Joint Uplink-Downlink Resource Allocation for OFDMA-URLLC MEC Systems

Walid R. Ghanem, Vahid Jamali, Qiuyu Zhang, and Robert Schober
Friedrich-Alexander-University Erlangen-Nuremberg, Germany

Abstract—In this paper, we study resource allocation algorithm design for multiuser orthogonal frequency division multiple access (OFDMA) ultra-reliable low latency communication (URLLC) in mobile edge computing (MEC) systems. To achieve the stringent end-to-end delay and reliability requirements of URLLC MEC systems, we propose joint uplink-downlink resource allocation and finite blocklength transmission. Furthermore, we propose a partial time overlap between the uplink and downlink frames to minimize the end-to-end delay, which introduces new time causality constraints. Then, the proposed resource allocation algorithm is formulated as an optimization problem for minimization of the total weighted transmit power of the network under constraints on the minimum quality-of-service regarding the number of computed URLLC user bits within the maximum allowable computing time, i.e., the end-to-end delay of a computation task. Due to the non-convexity of the optimization problem, finding the globally optimal solution entails a high computational complexity which is not tolerable for real-time applications. Therefore, a low-complexity algorithm based on successive convex approximation is proposed to find a high-quality sub-optimal solution. Our simulation results show that the proposed resource allocation algorithm design facilitates the application of URLLC in MEC systems, and yields significant power savings compared to a benchmark scheme.

I. INTRODUCTION

Future wireless communication networks have several system design objectives including high data rates, reduced latency, and massive device connectivity. One important objective is to enable ultra-reliable low latency communication (URLLC). URLLC will be widely adopted for mission-critical applications such as remote surgery, factory automation, autonomous driving, tactile Internet, and augmented reality to applications such as remote surgery, factory automation, autonomous driving, tactile Internet, and augmented reality (URLLC). URLLC will be widely adopted for mission-critical applications such as remote surgery, factory automation, autonomous driving, tactile Internet, and augmented reality (URLLC). URLLC will be widely adopted for mission-critical applications such as remote surgery, factory automation, autonomous driving, tactile Internet, and augmented reality (URLLC).

A promising solution to enable efficient and fast computation for URLLC devices is mobile edge computing (MEC). MEC enhances the battery lifetime and reduces the power consumption of users with delay-sensitive tasks. By offloading these tasks to nearby MEC servers, the power consumption and computation time at the local users can be considerably reduced at the expense of the power required for the data transmission for offloading. Thus, efficient resource allocation algorithm design is paramount for MEC for optimization of the available resources (e.g., power and bandwidth) while guaranteeing the maximum delay for the computation tasks. Existing resource allocation algorithms for MEC designs, such as [7], [8], were designed based on Shannon’s capacity formula. In particular, the authors in [7] studied energy-efficient resource allocation for MEC, while computation rate maximization was considered in [8]. However, if the resource allocation design for URLLC MEC systems is based on Shannon’s capacity formula, the reliability of the offloading and downloading processes cannot be guaranteed. To cope with this issue, recent works applied finite blocklength transmission (FBT) [9] for resource allocation algorithm design for URLLC MEC systems. In particular, the authors in [10] studied binary offloading in single-carrier TDMA systems. However, single-carrier systems suffer from poor spectrum utilization and require complex equalization at the receiver. In [11], the authors investigated the minimization of the normalized energy consumption for OFDMA. However, the algorithm proposed in [11] assumes that the channel gain is identical for different sub-carriers which may not be realistic for broadband wireless channels. Moreover, the resource allocation algorithms proposed in [11] are based on a simplified version of the general expression for the achievable rate for FBT [9]. Furthermore, the existing MEC designs, such as [7], [12], do not take into account the size of the computation result of the tasks and do not consider the communication resources consumed for downloading of the processed data by the users. Nevertheless, the size of the processed data can be large for applications such as augmented reality URLLC. To the best of the authors’ knowledge, joint uplink-downlink resource allocation for OFDMA-URLLC MEC systems has not been considered in the literature, yet.

