

QoS-Aware Base-Station Selections for Distributed MIMO Links in Broadband Wireless Networks

Qinghe Du, *Student Member, IEEE*, and Xi Zhang, *Senior Member, IEEE*

Abstract—The distributed multiple-input-multiple-output (MIMO) techniques across multiple cooperative base stations (BS) can significantly enhance the capability of the broadband wireless networks in terms of quality-of-service (QoS) provisioning for wireless data transmissions. However, the computational complexity and the interfering range of the distributed MIMO systems also increase rapidly as the number of cooperative BS's increases. In this paper, we propose the QoS-aware BS-selection schemes for the *distributed wireless MIMO links*, which aim at minimizing the BS usages and reducing the interfering range, while satisfying diverse statistical delay-QoS constraints characterized by the delay-bound violation probability and the effective capacity technique. In particular, based on the channel state information (CSI) and QoS requirements, a subset of BS with variable cardinality for the distributed MIMO transmission is dynamically selected, where the selections are controlled by a central server. For the *single-user* scenario, we develop two optimization frameworks, respectively, to derive the efficient BS-selection schemes and the corresponding resource allocation algorithms. One framework uses the incremental BS-selection and time-sharing (IBS-TS) strategies, and the other employs the ordered-gain based BS-selection and probabilistic transmissions (OGBS-PT). The IBS-TS framework can yield better performance, while the scheme developed under the OGBS-PT framework is easier to implement. For the *multi-user* scenario, we propose the optimization framework applying the priority BS-selection, block-diagonalization precoding, and probabilistic transmission (PBS-BD-PT) techniques. We also propose the optimization framework applying the priority BS-selection, time-division-multiple-access, and probabilistic transmission (PBS-TDMA-PT) techniques. We derive the optimal transmission schemes for all the aforementioned frameworks, respectively. Also conducted is a set of simulation evaluations which compare our proposed schemes with several baseline schemes and show the impact of the delay-QoS requirements, transmit power, and traffic loads on the performances of BS selections for distributed MIMO systems.

Index Terms—Distributed MIMO, broadband wireless networks, statistical QoS provisioning, wireless fading channels.

I. INTRODUCTION

IN ORDER to increase the coverage of broadband wireless networks, distributed multiple-input-multiple-output (MIMO) techniques, where multiple location-independent base stations (BS) cooperatively transmit data to mobile users, have attracted more and more research attentions [1]–[4]. In

particular, the distributed MIMO techniques can effectively organize multiple location-independent BS's to form the distributed MIMO links connecting with mobile users, while not requiring too many multi-antennas, which are expensive, equipped at individual BS's. Like the conventional centralized MIMO system [5]–[7], the distributed MIMO system can significantly enhance the capability of the broadband wireless networks in terms of the quality-of-service (QoS) provisioning for wireless transmissions as compared to the single antenna system. However, the distributed nature for cooperative multi-BS transmissions also imposes many new challenges in wide-band wireless communications, which are not encountered in the centralized MIMO systems.

First, the cooperative distributed transmissions cause the severe difficulty for synchronization among multiple location-independent BS transmitters. Second, as the number of cooperative BS's increases, the computational complexity for MIMO signal processing and coding also grow rapidly. Third, because the coordinated BS's are located at different geographical positions, the cooperative communications in fact enlarge the interfering areas for the used spectrum, thus drastically degrading the frequency-reuse efficiency in the spatial domain. Finally, many wide-band transmissions are sensitive to the delay, and thus we need to design QoS-aware distributed MIMO techniques, such that the scarce wireless resources can be more efficiently utilized.

Towards the above issues, many research works on distributed MIMO transmissions have been proposed recently. The feasibility of transmit beamforming with efficient synchronization techniques over distributed MIMO link has been demonstrated through experimental tests and theoretical analyses [2], [3], suggesting that complicated MIMO signal processing techniques are promising to implement in realistic systems. While the antenna selection [6], [7] is an effective approach to reduce the complexity for centralized MIMO systems, which can be also extended to distributed MIMO systems for the BS selection. It is clear that the BS-selection techniques can significantly decrease the processing complexity, while still achieving high throughput gain over the single BS transmission. Also, it is desirable to minimize the number of selected BS's through BS-selection techniques, which can effectively decrease the interfering range and thus improve the frequency-reuse efficiency of the entire wireless network. Most previous research works for BS/antenna selections mainly focused on the scenarios of selecting a subset of BS's/antennas with the fixed cardinality [4], [6], [7]. However, it is evident that based on the wireless-channel status, BS-subset selections with dynamically adjusted cardinality can further decrease the BS

Manuscript received 1 June 2010; revised 1 December 2010. The research reported in this paper was supported in part by the U.S. National Science Foundation CAREER Award under Grant ECS-0348694.

