IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 18, 2020, accepted April 30, 2020, date of publication May 11, 2020, date of current version May 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993597

A Hybrid-Grant Random Access Scheme in Massive MIMO Systems for IoT

QI ZHANG[1], (Member, IEEE), SHI JIN[2], (Senior Member, IEEE), AND HONGBO ZHU[1]
[1]Jiangsu Key Laboratory of Wireless Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2]National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China

Corresponding author: Hongbo Zhu (zhuhb@njupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61801244, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180754, and in part by the Initial Scientific Research Foundation of NJUPT under Grant NY218103. The work of Shi Jin was supported in part by the National Key Research and Development Program under Grant 2018YFA0701602, and in part by the National Science Foundation of China under Grant 61941104.

ABSTRACT The grant-free random access (RA) can minimize the access delay but also brings severe data transmission interferences. To overcome this defect, we propose a new RA scheme that inserts a base station (BS) broadcasting message after user equipments (UEs) transmitting pilots. In this scheme, UEs can determine whether they have colliders by resolving the broadcasting message, and only non-colliding UEs can transmit data in the following step while colliding UEs keep silent. By doing this, the data interferences from colliding UEs are eliminated without costing much extra time. Since this BS broadcasting message is also used in the legacy grant-based RA, we call the new RA scheme as a hybrid-grant RA. We investigate the hybrid-grant RA in massive multiple-input multiple-output (MIMO) systems and obtain a tight closed-form approximation of the spectral efficiency with maximum-ratio-combing (MRC) and zero-forcing (ZF) receivers, respectively. Via simulation, we find that our proposed hybrid-grant RA can obtain a significant gain on the spectral efficiency compared with grant-free RA, especially for ZF receivers. In particular, this gain grows rapidly as the UE number goes up, which means the hybrid-grant RA is more suitable for the system with large amount of UEs and it is a typical scenario in future communications networks. Moreover, we also analyze the optimal pilot length and UE activation probability that maximize the spectral efficiency, which can be used as references for the practical application of the proposed hybrid-grant RA.

INDEX TERMS IoT, massive MIMO, random access.

I. INTRODUCTION

The massive multiple-input multiple-output (MIMO) system which employs hundreds of antennas at the BS to serve tens of users simultaneously in the same time-frequency resource has been regarded as an essential technique of the fifth generation (5G) wireless systems [1], [2]. The large size of transmit antenna array not only improves the system capacity significantly [3], [4], but also averages out the effect of fast channel fading and provides extremely sharp beamforming concentrated into small areas [5], [6]. Aside from these, the huge degrees-of-freedom offered by massive MIMO also reduce the transmit power [7]. Due to the limited amount of user equipments (UEs), the conventional massive MIMO usually considers fully-loaded access of all UEs. However, this situation will change in Internet-of-Things (IoT) systemS. IoT intends to make everyday objects connected and smart by deploying great amount of machines that are typically wireless, such as sensors [8], [9]. The autonomous communications among machines in IoT creates a new frontier of wireless communications and networks, called machine-to-machine (M2M) communications. Millions of devices are employed in M2M which leads to the number of terminals much larger than the available pilots. Therefore, the access of the massive devices becomes a key issue in M2M communications [10], [11].

Random access (RA) has been studied extensively in long term evolution (LTE). The legacy grant-based RA briefly includes four steps: UE request, BS acknowledge, colliding UE retransmit pilot, and BS grant admission [12]. Using the advantage of massive MIMO like high spatial resolution and channel hardening to improve the performance of grant-based RA has been studied in [13], [14]. However, due to the complicated signaling and several possible iterations between the

The associate editor coordinating the review of this manuscript and approving it for publication was Liang Yang.

BS acknowledgment and UE retransmission, the grant-based RA has a relative long waiting time before the data transmission. Therefore, it cannot meet the demand of short delay in M2M communications [15], [16]. Given this, the grant-free RA with low signaling overhead attracts much attentions. In grant-free RA, the request-grant procedure is removed and UEs transmits directly the randomly-select pilot along with data [17]–[19]. By doing this, the access delay is minimized, but the transmission interference is also enhanced. The colliding UEs which select the same pilots cannot be detected by the BS, but they transmit data together with non-colliding UEs. Therefore, their data transmissions have no positive effect on the system performance but only bring interferences to the data transmission of other UEs, especially when the UE number is large. A new random access mechanism is needed.

In this paper, we modify the grant-free RA by inserting a BS broadcasting message after the UE transmitting pilot sequences. This message contains the identifications (IDs) of all non-colliding UEs, and UEs can determine whether it has colliders by resolving the message. Only non-colliding UEs can transmit data in the following time while colliding UEs keep silent. By doing this, the data transmission interferences from colliding UEs are eliminated. Since the BS broadcasting message is also used in the legacy grant-based RA, the proposed new RA scheme can be regarded as a combination of grant-based and grant-free RA. Therefore, we call it hybrid-grant RA. In this paper, we investigate the performance of the hybrid-grant RA in massive MIMO systems. After taking into account the extra time consumed by the BS broadcasting, we obtain a tight closed-form approximation of the spectral efficiency with maximum-ratio-combing (MRC) and zero-forcing (ZF) receivers, respectively. For comparison, we also give the spectral efficiency using the grant-free RA, and compare these two schemes from multiple aspects. In particular, we find that our proposed hybrid-grant RA can obtain a significant gain on the spectral efficiency, and the gain with ZF receivers is more remarkable. This gain grows rapidly as the UE number goes up. Hence, the proposed hybrid-grant RA is more suitable for the system with large amount of UEs, which is a typical application in the future communications networks. Moreover, we also give the optimal pilot length and UE activation probability that maximize the spectral efficiency using hybrid-grant RA, which can be used as references for practical configuration of hybrid-grant RA.

The remainder of this paper is organized as follows. Section II describe the system model and the proposed hybrid-grant RA. In Section III, we derive a tight approximation of the spectral efficiency, and the spectral efficiency with grant-free RA is also given for comparison. In Section V, we provide numerical results to validate the analytical results and further study the performance of the hybrid-grant RA. Finally, Section VI summarizes the main results of this paper.

Notation-Throughout the paper, vectors are expressed in lowercase boldface letters while matrices are denoted by uppercase boldface letters. We use $\mathbf{X}^H$ to denote the conjugate-transpose of $\mathbf{X}$, and use $[\mathbf{X}]_{ij}$ to denote the (i, j)-th entry of $\mathbf{X}$. Finally, $\|\cdot\|$ is the Euclidean norm.

II. SYSTEM MODEL AND RANDOM ACCESS

In this section, we describe the system model and introduce the proposed hybrid-grant RA in detail.

A. SYSTEM MODEL

Consider a single-cell multiuser MIMO system, where the BS is equipped with M antennas and N single-antenna UEs are uniformly distributed in the cell which are denoted from UE 1 to UE N. UEs transmit their signals to the BS in the same time-frequency channel. The $M \times 1$ channel vector between UE n and the BS is

$$\mathbf{g}_n = \mathbf{h}_n \sqrt{\beta_n}, \tag{1}$$

where $\mathbf{h} \sim \mathcal{CN}(0, \mathbf{I}_M)$ is the $M \times 1$ small-scale fading vector between UE n and the BS, and β_n is the large-scale fading coefficient between UE n and the BS. Note that β_n models both path loss and shadow fading and is assumed to be constant across the BS antenna array.

In each RA slot, every UE decides randomly and independently whether or not to transmit data to the BS and the activation probability is p_{a}. See Fig. 1. Each active UE randomly select a pilot sequence from the predefined pilot pools and transmit it to the BS. A total number of τ orthogonal pilot sequences are available, where each pilot sequence is τ symbols long. The channel coherence interval is T symbols long and $T > \tau$. If several UEs select the same pilots, we say a collision occurs. UEs that select the same pilots are called colliding UEs, while UEs that select pilots different from the pilots selected by other UEs are called non-colliding UEs. Colliding UEs call each other as colliders. Only non-colliding UEs can be detected and estimated by the BS.