Motivated by the above discussion, in this paper, we propose a novel power-efficient joint uplink-downlink resource allocation algorithm design for multiuser OFDMA-URLLC MEC systems. To reduce the end-to-end delay of the uplink and downlink transmission while efficiently exploiting the avail-
able spectrum, we propose a partial time overlap between the uplink and downlink frames which introduces new causality constraints. Then, the resource allocation algorithm design is formulated as an optimization problem for the minimization of the total weighted power consumed by the base station (BS) and the users subject to QoS constraints for the URLLC users. The QoS constraints include the minimum required number of bits computed within the maximum allowable time for computation, i.e., the maximum end-to-end delay of each user. The formulated optimization problem is a non-convex mixed-integer problem that is difficult to solve globally. Thus, we develop a low-complexity sub-optimal algorithm based on successive convex approximation (SCA) in order to find a locally optimal solution.

**Notation:** Lower-case letters \( x \) refer to scalar numbers, while bold lower-case letters \( \mathbf{x} \) represent vectors. \( (-)^T \) denotes the transpose operator. \( \mathbb{R}^{N \times 1} \) represents the set of all \( N \times 1 \) vectors with real valued entries. The circularly symmetric complex Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \) is denoted by \( \mathcal{CN}(\mu, \sigma^2) \), \( \sim \) stands for “distributed as”, and \( \mathbb{E}\{\cdot\} \) denotes statistical expectation. \( \nabla_x f(x) \) denotes the gradient vector of function \( f(x) \) and its elements are the partial derivatives of \( f(x) \).

## II. System and Channel Models

In this section, we present the considered system and channel models for OFDMA-URLLC MEC systems.

### A. System Model

We consider a single-cell multiuser MEC system which comprises a BS and \( K \) URLLC users indexed by \( k = \{1, \ldots, K\} \), cf. Fig. 1. All transceivers have single antennas. The system employs frequency division duplex (FDD).  

Thereby, the total bandwidth \( W \) is divided into two bands for uplink and downlink having bandwidths \( W^u \) and \( W^d \), respectively. The bandwidths for uplink and downlink are further divided into \( M^u \) and \( M^d \) orthogonal sub-carriers indexed by \( m^u = \{1, \ldots, M^u\} \) and \( m^d = \{1, \ldots, M^d\} \), respectively. The bandwidth of each sub-carrier is \( B_s \). Thus, the symbol duration is \( T_s = \frac{W^u}{M^u} \). The uplink and downlink frames are divided into \( N^u \) time slots indexed by \( n^u = \{1, \ldots, N^u\} \) and \( N^d \) time slots indexed by \( n^d = \{1, \ldots, N^d\} \), respectively. Moreover, each time slot contains one OFDM symbol. The downlink transmission starts after \( \tau \) time slots. Thus, uplink and downlink transmission overlap in \( O = N^u - \tau \) time slots. The value of \( \tau \) is a design parameter. On the one hand, if \( \tau \) is chosen too small, the users’ information bits to be computed may not have yet arrived at the BS and hence the downlink resource is wasted. On the other hand, if \( \tau \) is chosen too large, the computed bits at the BS have to wait before being transmitted to the users, which increases the end-to-end delay, see Fig. 1. Each user has one computation task \((B_k, D_k)\) that needs to be processed, where \( B_k \) is the task length in bits and \( D_k \) is the required time for computation in time slots.

\(^1\)In FDD systems, different frequency bands are assigned to uplink and downlink.

![Figure 1: Multiuser MEC system with a single BS with an edge server and \( K \) URLLC users.](image)

Moreover, we assume that all users offload their tasks to the MEC server. The maximum transmit power of the BS is \( P_{\text{max}} \), while the maximum transmit power of each user in the uplink is \( P_{\text{e, max}} \).

In order to facilitate the presentation, in the following, we use superscript \( j \in \{u, d\} \) to denote uplink \( u \) and downlink \( d \).

**Remark 1.** The power and time consumed for channel estimation and resource allocation are constant and will not affect the validity of the proposed resource allocation algorithm. For simplicity of illustration, they are neglected in this paper. Furthermore, perfect channel state information (CSI) is assumed to be available at the BS for resource allocation design to obtain a performance upper bound for OFDMA-URLLC MEC systems.