The authors are with the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mails: duqinghe@tamu.edu; xizhang@ece.tamu.edu).

Digital Object Identifier 10.1109/JSAC.2011.110602.

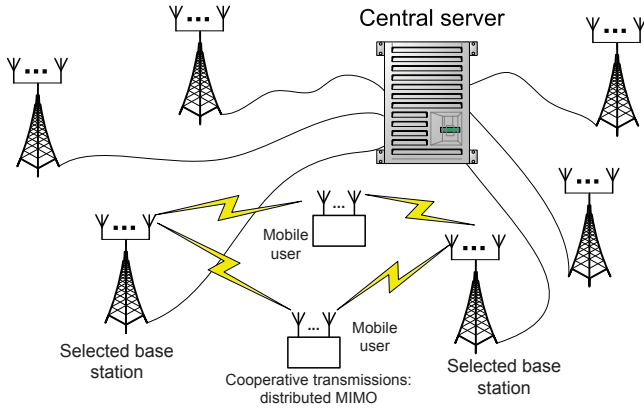


Fig. 1. System model of a wireless distributed MIMO system for downlink transmissions.

usage. More importantly, how to efficiently support diverse delay-QoS requirements through BS-selection in distributed MIMO systems still remains as a widely cited open problem.

To overcome the aforementioned problems, we propose the QoS-aware BS-selection schemes for the distributed wireless MIMO links, which aim at minimizing the BS usages and reducing the interfering range, while satisfying diverse statistical delay-QoS requirements. In particular, based on the channel state information (CSI) and QoS requirements, the subset of BS with variable cardinality for the distributed MIMO transmission is dynamically selected, where the selections are controlled by a central server. For the *single-user* scenario, we consider the optimization framework which uses the incremental BS-selection and time-sharing (IBS-TS) strategies, and study another framework which employs the ordered-gain based BS-selection and probabilistic-transmissions (OGBS-PT) techniques. For the *multi-user* scenario, we propose the optimization framework applying the priority BS-selection, block-diagonalization multiple-access, and probabilistic transmission (PBS-BD-PT) techniques. We also propose the optimization framework applying the priority BS-selection, time-division-multiple-access, and probabilistic transmission (PBS-TDMA-PT) techniques. We derive the optimal transmission schemes for the above frameworks, respectively, and conduct comparative analyses with the baseline schemes through simulations.

The rest of this paper is organized as follows. Section II describes the system model for distributed MIMO transmissions. Section III introduces the statistical QoS guarantees and the concept of effective capacity. Section IV proposes the optimization frameworks for QoS-aware BS selections of the single-user case and develops their corresponding optimal solutions. Section V develops the optimization frameworks for multi-user cases and derives their optimal solutions. Section VI simulates our proposed schemes. The paper concludes with Section VII.

II. SYSTEM MODEL

A. System Architecture

We concentrate on the wireless *distributed MIMO* system for downlink transmissions depicted in Fig. 1, which consists

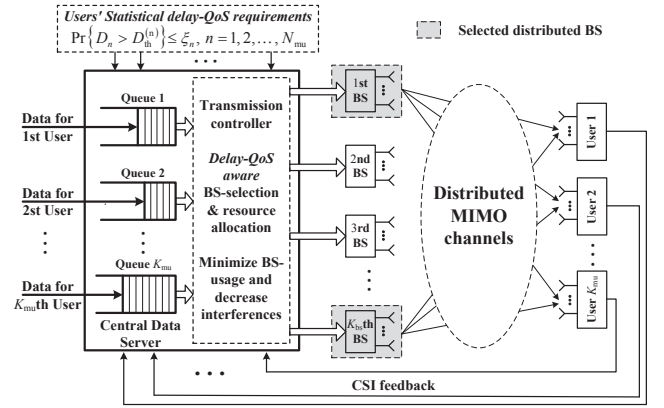


Fig. 2. Our proposed QoS-aware BS-selection framework.

of K_{bs} distributed BS's, K_{mu} mobile users, and one central server. The m th BS has M_m transmit antennas for $m = 1, 2, \dots, K_{bs}$ and the n th mobile user has N_n receive antennas for $n = 1, 2, \dots, K_{mu}$. All distributed BS's are connected to the central server through high-speed optical connections. The data to be delivered to the n th mobile user, $n = 1, 2, \dots, K_{mu}$, arrives at the central server with a constant rate, which is denoted by \bar{C}_n . Then, the central server dynamically controls these distributed BS's to cooperatively transmit data to the mobile users under the specified delay-QoS requirements.