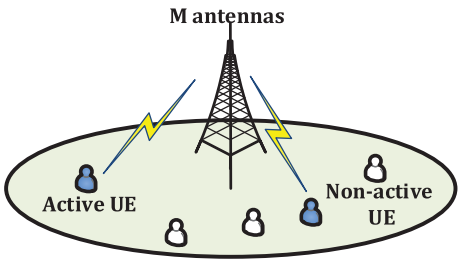


FIGURE 1. Illustration of the system model.

B. PROPOSED HYBRID-GRANT RA

In the grant-free RA, active UEs directly send their data to the BS after transmitting pilots. By doing this, the access delay can be minimized. The procedure and the time allocated for each step in a channel coherence interval of grant-free RA is shown in Fig. 2 and 3, respectively.

However, the shortcoming of grant-free RA is also obvious. Since colliding UEs cannot be detected and estimated by the BS, the data transmissions from them have no positive

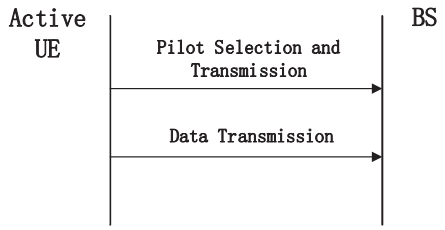


FIGURE 2. Procedure of the grant-free RA.

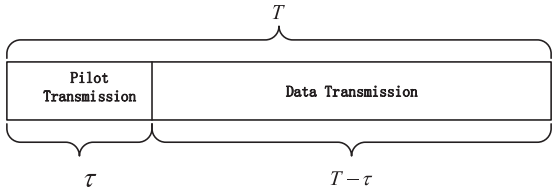


FIGURE 3. Time allocation for each step of the grant-free RA.

effect on the system performance but bring interferences to non-colliding UEs. The interference become more severe as the number of UEs goes up. To fix this problem, we propose a new RA scheme which inserts a BS broadcasting message after active UEs transmitting pilot. UEs can determine whether they have colliders by resolving this message and only non-colliding UEs are allowed to transmit data. Since this BS broadcasting message is also used in the legacy grant-based RA, the new RA scheme can be regarded as a combination of grant-free and grant-based RA. Therefore, we call it hybrid-grant RA.

The detailed procedure of hybrid-grant RA is described in Fig. 4. There are three steps. In Step I, each active UE randomly selects a pilot sequences from the predefined pilot pools and then transmit it to the BS. In Step II, after receiving pilots, the BS detects non-colliding UEs and estimates their channels. Then, it broadcasts a message which contains the IDs of non-colliding UEs. In Step III, UEs that can match their IDs from the message transmit data to the BS, while other UEs keep silent. In hybrid-grant RA, the data interferences from colliding UEs in grant-free RA are eliminated and the system performance can be improved.

Although the BS broadcasting message that contains collison information is also used in the grant-based RA, we can note the difference. In grant-based RA, after resolving the broadcasting message, the colliding UEs will randomly select pilot sequences again (perhaps from a different pilot set) and send them to the BS. Then, another round of message broadcasting from the BS as well as the pilot retransmission from colliding UEs may occur until the access requirement is satisfied. Due to the complicated signaling and several possible iterations between BS broadcasting and UE retransmission, the grant-based RA has a relative long waiting time before data transmission. Given that, our proposed access mechanism only uses the BS broadcasting message to prevent the data transmission from colliding UEs but does not allow them to reselect and retransmit pilots. Therefore, the interference can be reduced without costing much extra time.

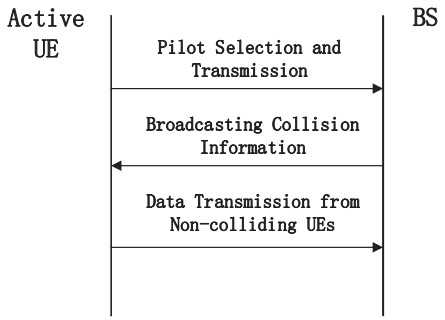


FIGURE 4. Procedure of the hybrid-grant RA.

Assume that the broadcasting message in Step II takes μ symbols. Then, the time allocation for each step in a channel coherence interval is given in Fig. 5.

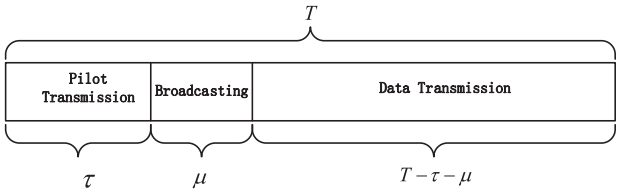


FIGURE 5. Time allocation for each step of the hybrid-grant RA.

III. PERFORMANCE ANALYSIS

In this section, we derive the spectral efficiencies using hybrid-grant RA with MRC and ZF receivers, respectively, and the spectral efficiency using grant-free RA is also given for comparison.

A. CHANNEL ESTIMATION

We use F_{a} to denote the set of active UEs and the number of UEs in F_{a} is N_{a}. The pilots sent by UE n can be stacked into a $\tau \times 1$ vector, denoted as $\sqrt{\tau}\boldsymbol{\phi}_n$, where $\boldsymbol{\phi}_n^H \boldsymbol{\phi}_n = 1$ and $\boldsymbol{\phi}_{n_1}^H \boldsymbol{\phi}_{n_2} = 0$ for $n_1 \neq n_2$. As such, the received $M \times \tau$ noisy pilots matrix at the BS is

$$\mathbf{Y} = \sum_{n \in F_{\mathrm{a}}} \sqrt{\tau P_n} \mathbf{g}_n \boldsymbol{\phi}_n^T + \boldsymbol{\Omega}, \tag{2}$$

where P_n is the transmit power of UE n and Ω is the additive white Gaussian noise (AWGN) matrix. In Assume that the pilot sent by UE n does not collide with other pilots. Then, the BS can detect this UE and estimate its channel by multiplying $\boldsymbol{\phi}_n^*$ as follows

$$\mathbf{y}_n = \mathbf{Y}\boldsymbol{\phi}_n = \sqrt{\tau P_n}\mathbf{g}_n + \boldsymbol{\omega}, \tag{3}$$

where $\boldsymbol{\omega} = \boldsymbol{\Omega}\boldsymbol{\phi}_n^*$. Since $\boldsymbol{\phi}_n^H \boldsymbol{\phi}_n = 1$, elements of $\boldsymbol{\omega}$ has the same distribution as that of $\boldsymbol{\Omega}$. That is, $\boldsymbol{\omega} \sim \mathcal{CN}(0, \mathbf{I}_M)$. With (3), we can get the minimum-mean-square-estimation (MMSE) of $\mathbf{g}_n$ as

$$\hat{\mathbf{g}}_n = \eta \left(\mathbf{g}_n + \frac{1}{\sqrt{\tau P_n}} \boldsymbol{\omega} \right), \tag{4}$$

where $\eta = \tau\lambda/(\tau\lambda+1)$, while $\lambda = P_i\beta_i$ (for any i) is the uniform product of each UE's transmit power and its large-scale fading coefficient. λ is usually defined as the criterion for power control to compensate the large-scale fading among different UEs.

