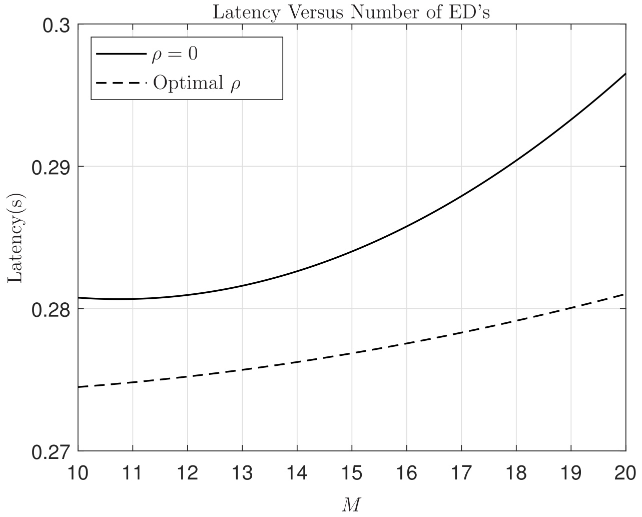


FIGURE 6. Latency as a function of total data size Z, with $E_m^{\max} = 2$J and $M = 10$ and $M = 20$, for different splitting factors ρ, where $\rho = 0$ corresponds to the NOMA-interfaced WFL SotA method of [13].

a clustered setting. However, the results are also consistent with the following insights regarding the CWFL problem. First, notice that the larger the cluster size, the larger the diversity of information gathered by the federated learning procedure, such that the greater is the amount of extrinsic information contained in the locally trained model with the greatest Hamming distance to the global model. Secondly, as shown in Appendix A, the time to select such a "most representative local model" via Dutch Auction reduces with the cluster size M. Thirdly, the larger the cluster size, the greater the beamforming gain resulting from the collaborative transmission of the elected locally common model optimized by the precoding vector $\mathbf{p}_\mathrm{c}$. Fourth and finally, under all the above, the larger the cluster the smaller is the significance of the additional information offered by each of the distinct ED individually, since there are more EDs contributing to training the model. The results motivate further investigation in which the Dutch Auction method and RSMA are integrated with the actual training of FL models, which shall be pursued in a follow up work.

VI. CONCLUSION

We considered the latency minimization of WFL schemes, proposing an optimization scheme and an RSMA-base uplink transmission interface in order to reduce the idleness in the local learning and uploading stages of each FL round. The new method is suitable for CWFL paradigm, where EDs are locally concentrated, such that they can exchange information in a D2D fashion and one of the locally trained models can be elected at each round as locally common and therefore be used in the construction of the common message component of the RSMA scheme. The optimization problem formulated captures the interplay between optimum processing and transmission time, as well as the precoding coefficients applied by the clustered EDs during uplink. The non-convexity of the problem, resulting from constraints built around the SINRs at the EDs, was then circumvented by the FP technique, allowing for its efficient solution. The proposed RSMA interface, which includes a preceding NOMA-based architecture of [13] as a special case, is shown to significantly outperform the latter, with gains in latency reduction increasing with the number of EDs in the cluster is sufficiently large. As a future work, the interplay between local model selection and the convergence of the FL scheme will be investigated.

.

APPENDIX A: TIME TO ELECT $\mathcal{L}_\mathrm{c}^{(i)}$ VIA DUTCH AUCTION

Let Z denote the length of the binary codewords encoding the FL models $\mathcal{L}^{(i)}$ and $\mathcal{L}_m^{(i)}$. Let the Hamming distance between any two models be denoted by $\mathcal{B}(\mathcal{L}_p^{(i)}, \mathcal{L}_q^{(i)})$, and β denote the average number of common bits between the encodings of the models $\mathcal{L}_m^{(i)}$, normalized by Z, that is

$$\beta \triangleq \frac{Z - \mathbb{E}\left[\mathcal{B}(\mathcal{L}_p^{(i)}, \mathcal{L}_q^{(i)})\right]}{Z}, \tag{30}$$

