ABSTRACT Wireless resource virtualization scheme has been proposed to satisfy the quality of service (QoS) requirement of multi-service traffic flows on multi-dimensional resources. However, it is a challenge to design an appropriate virtual resource allocation strategy for multi-dimensional resource to meet the delay demand of multi-service. In this paper, we study the virtual resource allocation method in TDD-F-OFDM system with MU-MIMO and formulate the multi-service virtual resource optimization problem as a cross-layer problem. Then, we develop an effective dynamic algorithm with an iterative dual update to meet the optimization targets of the Media Access Control (MAC) layer and physical layer. Finally, the simulation results show that our proposed virtual resource allocation algorithm not only achieves better spectral efficiency and flow access ratio, but also guarantees the delay requirements of multi-service flows.

INDEX TERMS Resource virtualization, cross-layer allocation, QoS guarantee, TDD-F-OFDM.

I. INTRODUCTION

Recently, the rapid growth of traffic demand and application proliferation from multi-service creates irresistible challenges for wireless networks to ensure the quality of service (QoS) and quality of experience of subscribers [1]. Meanwhile, wireless communication networks are undergoing their next evolutionary step. The next generation networks are envisioned to provide a flexible, scalable, agile, and programmable network platform over which different services with varying requirements can be deployed and managed within strict performance bounds. To support multiple on-demand services over fixed communication networks, network operators must allow flexible customization and fast provision of their network resources. However, due to the inherent broadcast nature of wireless communications and stochastic fluctuation of wireless channel quality, radio resource abstraction and isolation is not straightforward. Another significant challenge of wireless network virtualization is resource allocation, which decides how to embed a virtual wireless network on physical networks.

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Tsun Cheng.
and allocation, this paper has the following major contributions:

- To propose a mechanism of multi-dimensional resources virtualization and allocation for required delay bounded QoS provisioning.
- The virtual resource allocation problem is decomposed into two sub-problems of the physical layer and the MAC layer.
- To propose an efficient dynamic cross-layer iterative algorithm to solve the problem of resource reservation and traffic access.

The rest of this paper is organized as follows. We give a review of the related research work in Section II. Section III describes the system model in terms of TDD-F-OFDM system based on MU-MIMO and resource measure. Section IV formulates the delay-bounded QoS guarantee and optimization problem. Section V presents the algorithmic solutions to the formulated problems. Section VI presents the simulation results. The conclusions are stated in Section VII. Here are some notations to be used in this paper:

- $P(\cdot)$: probability operation.
- $I_m$: $m \times m$ identity matrix.
- $(\cdot)^T$, $(\cdot)^H$: Transposition, Hermitian, respectively.
- $E(\cdot)$: expectation operation.

Throughout this paper, all matrices and sets are denoted by capital letters in boldface, and vectors are denoted by lowercase letters in boldface.

II. RELATED WORK

In this Section, we introduce some work related to the research in this article, including wireless resource virtualization, MU-MIMO, F-OFDM technology, and QoS exponent.

A. WIRELESS RESOURCE VIRTUALIZATION

Wireless resources virtualization are created by slicing wireless network infrastructure and spectrum resources into multiple virtual slices. Wireless network infrastructures are sliced to create virtual resource, which can offer customized services to virtual mobile networks (VMNs) by different schedulers in a secure and isolated manner. Therefore, wireless resource virtualization has garnered increasing attention in recent times and some of the literature investigated this idea from different aspects [6]–[14]. In [6], Moubayed et al. study the virtual resource allocation for D2D communications under laying virtual LTE networks, which take both the sum rate of the users and the cost of the service providers into consideration, and formulate the virtual resource allocation as a 0-1 integer programming. Zimmo et al. [7] introduce a coordination mechanism between the LTE and WiFi by proposing a scheduler that involves time-domain resource virtualization as a cornerstone of the coexistence mechanism. In [8], wireless network infrastructures and RF bands are reconfigured and divided to sublease to MVNOs based on the service level agreements between them so that MVNOs get enough wireless resources to serve their users and meet their QoS requirements. Nakauchi et al. [9] propose an airtime-based resource control technique for wireless network virtualization, in which wireless network resources are allocated among competing virtual networks while keeping their programmability. The above documents all virtualize single-dimensional resources, and virtualization of multi-dimensional resources are not considered. In multidimensional resources, Liu and Han [10] demonstrate the Themis protocol and its system implementation that realizes cross-domain resource orchestration and virtualization in cellular computing networks. In [11], the optimal resource allocation schemes in the time, power, and spatial domains are obtained by solving the uplink sum-throughput maximization and uplink fair-throughput maximization problems. Although the above literature considers the allocation of multi-dimensional resources, the QoS constraint of the multi-service flows is ignored. In multi-service, Hu et al. [12] study the service oriented resource allocation issue in virtualization environment and proposed a novel service oriented framework considering resource and QoS demands. In [13], Ansah and Meer propose a demand-based virtual link embedding approach for multi-service mapping in programmable industrial networks. Han et al. [14] considering dynamic traffic arrivals and time-varying channel conditions, the virtualized resource management in the multi-service coexisting LTE-D2D networks is formulated into a multi-objective stochastic optimization problem. Though the above research works considered multi-service flows and QoS demands, cross-layer analysis is neglected widely. To solve the above problems, we study the virtual resource allocation method in TDD-F-OFDM system with MU-MIMO and formulate the multi-service virtual resource optimization problem as a cross-layer problem. We formulate the multi-service with different QoS requirement virtual resource optimization problem as a cross-layer problem. In detail, we cut traffic flows with different QoS requirements into different virtual slices at the MAC layer, and utilize resource virtualization technology to measure and schedule time-frequency resources at the physical layer. In summary, we propose a framework for multi-dimensional resource virtualization and allocation, taking into account the QoS requirements of multi-service flows, which can be extended to other possible resource forms.

B. MU-MIMO

MU-MIMO is also called virtual MIMO (VMIMO), which is defined in 3GPP LTE has attracted substantial interest because it alleviates the requirement of multi-antennas for user equipment by generating a virtual MIMO channel between a base station and multiple terminal stations. The MIMO technology can greatly increase system capacity by spreading the spatial resources without increasing system bandwidth. Compared with a conventional MIMO system, MU-MIMO can obtain additional multi-user diversity gain by grouping users using well-designed strategies. At present, there are a lot of researches on the resource allocation scheme of MU-MIMO system, which are based
on sub-carrier allocation, correlation grouping and channel state estimation [3], [4], [15]–[19]. Specifically, Tseng et al. derive a new optimal condition for sub-carrier assignment for infinitesimal bandwidth increment, and extend to a MU-MIMO uplink transmission based on OFDMA transmission systems to improve the information rate and the traffic quality [3]. In paper [4], the author proposes a user scheduling and resource allocation scheme based on reinforcement learning to match the schedule between the RBs and the users in MU-MIMO system. Lu et al. study the joint dynamic user grouping and RB allocation in uplink multi-cell MU-MIMO-SC-FDMA systems, meanwhile, the corresponding algorithms are proposed to reduce the complexity of resource allocation [15], [16]. In order to meet the service quality of the traffic flows, the researchers in paper [17]–[19] study the QoS index in the resource allocation of the MU-MIMO system. In our paper, we consider multi-dimensional resources virtualization and allocation including spatial and time-frequency resources. In spatial dimension, MU-MIMO can obtain additional multi-user diversity gain by using well-designed user grouping strategies. Therefore, we use the combination of resource virtualization and MU-MIMO to jointly consider multi-dimensional resources, which is a mechanism of multidimensional resources virtualization and allocation and can be extended to other possible resource forms.

C. F-OFDM
The traditional OFDM technology sets the same parameters for the entire frequency band of the system, and the out-of-band leakage is serious, making it difficult to support the diversified service requests in 5G [20]. Filter OFDM technology is referred to as F-OFDM technology, which was formally put forward in 2015 [2]. F-OFDM technology can divide the frequency band in the system into multiple sub-bands, and each sub-band configures different parameters according to the type of service, which overcomes the limitation of setting only one parameter for the entire frequency band in the OFDM system and improves the spectrum utilization [21]. In [22], Ferreira et al. present a resource allocation scheme in wireless networks based on F-OFDM, aiming at optimizing the user data delay. In paper [23], Gonzaga Ferreira and Teles Vieira propose a resource allocation algorithm for F-OFDM wireless communication systems. In [24], to meet target bit error rate (BER) requirements of all users and maximize the spectrum efficiency, Yang et al. propose a filter parameter and guard bandwidth optimization model to design related parameters dynamically for uplink asynchronous F-OFDM system.