B. SPECTRAL EFFICIENCY

After detecting the non-colliding UEs, the BS broadcasts a message that takes μ symbols, and UEs can determine wether it has colliders or not by resolving this message. Only non-colliding UEs can transmit data while colliding UEs keep silent. We use F_{non} to denote the set of non-colliding UEs and the number of UEs in F_{non} is N_{non}. The estimated channel matrix between non-colliding UEs and the BS is $\hat{\mathbf{G}} = \left\{\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_n, \ldots | n \in F_{\text{non}}\right\}$. The received $M \times 1$ data signal matrix at the BS is

$$\mathbf{r} = \sum_{n\in F_{\text{non}}} \sqrt{P_n}\mathbf{g}_n x_n + \boldsymbol{\theta}, \tag{5}$$

where x_n is the data symbol transmitted by UE n with $\mathbb{E}\{|x_n|^2\} = 1$, and $\boldsymbol{\theta} \sim \mathcal{CN}(0, \mathbf{I}_M)$ is the AWGN vector. Let $\mathbf{A}$ denote the receiver matrix. Then, after linear reception, we can get that

$$\mathbf{c} = \mathbf{A}^H\mathbf{r}. \tag{6}$$

Next, we analyze the uplink rate with MRC and ZF receivers, respectively.

1) MRC

For MRC receivers, $\mathbf{A} = \hat{\mathbf{G}}$. Assume that UE n does not collide with other UEs. Substituting $\mathbf{A} = \mathbf{G}$ into (6), we can get the detected data symbol for UE n as

$$c_n^{\text{MRC}} = \sum_{i\in F_{\text{non}}} \sqrt{P_i}\hat{\mathbf{g}}_n^H\mathbf{g}_i x_i + \hat{\mathbf{g}}_n^H\boldsymbol{\theta}, \tag{7}$$

and it can be further written as

$$c_n^{\text{MRC}} = \sum_{i\in F_{\text{non}}} \sqrt{P_i}\hat{\mathbf{g}}_n^H(\hat{\mathbf{g}}_i - \tilde{\mathbf{g}}_i)x_i + \hat{\mathbf{g}}_n^H\boldsymbol{\theta}, \tag{8}$$

where $\tilde{\mathbf{g}}_i = \hat{\mathbf{g}}_i - \mathbf{g}_i$ is the channel estimation error which is independent on $\hat{\mathbf{g}}_i$ according to the property of MMSE and $\tilde{\mathbf{g}}_i \sim \mathcal{CN}(0, \beta_i(1-\eta)\mathbf{I}_M)$. By treating the uncorrelated interference and noise in (8) as independent Gaussian noise, we can get the ergodic achievable uplink rate of UE n as

$$R_n^{\text{MRC}} = \mathbb{E}\left\{\log_2\left(1+\frac{P_n\left|\hat{\mathbf{g}}_n^H\hat{\mathbf{g}}_n\right|^2}{\sum\limits_{i\in F_{\text{non}}^{!n}} P_i\left|\hat{\mathbf{g}}_n^H\hat{\mathbf{g}}_i\right|^2 + \sum\limits_{i\in F_{\text{non}}} P_i\left|\hat{\mathbf{g}}_n^H\tilde{\mathbf{g}}_i\right|^2 + \left\|\hat{\mathbf{g}}_n\right\|^2}\right)\right\}, \tag{9}$$

where $F_{\text{non}}^{!n}$ is the set that excludes n from F_{non}. Since the data transmission only takes part of the resource in the whole coherence time, according to Fig. 5, we can get the spectral efficiency of the system as

$$S^{\text{MRC}} = (1-\frac{\tau+\mu}{T})\sum_{n\in F_{\text{non}}} R_n^{\text{MRC}} \tag{10}$$

The spectral efficiency in (10) is conditioned on a specific UE activation and pilot selection. The unconditioned spectral efficiency is given as

$$\bar{S}^{\text{MRC}} = \mathbb{E}\left\{S^{\text{MRC}}\right\}, \tag{11}$$

The following theorem presents the closed-form approximation of (11).

Theorem 1: For MRC receivers, the closed-form approximation of spectral efficiency is

$$\bar{S}^{\text{MRC}} \approx \bar{N}_{\text{non}}(1-\frac{\tau+\mu}{T})\log_2\left(1+\frac{(M+1)\eta\lambda}{\bar{N}_{\text{non}}\lambda - \eta\lambda + 1}\right), \tag{12}$$

where

$$\bar{N}_{\text{non}} = \sum_{n=1}^{N} n\binom{N}{n}p_{\text{a}}^n(1-p_{\text{a}})^{N-n}(1-\frac{1}{\tau})^{n-1} \tag{13}$$

is the average number of non-colliding UEs.

Proof: See Appendix A. □

Theorem 1 provides an analytical metric to evaluate the performance of the hybrid-grant RA with MRC receivers, and its tightness will be validated in Section V. To apply the hybrid-grant RA in practice, we are interested in the optimal τ and p_{a} that maximize $\bar{S}^{\text{MRC}}$, where p_{a} is adjustable since we can set a random backoff time for each UE. However, since $\bar{S}^{\text{MRC}}$ does not monotonously change with both τ and p_{a}, the optimal τ and p_{a} cannot be got straightforwardly. The precise behavior of $\bar{S}^{\text{MRC}}$ with respect to τ, p_{a} and N will be investigated in Section V, and we will use a low-complexity algorithm to get the combination of optimal τ and p_{a} in Section IV.

If UEs are always active, $p_{\text{a}} = 1$. The following corollary gives the spectral efficiency under this special case.

Corollary 1: When UEs are always active, i.e., $p_{\text{a}} = 1$, the approximation of the spectral efficiency with MRC receivers becomes

$$\bar{S}_{p_{\text{a}}=1}^{\text{MRC}} \approx N(1-\frac{\tau+\mu}{T})\left(1-\frac{1}{\tau}\right)^{N-1} \times \log_2\left(1+\frac{(M+1)\eta\lambda}{N\lambda\left(1-\frac{1}{\tau}\right)^{N-1} - \eta\lambda + 1}\right). \tag{14}$$

Proof: When $p_{\text{a}} = 1$, we have that

$$\bar{N}_{\text{non}} = N(1-\frac{1}{\tau})^{N-1}. \tag{15}$$

Then, (14) is got by substituting (15) into (12). □

The spectral efficiency in Corollary 1 applies to the system with heavy traffic, where the service requirements arrival frequently in the UE and thus the activation probability can be regarded as 1.

2) ZF

For ZF receivers, $\mathbf{A} = \hat{\mathbf{G}}\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}$. Assume that UE n does not collide with other UEs, and $\hat{\mathbf{g}}_n$ is the n'-th column of $\hat{\mathbf{G}}$. Then, the detected data symbol of UE n is

$$c_n^{\mathrm{ZF}} = \sum_{i\in F_{\mathrm{non}}} \sqrt{P_i}\mathbf{a}_{n'}^H(\hat{\mathbf{g}}_i - \tilde{\mathbf{g}}_i)x_i + \mathbf{a}_{n'}^H\boldsymbol{\theta}, \tag{16}$$

where $\mathbf{a}_{n'}$ is the n'-th column of $\mathbf{A}$. Then, following the same procedure as (9), we can get the achievable uplink rate of UE n with ZF receivers as

$$R_n^{\mathrm{ZF}} = \mathbb{E}\left\{\log_2\left(1 + \frac{P_n}{\sum\limits_{i\in F_{\mathrm{non}}} P_i\left|\mathbf{a}_{n'}^H\tilde{\mathbf{g}}_i\right|^2 + \|\mathbf{a}_{n'}\|^2}\right)\right\}. \tag{17}$$

Following the same procedure as (11), we can get the unconditioned spectral efficiency of the system for ZF receivers as

$$\bar{S}^{\mathrm{ZF}} = (1 - \frac{\tau+\mu}{T})\mathbb{E}\left\{\sum_{n\in F_{\mathrm{non}}} R_n^{\mathrm{ZF}}\right\}. \tag{18}$$

The following theorem presents the closed-form approximation of (18).