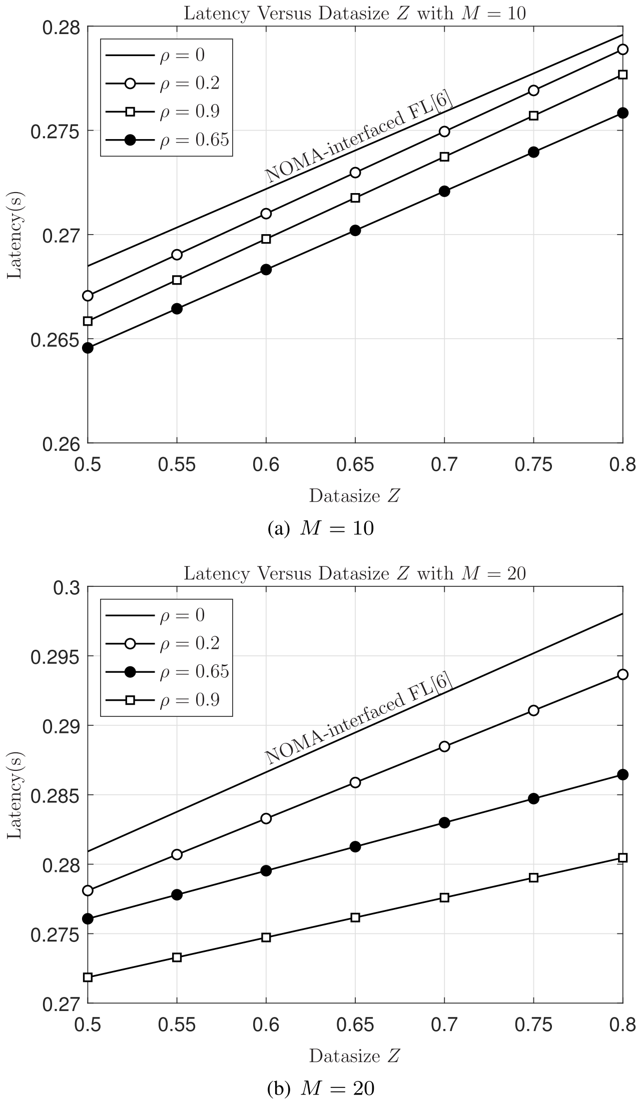
where the expectation is to be taken among all pairs of non-equal indices $(p, q) \in \mathcal{M}$, as well as over multiple FL cycles.

Next, consider the Hamming distances between local and central models, which can be described as

$$b_m^{(i)} \triangleq \mathcal{B}(\mathcal{L}_m^{(i)}, \mathcal{L}^{(i)}), \tag{31}$$

and whose ascending order will hereafter be considered, for convenience and without of generality, the order order of the indices m themselves, such that $b_1^{(i)} \leq b_2^{(i)} \leq \cdots \leq b_M^{(i)}$.

From the above, and given that the local models $\mathcal{L}_m^{(i)}$ are trained with private and distinct datasets, it follows that $b_m^{(i)}$ can be described as uniformly-distributed random variables in the interval $[0, (1 - \beta)Z]$, as illustrated in Fig. 8, with the time $t_\mathrm{c}^{(i)}$ required to elect the locally common model being proportional to the distance between b_M and the interval



(a) $M = 10$

(b) $M = 20$

FIGURE 7. Latency as a function of total data size Z, with $E_m^{\max} = 2$J and $M = 10$ and $M = 20$, for different splitting factors ρ, where $\rho = 0$ corresponds to the NOMA-interfaced WFL SotA method of [13].