D. QoS EXPONENT
With the gradual diversification of business, more and more new services with strict QoS requirements in terms of low delays have emerged. Users’ requirements can be classified according to different QoS exponent, and services can also be classified in MAC layer. In this way, virtual slices based on service types can represent different QoS requirements. Different QoS groups correspond to different virtual slices, which can provide more flexible services.

EC and EB [25], have been proposed to characterize the effect of delay on the system. EB gives the minimum channel transmission rate required for a given service request process under the premise of meeting QoS requirements. In [26], [27], EB has been developed to model the statistical behaviour of traffic. As the dual concept of EB, EC is defined as the maximum service arrival rate that can be supported by a given channel service process under the premise of meeting QoS requirements. In [28], Musavian and Ni use the ratio of EC and the total expenditure power, including the transmission power and the circuit power to measure link-layer energy efficiency. Chen et al. [29] propose a resource allocation strategy based on effective capacity under the scenario of secondary user’s parallel transmission. Since EB and EC mentioned above facilitate capturing the delay-bounded constraint of wireless link without going into complex queuing analysis, we use these dual concepts to characterize the delay-bounded constraints in our cross-layer virtual resource allocation scheme.

III. SYSTEM MODEL
In this section, we introduce the MU-MIMO TDD-F-OFDM system and wireless virtual network (WVN) model at first. Then, after introducing the sub-frame configuration factor of uplink and downlink, we derive the reserve resources of uplink and downlink for TDD-F-OFDM system respectively. Finally, we introduce the concept of user grouping and resource allocation.

A. SYSTEM MODEL
In the wireless network virtualization (WVN), BS and users with different services are partitioned into slices, and each of them represents a VMN respectively. VMNs with different delay-bounded share physical layer infrastructure and resources.

As depicted in Fig.1, a PMN is split into S VMNs which support different services with different delay-bounded requirements. Obviously, WVN decouples the physical supply process and the service provisioning process, can abstract, isolate, and share the physical infrastructure network
equipment. In this paper, slice isolation is implemented at the MAC layer with virtual link rate.

As depicted in Fig.1, we mainly study a MU-MIMO TDD-F-OFDM system including one BS equipped with $N_t$ receiving antennas and $N_f$ users with single transmitting antenna. Each user group has a maximum of $N_{max}$ users, and $N_{max} \leq N_f$. Multiple users in one user group and BS form MU-MIMO system. The channel mode of MU-MIMO can be expressed as follows:

$$\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_{N_r}
\end{bmatrix} = \begin{bmatrix}
h_{1,1} & h_{1,2} & \cdots & h_{1,N_t} \\
h_{2,1} & h_{2,2} & \cdots & h_{2,N_t} \\
\vdots & \vdots & \ddots & \vdots \\
h_{N_r,1} & h_{N_r,2} & \cdots & h_{N_r,N_t}
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_{N_t}
\end{bmatrix}$$

where $h_{i,j}$ denotes the channel gain between the $i$-th receiving antenna and the $j$-th transmitting antenna, and $n_j$ ($i = 1, 2, \ldots, N_r$) indicates a transmission noise signal.

Due to the relative delay of multi-path, flat fading channels cannot be directly applied to the actual broadband wireless communication system, where channels are with frequency-selective fading channels. In this paper, we use F-OFDM technology to divide the system bandwidth into several sub-bands with the sub-carrier as the sub-channel in order to adapt to the traffic flow with different QoS exponent. According to relevant papers [22], [23], [30], [31], we assume that the fading gains between all antenna pairs are independent and identically distributed (i.i.d.) Rayleigh fading, and vary according to a Gauss-Markov model. Using this model, the relationship between the channel matrices of $g$-th user group among adjacent sub-frames is as follows:

$$\text{vec}(\mathbf{H}(l)) = \sqrt{\alpha} \text{vec}(\mathbf{H}(l-1)) + \sqrt{1-\alpha} \mathbf{u}(l),$$

(2)

where $\mathbf{H}(l)$ denotes the channel matrix of the $l$-th sub-frame slot in one scheduling period, $0 \leq \alpha \leq 1$ is the channel de-correlation coefficient that satisfies $\alpha^{T_c/(2T_r)} = r_{hh}(T_c)$, where $r_{hh}(t)$ denotes the time autocorrelation function, $T_c$ denotes the channel coherence time. In addition, $\mathbf{u}(l)$ is independent of $\mathbf{H}(l-1)$ and it is a column vector with a dimension of $N_t N_f \times 1$. $\mathbf{u}(l)$ and $\mathbf{H}(l)$ have i.i.d. complex Gaussian distribution $CN (0, 1)$ respectively.

In addition, let $\Gamma = [\Gamma_1, \Gamma_2, \cdots, \Gamma_{Q+1}]^T$ represent the thresholds of the element $h$ of the matrix $\mathbf{H}$ in increasing order with $\Gamma_1 = 0$ and $\Gamma_{Q+1} = \infty$. $h$ is in state $q$ if $\Gamma_q \leq h < \Gamma_{q+1}$, expressed as $h_q = \Gamma_q$. Thus, $\mathbf{H}$ can be designed as follows:

$$\mathbf{H} = \begin{bmatrix}
h_{1,1} & h_{1,2} & \cdots & h_{1,N_t} \\
h_{2,1} & h_{2,2} & \cdots & h_{2,N_t} \\
\vdots & \vdots & \ddots & \vdots \\
h_{N_r,1} & h_{N_r,2} & \cdots & h_{N_r,N_t}
\end{bmatrix},$$

(3)

where $h_{i,j}$ denotes the channel gain between the $i$-th receiving antenna and the $j$-th transmitting antenna, and values from the set $\{h_q | q = 1, 2, \ldots, Q\}$. Thus, we can get the state space of matrix $\mathbf{H}$, that is $\mathbf{H} \in \{\mathbf{H}_q | q = 1, 2, \ldots, Q^{N_t N_f}\}$, where $\mathbf{H}_q$ denotes the $d$-th state of the matrix $\mathbf{H}$. Specifically, $\mathbf{H}_q$ is described as follows:

$$\mathbf{H}_d = \begin{bmatrix}
h_{d_1} & h_{d_1+N_t+1} & \cdots & h_{d_1+(N_f-1)+1} \\
h_{d_2} & h_{d_2+N_t+2} & \cdots & h_{d_2+(N_f-1)+2} \\
\vdots & \vdots & \ddots & \vdots \\
h_{d_N_t} & h_{d_N_t+N_f} & \cdots & h_{d_N_t+N_f N_t}
\end{bmatrix},$$

(4)

where $\{h_{d_i} | i = 1, 2, \ldots, N_f N_t\}$ values from $\{h_q | q = 1, 2, \ldots, Q\}$.

When the initial state information in one scheduling period $\mathbf{H}(1)$ is given, and all elements of $\mathbf{H}(1)$ are $h_1$, we can compute the probability distribution of $\mathbf{H}(2)$ according to Eq. (1). The specific process can be designed as follows:

$$P(\mathbf{H}(2) = \mathbf{H}_d) = \int_{h_d}^{h_{d+1}} f_h(h_1)dh \int_{h_d}^{h_{d+1}} f_h(h_1)dh \cdots \int_{h_d}^{h_{d+1}} f_h(h_1)dh$$

(5)

where $\int_{a}^{b} f(h)dh$ denotes the integral operation of the function $f(h)$ in the interval $[a, b]$, $f_h(h_q)$ denotes the probability density function of $h$ that follows $CN (\sqrt{\alpha} h_q, 1 - \alpha)$.

In this way, we can get the probability distribution of $\mathbf{H}$ in all sub-frame slots in one scheduling period $\{\mathbf{H}(l) | l = 1, 2, \ldots, mL\}$.

**B. Link Rate for Uplink and Downlink**

In the TDD technology, the sub-frame configuration factor is defined as the ratio of the uplink sub-frames to all the sub-frames in this paper. A sub-frame configuration factor of uplink and downlink is given in this system, and the traffic access problem of uplink and downlink is studied. The BS can obtain CSI by using TDD technology, we assume all subcarriers have the same CSI in one RB. Thus, we can obtain the CSI in one RB by taking the average of the CSIs of the subcarriers within the RB.