Theorem 2: For ZF receivers, the closed-form approximation of spectral efficiency is

$$\bar{S}^{\mathrm{ZF}} \approx \bar{N}_{\mathrm{non}}(1-\frac{\tau+\mu}{T})\log_2\left(1+\frac{(M-\bar{N}_{\mathrm{non}})\eta\lambda}{\bar{N}_{\mathrm{non}}\lambda(1-\eta)+1}\right). \tag{19}$$

Proof: See Appendix B. □

The tightness of (19) will be validated in Section V, and the behavior of $\bar{S}^{\mathrm{ZF}}$ with respect to τ, p_{a} and N will be investigated in the same section as well. Moreover, to facilitate the practical application of the hybrid-grant RA with ZF receivers, we also obtain the optimal τ and p_{a} that maximize $\bar{S}^{\mathrm{ZF}}$ with a low-complexity algorithm in Section IV. The following corollary gives the spectral efficiency under the special case with $p_{\mathrm{a}} = 1$.

Corollary 2: When UEs are always active, i.e., $p_{\mathrm{a}} = 1$, the approximation of the spectral efficiency with ZF receivers becomes

$$\bar{S}^{\mathrm{ZF}}_{p_{\mathrm{a}}=1} \approx N(1-\frac{\tau+\mu}{T})\left(1-\frac{1}{\tau}\right)^{N-1} \times \log_2\left(1+\frac{\left[M\text{-}N\left(1-\frac{1}{\tau}\right)^{N-1}\right]\eta\lambda}{N\lambda\left(1-\frac{1}{\tau}\right)^{N-1}(1-\eta)+1}\right). \tag{20}$$

Proof: The proof is omitted because it is similar with the proof of Corollary 1. □

The result in Corollary 2 can be used in the system with heavy traffic.

Now, we have derived the spectral efficiencies of the proposed hybrid-grant RA with both MRC and ZF receivers. To show the effectiveness of this new scheme, we next give the spectral efficiency using the grant-free RA for comparison.

3) COMPARISON

In grant-free RA, active UEs transmit their data directly after the random selecting and transmitting pilots, no matter they have colliders or not.

Assume that UE n does not collide with other UEs. Since the BS still can only detect and estimate non-colliding UEs, the estimation of $\mathbf{g}_n$ is the same as (4). The received $M \times 1$ data signal matrix at the BS is

$$\mathbf{r}^{\mathrm{f}} = \sum_{n=1}^{N}\sqrt{P_n}\mathbf{g}_n x_n + \boldsymbol{\theta}^{\mathrm{f}}, \tag{21}$$

where $\boldsymbol{\theta}^{\mathrm{f}} \sim \mathcal{CN}(0, \mathbf{I}_M)$ is the AWGN vector. Then, after linear reception, we can get that

$$\mathbf{c}^{\mathrm{f}} = \mathbf{A}^H\mathbf{r}^{\mathrm{f}}. \tag{22}$$

The spectral efficiency in the following still given for MRC and ZF receivers, respectively.

For MRC receivers, the detected data symbol for UE n is

$$c_n^{\mathrm{f,MRC}} = \sum_{i\in F_{\mathrm{non}}}\sqrt{P_i}\hat{\mathbf{g}}_n^H(\hat{\mathbf{g}}_i - \tilde{\mathbf{g}}_i)x_i + \sum_{j\in F_{\mathrm{col}}}\sqrt{P_j}\hat{\mathbf{g}}_n^H\mathbf{g}_j x_j + \hat{\mathbf{g}}_n^H\boldsymbol{\theta}^{\mathrm{f}}, \tag{23}$$

where F_{col} is the set of colliding UEs. Then, by treating the uncorrelated interference and noise in (23) as independent Gaussian noise, we can get the ergodic achievable uplink rate of UE n as (24), as shown at the bottom of the next page.

In grant-free RA, according to Fig. 3, the data transmission takes a $\frac{T-\tau}{T}$ fraction of the whole channel coherence time. Therefore, the unconditioned spectral efficiency of the system is

$$\bar{S}^{\mathrm{f,MRC}} = (1-\frac{\tau}{T})\mathbb{E}\left\{\sum_{n\in F_{\mathrm{non}}} R_n^{\mathrm{f,MRC}}\right\}. \tag{25}$$

Following the same procedure as (12), we can get the closed-form approximation of (25) as follows

$$\bar{S}^{\mathrm{f,MRC}} \approx \bar{N}_{\mathrm{non}}(1-\frac{\tau}{T})\log_2\left(1+\frac{(M+1)\eta\lambda}{\bar{N}_{\mathrm{a}}\lambda-\eta\lambda+1}\right), \tag{26}$$

where $\bar{N}_{\mathrm{a}} = Np_{\mathrm{a}}$ is the average number of active UEs.

For ZF receivers, the detected data symbol for UE n is

$$c_n^{\mathrm{f,ZF}} = \sum_{i\in F_{\mathrm{non}}}\sqrt{P_i}\mathbf{a}_{n'}^H(\hat{\mathbf{g}}_i - \tilde{\mathbf{g}}_i) + \sum_{j\in F_{\mathrm{col}}}\sqrt{P_j}\mathbf{a}_{n'}^H\mathbf{g}_j x_j + \mathbf{a}_{n'}^H\boldsymbol{\theta}^{\mathrm{f}}. \tag{27}$$

Then, the ergodic achievable uplink rate of UE n is

$$R_n^{\mathrm{f,ZF}} = \mathbb{E}\left\{\log_2\left(1+\frac{P_n}{\sum\limits_{i\in F_{\mathrm{non}}}P_i\left|\mathbf{a}_{n'}^H\tilde{\mathbf{g}}_i\right|^2 + \sum\limits_{j\in F_{\mathrm{col}}}P_j\left|\mathbf{a}_{n'}^H\mathbf{g}_j\right|^2 + \|\mathbf{a}_{n'}\|^2}\right)\right\}. \tag{28}$$

Therefore, the unconditioned spectral efficiency of the system is

$$\bar{S}^{\mathrm{f,ZF}} = (1-\frac{\tau}{T})\mathbb{E}\left\{\sum_{n\in F_{\mathrm{non}}} R_n^{\mathrm{ZF,c}}\right\}. \tag{29}$$

Following the same procedure as (19), we can get the closed-form approximation of (29) as follows

$$\bar{S}_n^{\mathrm{f,ZF}} \approx \bar{N}_{\mathrm{non}}(1-\frac{\tau}{T})\log_2\left(1+\frac{(M-\bar{N}_{\mathrm{non}})\eta\lambda}{\bar{N}_{\mathrm{a}}\lambda-\bar{N}_{\mathrm{non}}\eta\lambda+1}\right). \tag{30}$$

Comparing (12) and (19) with (26) and (30), respectively, we can find that our proposed hybrid-grant RA can get a larger achievable rate due to the elimination of interferences from colliding UEs, but it diminishes the time used for data transmission. On the other hand, the grant-free RA can provide more time for data transmission, but the achievable rate is impaired by the data interferences from colliding UEs. We will compare these two schemes comprehensively and reveal the superiority of the hybrid-grant RA with numerical results in Section V.

To apply the hybrid-grant RA in practice, we are interested in the optimal τ and p_{a} that maximize the spectral efficiency, which will be given in the next section.

IV. OPTIMAL SYSTEM CONFIGURATIONS

To facilitate the practical application of the hybrid-grant RA, we aim to find the optimal pilot length and UE activation probability that maximize the spectral efficiency using hybrid-grant RA. Note that the UE activation probability is adjustable, since we can set a random backoff time for each UE to lower its activation probability. Hence, the optimal UE activation probability is upper bounded by the original activation probability. We give the optimal system configurations for MRC and ZF receivers, respectively.