### B. Uplink and Downlink Channel Models

In the following, we introduce the uplink and downlink channel models for OFDMA-URLLC MEC systems. We assume that the channel gains of all users for all sub-carriers are constant during uplink and downlink transmission. In the uplink, the signal received at the BS from user \( k \) on sub-carrier \( m^u \) in time slot \( n^u \) is given as follows:

\[
y_k^u[m^u, n^u] = h_k^u[m^u]x_k^u[m^u, n^u] + z_{BS}^u[m^u, n^u],
\]

where \( x_k^u[m^u, n^u] \) denotes the symbol transmitted by user \( k \) on sub-carrier \( m^u \) in time slot \( n^u \) to the BS. Moreover, \( z_{BS}^u[m^u, n^u] \sim \mathcal{CN}(0, \sigma^2) \) denotes the noise at the BS, and \( h_k^u[m^u] \) represents the complex channel coefficient between user \( k \) and the BS on sub-carrier \( m^u \). Moreover, for future use, we define the signal-to-noise ratio (SNR) of user \( k \)’s signal at the input of the BS’s receiver on sub-carrier \( m^u \) in time slot \( n^u \) as follows:

\[
\gamma_k^u[m^u, n^u] = \left| h_k^u[m^u] \right|^2 \frac{P_k^u[m^u, n^u]}{\sigma^2},
\]

where \( P_k^u[m^u, n^u] = \mathbb{E}\{\left| x_k^u[m^u, n^u]\right|^2\} \) is the uplink transmitted power of user \( k \) on sub-carrier \( m^u \) in time slot \( n^u \), and \( g_k^u[m^u] = \left| h_k^u[m^u] \right|^2 \). A similar channel model is adopted for downlink transmission and the corresponding SNR at user \( k \) on sub-carrier \( m^d \) in time slot \( n^d \) is denoted by \( \gamma_k^d[m^d, n^d] \).

### C. Achievable Rate for FBT

Shannon’s capacity theorem, on which most conventional resource allocation designs are based, applies to the asymptotic
case where the packet length approaches infinity and the decoding error probability goes to zero [13]. Thus, it cannot be used for resource allocation design for URLLC systems, as URLLC systems have to employ short packets to achieve low latency, which makes decoding errors unavoidable. For the performance evaluation of FBT, the so-called normal approximation for short packet transmission was developed in [14]. For parallel complex AWGN channels, the maximum number of bits $\Psi$ conveyed in a packet comprising $L$ symbols can be approximated as follows [14, Eq. (4.277)], [15, Fig. 1]:

$$\Psi = \frac{1}{\log(1 + \gamma[l]) - aQ^{-1}(\epsilon)} \sum_{l=1}^{L} V[l], \quad (3)$$

where $\epsilon$ is the decoding packet error probability, and $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function with $Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt$. $V[l] = (1 - (1 + \gamma[l])^{-2})$ and $\gamma[l]$ are the channel dispersion [14] and the SNR of the $l$-th symbol, respectively, and $a = \log_2(\epsilon)$.

In this paper, we base the joint uplink-downlink resource allocation algorithm design for OFDMA-URLLC MEC systems on (3). By allocating several resource blocks from the available resources to a given user, the number of offloaded and downloaded bits of the user can be adjusted.

### III. Problem Formulation

In this section, we explain the offloading and downloading process and introduce the QoS requirements of the URLLC MEC users. Moreover, we formulate the proposed resource allocation optimization problem.