For the case of $K_{mu} = 1$, the distributed BS's and the mobile user form a single wireless MIMO link; when $K_{mu} \geq 2$, the distributed BS's and the mobile users form the broadcast MIMO link for data transmissions. The wireless fading channels between the m th BS and the n th mobile user is modeled by an $N_n \times M_m$ matrix $\mathbf{H}_{n,m}$. The element at the i th row and j th column of $\mathbf{H}_{n,m}$, denoted by $(\mathbf{H}_{n,m})_{i,j}$, is the complex channel gain between the i th receive antenna of n th mobile user and the j th transmit antenna of the m th BS. All elements of $\mathbf{H}_{n,m}$ are independent and circularly symmetric complex Gaussian random variables with zero mean and the variance equal to $\bar{h}_{n,m}$. Also, the instantaneous aggregate power gain of the MIMO link between the n th mobile user and the m th BS, denoted by $\gamma_{n,m}$, is defined by

$$\gamma_{n,m} \triangleq \frac{1}{M_m} \sum_{i=1}^{N_n} \sum_{j=1}^{M_m} |(\mathbf{H}_{n,m})_{i,j}|^2. \quad (1)$$

We define $\mathbf{H}_n \triangleq [\mathbf{H}_{n,1} \ \mathbf{H}_{n,2} \ \dots \ \mathbf{H}_{n,K_{bs}}]$ as the CSI for the n th mobile user for $n = 1, 2, \dots, K_{mu}$. The matrix \mathbf{H}_n follows the independent block-fading model, where \mathbf{H}_n does not change within a time period with the fixed length T , called a time frame, but varies independently from one frame to the other frame. Furthermore, we define $\mathbf{H} \triangleq [\mathbf{H}_1^T \ \mathbf{H}_2^T \ \dots \ \mathbf{H}_{K_{bs}}^T]^T$, representing a fading state of the entire distributed MIMO system, where the superscript τ denotes the transpose operation on a matrix or a vector.

Under the aforementioned model, our proposed BS-selection framework is illustrated in Fig. 2. As shown in Fig. 2, based on CSI feedback from the users, the central data server will dynamically select a subset of BS's and then use all transmit antennas of the selected BS's to construct the distributed channels to transmit data to the K_{mu} users.

Our BS-selection strategies and the corresponding resource allocation algorithms depends not only on the CSI, but also the *statistical delay-QoS constraints* (to be detailed on Section II-B and Section III) for the incoming traffics. Although the BS/antenna selection techniques for distributed/centralized MIMO systems have been extensively studied, most existing works [4], [6], [7] focused on maximizing the capacity or minimizing the error rate given the specified BS/antenna subset cardinality. In contrast, our work in this paper aims at tackling the following new challenges for distributed MIMO system. 1). Our BS-selection and the associated resource allocation algorithms need to guarantee the specified delay-QoS requirements for the incoming traffics. 2). The cardinality of selected BS-subset vary with CSI and delay QoS, which is more flexible and efficient than the selection schemes with the fixed subset cardinality. 3). We aim at using the minimum average number of BS's to support the incoming traffic with delay-QoS guarantees, which also decreases the interferences to the entire network.

B. The Delay QoS Requirements

The central data server maintains a queue for the incoming traffic to each mobile user. We mainly focus on the queueing delay in this paper because the wireless channel is the major bottleneck for high-rate wireless transmissions. Since it is usually unrealistic to guarantee the hard delay bound over the highly time-varying wireless channels, we employ the statistical metric, namely, the *delay-bound violation probability*, to characterize the diverse delay QoS requirements. Specifically, for the n th mobile user, the probability of violating a specified delay bound, denoted by $D_{\text{th}}^{(n)}$, cannot exceed a given threshold ξ_n . That is, the inequality

$$\Pr \left\{ D_n > D_{\text{th}}^{(n)} \right\} \leq \xi_n, \quad n = 1, 2, \dots, N_{\text{mu}}, \quad (2)$$

needs to hold, where D_n denotes the queueing delay in the n th mobile user's queueing system.

C. Performance Metrics and Design Objective

We denote by L the cardinality of the selected BS subset (the number of selected BS's) for the distributed MIMO transmission in a fading state. Then, we denote the expectation of L by \bar{L} and call \bar{L} the *average BS usage*. As mentioned in Section II-A, our major objective is to minimize \bar{L} through dynamic BS selection while guaranteeing the delay QoS constraint specified by Eq. (2). Besides the average BS usage, we also need to evaluate the *average interfering range* affected by the distributed MIMO transmission. The instantaneous interfering range, denoted by A , is defined as the area of the region where the average received power under the current MIMO transmission is larger than a certain threshold denoted by σ_{th}^2 . The average interfering area is then defined as the expectation $\mathbb{E}\{A\}$ over all fading states. Clearly, minimizing \bar{L} can not only reduce implementation complexity, but also decrease the average interfering range affected by the transmit power.