A. MRC

For MRC receivers, the spectral efficiency is given in (12). To get the optimal pilot length and UE activation probability that maximize the spectral efficiency, we need to solve the following optimization problem

$$\begin{aligned}(\overset{*}{\tau}^{\mathrm{MRC}}, \overset{*}{p}_{\mathrm{a}}^{\mathrm{MRC}}) = \max_{\tau,p_{\mathrm{a}}}\ & \bar{S}^{\mathrm{MRC}} \\ \text{s.t. } & 1 \le \tau < T-\mu \\ & 0 < p_{\mathrm{a}} \le p_{\mathrm{a}}^{\max},\end{aligned} \tag{31}$$

where $\overset{*}{\tau}^{\mathrm{MRC}}$ and $\overset{*}{p}_{\mathrm{a}}^{\mathrm{MRC}}$ denote the optimal τ and p_{a} with MRC receivers, respectively. $p_{\mathrm{a}}^{\max}$ is the maximum activation probability of UEs, i.e., the original UE activation probability without manual backoff. The upper bound of τ is got from that $\tau+\mu<T$, which means at least one time resource should be left for the data transmission.

After observing (31), we find that the objective function $\bar{S}^{\mathrm{MRC}}$ is neither concave or convex with respect to τ and p_{a}, thus the classical Karush-Kuhn-Tucker (KKT) conditions cannot be adopted. However, since τ is an upper-bounded integer, we can exhaustively search all its feasible values. Then, with a fixed τ, (31) reduces to a single-variable problem that optimizes p_{a}. To solve this single-variable problem, we employ the Majorization-Minimization (MM) Algorithm which can get the desired solution iteratively [20]. The MM Algorithm briefly contains two steps: first, we approximate the objective function by its second-order taylor expansion to turn the targeted optimization problem into a quadratic one, which can be solved with much lower complexity; secondly, we successively maximize the objective function until the sequence of iterative solutions converge to the optimal. After getting the optimal p_{a} for each value of τ, the final optimal τ and p_{a} are the combination that yield the largest spectral efficiency. The detailed procedure is described in Algorithm 1. The solution of (31) is summarized in the following theorem.

Theorem 3: The optimal τ and p_{a} that maximizes the spectral efficiency with MRC receivers, i.e., the solution of the optimization problem (31) *are the output of Algorithm 1.*

From Theorem 3, we can get the optimal τ and p_{a} that maximizes the spectral efficiency with MRC receivers. The effectiveness of these parameters will be validated through numerical results by comparing with the system configured without optimizations.

B. ZF

For ZF receivers, the spectral efficiency is given in (19). To get the optimal pilot length and UE activation probability that maximize the spectral efficiency, we need to solve the following optimization problem

$$\begin{aligned}(\overset{*}{\tau}^{\mathrm{ZF}}, \overset{*}{p}_{\mathrm{a}}^{\mathrm{ZF}}) = \max_{\tau,p_{\mathrm{a}}}\ & \bar{S}^{\mathrm{ZF}} \\ \text{s.t. } & 1 \le \tau < T \\ & p_{\mathrm{a}} \le p_{\mathrm{a}}^{\max},\end{aligned} \tag{32}$$

After observing (32), we find that the objective function $\bar{S}^{\mathrm{ZF}}$ is also not concave or convex with respect to τ and p_{a}. Therefore, we still use the MM Algorithm to solve the optimal p_{a} for each feasible value of τ, and select the τ and p_{a} that maximize the spectral efficiency as the final output.

[1]This is quadratic optimization problem and can be easily solved.

$$R_n^{\mathrm{f,MRC}} = \mathbb{E}\left\{\log_2\left(1+\frac{P_n\left|\hat{\mathbf{g}}_n^H\hat{\mathbf{g}}_n\right|^2}{\sum\limits_{i\in F_{\mathrm{non}}^{!n}} P_i\left|\hat{\mathbf{g}}_n^H\hat{\mathbf{g}}_i\right|^2+\sum\limits_{i\in F_{\mathrm{non}}} P_i\left|\hat{\mathbf{g}}_n^H\tilde{\mathbf{g}}_i\right|^2+\sum\limits_{j\in F_{\mathrm{col}}} P_j\left|\hat{\mathbf{g}}_n^H\mathbf{g}_j\right|^2+\left\|\hat{\mathbf{g}}_n\right\|^2}\right)\right\}. \tag{24}$$

Algorithm 1 Solving the Optimization Problem (31)

Initializ: $\tau_0 = 1$, $\chi = 0$, any feasible $p_{\mathrm{a}}^{(0)}$, maximum error $\epsilon > 0$

repeat
  $k = 0$, $C = \bar{S}^{\mathrm{MRC}}_{|\tau=\tau_0}$
  repeat
    $\Delta_1 = \frac{dC}{dp_{\mathrm{a}}}_{|p_{\mathrm{a}}=p_{\mathrm{a}}^{(k)}}$, $\Delta_2 = \frac{d^2C}{dp_{\mathrm{a}}^2}_{|p_{\mathrm{a}}=p_{\mathrm{a}}^{(k)}}$
    $\mathcal{V} = C_{|p_{\mathrm{a}}=p_{\mathrm{a}}^{(k)}} + (p_{\mathrm{a}} - p_{\mathrm{a}}^{(k)})\Delta_1 + \frac{1}{2}(p_{\mathrm{a}} - p_{\mathrm{a}}^{(k)})^2\Delta_2$
    $p_{\mathrm{a}}^{(k+1)} = \max_{p_{\mathrm{a}}} \quad \mathcal{V}$ [1]
      s.t. $\quad 1 \le p_{\mathrm{a}} \le p_{\mathrm{a}}^{\max}$
    $k = k + 1$
  until $\left| p_{\mathrm{a}}^{(k)} - p_{\mathrm{a}}^{(k-1)} \right| < \epsilon$
  $C = \bar{S}^{\mathrm{MRC}}_{|\tau=\tau_0, p_{\mathrm{a}}=p_{\mathrm{a}}^{(k)}}$
  if $C > \chi$ **then**
    $\chi = C$, $\tau^{*\mathrm{MRC}} = \tau_0$, $p_{\mathrm{a}}^{*\mathrm{MRC}} = p_{\mathrm{a}}^{(k)}$
  end if
  $\tau_0 = \tau_0 + 1$
until $\tau_0 \ge T - \mu$

The detailed procedure is similar as Algorithm 1 but a replacement of $\bar{S}^{\mathrm{MRC}}$ by $\bar{S}^{\mathrm{ZF}}$ is needed. The solution of the problem (32) is summarized in the following theorem.

Theorem 4: The optimal τ and p_{a} that maximize the spectral efficiency with ZF receivers, i.e., the solution of the optimization problem (32) *are the output of the algorithm modified from Algorithm 1 by replacing $\bar{S}^{\mathrm{MRC}}$, $\tau^{*\mathrm{MRC}}$, $p_{\mathrm{a}}^{*\mathrm{MRC}}$ with $\bar{S}^{\mathrm{ZF}}$, $\tau^{*\mathrm{ZF}}$ and $p_{\mathrm{a}}^{*\mathrm{ZF}}$, respectively.*

The effectiveness of the optimal results got from Theorem 4 will be justified in Section V by comparing with the system configured with non-optimized parameters. Moreover, the performance of the optimal parameters with MRC and ZF receivers will also have a comparison.

V. NUMERICAL RESULTS

In this section, we validate the accuracy of the spectral efficiency in Theorem 1 and 2, and show the effectiveness of our proposed access mechanism. We set $\lambda = 1$ which means the large-scale fading is completely compensated. The following parameters were chosen according to the LTE standard: an OFDM symbol interval of $T_s = 500/7 \approx 71.4\ \mu s$, a subcarrier spacing of $\Delta f = 15$ kHz, a useful symbol duration $T_u = 1/\Delta f \approx 66.7\ \mu s$. We choose the channel coherence time to be $T_c = 1$ ms. As a result, the coherence time of the channel becomes $T = T_c T_u/[T_s(T_s - T_u)] = 196$ symbols. According to 3GPP [12], the BS broadcasts the identifies of non-colliding UEs in a message with fixed size of 48 bits. Assume that QPSK modulation is adopted, so each OFDM symbol contains 6 bits. Hence, we set the BS broadcasting message takes $\mu = 8$ symbols.