#### A. Offloading and Downloading

The edge computing process is performed as follows. First, each user offloads its data to the edge server in the uplink. Subsequently, the edge server processes this data and sends the results back in the downlink to the user. Thus, uplink and downlink should satisfy the following constraints:

**C1:** $\Psi_k^u(s_k^u, p_k^u) \geq B_k, \forall k$, **C2:** $\Psi_k^d(s_k^d, p_k^d) \geq \Gamma_k B_k, \forall k$, \hspace{1cm} (4)

where

$$\Psi_k^u(s_k^u, p_k^u) = C_k^u(s_k^u, p_k^u) - V_k^u(s_k^u, p_k^u), \hspace{1cm} (5)$$

and

$$C_k^u(s_k^u, p_k^u) = \sum_{m=1}^{M_k^u} \sum_{n=1}^{N_k^u} s_k^u[m^u, n^u] \log_2(1 + \gamma_k^u[m^u, n^u]), \hspace{1cm} (6)$$

$$V_k^u(s_k^u, p_k^u) = aQ^{-1}(\epsilon_k^u) \sum_{m=1}^{M_k^u} \sum_{n=1}^{N_k^u} s_k^u[m^u, n^u] V_k^u[m^u, n^u]. \hspace{1cm} (7)$$

Here, $s_k^u[m^u, n^u] = \{0, 1\}, \forall m^u, n^u$, are the sub-carrier assignment indicators. If sub-carrier $m^u$ in time slot $n^u$ is assigned to user $k$, we have $s_k^u[m^u, n^u] = 1$, otherwise $s_k^u[m^u, n^u] = 0$. Furthermore, we assume that each sub-carrier is allocated to at most one user to avoid multiple access interference.

$p_k^u[m^u, n^u]$ is the power allocated to user $k$ on sub-carrier $m^u$ in time slot $n^u$. $s_k^d$ and $p_k^d$ are the collections of optimization variables $s_k^d[m^d, n^d], \forall m^d, n^d$, and $p_k^d[m^d, n^d], \forall m^d, n^d, \forall j$, respectively, and $V_k^d[m^d, n^d] = (1 - (1 + \gamma_k^d[m^d, n^d])^{-2})$.

Constraints C1 and C2 guarantee for user $k$ the transmission of $B_k$ bits in the uplink and $\Gamma_k B_k$ bits in the downlink, respectively. Moreover, $\Gamma_k \neq \emptyset$, is the ratio of the sizes of the computation results and the offloaded task. The value of $\Gamma_k$ depends on the application type, e.g., $\Gamma_k > 1$ is expected for augmented reality applications. [16].

#### B. Causality and Delay

In the following, we explain the causality and delay constraints.

1) **Causality:** According to Fig. 1, downlink transmission cannot start for a given user before all data of this user has been received at the BS via the uplink. Thus, we impose the following causality constraints:\n
$$C3: s_k^d[m^d, \tau + o] + s_k^d[m^d, n^d] \leq 1,$$

$$\forall o \in \{1, \ldots, \bar{O}\}, \forall k, \forall m^u, \forall n^d = \{1, \ldots, o\}, \forall m^d. \hspace{1cm} (8)$$

This constraint ensures that the downlink transmission for a particular user cannot start before its data has arrived at the BS.

2) **Delay:** The delay of a computation task is limited by requiring the downlink transmission to be finished before $D_k - \tau$ time slots as follows:

$$C4: s_k^d[m^d, n^d] = 0, \forall n^d \geq D_k - \tau. \hspace{1cm} (9)$$

The total latency of a computation task is determined by $D_k$ and $\tau$. Note that the values of $D_k$ and $\tau$ are known for resource allocation.

#### C. Optimization Problem Formulation

In the following, we formulate the resource allocation design problem with the objective to minimize the total weighted network power consumption, while satisfying the latency requirements for the users’ task computation. In particular, we optimize the power and sub-carrier assignments in uplink and downlink. To this end, the optimization problem is formulated as follows:

$$\min_{s^u, s^d, p^u, p^d} \sum_{k=1}^{K} \sum_{m=1}^{M_k^u} \sum_{n=1}^{N_k^u} \sum_{m=1}^{M_k^d} \sum_{n=1}^{N_k^d} \sum_{m=1}^{M_k^u} \sum_{n=1}^{N_k^u} s_k^u[m^u, n^u] p_k^u[m^u, n^u] \hspace{1cm} (10)$$

$$+ \sum_{k=1}^{K} \sum_{m=1}^{M_k^d} \sum_{n=1}^{N_k^d} \sum_{m=1}^{M_k^d} \sum_{n=1}^{N_k^d} s_k^d[m^d, n^d] p_k^d[m^d, n^d]$$

s.t. **C1** - **C4**, **C5:** $\sum_{k=1}^{K} s_k^u[m^u, n^u] \leq 1, \forall m^u, n^u,$