![FIGURE 2. Uplink and downlink sub-frame configuration of L frames in one scheduling slot.](image-url)

As shown in Fig.2, we present a scheme of uplink and downlink sub-frame configuration of $L$ frames in a scheduling slot. In this case where a certain frame configuration is fixed, if the initial CSI of the scheduled slot is known, we can predict the CSI of the subsequent multiple sub-frames. Assuming that the channel state transfers $L$ times in one scheduling period, thus one scheduling period $T_c$ consists of $L$ frame slots $T_p$, that is, $T_c = L T_p$. In addition, we assume that one frame contains $m$ sub-frames, and one scheduling period includes $C$ uplink sub-frames. Then, we can obtain...
the proportion of uplink sub-frames in one scheduling period is \( \mu = \frac{C}{mL} \). In LTE standard, one time slot is 0.5 milliseconds, two time slots form a 1 millisecond subframe, and 10 subframes form a scheduling time slot with a duration of 10 milliseconds. Furthermore, there are \( mL - C \) downlink sub-frames in one scheduling period and the proportion of downlink sub-frames is \( 1 - \mu = \frac{ml-C}{ml} \). In TDD-F-OFDM system, the uplink data and the downlink data are transmitted on different sub-frames in a same frame and we fixed the subframe configuration factor \( \mu = 0.5 \).

There are \( C \) uplink sub-frames and \( mL - C \) downlink sub-frames, which are recorded as sets \( \mathbf{Z}_U \) and \( \mathbf{Z}_D \), respectively. According to Eq.(5) and Markov model, we can get the uplink and downlink channel statistics.

On the uplink side of the BS, we adopt Minimum Mean Square Error (MMSE) channel estimation techniques for fading channel models. Assuming that the channel matrices corresponding to the \( C \) sub-frames in the sub-frame set \( \mathbf{Z}_U \) of the uplink are \( \mathbf{H}_U(1), \ldots, \mathbf{H}_U(c), \ldots, \mathbf{H}_U(C) \), after getting the channel matrix \( \mathbf{H}_U(c) = \mathbf{H}_d \) of the \( c \)-th sub-frame, we can calculate the SINR of user-\( k \) after as:

\[
\text{SINR}^U_{k,d} = \frac{E_i}{\sigma^2 \left( \left( \mathbf{H}_d \right)^H \mathbf{H}_d + \frac{P_t}{P_i} \mathbf{I}_N \right)^{-1}} - 1, \quad (6)
\]

where \( \text{SINR}^U_{k,d} \) denotes signal to interference plus noise ratio (SINR) of the \( k \)-th user in the \( d \)-th channel matrix in the uplink, \( \sigma^2 \) denotes the spectral density power of noise, \( P_t \) denotes the transmitting power of each user signal, which is normalized in calculation.

According to Shannon theory, we can calculate the corresponding capacity as follows:

\[
r^U_{k,d} = \log_2 \left( 1 + \text{SINR}^U_{k,d} \right). \quad (7)
\]

In MU-MIMO system, user grouping is another way to use the available spatial resources. Furthermore, the channel capacity of the \( c \)-th uplink sub-frame can be expressed as:

\[
r^U(c) = \sum_{k \in \Omega_c} \sum_{d=1}^{Q^N_N} r^U_{k,d} P \left( \mathbf{H}_U(c) = \mathbf{H}_d \right). \quad (8)
\]

Once we have obtained the channel capacity of each uplink sub-frame in a scheduling slot, the capacity of \( C \) uplink sub-frames in the scheduling slot can be calculated as follows:

\[
R^U = \frac{r^U(1) + \cdots + r^U(c) \cdots + r^U(C)}{C}. \quad (9)
\]

Further, according to the Markov process, the probability distribution of channel capacity of the uplink sub-frame is expressed as follows:

\[
P \left( R^U = r^U \left| \mathbf{H}_U(1) \right. \right) = \sum_{E^I} \sum_{\xi = 1}^{\left| E^I \right|} P \left( r^U(1) = r^U_{E^I}(1) \right) P \left( r^U(2) = r^U_{E^I}(2) \right) \mathbf{H}_U(1)
\]

\[
\cdots P \left( r^U(C) = CR^U - \sum_{i=1}^{C-1} r^U_{E^I}(c) \right) r^U(C - 1) = r^U_{E^I}(C - 1), \quad (10)
\]

where \( \xi \) denotes the number index of state combinations, \( |E| \) denotes the total number of the combinations, \( r^U_{E^I}(c) \) denotes the sub-channel capacity of the \( \xi \)-th uplink sub-frame of the \( E \)-th combination in one scheduling period and takes value from the capacity corresponding to all states, \( R^U_{E^I} \) represents the \( j \)-th value of the uplink sub-channel channel capacity in a scheduling slot, \( 1 \leq j \leq J \).

In this paper, we take the mean value of the link rates as its logic representation, that is, the service rate of resources on the link. Therefore, under the condition of given initial CSI, the average link rate in uplink in one scheduling period can be obtained as follows [13]:

\[
E \left[ R^U \right] = \sum_{j=1}^{J} P \left( R^U = R^U_{E^I} \right) r^U(1) R^U_{E^I}. \quad (11)
\]

Assuming that the channel matrices corresponding to the \( mL - C \) downlink sub-frames in the sub-frame set \( \mathbf{Z}_D \) are \( \mathbf{H}_D(1), \mathbf{H}_D(2), \ldots, \mathbf{H}_D(ml-C) \). Meanwhile, the MMSE equalization is adopted in the transmitting terminal of the downlink. Furthermore, according to the MMSE equalization, the SINR of the \( u \)-th user in the \( e \)-th downlink sub-frame can be expressed as:

\[
\text{SINR}^D_{k,e} = \frac{\left| \mathbf{h}_k(e) \mathbf{f}_k(e) \right|^2 \rho}{\sum_{i=1, i \neq k}^{N^U} \left| \mathbf{h}_k(e) \mathbf{f}_i(e) \right|^2 \rho + \text{tr} \left( \mathbf{F}_{\text{MMSE}}(e)^H \mathbf{F}_{\text{MMSE}}(e) \right)}, \quad (12)
\]

where \( \mathbf{F}_{\text{MMSE}}(e) = [\mathbf{f}_1(e), \mathbf{f}_2(e), \ldots, \mathbf{f}_{N^U}(e)] \) is precoding matrix, \( \mathbf{f}_1(e), \ldots, \mathbf{f}_u(e), \ldots, \mathbf{f}_{N^U}(e) \) are precoding vectors for each signal of users in the \( e \)-th downlink sub-frame, respectively. \( \mathbf{h}_k(e) \) denotes the downlink channel vector of the \( u \)-th user in the \( e \)-th downlink sub-frame. \( \rho \) denotes the overall SNR of transmission after power normalization.

Correspondingly, when \( \mathbf{H}_D(e) \) can be replaced by \( (\mathbf{H}_D)^T \), the average link rate in downlink in one scheduling period can be obtained based on the Eq.(11) as follows:

\[
E \left[ R^D \right] = \sum_{i=1}^{J} P \left( R^D = R^D_{E^I} \right) \mathbf{H}_U(1)^H R^D_{E^I}. \quad (13)
\]

C. USER GROUPING AND RESOURCE ALLOCATION

In multi-user groups, the set of user group is defined as \( \Omega = \{ \Omega_1, \Omega_2, \ldots, \Omega_g, \ldots, \Omega_{m+1} \} \), and the \( g \)-th user group can be expressed as \( \Omega_g = \{ u_1, u_2, \ldots, u_m \} \). One BS serves users within its coverage, and antennas of the BS and the grouped users form the MU-MIMO system, which can obtain additional multi-user diversity gain by using well-designed user grouping strategies.