A. PERFORMANCE VALIDATION

In Fig. 6, the simulated spectral efficiency in (11), (18), (25) and (29) are compared with the analytical approximation in (12), (19), (26) and (30), respectively. We can see that the simulation results and analytical approximation have a close match, thus verifies our analytical results. Moreover, it can be seen that the proposed hybrid-grant RA can get a remarkable gain on the spectral efficiency over the grant-free RA, for both MRC and ZF receivers. It is also found that the performance gap between MRC and ZF is more obvious in the hybrid-grant RA, which also reveals the superiority of the hybrid-grant RA since ZF outperforms more apparently in the high SINR region. Due to the tightness between the simulations and analysis, we will use the latter for our following investigations.

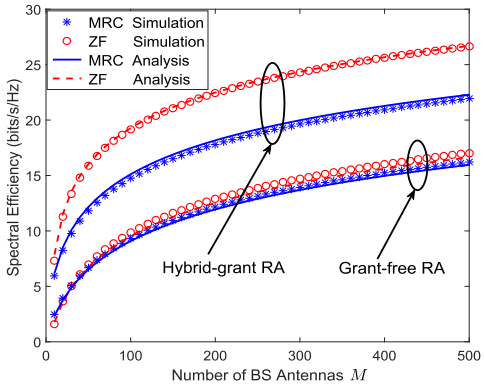


FIGURE 6. Spectral efficiency vs. BS antennas number M, where $\tau = 10$, $N = 30$ and $p_{\mathrm{a}} = 0.5$.

Fig. 7 shows the spectral efficiency vs. UE number N with different τ. We can see that the spectral efficiency first increases and then decreases as N grows. This is because when N is small, there are sufficient orthogonal pilots for all UEs, thus the mean number of non-colliding UEs $\bar{N}_{\mathrm{non}}$ increases as N grows; while when N exceeds a critical point, more and more UEs need to compete for the limited pilots, thus $\bar{N}_{\mathrm{non}}$ reduces as N grows. The spectral efficiency is monotonously increasing with $\bar{N}_{\mathrm{non}}$. Therefore, the spectral efficiency first goes up and then decays as N goes up. The critical point of N is closely related to τ, and we can observe that bigger τ gives a bigger critical point. Moreover, we can also find that the hybrid-grant RA can significantly improve the spectral efficiency compared with the grant-free RA. To demonstrate this improvement more precisely, we define the relative gain. Its expression with MRC receivers is defined as

$$\Delta^{\mathrm{MRC}} = \frac{\bar{S}^{\mathrm{MRC}} - \bar{S}^{\mathrm{f,MRC}}}{\bar{S}^{\mathrm{f,MRC}}}, \tag{33}$$

and that with ZF receivers is defined in the same way.

Fig. 8 shows the relative gain vs. UE number N. We can see that the relative gain increases significantly as N grows, and the gain with ZF receivers is more remarkable. When N is very small, like 10, pilots are sufficient for all UEs and there is barely no colliding UEs. Hence, the advantage of

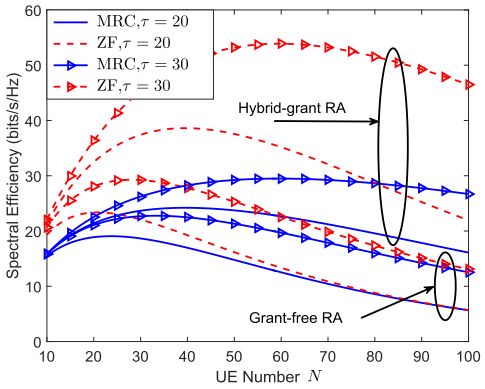


FIGURE 7. Spectral efficiency vs. UE number N, where $M = 100$ and $p_{\mathrm{a}} = 0.5$.

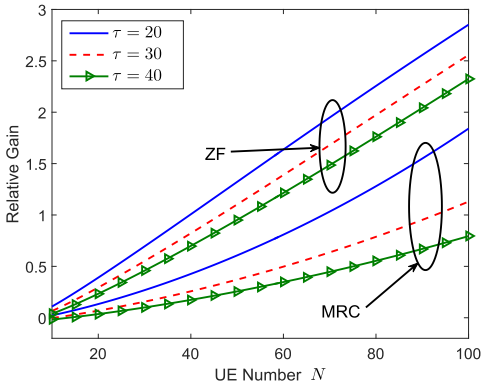


FIGURE 8. Relative gain vs. UE number N, where $M = 100$ and $p_{\mathrm{a}} = 0.5$.

the hybrid-grant RA which comes from blocking the data transmission of colliding UEs vanishes. Meanwhile, the extra time cost by the BS broadcasting makes the hybrid-grant RA even worse than the grant-free RA. Therefore, the relative gain for $N = 10$ is negative. However, as N grows, the superiority of the hybrid-grant RA begins to show out and grows rapidly. For $\tau = 20$, when $N = 40$, the hybrid-grant RA can improve the spectral efficiency %100 with ZF receivers, i.e., double the spectral efficiency, and improve the spectral efficiency %40 with MRC receivers. When N goes up to 80, the hybrid-grant RA can triple the spectral efficiency with ZF receivers and improve the spectral efficiency %120 with MRC receivers. There observations reveal that the proposed hybrid-grant RA is highly effective on boosting up the spectral efficiency, especially when N is large which is typical in future communications networks. From this figure, we also find that the relative gain reduces as τ increases. This is because more pilots yields less colliding UEs, so the gain of the hybrid-grant RA abates. After verifying the effectiveness of the proposed hybrid-grant RA scheme, we next analyze the performance of it more deeply.

Fig. 9 shows the spectral efficiency using hybrid-grant RA vs. pilot length τ. We can see that the spectral efficiency first increases and then reduce as τ grows. This is because at the beginning of increasing τ, more pilots can yield more non-colliding UEs and further promote the spectral

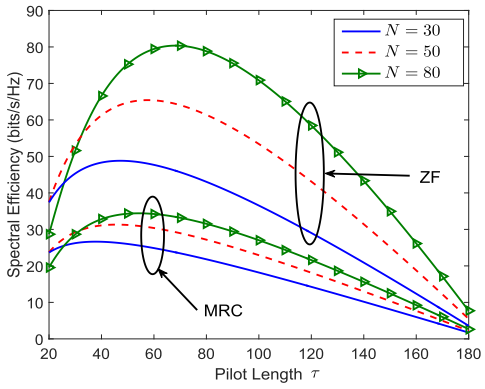


FIGURE 9. Spectral efficiency using hybrid-grant RA vs. pilot length τ, where $M = 100$ and $p_{\mathrm{a}} = 0.5$.

efficiency. When pilots go up to be sufficient for all UEs, increasing τ cannot give more non-colliding UEs but retrench the time used for data transmission. Therefore, the spectral efficiency has a decline when τ exceeds the critical value. The critical value is the optimal τ that maximizes the spectral efficiency.

Fig. 10 shows the spectral efficiency using hybrid-grant RA vs. UE activation probability p_{a}. We can observe that the increment of the spectral efficiency is followed by a decline. This is because when p_{a} is small, pilots are enough for all the UEs that try to access. As p_{a} grows, the pilot resources become lacking. Hence, if p_{a} continues to grow, more and more UEs compete for deficient pilots, which results in the decrease of the number of non-colliding UEs and further reduce the spectral efficiency. Therefore, there exists an optimal p_{a} that maximize the spectral efficiency. The optimal τ and p_{a} will be investigated in the next subsection.