**C6:** $s_k^u[m^u, n^u] \in \{0, 1\}, \forall k, m^u, n^u,$

$3$In this paper, we neglect the computation time and power consumption at the edge server, and we only focus on uplink and downlink transmission. This model is valid when the edge server has sufficient processing and computation resources to carry out the small tasks of URLLC users.
In this section, we first transform the problem in (10) into a more tractable equivalent form. In particular, we first employ the Big-M formulation. Then, we use the difference of convex programming and SCA approaches in order to solve the optimization problem in (10) iteratively. The main steps of the proposed low-complexity algorithm are summarized in Fig. 2.

IV. SOLUTION OF THE PROBLEM

In this section, we first transform the problem in (10) into a more tractable equivalent form. In particular, we first employ the Big-M formulation. Then, we use the difference of convex programming and SCA approaches in order to solve the optimization problem in (10) iteratively. The main steps of the proposed low-complexity algorithm are summarized in Fig. 2.

A. Problem Transformation

To deal with the non-convex product terms in optimization problem (10), the Big-M method is employed [17].

Step 1 (Big-M Formulation)

Let us introduce new optimization variables as

$$\bar{p}_k^d[m^j, n^j] = s_k^d[m^j, n^j]p_k^d[m^j, n^j], \forall k, m^j, n^j, \forall j.$$  
(11)

Now, we decompose the product terms above using the Big-M formulation (McCormick envelopes) and impose the following additional constraints [17]:

$$C_{13} : \bar{p}_k^d[m^j, n^j] \leq P_{k, \text{max}}s_k^d[m^j, n^j], \forall k, m^j, n^j,$$
(12)

$$C_{14} : \bar{p}_k^d[m^j, n^j] \leq \bar{p}_k^d[m^j, n^j], \forall k, m^j, n^j,$$
(13)

$$C_{15} : \bar{p}_k^d[m^j, n^j] \geq \bar{p}_k^d[m^j, n^j] - (1 - s_k^d[m^j, n^j])P_{k, \text{max}}, \forall k, m^j, n^j,$$
(14)

$$C_{16} : \bar{p}_k^d[m^j, n^j] \geq 0, \forall k, m^j, n^j,$$
(15)

$$C_{17} : \bar{p}_k^d[m^j, n^j] \leq \max_{k'}s_k^d[m^j, n^j], \forall k, m^j, n^j,$$
(16)

$$C_{18} : \bar{p}_k^d[m^j, n^j] \leq \bar{p}_k^d[m^j, n^j], \forall k, m^j, n^j,$$
(17)

$$C_{19} : \bar{p}_k^d[m^j, n^j] \geq \bar{p}_k^d[m^j, n^j] - (1 - s_k^d[m^j, n^j])P_{k, \text{max}}, \forall k, m^j, n^j,$$
(18)

$$C_{20} : \bar{p}_k^d[m^j, n^j] \geq 0, \forall k, m^j, n^j.$$  
(19)

The non-convex product terms

$$s_k^d[m^j, n^j]p_k^d[m^j, n^j], \forall k, m^j, n^j, \forall j$$ in (11) are transformed into a set of convex linear inequalities. Note that constraints C13-C20 do not change the feasible set. Now, optimization problem (10) is transformed into the following equivalent form:

$$\min_{s^u, p^u, s^d, p^d} \sum_{k=1}^{K} \sum_{m^u=1}^{M^u} \sum_{n^u=1}^{N^u} \bar{p}_k^d[m^j, n^j]$$

s.t. $C_1$ $C_2$ $C_3$ $C_4$ $C_5$ $C_6$ $C_7$ $C_8$ $C_9$ $C_{10}$ $C_{11}$ $C_{12}$

where

$$C_k^d(\bar{p}_k^d) = \sum_{m^j=1}^{M^j} \sum_{n^j=1}^{N^j} \log_2(1 + \gamma_k^d[m^j, n^j]),$$
(21)

$$V_k^d(\bar{p}_k^d) = aQ^{-1}(\epsilon_k^d) \sqrt{\sum_{m^j=1}^{M^j} \sum_{n^j=1}^{N^j} V_k^d[m^j, n^j]},$$
(22)