For the TDD-F-OFDM system, adjacent time-frequency RBs are available to be collected to user groups. Since the
resources used for allocation in the uplink are allocated to the users continuously and the joint uplink and downlink resource allocation of the TDD system is considered, the downlink resources are allocated to users continuously [25], [32], [33]. The system bandwidth is divided into B sub-bands, and each sub-band is configured with different parameters. Assuming that the b-th sub-band has Nb RBs with the same parameters, thus there are NRB = \sum_{b=1}^{B} Nb RBs in the system. The corresponding RB pattern matrix T is designed as follows:

\[ T = \text{diag}(T_1, T_2, \ldots, T_b, T_B), \]

where \( T_b \) represents the resource mode matrix of the RB b on the b-th sub-band. Under the condition that all RBs are continuously allocated, the corresponding matrix \( T_b \) is expressed as follows:

\[
T_b = \begin{bmatrix}
1 & \cdots & 0 & \cdots & 1 & \cdots & 0 & \cdots & 1 \\
0 & \cdots & 1 & \cdots & 0 & \cdots & 1 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & \cdots & 1 & \cdots & 1 & \cdots & N_b - 1 \\
0 & \cdots & 1 & \cdots & 1 & \cdots & 1 & \cdots & N_b \\
\end{bmatrix},
\]

(15)

where each element indicates whether the RB is involved in the RB pattern (1) or not (0). The set of RB pattern on the b-th sub-band is defined as \( \Gamma_b \). The total number of modes included in \( T_b \) is \( W_b = N_b + (N_b - 1) + \cdots + 1 = \frac{N_b (N_b + 1)}{2} \), and the total number of modes included in the RB pattern matrix \( T \) of the entire system is \( W = \sum_{b=1}^{B} W_b = \sum_{b=1}^{B} \frac{N_b (N_b + 1)}{2} \) [33].

After obtaining the RB patterns, we build a metric matrix to calculate the resources available to each user on the RB pattern under the principle of on-demand allocation. We calculate the metric value of each RB pattern occupied by each user. Further, we take the p-th RB pattern occupied by the i-th user group as an example, considering the configuration factor of uplink and downlink, the corresponding metric matrix of each user can be expressed as

\[ L_{p,g}^u = \log \left( \sum_{p \in \Gamma_b} \left( \frac{\mu E_p^{R_B^G}}{E_B^G} + \frac{(1-\mu) E_p^{R_B^D}}{E_B^D} \right) \right). \]

Then the metric matrix on user can be expressed as

user index 1 2 \ldots I RB pattern index

\[
L_{q,i}^u = \begin{bmatrix}
L_{1,1}^u & L_{1,2}^u & \cdots & L_{1,I}^u \\
L_{2,1}^u & L_{2,2}^u & \cdots & L_{2,I}^u \\
\vdots & \vdots & \ddots & \vdots \\
L_{Z,1}^u & L_{Z,2}^u & \cdots & L_{Z,I}^u \\
\end{bmatrix},
\]

(16)

In the process of problem formulation, since EB and EC mentioned in section II facilitate capturing the delay-bounded constraint of wireless link without going into complex queuing analysis, we use these dual concepts to characterize the delay-bounded constraints in our paper. Considering that the allocation of the reserved resources is a dynamic process, to simplify the solution of the problem, we split the resource allocation problem into two sub-problems, including the traffic flow access problem at the MAC layer and the resource allocation problem at the physical layer respectively.

In addition, we take user groups fairness into account, the spectral efficiency of each user group should be weighted by a fairness factor. An effective approach to get the fairness factor is using proportional scheduling. Considering the configuration factor of uplink and downlink, we calculate the fairness factor of g-th user groups in a scheduling slot is as follows:

\[
\lambda_g = \frac{\mu E_p^{R_B^G}}{EB_g^U} + \frac{(1-\mu) E_p^{R_B^D}}{EB_g^D}. \]

(17)

To evaluate fairness performance of user groups, we adopt the Jain’s fairness index [34] as follows:

\[
JF = \left( \frac{\sum_{g=1}^{\Omega} \lambda_g}{|\Omega|} \right)^2 \sum_{g=1}^{\Omega} \lambda_g^2. \]

(18)

Jain’s fairness index ranges from 0-1, the closer the fairness factor is to 1, the better the fairness among user groups.

IV. PROBLEM FORMULATION

Firstly, we introduce the Delay-bounded QoS exponent and the definition of EB and EC is given based on QoS exponent. Then, we formulate a cross-layer virtual resource allocation problem combining with the MAC layer and the physical layer. For the MAC layer, a traffic flow access method with the overall benefit as the optimization goal is established according to the resources obtained by the virtual slice. For the physical layer, a resource reservation problem is established to solve the corresponding EB requirement according to the uplink and downlink services of users on the principle of on-demand distribution and proportional fairness.

A. DELAY-BOUNDQD QoS GUARANTEE

The link rates are scheduled to different delay-sensitive traffic flows with customized delay-bounded QoS provisioning. Based on the Markov channel model and QoS exponent method, the resources reserved on the link are abstracted by average rate to hide the random characteristics of the fading channel. Therefore, EB can be used for link rate splitting and sharing. Correspondingly, to link service rate, EC is used for giving the actual service capability with specific QoS exponent.

We assume that there are K users in TDD-F-OFDM system, and the user’s traffic request is not a special service but a service type. In this paper, the set of the traffic flows of the k-th user in the s-th VMN, is defined as \( F = \{ f_{s,k} \mid k = 1, \ldots, K_s \} \).
and we assume that the arrival of traffic flow follows Poisson process. For the uplink, the QoS character of the traffic flow request $f_{s,k}^U$ can be described by $\left\{ \lambda_{s,k}^U, D_{s,k}^{\max}, P_{s,k}^{\text{delay}} \right\}$, where $\lambda_{s,k}^U$ is average arrival rate of in uplink, $D_{s,k}^{\max}$ denotes the maximum delay-bounded constraint of traffic flows, $P_{s,k}^{\text{delay}}$ is the delay-outage probability of $f_{s,k}$, which shows the probability that the delay exceeds the maximum delay bound. Let $Y_{s,k}$ denotes the access variable of traffic flow $f_{s,k}$. Considering user-$k$ in the $s$-th slice obtains the channel service rate $E\left[ R_s^U \right]$, we calculate the QoS exponent $\theta_{s,k}^U$ as follows:

$$\theta_{s,k}^U \approx -\frac{1}{Y_{s,k}^U \lambda_{s,k}^U} \ln \left( \frac{P_{s,k}^{\text{delay}} E\left[ R_s^U \right]}{Y_{s,k}^U \lambda_{s,k}^U} \right).$$

Furthermore, since $f_{s,k}^U$ follows Poisson process, we can calculate the EB of the uplink in one scheduling period as follows [32]:

$$\tilde{E}B_{s,k}^U = \frac{U_{s,k}^U \lambda_{s,k}^U \left( e^{\theta_{s,k}^U} - 1 \right)}{\theta_{s,k}^U \ln 2}.$$  (20)

Similarly, we can obtain the QoS exponent $\theta_{s,k}^D$ and $\tilde{E}B_{s,k}^D$ in one scheduling period of the downlink traffic flow $f_{s,k}^D$ expressed as

$$\theta_{s,k}^D \approx -\frac{1}{Y_{s,k}^D \lambda_{s,k}^D} \ln \left( \frac{P_{s,k}^{\text{delay}} E\left[ R_s^D \right]}{Y_{s,k}^D \lambda_{s,k}^D} \right).$$

and

$$\tilde{E}B_{s,k}^D = \frac{U_{s,k}^D \lambda_{s,k}^D \left( e^{\theta_{s,k}^D} - 1 \right)}{\theta_{s,k}^D \ln 2}.$$  (22)

As the dual concept of the EB source model, the EC is defined as the maximum constant arrival rate that a given channel service process can support in order to guarantee the QoS requirement specified by QoS exponent $\theta$ [33]. After obtaining the channel resources of the uplink and downlink of each user, we can calculate EC as follows [35]:

$$EC_{s,k}^U = -\frac{1}{\theta_{s,k}^U} \log E\left( e^{\theta_{s,k}^U R_s^U} T_s^U \right),$$

and

$$EC_{s,k}^D = -\frac{1}{\theta_{s,k}^D} \log E\left( e^{\theta_{s,k}^D R_s^D} T_s^D \right),$$

satisfying the constraints $EC_{s,k}^U \geq Y_{s,k}^U \lambda_{s,k}^U$ and $EC_{s,k}^D \geq Y_{s,k}^D \lambda_{s,k}^D$ respectively.