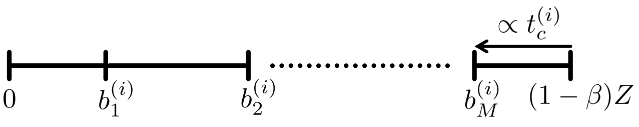


FIGURE 8. Changing ρ for different cost functions.

upper-limit $(1-\beta)Z$. In other words, at each i-th cycle, the random placement of $b_m^{(i)}$ in the interval $[0, (1-\beta)Z]$ is akin to a Poisson arrival process [46] with density

$$\lambda \triangleq \frac{M}{(1-\beta)Z}. \tag{32}$$

It follows that the distances between consecutive $b_m^{(i)}$ themselves are akin to inter-arrival variables, known to follow an exponential distribution with exponent coefficient given by the aforementioned density. In turn, assuming that $t_c^{(i)}$ is proportional to the last "inter-arrival" as illustrated in Fig. 8, we have

$$t_c^{(i)} \sim \frac{\lambda}{\alpha} e^{-\frac{\lambda}{\alpha} t}, \tag{33}$$

where α is a proportionality constant.

From the above, we finally obtain that the expected time of closure of a DA-based election of $\mathcal{L}_c^{(i)}$ is given by

$$\bar{t}_c^{(i)} = \frac{\alpha}{\lambda} = \frac{\alpha(1-\beta)Z}{M}. \tag{34}$$

A trivial inspection of equation (34) yields two important conclusions, namely, that the average time of closure of the DA-based election mechanism, decreases with the number of EDs in the cluster as well as with the similarity among private models, *i.e.*, $\bar{t}_c^{(i)} \to 0$, both with $M \to \infty$ and with $\beta \to 1$.

REFERENCES

[1] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "FedHealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Jul. 2020.

[2] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 46–51, Jun. 2020.

[3] J. Zhou, Q. Lu, W. Dai, and E. Herrera-Viedma, "Guest editorial: Federated learning for industrial IoT in Industry 4.0," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 8438–8441, Dec. 2021.

[4] L. Zhang, G. Feng, S. Qin, Y. Sun, and B. Cao, "Access control for ambient backscatter enhanced wireless Internet of Things," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5614–5628, Jul. 2022.

[5] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless federated learning (WFL) for 6G networks—Part I: Research challenges and future trends," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 3–7, Jan. 2022.

[6] O. Obulesu, M. Mahendra, and M. ThrilokReddy, "Machine learning techniques and tools: A survey," in *Proc. Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2018, pp. 605–611.

[7] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2017, pp. 1273–1282.

[8] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

[9] H. Hellström, V. Fodor, and C. Fischione, "Federated learning over-the-air by retransmissions," *IEEE Trans. Wireless Commun.*, early access, Apr. 26, 2023, doi: 10.1109/TWC.2023.3268742.

[10] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2019, pp. 1387–1395.

[11] Z. Yang, M. Chen, W. Saad, C. S. Hong, M. Shikh-Bahaei, H. V. Poor, and S. Cui, "Delay minimization for federated learning over wireless communication networks," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1–7.

[12] N. Huang, T. Wang, Y. Wu, S. Bi, L. Qian, and B. Lin, "Delay minimization for intelligent reflecting surface assisted federated learning," *China Commun.*, vol. 19, no. 4, pp. 216–229, Apr. 2022.

[13] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless federated learning (WFL) for 6G networks—Part II: The compute-then-transmit NOMA paradigm," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 8–12, Jan. 2022.

[14] C. Liu, T. J. Chua, and J. Zhao, "Time minimization in hierarchical federated learning," in *Proc. IEEE/ACM 7th Symp. Edge Comput. (SEC)*, Dec. 2022, pp. 96–106.

[15] Q. Ju and Y. Zhang, "Adaptive clustering for Internet of Battery-less Things," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[16] Y. Kim, E. A. Hakim, J. Haraldson, H. Eriksson, J. M. B. da Silva, and C. Fischione, "Dynamic clustering in federated learning," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[17] J. Wang, Z. Zhao, W. Hong, T. Q. S. Quek, and Z. Ding, "Clustered federated learning with model integration for non-IID data in wireless networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 1634–1639.

[18] F. Sattler, K.-R. Müller, T. Wiegand, and W. Samek, "On the Byzantine robustness of clustered federated learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8861–8865.