For more details on the big M-formulation, please refer to [17, Section 2.3].
Lemma 1. For sufficiently large constant values $\eta_1$ and $\eta_2$, the optimization problem in (20) is equivalent to the following problem:

\[
\begin{align*}
\text{minimize} & \quad \Phi(\tilde{p}^u, \tilde{p}^d) + \eta_1(E^u - H^u) + \eta_2(E^d - H^d) \\
\text{s.t.} & \quad C1 - C5, C6a, C7 - C9, C10a, C11 - C20,
\end{align*}
\]

where $\Phi(\tilde{p}^u, \tilde{p}^d)$ is the objective function of problem (20).

Proof. The proof follows similar steps as corresponding proofs in [5], [18], [19]. Interested readers are referred to [21] which is an extended version of this paper and includes the full proof.

The only remaining sources of non-convexity are the structure of the achievable rate for FBT and the non-convex objective function. In the following, we employ SCA to approximate problem (29) by a convex problem. Subsequently, we propose an iterative algorithm to find a low-complexity solution to problem (29).

Algorithm 1 Successive Convex Approximation

1: Initialize: Random initial points $s^{u(1)}$, $s^{d(1)}$, $\tilde{p}^{u(1)}$, $\tilde{p}^{d(1)}$, set iteration index $i = 1$, maximum number of iterations $I_{\text{max}}$, and initial penalty factors, $\eta_1 > 0$ and $\eta_2 > 0$.

2: Repeat

3: Solve convex problem (34) for given $s^{u(i)}$, $s^{d(i)}$, $\tilde{p}^{u(i)}$, $\tilde{p}^{d(i)}$, and store the intermediate solutions $s^u$, $s^d$, $\tilde{p}^u$, $\tilde{p}^d$.

4: Set $i = i + 1$ and update $s^{u(i)} = s^u$, $s^{d(i)} = s^d$, $\tilde{p}^{u(i)} = \tilde{p}^u$, $\tilde{p}^{d(i)} = \tilde{p}^d$.

6: Until convergence or $i = I_{\text{max}}$.

7: Output: $s^{u*} = s^u$, $s^{d*} = s^d$, $\tilde{p}^{u*} = \tilde{p}^u$, $\tilde{p}^{d*} = \tilde{p}^d$.

C. Successive Convex Approximation

Step 3: In order to cope with the remaining non-convexity of (29), we employ the Taylor series approximation to approximate the non-convex parts of the objective function and constraints $C1$ and $C2$. Since $H^j(s^j), \forall j$, and $-\tilde{V}_k^j(\tilde{p}_k^j), \forall j$, are differentiable convex functions, then for any feasible points $s^{i(j)}$, $\tilde{p}_k^{i(j)}$, $\forall j$, the following inequalities hold:

\[
\begin{align*}
H^j(s^j) & \geq \tilde{H}_k^j(s^j) = H^j(s^{i(j)}) + \nabla s^j H^j(s^{i(j)})^T(s^j - s^{i(j)}), \forall j, \\
\tilde{V}_k^j(\tilde{p}_k^j) & \leq \tilde{V}_k^j(\tilde{p}_k^j, \tilde{p}_k^{i(j)}) = \tilde{V}_k^j(\tilde{p}_k^{i(j)}) + \nabla \tilde{p}_k \tilde{V}_k^j(\tilde{p}_k^{i(j)})^T(\tilde{p}_k^j - \tilde{p}_k^{i(j)}), \forall j.
\end{align*}
\]