**B. TRAFFIC FLOW ACCESS PROBLEM MODEL AT THE MAC LAYER**

At the MAC layer, service flows are classified by us, that is, service flows are divided into different slices $s$ according to different QoS exponent. Considering the relevant contents in economics [36] and assuming that the number of users on $s$-th virtual slice is $K_s$, we define the pricing function at the MAC layer, and the utility function at the MAC layer is designed as follows:

$$U^{\text{MAC}}(Y) = \sum_{s=1}^{S} \left[ \sum_{k=1}^{K_s} \left( 1 - \beta \tilde{E}B_{s,k}^U \right) Y_{s,k}^U \lambda_{s,k}^U \right] + \sum_{k=1}^{K_s} \left( 1 - \beta \tilde{E}B_{s,k}^D \right) Y_{s,k}^D \lambda_{s,k}^D,$$  (25)

where $1 - \beta \tilde{E}B_{s,k}^U$ and $1 - \beta \tilde{E}B_{s,k}^D$ represent the pricing functions of the uplink and the downlink respectively, where the pricing parameter $\beta > 0$.

Then, combined with Eq.(20) and Eq.(22), the utility function at the MAC layer can be expressed as:

$$U^{\text{MAC}}(Y) = \sum_{s=1}^{S} \left[ \sum_{k=1}^{K_s} \left( 1 - \frac{Y_{s,k}^U \lambda_{s,k}^U \left( e^{\theta_{s,k}^U} - 1 \right)}{\theta_{s,k}^U \ln 2} \right) Y_{s,k}^U \lambda_{s,k}^U \right] + \sum_{k=1}^{K_s} \left( 1 - \frac{Y_{s,k}^D \lambda_{s,k}^D \left( e^{\theta_{s,k}^D} - 1 \right)}{\theta_{s,k}^D \ln 2} \right) Y_{s,k}^D \lambda_{s,k}^D.$$  (26)

Furthermore, optimized by profit maximization, the optimization problem of the traffic flow access at the MAC layer can be designed as follows:

$$\arg \max_{Y} U^{\text{MAC}} \left\{ \begin{array}{l} \text{BC1} : EC_{s,k}^U \geq Y_{s,k}^U \lambda_{s,k}^U \quad \forall k = 1, \ldots, K; \forall s = 1, \ldots, S \\ \text{BC2} : EC_{s,k}^D \geq Y_{s,k}^D \lambda_{s,k}^D \quad \forall k = 1, \ldots, K; \forall s = 1, \ldots, S \\ \text{BC3} : 0 \leq Y_{s,k}^U \leq 1 \quad \forall k = 1, \ldots, K; \forall s = 1, \ldots, S \\ \text{BC4} : 0 \leq Y_{s,k}^D \leq 1 \quad \forall k = 1, \ldots, K; \forall s = 1, \ldots, S \end{array} \right.$$  (27)

BC1 and BC2 are to ensure the traffic flows are no more than the corresponding EC constraints, BC3 and BC4 are to ensure both $Y_{s,k}^U$ and $Y_{s,k}^D$ value from 0 to 1.

**C. RESOURCE ALLOCATION PROBLEM MODEL AT THE PHYSICAL LAYER**

According to Eq.(11) and Eq.(13), we can calculate the uplink sub-frame transmission rate $E\left[ R_{w,g,k}^U \right]$ and the downlink sub-frame transmission rate $E\left[ R_{w,g,k}^D \right]$ of the $w$-th allocated mode to the user-$k$ in the $g$-th user group.

To support more traffic access, the resource reservation at the physical layer should satisfy different traffic requirements of user as far as possible. Considering the on demand and proportional fairness principle, the utility function at the physical layer is designed as follows:

$$U^{\text{PHY}}(\delta) = \frac{B}{b=1} \sum_{b=1}^{B} \sum_{k=1}^{K} \log \left( \sum_{w=1}^{W_b} \sum_{g=1}^{\Omega} \left( \mu_{w,g,k} E\left[ R_{w,g,k}^U \right] \right) \right) \left( 1 - \frac{E\left[ R_{w,g,k}^D \right]}{E\left[ R_{w,g,k}^D \right]} \right) \delta_{w,g,k}.$$  (28)
where $\delta_{w,g,k}$ is the assignment index which indicates whether user-$k$ in the $g$-th user group occupies the $w$-th allocated mode or not.

Furthermore, the optimization problem of the resource reservation at the physical layer can be designed as follows:

$$
\arg \max_\delta U_{PHY}(\delta)
$$

s.t.

\begin{align*}
\text{AC1: } & A_1 \delta = 1_{N_{RB} \times 1} \\
\text{AC2: } & \sum_{w=1}^{W} \sum_{k=1}^{K} \delta_{w,g,k} = |\Omega_g| \quad \forall g = 1, \cdots, |\Omega| \\
\text{AC3: } & \delta_{w,g,k} \in \{0, 1\} \quad \forall k = 1, \cdots, K; \\
& w = 1, \cdots, W; \ g = 1, \cdots, |\Omega|.
\end{align*}

(29)

The constraint AC1 is to ensure that each RB can only be allocated to one user group, and $A_1 = T_{N_{RB} \times W} \otimes 1_{1 \times K} |\Omega|$; AC2 is to ensure each user group can obtain one RB pattern, and $|\Omega_g|$ is the number of user group, AC3 indicates whether the $w$-th RB allocation pattern is assigned to user-$k$ in the $g$-th group (value 1) or not (value 0).

V. DYNAMIC ALGORITHM FOR RESOURCE RESERVATION AND TRAFFIC ACCESS

In this section, the corresponding algorithms are proposed according to the problem of resource reservation at the physical layer and the problem of traffic flow access at the MAC layer. The physical layer uses the configuration factor of uplink and downlink to solve the resource reservation variables according to the accessed traffic flow at the MAC layer, while the MAC layer solves the traffic flow access variables based on the resources allocation variables obtained from the physical layer. Furthermore, since the results of the two problems depend on the results of each other, a dynamic cross-layer iterative algorithm is proposed for the final steady state result.

A. ALGORITHM FOR RESOURCE RESERVATION AT PHYSICAL LAYER

The optimization problem (29) is a typical binary integer programming problem, which is suitable for solving with the iterative Hungarian algorithm (IHA) [15]. However, due to the large scale of the proposed model, we use the fast unfolding algorithm (FUA) [37] to divide the large network topology composed of users and resource patterns in problem (29) into multiple small communities. Then, in each small community, we use IHA to solve the final resource allocation and user grouping results.

Set $B$ sub-bands in the system, thus the RB type of the $b$-th sub-band can be expressed as $RB_b$, $b \in [1, B]$, and there are $N_b$ resource blocks in $RB_b$. For each RB type, we construct the complete RB pattern [16]. $2^{N_b-1}$ RB patterns can be obtained in $RB_b$, thus there are $N_{RP} = 2^r \left( \sum_{b=1}^{B} N_b - S \right)$ RB patterns in the system. For all complete RB pattern subsets, we perform the same operation in the following. Specifically, we consider all possible matches between all users and RB patterns in each subset as a bipartite graph.

As shown in Fig.3, both the users and the RB patterns are considered as nodes of the weighted network topology map, where there are $K$ users and $N$ RB patterns. If the $i$-th node represents a user and the $j$-th node represents a RB pattern, the weight between the $i$-th node and the $j$-th node can be expressed as $F_{i,j} = \frac{\mu(E^{U}_{i,j})}{E^{U}_i} + \frac{1-\mu(E^{D}_{i,j})}{E^{D}_j}$.

After running the FUA, the bipartite graph is divided into multiple sub-graphs, and we can obtain sub-graphs that make the modularity the most, where there are $Q$ sub-graphs. For a sub-graph, we assume that it contains $X$ users and $Z$ RB patterns.

According to the user grouping matrix, we can obtain all possible user groups as $\hat{\Omega} = \left\{ \Omega^{(1)}, \cdots, \Omega^{(x)}, \cdots, \Omega^{(X)} \right\}$, where $\hat{\Omega}^{(x)}$ denotes the user set including $x$ users, $|\hat{\Omega}^{(x)}| = C^X_X$, and $|\hat{\Omega}| = \sum_{x=1}^{X} |\hat{\Omega}^{(x)}| = \sum_{x=1}^{X} C^X_X$.