[19] J. Ma, G. Long, T. Zhou, J. Jiang, and C. Zhang, "On the convergence of clustered federated learning," 2022, *arXiv:2202.06187*.

[20] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 8076–8091, Dec. 2022.

[21] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

[22] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access: A new frontier for the PHY layer of 6G," in *Proc. IEEE 92nd Veh. Technol. Conf. (VTC-Fall)*, Nov. 2020, pp. 1–7.

[23] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 349–353, Mar. 2020.

[24] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2073–2126, 4th Quart., 2022.

[25] Y. Mao, B. Clerckx, and V. O. K. Li, "Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018, pp. 1–5.

[26] A. Schröder, M. Röper, D. Wübben, B. Matthiesen, P. Popovski, and A. Dekorsy, "A comparison between RSMA, SDMA, and OMA in multibeam LEO satellite systems," in *Proc. 26th Int. ITG Workshop Smart Antennas, 13th Conf. Syst., Commun., Coding*, 2023, pp. 1–6.

[27] B. Lee and W. Shin, "Max-min fairness precoder design for rate-splitting multiple access: Impact of imperfect channel knowledge," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 1355–1359, Jan. 2023.

[28] X. Lyu, S. Aditya, J. Kim, and B. Clerckx, "A prototype implementation of rate splitting multiple access using software-defined radios," 2023, *arXiv:2305.07361*.

[29] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[30] K. Shen and W. Yu, "Fractional programming for communication systems—Part II: Uplink scheduling via matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, May 2018.

[31] A. Shokrollahi, "Raptor codes," in *Proc. IEEE Inf. Theory Workshop Inf. Theory Wireless Netw.*, Jul. 2007, pp. 2551–2567.

[32] R. Mudumbai, G. Barriac, and U. Madhow, "On the feasibility of distributed beamforming in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1754–1763, May 2007.

[33] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 174–180, Oct. 2019.

[34] L. Li, P. Shi, X. Fu, P. Chen, T. Zhong, and J. Kong, "Three-dimensional tradeoffs for consensus algorithms: A review," *IEEE Trans. Netw. Service Manag.*, vol. 19, no. 2, pp. 1216–1228, Jun. 2022.

[35] C. Ma, J. Li, M. Ding, L. Shi, T. Wang, Z. Han, and H. V. Poor, "When federated learning meets blockchain: A new distributed learning paradigm," 2020, *arXiv:2009.09338*.

[36] A. Paz, H. R. Galeana, S. Schmid, U. Schmid, and K. Winkler, "Time complexity of consensus in dynamic networks under oblivious message adversaries," 2022, *arXiv:2202.12397*.

[37] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 659–672, Mar. 2006.

[38] K.-H. Liu and H.-H. Chen, "Probabilistic relay selection for fast selection cooperation in half-duplex wireless networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2009, pp. 1–6.

[39] M. E. Eltayeb, K. Elkhalil, H. R. Bahrami, and T. Y. Al-Naffouri, "Opportunistic relay selection with limited feedback," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2885–2898, Aug. 2015.

[40] E. Pereira, J. Reis, G. Goncalvesl, L. P. Reis, and A. P. Rocha, "Dutch auction based approach for task/resource allocation," in *Innovations in Mechatronics Engineering*, J. Machado, F. Soares, J. Trojanowska, and S. Yildirim, Eds. Cham, Switzerland: Springer, 2022, pp. 322–333.

[41] N. C. Luong, D. T. Hoang, P. Wang, D. Niyato, and Z. Han, "Applications of economic and pricing models for wireless network security: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2735–2767, 4th Quart., 2017.

[42] C. Lima and G. T. F. de Abreu, "Game-theoretical relay selection strategy for geographic routing in multi-hop WSNs," in *Proc. 5th Workshop Positioning, Navigat. Commun.*, Mar. 2008, pp. 277–283.