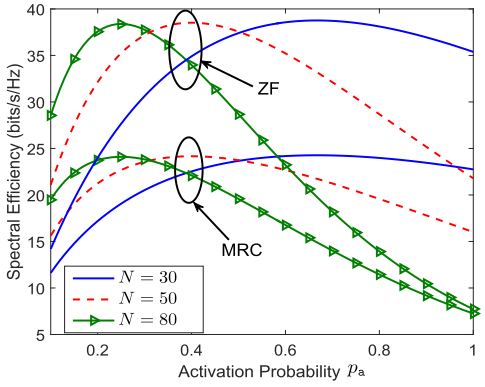
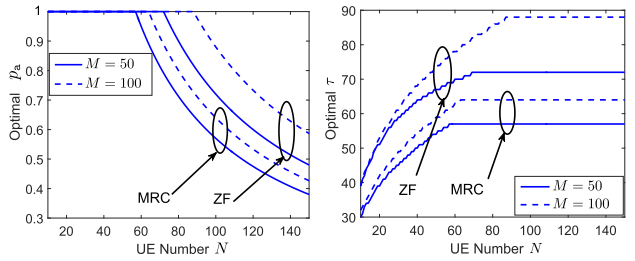


FIGURE 10. Spectral efficiency using hybrid-grant RA vs. activation probability p_{a}, where $M = 100$ and $\tau = 20$.

B. OPTIMAL PARAMETERS

In this subsection, we investigate the optimal τ and p_{a} got from Theorem 3 and 4. Assume that $p_{\mathrm{a}}^{\max} = 1$.

Fig. 11 shows the optimal τ and p_{a} vs. UE number N. We can see that when N is small, the optimal $p_{\mathrm{a}} = p_{\mathrm{a}}^{\max}$, and when N is bigger than a threshold, the optimal p_{a}



(a) optimal p_a (b) optimal τ

FIGURE 11. Optimal parameters vs. UE number *N*.

begins to decrease as N grows. Meanwhile, the optimal τ keeps increasing as N grows until reaches the peak value and remains constant afterwards. This is because increasing pilot can benefit the spectral efficiency until it satisfies the access requirements of all UEs, and after that, increasing pilots instead impair the spectral efficiency since the time used for the data transmission is retrenched. Therefore, as N grows, more pilots are needed to fullfill their access requirements, but pilots cannot increase infinitely. The platform in Fig. 11(b) indicates that when τ reaches a peak value, the impairment brought by large τ is remarkable. In the meantime, too many UEs that participate in the pilot selection will cut down the number of non-colliding UEs, so the active UEs should decreases as N grows which is controlled by p_a. τ and p_a interact with each other. The results in Fig. 11 find the balance between them which gives the maximum spectral efficiency. Moreover, we also note that the optimal τ and p_a with bigger M and ZF receivers are larger than that with smaller M and MRC receivers, respectively. This is because the larger antenna array and ZF receivers are both more capable to eliminate the interferences among UEs. Hence, they can use more pilots to permit more UEs into the transmission for better transmission performance.

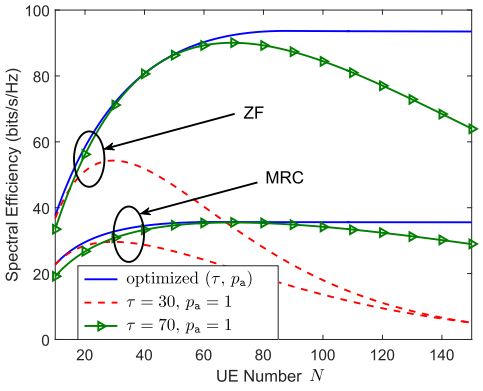


FIGURE 12. Spectral efficiency with optimized and non-optimized parameters and where $M = 100$.

Fig. 12 compares the spectral efficiency with optimized and non-optimized parameters. The optimized parameters is got from Theorem 3 and 4, and p_a in the non-optimized parameters is set equal to p_a^{max}, i.e., the original activation probability without backoff. We can see that compared with the non-optimized parameters, the optimized parameters can lead to a remarkable gain on the spectral efficiency, especially in contrast with the spectral efficiency using a small τ. In particular, as the UE number grows, the spectral efficiency using non-optimized parameters declines after it reaches the peak value, while that using optimized parameters keeps constant after the peak value. Therefore, the gain got by optimized parameters becomes more significant for larger N.

VI. CONCLUSION

In this paper, we proposed a new RA scheme which inserts a BS broadcasting message including the IDs of non-colliding UEs into the grant-free RA, with which UEs could determine whether they had colliders. Only non-colliding UEs could transmit data in the following step while colliding UEs keep silent. By doing this, the data interference from colliding UEs could be eliminated without costing much extra time. Since the BS broadcasting message was also used in the grant-based RA, the new RA scheme could be regarded as a combination of grant-free and grant-based RA. Hence, we called it as hybrid-grant RA. We investigated the hybrid-RA in massive MIMO systems. A tight closed-form approximation of the spectral efficiency with MRC and ZF receivers are obtianed, respectively. By comparing with the grant-free RA, we found that our proposed hybrid-grant RA could improve the spectral efficiency significantly, especially for ZF receivers. This improvement grew rapidly as the UE number increased, which meant that the hybrid-grant RA was more suitable for the future network with large amount of UEs. Moreover, we also gave the optimal pilot length and UE activation probability that maximize the spectral efficiency using hybrid-grant RA which could be used as references for practical applications of the hybrid-grant RA.

APPENDIX A
PROOF OF THEOREM 1

Applying the approximation in [21, Lemma 1] into (9), we can get that (34), as shown at the top of the next page, where $\mathbb{E}_s$ denote the average over all possible UE activations and pilot selections, and $\mathbb{E}_h$ denote the average over the small-scale fading. Then, with some basic algebraic operations, we can get that

$$\begin{aligned}\bar{S}^{\mathrm{MRC}} &\approx (1-\frac{\tau+\mu}{T})\mathbb{E}_s\left\{N_{\mathrm{non}}\log_2\left(1+\frac{(M+1)\eta\lambda}{N_{\mathrm{non}}\lambda-\eta\lambda+1}\right)\right\}\\ &\overset{(a)}{\leq}(1-\frac{\tau+\mu}{T})\mathbb{E}_s\{N_{\mathrm{non}}\}\log_2\left(1+\frac{(M+1)\eta\lambda}{\lambda\mathbb{E}_s\{N_{\mathrm{non}}\}-\eta\lambda+1}\right),\end{aligned} \quad (35)$$

where (a) is from the Jensen's inequality. According to the law of total expectation, we know that

$$\mathbb{E}_s\{N_{\mathrm{non}}\}=\sum_{n=1}^{N}\mathbb{P}[N_a=n]\,\mathbb{E}_s\{N_{\mathrm{non}}|N_a=n\}, \quad (36)$$

where $\mathbb{P}[x]$ is the probability of x. Conditioned on $N_a = n$, the probability that a UE does not collide with other UEs is

$$p_{\mathrm{non}}=(1-\frac{1}{\tau})^{n-1} \quad (37)$$

$$\bar{S}^{\mathrm{MRC}} \approx (1-\frac{\tau+\mu}{T})\mathbb{E}_{\mathrm{s}}\left\{\sum_{n\in F_{\mathrm{non}}}\log_2\left(1+\frac{P_n\mathbb{E}_{\mathrm{h}}\left\{\left|\hat{\mathbf{g}}_n^H\hat{\mathbf{g}}_n\right|^2\right\}}{\sum\limits_{i\in F_{\mathrm{non}}^{!n}}P_i\mathbb{E}_{\mathrm{h}}\left\{\left|\hat{\mathbf{g}}_n^H\hat{\mathbf{g}}_i\right|^2\right\}+\sum\limits_{i\in F_{\mathrm{non}}}P_i\mathbb{E}_{\mathrm{h}}\left\{\left|\hat{\mathbf{g}}_n^H\tilde{\mathbf{g}}_i\right|^2\right\}+\mathbb{E}_{\mathrm{h}}\left\{\left\|\hat{\mathbf{g}}_n\right\|^2\right\}}\right)\right\}, \tag{34}$$

Therefore, the conditioned mean number of non-colliding UEs is np_{non}. Moreover, since N_{a} has binomial distribution, we know that

$$\mathbb{P}\left[N_{\mathrm{a}}=n\right]=\binom{N}{n}p_{\mathrm{a}}^n(1-p_{\mathrm{a}})^{N-n}. \tag{38}$$

Applying these results into (36) gives that $\mathbb{E}_{\mathrm{s}}\left\{N_{\mathrm{non}}\right\}=\bar{N}_{\mathrm{non}}$ in (13). Then, the desired result in Theorem 1 follows by substituting (13) into (35).