In addition, we can calculate the metric value of each RB pattern occupied by each user group. Similar to Eq.(28), take the $p$-th RB pattern occupied by the $g$-th user group as an example, considering the configuration factor of uplink and downlink, the corresponding metric can be expressed as $L_{p,g} = \sum_{i \in \hat{\Omega}_g} \sum_{j = 1}^{Z} \left( \sum_{p \in E^{U}_i} \frac{\mu(E^{U}_{p,j})}{E^{U}_i} + \frac{(1-\mu(E^{D}_{p,j}))}{E^{D}_j} \right)$. Then the overall matrix $L_{q}$ is

$$
L_{q} = \begin{bmatrix}
\hat{L}_{1,1} & \hat{L}_{1,2} & \cdots & \hat{L}_{1,|\hat{\Omega}|} \\
\hat{L}_{2,1} & \hat{L}_{2,2} & \cdots & \hat{L}_{2,|\hat{\Omega}|} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{L}_{Z,1} & \hat{L}_{Z,2} & \cdots & \hat{L}_{Z,|\hat{\Omega}|}
\end{bmatrix}
$$

(30)

The specific steps can be described as follows:

**Algorithm 1 Joint Uplink and Downlink Resource Reservation Algorithm Based on Fast Unfolding**

Step 1:
- Initialize $t = 0$, $v = 0$;

Step 2:
- Generate the complete RB pattern set, the total number is $N_{RP}$;

Step 3:
Algorithm 2 Traffic Flow Access Control Algorithm

Step 1: Initialize $\tau = 1$, $\nu = 1$, $Y^U_{s,k}(1) = 0$ and $Y^D_{s,k}(1) = 0$;

Step 2: Update $Y^U_{s,k}(\tau + 1)$ and $EC^U_{s,k}$ by Eq. (31) and Eq. (20) respectively;

Step 3: If $EC^U_{s,k} \geq Y^U_{s,k}(\tau + 1)$ is satisfied, Go to Step 4; Else Go to Step 6;

Step 4: If $Y^U_{s,k} \leq 1$ is satisfied, Go to Step 5; Else Go to Step 6;

Step 5: If $Y^U_{s,k}(\tau) = Y^U_{s,k}(\tau + 1)$ is satisfied, Go to Step 6; Else Let $\tau = \tau + 1$, Go to Step 2;

Step 6: Obtain $Y^U_{s,k} = Y^U_{s,k}(\tau)$;

Step 7: Update $Y^D_{s,k}(v + 1)$ and $EC^D_{s,k}$ by Eq. (32) and Eq. (21) respectively;

Step 8: If $EC^D_{s,k} \geq Y^D_{s,k}(v + 1)$ is satisfied, Go to Step 9; Else Go to Step 11;

Step 9: If $Y^D_{s,k} \leq 1$ is satisfied, Go to Step 10; Else Go to Step 11;

Step 10: If $Y^D_{s,k}(v) = Y^D_{s,k}(v + 1)$ is satisfied, Go to Step 11; Else Let $v = v + 1$, Go to Step 7;

Step 11: Obtain $Y^D_{s,k} = Y^D_{s,k}(v)$;

**C. DYNAMIC CROSS-LAYER ITERATIVE ALGORITHM**

We study multi-dimensional resources virtualization and allocation for required delay-bounded QoS provisioning, which involves cross-layer iteration between the MAC layer and the physical layer. Fig. 4 shows the cross-layer interaction process between MAC layer and physical layer in one scheduling period in detail. At MAC layer, traffic flows with different QoS exponent are cut into different slices. Each slice set the traffic flow access ratio $Y_s$, and present EB requirement of MAC layer to the physical layer. At physical layer, time-frequency resources are allocated to maximize utility function $U^{PHY}$ and we can get allocation vector $\delta$ and the EC constraint of each traffic flow. Then, each slice adjust the traffic flow access ratio $Y_s$ to maximize utility function $U^{MAX}$, update EB requirement and transfer the EB requirement to the physical layer. The iterative processes are repeated until MAC layer and physical layer achieve the balance of their own optimal strategy.

The resource reservation of the physical layer and the traffic flow access of the MAC layer interact with each other, thus the overall dynamic interaction algorithm and the message exchange are presented in Algorithm 3, which is described as follows:

Algorithm 3 Dynamic Cross-layer Iterative Algorithm

Step 1: Initialize $\lambda^U_k, \lambda^D_k, D^\text{max}_k, P^\text{delay}_k, (Y^U_k)^{(1)} = 1, (Y^D_k)^{(1)} = 1, E\left[R^U_{b,k}\right]^{(1)} = \frac{1}{N_b} \sum_{n=1}^{N_b} \sum_{g=1}^{E} E\left[R^U_{n,g,k}\right]$,
the search scale and reduce the computational complexity of the optimization model. For the Algorithm 1 of the physical layer, at first, the FUA algorithm is used to solve the problem by dividing the overall users and RBs into \( L \) communities. Assuming that there are \( K_L \) users and \( N_L \) RB patterns in the largest communities, we can compute the number of all possible user groups is \( \sum_{i=1}^{K_L} C_i^k \). Further, we use IHA to obtain the best match between user groups and RB patterns, and the time complexity in the largest community can be expressed as \( O(K_L N_L \sum_{i=1}^{K_L} C_i^k) \). Therefore, at the worst case, the time complexity of the Algorithm 1 can be expressed as \( O(L K_L N_L \sum_{i=1}^{K_L} C_i^k) \), which is less than \( \left( \frac{N(N+1)}{2} \right)^K \) in large-scale case.

In the traffic flow access problem model, \( U^{MAC}(Y) \) is a convex function with respect to \( Y_{s,k}^U \) and \( Y_{s,k}^D \); so, we use the gradient iteration to compute the extreme point of the problem. The smaller the iteration step, the more the number of iterations, the closer the result of the iteration is to the optimal value of the problem. In order to solve the traffic flow access problem in a limited time, we adopt a compromise method in Algorithm 2. In the case of determining a suitable iteration step size, the sub-optimal solution is obtained under a limited number of iterations. The idea of the linear search method with the time complexity is \( O(n) \) used for the Algorithm 2 of the MAC layer. Assuming that constraint Eq.(27) can be satisfied after \( KSn + KS \) iterations, the main execution times of the program are \( KSn + KS \). Thus the time complexity of the Algorithm 2 can be expressed as \( O(n) \).

For the Algorithm 3, the cross layer iterative process can reach equilibrium within a limited number of times.

VI. PERFORMANCE EVALUATION AND ANALYSIS

In this section, the proposed algorithms are used to simulate the traffic flow access performance of the uplink and downlink in the system. The resource reservation method of the physical layer and the traffic access problem of the MAC layer are solved by Algorithm 1 and Algorithm 2, respectively. In addition, we use Algorithm 3 to iterated the optimization results of the two problems interactively.

A. SIMULATION SETTING

We conduct the simulations based on TDD-F-OFDM system, and the configuration scheme of TDD uplink and downlink sub-frames extends the configuration scheme in the LTE system. For comparison, we consider two systems shown in Fig.5. System 1 and System 2 have the same system bandwidth, resource configuration and number of users. The difference is that the sub-band composition of the two systems is different and the detailed configuration of the sub-band is shown in Table 1. In system 1, all sub-bands are configured as sub-band 1 in Table 2. In system 2, two different sub-bands are designed as sub-band 1 and sub-band 2 in Table 2. For sub-band 1, we use the existing parameter settings of the LTE system.
TABLE 1. System simulation parameters.

| Parameter                  | System 1 | System 2 |
|----------------------------|----------|----------|
| Carrier frequency          | 20Hz     | 20Hz     |
| Sub-carrier Interval(kHz)  | 15       | 30       |
| Maximum Doppler shift      | 10Hz     | 10Hz     |
| System bandwidth           | 3MHz     | 3MHz     |
| Number of RBs              | 15       | 15       |
| Number of users            | 9        | 9        |

TABLE 2. Sub-band parameters.

| Parameter                  | Sub-band 1 | Sub-band 2 |
|----------------------------|------------|------------|
| Antenna Configuration      | MIMO       | MIMO       |
| Sampling Rate(Mb/s)        | 30.72      | 30.72      |
| Sub-band Interval(KHz)     | 15         | 30         |
| N-point FFT                | 2048       | 1024       |
| Symbol Period ($\mu$s)     | 66.67      | 33.33      |
| TTI (ms)                   | 1          | 0.2        |
| Number of Symbols /TTI     | 14         | 25         |
| CP                         | Symbol 1:160 point Symbols 2:144 point | Symbol 1:224 point Symbols 2:256 point |

FIGURE 5. Resource mapping for sub-band 1 and sub-band 2.