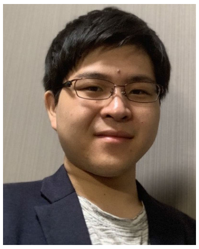
[43] W. T. Toor, J.-B. Seo, and H. Jin, "Practical splitting algorithm for multichannel slotted random access systems," *IEEE Trans. Mobile Comput.*, vol. 19, no. 12, pp. 2863–2873, Dec. 2020.

[44] A. Bekasiewicz and S. Koziel, "Reliable multistage optimization of antennas for multiple performance figures in highly dimensional parameter spaces," *IEEE Antennas Wireless Propag. Lett.*, vol. 18, no. 7, pp. 1522–1526, Jul. 2019.

[45] Y. Tian, Y. Zhang, Y. Su, X. Zhang, K. C. Tan, and Y. Jin, "Balancing objective optimization and constraint satisfaction in constrained evolutionary multiobjective optimization," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9559–9572, Sep. 2022.

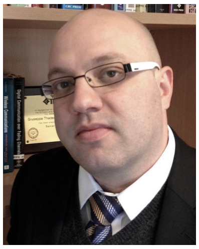
[46] G. Last and M. Penrose, *Lectures on the Poisson Process*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

NICLAS FÜHRLING (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and computer engineering from Constructor University, Bremen, Germany, in 2022. He is currently pursuing the M.Sc. degree in electrical engineering with the University of Bremen, with a focus on communication and information technology, while working on a research project at Constructor University, focusing on 6G connectivity. His research interests include wireless communications and signal processing.



KENGO ANDO (Graduate Student Member, IEEE) received the B.E. and M.E. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2020 and 2022, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Constructor University, Germany. His research interests include wireless communications and signal processing. He was a recipient of the YKK Graduate Fellowship for master's students from the Yoshida Scholarship Foundation, Japan, (2020–2022), and the Fellowship for Overseas Study from the KDDI Foundation, Japan, in 2022.

HYEON SEOK ROU (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and computer engineering from Constructor University, Germany, in 2021, where he is currently pursuing the Ph.D. degree in electrical engineering, funded by a research project from the Wireless Communications Technologies Group, Continental A.G. His research interests include joint communications and sensing (JCAS), Bayesian statistics, multi-dimensional modulation schemes, and mmWave/sub-THz MIMO wireless communications.



GIUSEPPE THADEU FREITAS DE ABREU (Senior Member, IEEE) received the B.Eng. degree in electrical engineering and the Latu Sensu degree in telecommunications engineering from Universidade Federal da Bahia (UFBA), Salvador, Bahia, Brazil, in 1996 and 1997, respectively, and the M.Eng. and D.Eng. degrees in physics, electrical, and computer engineering from Yokohama National University, Japan, in March 2001 and March 2004, respectively. He was a Postdoctoral Fellow and later an Adjunct Professor (docent) of statistical signal processing and communications theory with the Department of Electrical and Information Engineering, University of Oulu, Finland, from 2004 to 2006 and from 2006 to 2011, respectively. From April 2015 to August 2018, he simultaneously held a full professorship with the Department of Computer and Electrical Engineering, Ritsumeikan University, Japan. Since 2011, he has been a Professor of electrical engineering with Constructor University, Bremen, Germany. His research interests include communications and signal processing, including communications theory, estimation theory, statistical modeling, wireless localization, cognitive radio, wireless security, MIMO systems, ultrawideband and millimeter wave communications, full-duplex and cognitive radio, compressive sensing, energy harvesting networks, random networks, and connected vehicles networks. He received the Uenohara Award from Tokyo University, in 2000, for his master's thesis. He was a co-recipient of the Best Paper Award at several international conferences. He was awarded the Prestigious JSPS, the Heiwa Nakajima, and the NICT Fellowships, in 2010, 2013, and 2015, respectively. He served as an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2009 to 2014, and IEEE TRANSACTIONS ON COMMUNICATIONS, from 2014 to 2017, and an Executive Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2017 to 2021. He is serving as an Editor for the IEEE Signal PROCESSING LETTERS and IEEE COMMUNICATIONS LETTERS.