The right hand sides of (30) and (31) are affine functions representing the global underestimation of $H^j(s^j), \forall j$, and $\tilde{V}_k^j(\tilde{p}_k^j), \forall j$, respectively, where $\nabla s^j H^j(s^{i(j)})^T(s^j - s^{i(j)})$ and $\nabla \tilde{p}_k \tilde{V}_k^j(\tilde{p}_k^{i(j)})$ are given on the top of the next page. By substituting the right hand sides of (30) and (31) into (29), we obtain the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \Phi(\tilde{p}^u, \tilde{p}^d) + \eta_1(E^u - H^u) + \eta_2(E^d - H^d) \\
\text{s.t.} & \quad C1 : C_k^{u(\tilde{p}_k^j)} - \tilde{V}_k^{u(\tilde{p}_k^j)} \geq B_k, \forall k, \\
& \quad C2 : C_k^{d(\tilde{p}_k^j)} - \tilde{V}_k^{d(\tilde{p}_k^j)} \geq G_k B_k, \forall k, \\
& \quad C3 - C5, C6a, C7 - C9, C10a, C11 - C20.
\end{align*}
\]

Optimization problem (34) is convex because the objective function is convex and the constraints span a convex set. Therefore, it can be efficiently solved by standard convex optimization solvers such as CVX [22]. Algorithm 1 summarizes the main steps to solve (29) in an iterative manner, where the solution of (34) in iteration $(i)$ is used as the initial point for the next iteration $(i + 1)$. The algorithm produces a sequence of improved feasible solutions until convergence to a local optimum point of problem (29) or equivalently problem (10) in polynomial time.

V. PERFORMANCE EVALUATION

In this section, we provide simulation results to evaluate the effectiveness of the proposed joint uplink-downlink resource
\[ \nabla_{\alpha_i} H^{j}(s^{(i)})^T(s^{j} - s^{(i)}) = \sum_{k=1}^{K} \sum_{m=1}^{M_j} \sum_{n=1}^{N_j} 2s_k^{j(i)} [m^j, n^j] \left( s_k^{j(i)} - s_k^{j(i)} [m^j, n^j] \right), \forall j, \] 
\[ \nabla_{p_k} V_k^{j}(d_k^{j(i)}) = \frac{aQ^{-1}(\epsilon_{k})}{\sqrt{\sum_{m=1}^{M_j} \sum_{n=1}^{N_j} V_k^{j(i)}[m^j, n^j]}} \left( \frac{\hat{q}_k^{[1]} (1+\hat{p}_k^{[1]} [1,1]g_k^{[1]} )^2}{\hat{q}_k^{[M]} (1+\hat{p}_k^{[M]} [M,N]g_k^{[M]} )^2} \right), \forall j. \] 

| Table I: Simulation Parameters |
|-----------------------------|
| Parameter                  | Value  |
| Total number of sub-carriers in uplink and downlink | 237,664 |
| Number of time slots in uplink and downlink $N^* = N^*$ | 4 |
| Bandwidth of each sub-carrier | 30 kHz |
| Noise power density         | $174$ dBm |
| Maximum BS transmit power, $P_{\text{max}}$ | 45 dBm |
| Maximum transmitted power of each user, $P_{\text{max}}$ | 23 dBm |
| Value of $T_s$, $\forall k$ | 1 |

We adopt the simulation parameters given in Table I, unless specified otherwise. In our simulations, a single cell is considered with inner and outer radii $r_1 = 50$ m and $r_2 = 100$ m, respectively. The BS is located at the center of the cell, and the users are randomly located between the inner and the outer radii. The user weights are set to $w_k = 1, \forall k$ for simplicity. The path loss is calculated as $35.3 + 37.6 \log_{10}(d_k)$ [23], where $d_k$ is the distance from the BS to user $k$. The values of the penalty factors are set to $\eta_1 = 10K P_{\text{max}}$ and $\eta_2 = 20 P_{\text{max}}$. The small scale fading gains between the BS and the users are modeled as independent and identically Rayleigh distributed. All simulation results are averaged over 100 realizations of the path loss and multipath fading.