APPENDIX B
PROOF OF THEOREM 2

We can write (17) as

$$R_n^{\mathrm{ZF}}=\mathbb{E}_{\mathrm{h}}\left\{\log_2\left(1+\frac{P_n}{\|\mathbf{a}_{n'}\|^2\left[P_i\sum\limits_{i\in F_{\mathrm{non}}}\beta_i(1-\eta_i)+1\right]}\right)\right\}$$
$$\geq\log_2\left(1+\frac{P_n}{\mathbb{E}_{\mathrm{h}}\left\{\|\mathbf{a}_{n'}\|^2\right\}\left[P_i\sum\limits_{i\in F_{\mathrm{non}}}\beta_i(1-\eta_i)+1\right]}\right) \tag{39}$$

Let $\zeta=1/\left[\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\right]_{n'n'}$. Then, since the covariance matrix of every row of $\hat{\mathbf{G}}$ is $\mathbf{V}=\mathtt{diag}\left[\eta\beta_1,\ldots,\eta\beta_n,\ldots\right]$, ζ is chi-squared distributed with probability density [22]

$$f(\zeta)=\frac{e^{-\zeta/\eta\beta_n}}{\eta\beta_n\Gamma(M-N_{\mathrm{non}}+1)}\left(\frac{\zeta}{\eta\beta_n}\right)^{M-N_{\mathrm{non}}},\quad \zeta\geq 0, \tag{40}$$

Therefore,

$$\mathbb{E}_{\mathrm{h}}\left\{\frac{1}{\zeta}\right\}=\frac{1}{\eta\beta_n(M-N_{\mathrm{non}})}. \tag{41}$$

Then, with

$$\|\mathbf{a}_{n'}\|^2=\left[\mathbf{A}^H\mathbf{A}\right]_{n'n'}=\left[\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\right]_{n'n'}, \tag{42}$$

we can get that $\mathbb{E}_{\mathrm{h}}\left\{\|\mathbf{a}_{n'}\|^2\right\}=1/\eta\beta_n(M-N_{\mathrm{non}})$. Plugging it into (39) and (18), we can get that

$$\bar{S}^{\mathrm{ZF}}\geq(1-\frac{\tau+\mu}{T})\mathbb{E}_{\mathrm{s}}\left\{N_{\mathrm{non}}\log_2\left(1+\frac{(M-N_{\mathrm{non}})\eta\lambda}{N_{\mathrm{non}}\lambda(1-\eta)+1}\right)\right\}$$
$$\overset{(a)}{\leq}(1-\frac{\tau+\mu}{T})$$
$$\times\mathbb{E}_{\mathrm{s}}\{N_{\mathrm{non}}\}\log_2\left(1+\frac{(M-\mathbb{E}_{\mathrm{s}}\left\{N_{\mathrm{non}}\right\})\eta\lambda}{\lambda(1-\eta)\mathbb{E}_{\mathrm{s}}\{N_{\mathrm{non}}\}+1}\right), \tag{43}$$

where (a) is from the Jensen's inequality. Then, the desired result in Theorem 2 can be got by substituting (13) into (43).

REFERENCES

[1] F. Rusek, D. Persson, B. Kiong Lau, E. G. Larsson, T. L. Marzetta, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[5] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[6] Q. Zhang, H. H. Yang, T. Q. S. Quek, and J. Lee, "Heterogeneous cellular networks with LoS and NLoS Transmissions—The role of massive MIMO and small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7996–8010, Dec. 2017.

[7] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[8] L. Atzori, A. Iera, and G. Morabito, "From, 'smart objects' to 'social objects': The next evolutionary step of the Internet of Things," *IEEE Commun. Mag.*, vol. 52, no. 1, pp. 97–105, Jan. 2014.

[9] C. S. Bontu, S. Periyalwar, and M. Pecen, "Wireless wide-area networks for Internet of Things: An air interface protocol for IoT and a simultaneous access channel for uplink IoT communication," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 54–63, Mar. 2014.

[10] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE–A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, 2nd Quart., 2015.

[11] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[12] *Medium Access Control (MAC) Protocol Specification*, Standard 3GPP TS 36.321 V13.3.0, 2011.

[13] O. Y. Bursalioglu, C. Wang, H. Papadopoulos, and G. Caire, "RRH based massive MIMO with, 'on fly' pilot contamination control," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.

[14] H. Han, X. Guo, and Y. Li, "A high throughput pilot allocation for M2M communication in crowded massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9572–9576, Oct. 2017.

[15] M. Wang, W. Yang, J. Zou, B. Ren, M. Hua, J. Zhang, and X. You, "Cellular machine-type comm unications: Physical challenges and solutions," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 126–135, Apr. 2016.

[16] K.-C. Chen and S.-Y. Lien, "Machine-to-machine communications: Technologies and challenges," *Ad Hoc Netw.*, vol. 18, pp. 3–23, Jul. 2014.

[17] A. T. Abebe and C. G. Kang, "Comprehensive grant-free random access for massive & low latency communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[18] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[19] J. Ding, D. Qu, H. Jiang, and T. Jiang, "Success probability of grant-free random access with massive MIMO," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 506–516, Feb. 2019.
[20] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statistician*, vol. 58, no. 1, pp. 30–37, Feb. 2004.
[21] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive MIMO systems with arbitrary-rank channel means," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 966–981, Oct. 2014.
[22] D. A. Gore, R. W. Heath, and A. J. Paulraj, "Transmit selection in spatial multiplexing systems," *IEEE Commun. Lett.*, vol. 6, no. 11, pp. 491–493, Nov. 2002.

QI ZHANG (Member, IEEE) received the B.S. and Ph.D. degrees in electrical and information engineering from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2010 and 2015, respectively. She was a Postdoctoral Research Fellow with the Singapore University of Technology and Design, from 2015 to 2017. She is currently with the Faculty of the Jiangsu Key Laboratory of Wireless Communications, Nanjing University of Posts and Telecommunications. Her research interests include massive MIMO systems, space-time wireless communications, heterogeneous cellular networks, and the Internet of Things.

SHI JIN (Senior Member, IEEE) received the B.S. degree in communications engineering from the Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with the Faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory. He serves as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE COMMUNICATIONS LETTERS, and *IET Communications*. He and his coauthors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and the 2010 Young Author Best Paper Award by the IEEE Signal Processing Society.

HONGBO ZHU received the bachelor's degree in telecommunications engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1982, and the Ph.D. degree in information and communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1996. He is currently working as a Professor with the Nanjing University of Posts and Telecommunications. He is also the Head of the Coordination Innovative Center of IoT Technology and Application, Jiangsu, which is the first governmental authorized Coordination Innovative Center of IoT in China. He also serves as a referee or an expert in multiple national organizations and committees. He has published more than 200 articles on information and communication area, such as the IEEE TRANSACTIONS. He is also leading a big group and multiple funds on IoT and wireless communications with current focus on architecture and enabling technologies for the Internet of Things.