TABLE 3. Parameters of users.

| User   | $\lambda_k^U$ (Kbit/s) | $\lambda_k^D$ (Kbit/s) | $D_k^{max}$ (s) | VMN |
|--------|------------------------|------------------------|-----------------|-----|
| User-1 | 100                    | 130                    | 0.1             | slice1 |
| User-2 | 140                    | 170                    |                 |     |
| User-3 | 180                    | 210                    |                 |     |
| User-4 | 110                    | 140                    | 0.05            | slice2 |
| User-5 | 150                    | 180                    |                 |     |
| User-6 | 160                    | 190                    |                 |     |
| User-7 | 120                    | 150                    |                 |     |
| User-8 | 170                    | 200                    | 0.01            | slice3 |
| User-9 | 110                    | 140                    |                 |     |

FIGURE 6. The spectral efficiency versus SNR with different strategies. ($\mu = 0.5$).

1) RESOURCE EFFICIENCY OF THE PMN

In this section, sub-band configuration of SC-FDMA system and F-OFDM system in Fig.5, and the traffic flows of each slice arrive as shown in Table 2. Specifically, we propose the scheme is based on the ODA strategy at the physical layer. For comparison purpose, the maximum capacity allocation(MCA) strategy and the average capacity allocation(ACA) strategy have been designed and implemented. In Fig.6, ODA1, ACA1, and MCA1 represent the corresponding simulation result curves for SC-FDMA system, and ODA2, ACA2, and MCA2 represent the corresponding simulation result curves of F-OFDM system.

Fig.6 simulates the variation of the spectral efficiency of the two system with three resource allocation strategies. From Fig.6, we can see that the spectral efficiency of F-OFDM systems is superior to the spectral efficiency of SC-FDMA systems when using the same strategy. Because F-OFDM systems utilize spectrum resources more rationally than SC-FDMA systems, and sub-band waveform parameters can be configured according to different requirements, the spectrum efficiency is improved. Meanwhile, MCA outperforms ODA and ACA, which is in accordance with the design goal of MCA strategy. ODA is between MCA and ACA, which is because that ODA mainly considers the traffic requirement to optimize the resource allocation. ACA makes no optimization to resource allocation, thus its performance is the worst.
For the traffic flow access performance with different $\mu$, the analysis results with different strategy are shown in the Fig.7.

It can be observed from Fig.7(a)-(d) that as the SNR increases, the traffic flow access ratio corresponding to the three strategies of ODA, MCA and ACA are increasing. This is because as the SNR increases, indicating that the channel conditions are better and the channel resources are more abundant, it is easier to meet the needs of the traffic flow. To show the impact of $\mu$ on the traffic flow access ratio, we simulate the traffic flow access rate versus SNR with $\mu = 0.3$, $\mu = 0.4$, and $\mu = 0.5$ in Fig.7(a)-(c). Then, it is found that while $\mu = 0.4$, traffic flow access ratio is maximum. In Fig.7(d), ODA strategy proposed in this paper allows more traffic flows access to the network than ACA and MCA with $\mu = 0.4$. This is because ODA considers the user requirement in the design of the utility function at the physical layer. However, the performance of ACA and MCA mainly depend on the actual situation of traffic flows, and in this simulation, ACA is better than MCA.

In Fig.8, fixed-delay 1, 2, and 3 are respectively represented as three kinds of fixed delay (0.1s, 0.05s, 0.01s), which are respectively resource allocation strategies for various users with delayed demands. We can see from the Fig.8, the virtual resource slicing strategy has a better traffic access rate according to the delayed requests of different users. Meanwhile, it can be seen from Fig.8 that under the same SNR, the higher the delay constraint, the higher the traffic flow access ratio.

2) DELAY AND PERFORMANCES OF VMN

Fig.9 plots the curve of the maximum delay of each slice versus SNR firstly, where the arrival traffic flows of all slices are consistent with that of slices in Table 3. Then, specific to the delay with SNR = 6dB and 10dB, the curves of cumulative distribution function (CDF) of delay in each slice are plotted. From Fig.9(a), we can see that with the increase of SNR, the max delay of each slice is getting smaller and smaller, and when the SNR is about 6dB, all slices meet the delay-bounded constraints basically. Meanwhile, it can be seen from Fig.9(b)(c) that when SNR = 6dB, the CDF of delay tends to be 1 near the delay-bounded constraint. When SNR = 10dB, the resource exceeds the traffic request, thus the delay performance is improved obviously.
We set transmitting SNR = 10dB and simulate the fairness index of user groups when $T_s$ scheduling slot ranges from 1 to 10. The results are shown in Fig. 10.

From Fig. 10, we can see that the fairness indicates of ODA is greater than 0.8 and close to 0.9. ODA outperforms MCA and ACA, which is in accordance with the design goal of ODA strategy. ACA is between ODA and MCA, which is because that ACA mainly considers the average resource allocation. MCA makes no optimization to fairness of resource allocation, thus its performance is the worst. So, when user group fairness is taken into account by using the proportional scheduling, the fairness of each user group in the system is guaranteed.

3) ALGORITHM CONVERGENCE ANALYSIS

In this section, we analysis the convergence speed of the gradient iteration method and the dynamic cross-layer iterative algorithm.

From Fig. 11, it can be seen that he initial access ratio of each slice is 0. According to formula (31) and (32), the traffic flow access ratio of slice can be stable after about 40 iterations.

From Fig. 12, we can see that the initial access rate of all virtual slices is 1 and under the constraints of access ratio and EC, the access rate of each slice is constant after 4 iterations.

VII. CONCLUSION

In this paper, a wireless resource virtualization and allocation algorithm in TDD-F-OFDM system with MU-MIMO is proposed. Through resource virtualization technology, the average link rate is obtained and EC for one slice scheduling period is calculated. Then, to reduce the complexity of cross-layer optimization problem, we decompose the problem into two sub-problems. One is resource allocation problem at the physical layer and the other is traffic flows access problem at the MAC layer. Finally, a dynamic algorithm is developed to solve the cross-layer optimization problem. Simulations results demonstrate that the proposed algorithm not only improve the spectrum efficiency and traffic flows access ratio, but also meet the delay requirement of multi-service flows flexibly.

REFERENCES

[1] M. Yang, Y. Li, D. Jin, L. Zeng, X. Wu, and A. V. Vasilakos, “Software-defined and virtualized future mobile and wireless networks: A survey,” Mobile Netw. Appl, vol. 20, no. 1, pp. 4–18, Feb. 2015.
[2] J. Abdoli, M. Jia, and J. Ma, “Filtered OFDM: A new waveform for future wireless systems,” in Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Stockholm, Sweden: IEEE, Jun./Jul. 2015, pp. 66–70.
[3] S.-M. Tseng and Y.-F. Chen, “Average PSNR optimized cross layer user grouping and resource allocation for uplink MU-MIMO OFDMA video communications,” IEEE Access, vol. 6, pp. 50559–50571, 2018.
[4] G. Bu and J. Jiang, “Reinforcement learning-based user scheduling and resource allocation for massive MU-MIMO system,” in Proc. IEEE/CIC Int. Conf. Commun. China (ICCC), Changchun, China, Aug. 2019, pp. 641–646.
[5] D. Wu and R. Negi, “Effective capacity: A wireless link model for support of quality of service,” IEEE Trans. Wireless Commun., vol. 24, no. 5, pp. 630–643, May 2003.
[6] A. Moubayed, A. Shami, and H. Luftiyya, “Power-aware wireless virtualized resource allocation with D2D communication underlaying LTE network,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Washington, DC, USA, Dec. 2016, pp. 1–6.
[7] S. Zimmo, A. Moubayed, A. Refaey, and A. Shami, “Coexistence of WiFi and LTE in the unlicensed band using time-domain virtualization,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.
[8] R. Xie, J. Wu, R. Wang, and T. Huang, “A game theoretic approach for hierarchical caching resource sharing in 5G networks with virtualization,” China Commun., vol. 16, no. 7, pp. 32–48, Jul. 2019.
[9] K. Nakauci, Y. Shojo, and N. Nishinaga, “Airtime-based resource control in wireless LANs for wireless network virtualization,” in Proc. 4th Int. Conf. Ubiquitous Future Netw. (ICUFN), Phuket, Thailand, Jul. 2012, pp. 166–169.