A. Performance Bound and Benchmark Scheme

We compare the performance of the proposed resource allocation algorithm with the following benchmark schemes:

- **Shannon’s capacity (SC):** To obtain an (unachievable) lower bound on the total network power consumption, Shannon’s capacity formula is adopted in problem (10), i.e., $V_k^{j}(s_k^{j(i)}, d_k^{j(i)}), \forall j$, is set to zero in constraints C1 and C2, respectively, and all other constraints are retained. The resulting optimization problem is solved using a modified version of the proposed algorithm.

- **Fixed sub-carrier assignment (FSA):** In this scheme, we fix the sub-carrier assignment. In fact, we divide the total number of sub-carriers among the users such that their delay and causality constraints are met. Then, we optimize the power allocated to the sub-carriers for the given channel realization. The resulting optimization problem is solved using the SCA method.

B. Simulation Results

In Fig. 3, we investigate the average system power consumption versus the size of the task of the URLLC users and study the impact of different delay requirements. For delay scenario $S_0$, none of the users has delay restrictions, i.e., $D_k = \tau + N^d = 7, \forall k$. In contrast, for delay scenario $S_1$, two users have strict delay constraints while the remaining users do not, i.e., $D_1 = D_2 = 5$ and $D_3 = D_4 = 7$. As expected, increasing the required number of transmitted bits leads to higher transmit powers. This is due to the fact that if more bits are to be transmitted in a given frame, higher SNRs are needed, and thus, the BS and the users have to increase the transmitted power. Furthermore, the proposed scheme leads to a substantially lower power consumption compared to the FSA scheme. This is due to the non-optimal sub-carrier allocation for the FSA scheme. Fig. 3 also reveals the impact of strict delay requirements. In particular, delay scenario $S_1$ leads to a higher power consumption compared to $S_0$ because the BS and the users are forced to allocate more power even if their channel conditions are poor to ensure their transmissions are completed with the desired delay. Furthermore, SC provides a lower bound for the required power consumption of OFDMA-URLLC MEC systems. However, SC cannot guarantee the required latency and reliability. This is due to the fact that, in this scheme, the performance loss incurred by FBT is not taken into account for resource allocation design, and thus the obtained resource allocation policies may not meet the QoS constraints.

In Fig. 4, we show the average system power consumption versus the packet error probability and study the impact of different delay requirements. As can be observed, for the proposed scheme and FSA, the average system power consumption is a monotonically decreasing function of the packet error probability. This is due to the fact that the complementary error function in the normal approximation is a monotonically decreasing function of $\epsilon$, and as a result, the impact of the dispersion part in the normal approximation decreases as $\epsilon$ increases. Fig. 4 also reveals the impact of delay constraints. In particular, delay scenario $S_1 = \{ D_1 = D_2 = D_3 = 5, D_4 = D_5 = 7 \}$ leads to a higher power consumption compared to $S_0 = \{ D_k = 7, \forall k \}$. This is due to the smaller feasible set of the optimization problem. Moreover, as can be seen, for SC, the power consumption is independent of the packet error probability. This is due to the fact that SC assumes that the decoding error probability is zero. Moreover, the gap between the proposed scheme and SC is the price to be paid for enforcing strict delay and reliability requirements to ensure URLLC.

VI. Conclusions

This paper studied the resource allocation algorithm design for OFDMA-URLLC MEC systems. To ensure the stringent end-to-end transmission delay and reliability requirements of URLLC, we proposed a joint uplink-downlink resource allocation scheme which takes into account FBT. Moreover, to min-
imize the end-to-end delay, we proposed a partial time overlap between the uplink and downlink frames which introduces new uplink-downlink causality constraints. The proposed resource allocation algorithm design was formulated as an optimization problem for minimization of the total weighted transmit power of the network under QoS constraints regarding the minimum required number of computed bits of the URLLC users within a maximum computing time, i.e., the end-to-end delay. Due to the non-convexity of the formulated problem, finding a global solution entails a prohibitive computational complexity. Thus, a low-complexity algorithm based on SCA was proposed to find a high-quality sub-optimal solution. Our simulation results showed that the proposed resource allocation algorithm design facilitates the application of URLLC in MEC systems, and achieves significant power savings compared to a benchmark scheme.