[10] Q. Liu and T. Han, “Demo abstract: Themis: Cross-domain resource orchestration and virtualization in cellular computing networks,” in Proc. IEEE 26th Int. Conf. Netw. Protocols (ICNP), Cambridge, U.K., Sep. 2018, pp. 257–258.

[11] J. Yang, J. Hu, K. Lv, Q. Yu, and K. Yang, “Multi-dimensional resource allocation for uplink throughput maximization in integrated data and energy communication networks,” IEEE Access, vol. 6, pp. 47163–47180, 2018.

[12] Y. Hu, H. Cao, S. Wu, F. Tian, and L. Yang, “A study on service oriented resource allocation in virtualization environment,” in Proc. IEEE 8th Global Conf. Consum. Electron. (GCCE), Osaka, Japan, Oct. 2019, pp. 117–118.

[13] F. Ansah and H. D. Meer, “DBVLEA: A demand-based approach to virtual link mapping for multi-service industrial applications,” in Proc. 15th Int. Conf. Netw. Service Manage. (CNSM), Halifax, NS, Canada, Oct. 2019, pp. 1–9.

[14] Y. Han, X. Tao, X. Zhang, and S. Jia, “Delay-aware resource management for multi-service coexisting LTE-D2D networks with wireless network virtualization,” IEEE Trans. Veh. Technol., vol. 69, no. 7, pp. 7339–7353, Jul. 2020.

[15] X. Lu, K. Yang, N. Fan, H. Guo, and H. Zhang, “Joint clustering of users and resources for multi-cell VMIMO-SC-FDMA uplink systems,” IEEE Trans. Veh. Technol., vol. 68, no. 2, pp. 1417–1430, Feb. 2019.

[16] X. Lu, Q. Ni, W. Li, and H. Zhang, “Dynamic user grouping and joint resource allocation with multi-cell cooperation for uplink virtual MIMO systems,” IEEE Trans. Wireless Commun., vol. 16, no. 6, pp. 3854–3869, Jun. 2017.

[17] S. Pan, Y. Yan, K. A. Bonsu, and W. Zhou, “Resource allocation algorithm for MU-MIMO systems with double-objective optimization under the existence of the rank deficient channel matrix,” IEEE Access, vol. 7, pp. 61307–61319, 2019.

[18] M. Manini, C. Gueguen, R. Legouabale, and X. Lagrange, “Study of MIMO channel matrices correlation to optimize resource allocation algorithms in multi-users 5G,” in Proc. 12th IFIP Wireless Mobile Netw. Conf. (WMNC), Paris, France, Sep. 2019, pp. 162–166.

[19] X. Lu, Q. Ni, D. Zhao, W. Cheng, and H. Zhang, “Resource virtualization for customized delay-bounded QoS provisioning in uplink VMIMO-SC-FDMA systems,” IEEE Trans. Commun., vol. 67, no. 4, pp. 2951–2967, Apr. 2019.

[20] F. A. P. de Figueiredo, N. F. T. Aniceto, J. Seki, I. Moerman, and G. Fraidenraich, “Comparing f-OFDM and OFDM performance for MIMO systems considering a 5G scenario,” in Proc. IEEE 2nd 5G World Forum (SGWF), Dresden, Germany, Oct. 2019, pp. 532–535.

[21] S. Wang, X. Ma, and Y. Chen, “Design and implementation of F-OFDM downstream filter,” in Proc. Int. Conf. Sensor Netw. Signal Process. (SNSS), Xi’an, China, Oct. 2018, pp. 195–199.

[22] M. V. G. Ferreira, F. H. T. Vieira, J. P. Félix, M. A. Ribeiro, S. G. Araujo, A. A. Cardoso, R. A. P. Franco, H. P. Correa, and P. A. V. Hamberger, “Delay optimization in resource allocation of f-OFDM wireless networks with subcarrier aggregation,” in Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE), Edmonton, AB, Canada, May 2019, pp. 1–4.

[23] M. V. G. Ferreira and F. H. T. Vieira, “Resource allocation in f-OFDM wireless networks with delay estimation using service curve and envelope process,” IEEE Latin Amer. Trans., vol. 18, no. 7, pp. 1222–1229, Jul. 2020.

[24] M. Yang, Y. Chen, and L. Du, “Interference analysis and filter parameters optimization for uplink asynchronous F-OFDM systems,” IEEE Access, vol. 7, pp. 48356–48370, 2019.

[25] C.-S. Chang and T. Zajic, “Effective bandwidths of departure processes from queues with time varying capacities,” in Proc. INFOCOM, Boston, MA, USA, vol. 3, Apr. 1995, pp. 1001–1009.

[26] M. Angulo, D. Torres-Roman, M. Turrubiertes, and D. Munoz-Rodriguez, “Admission control of a stateless network: Effective bandwidth approach,” in Proc. 15th Int. Conf. Electron., Commun. Comput. (CONIELECOMP), Puebla, Mexico, Feb./Mar. 2005, pp. 183–187.

[27] P. Teh and S. Hanly, “Effective bandwidths for wireless networks,” in Proc. GLOBECOM. IEEE Global Telecommun. Conf., San Antonio, TX, USA, Nov. 2001, pp. 584–588.

[28] L. Musavian and Q. Ni, “Effective capacity maximization with statistical delay and effective energy efficiency requirements,” IEEE Trans. Wireless Commun., vol. 14, no. 7, pp. 3824–3835, Jul. 2015.

[29] B. Chen, R. Qiu, and W. Xu, “Cognitive wireless network resource allocation strategy based on effective capacity,” in Proc. IEEE 20th Int. Conf. High Perform. Switching Routing (HPSR), Xi’an, China, May 2019, pp. 1–6.

[30] A. Thomas and A. K. Nair, “Null-subcarrier based channel estimation and mutual interference reduction in MIMO OFDM systems,” in Proc. Int. Conf. Wireless Commun. Signal Process. Network. (WISPNET), Chennai, India, Mar. 2019, pp. 275–278.

[31] A. Khare, K. Trivedi, and S. Dixit, “Effect of Doppler frequency and ber in FTT based OFDM system with Rayleigh fading channel,” in Proc. IEEE Students Conf. Electr., Electron. Comput. Sci., Bhopal, India, Mar. 2014, pp. 1–6.

[32] S. Imtiaz, G. S. Dahman, F. Rusek, and F. Tufvesson, “On the directional reciprocity of uplink and downlink channels in frequency division duplex systems,” in Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC), Washington, DC, USA, Sep. 2014, pp. 172–176.

[33] W. Cheng, X. Zhang, and H. Zhang, “Statistical-QoS driven energy-efficiency optimization over green 5G mobile wireless networks,” IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3902–3917, Dec. 2016.

[34] R. Jain, D. Chiu, and W. Hawe, “A quantitative measure of fairness and discrimination for resource allocation in shared systems,” Digit. Equip. Corp., Maynard, MA, USA, Tech. Rep. DEC-TR-301, Sep. 1984.

[35] W. Yu, L. Musavian, and Q. Ni, “Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications,” IEEE Trans. Wireless Commun., vol. 15, no. 5, pp. 3339–3353, May 2016.

[36] P. Samuelson, Economics: An Introductory Analysis. New York, NY, USA: McGraw-Hill, Dec. 1995, pp. 352–373.

[37] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks.” J. Stat. Mech., Theory Exp., vol. 2008, no. 10, 2008, Art. no. P10008.

[38] Evolved Universal Terrestrial Radio Access (E-UTRA), Physical Channel and Modulation (Release 10), document 3GPP TS 36.211, V10.5.0, 2009.
RUOXU CHEN was born in Shaanxi, China. He received the B.Sc. degree in electronic information science and technology from Sichuan University, Chengdu, China, in 2017. He is currently pursuing the M.Sc. degree in communication and information system with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi’an. His research interests include SC-FDMA, virtual MIMO, and resource allocation of wireless communications.

XIAOFENG LU received the B.Sc. degree from Sichuan University, Chengdu, China, in 1996, the M.Sc. degree from Hunan University, Changsha, China, in 1999, and the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2006. From 1999 to 2003, he was a Research and Development Engineer with the Wuhan Research Institute of Post and Telecommunications. He is currently a Full Professor with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi’an, China. His main research interests include broadband wireless communications and covering topics, such as resource allocation and virtualization, MU-MIMO, and OFDMA.