D. The Power Control Strategy

The transmit power of our distributed MIMO system varies with the number of selected BS's. In particular, given the number L of selected BS's, the total instantaneous transmitted power used for distributed MIMO transmissions is set as a constant equal to \mathcal{P}_L . Furthermore, \mathcal{P}_L linearly increases with L by using the strategy as follows:

$$\mathcal{P}_L = \mathcal{P}_{\text{ref}} + \kappa(L - 1), \quad L = 1, 2, \dots, K_{\text{bs}}, \quad (3)$$

where $\mathcal{P}_{\text{ref}} > 0$ is called the *reference power* and $\kappa \geq 0$ describes the power increasing rate against L . Also, we define $\mathcal{P}_L \triangleq 0$ for $L = 0$. The above power adaptation strategy is simple to implement, while the average transmit power can be effectively decreased through minimizing the average number of used BS's. In addition, Eq. (3) can upper-bound the instantaneous interferences and the interfering range over the entire network.

III. EFFECTIVE CAPACITY APPROACH FOR STATISTICAL DELAY-QoS GUARANTEES

In this paper, we apply the theory of statistical QoS Guarantees [8], [9], [13], [20] to integrate the constraint on delay-bound violation probability given by Eq. (2) into our BS selection design. Consider a stable dynamic discrete-time queueing system. The data arrival-rate and data service-rate of the queueing system are denoted by $C[k]$ and $R[k]$, respectively, where k is the index for the time frame with a fixed time-duration equal to T (as also described in Section II-A) and the units of $C[k]$ and $R[k]$ are both nats/frame. The $C[k]$ and $R[k]$ change from frame to frame and thus can be characterized as the time-varying processes. By using the asymptotic analyses based on the large deviation principal, the author of [8] showed that under the sufficient conditions, the queue-length process as a function of t , denoted by $Q[k]$, converges in distribution to a random variable Q satisfying the following equation:

$$-\lim_{Q \rightarrow \infty} \frac{\log(\Pr\{Q > Q_{\text{th}}\})}{Q_{\text{th}}} = \theta, \quad (4)$$

for a certain $\theta > 0$. A set of sufficient conditions given in [8] for Eq. (4) are summarized as follows. 1). Both $C[k]$ and $R[k]$ are stationary. 2). The Gartner-Ellis limits [8] for $\mathcal{C}[k]$ and $\mathcal{R}[k]$, denoted by $\Lambda_C(\theta)$ and $\Lambda_R(\theta)$, respectively, exist for all θ , where $\mathcal{C}[k] \triangleq \sum_{t=1}^k C[t]$ and $\mathcal{R}[k] \triangleq \sum_{t=1}^k R[t]$. 3). The processes $C[k]$ and $R[k]$ are independent. 4) There exists a certain $\theta > 0$ such that $\Lambda_C(\theta) = \Lambda_R(-\theta)$. For the details of the sufficient conditions for Eq. (4), please refer to [8]. It has been shown that Eq. (4) holds for many stable queueing systems with typical arrival/departure processes [8], [13], such as Markovian processes, auto-regressive processes, and capacity-achieving service processes over i.i.d. block-fading wireless channels.

Based on Eq. (4), the probability that the queue-length, denoted by Q , exceeding a given bound Q_{th} can be approximated [8] by

$$\Pr\{Q > Q_{\text{th}}\} \approx e^{-\theta Q_{\text{th}}}, \quad (5)$$

where $\theta > 0$ is a constant called QoS exponent. It is clear that the larger (smaller) θ implies the lower (higher) queue-length-bound violation probability. Furthermore, when the

delay bound becomes the main QoS metric of interests, the delay-bound violation probability can be approximated [9], [20] by

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta\varphi(\theta)D_{\text{th}}}, \quad (6)$$

where D and D_{th} denote the queueing delay and delay bound, respectively, and $\varphi(\theta)$ is known as the effective bandwidth [8] of the arrival-rate process under the given θ . When the arrival rate $C[k]$ is equal to a constant \bar{C} over all k and the departure rate $R[k]$ is time-varying, Eq. (6) can be written [9], [20], [21] as

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta\bar{C}D_{\text{th}}}. \quad (7)$$

Then, to upper-bound $\Pr\{D > D_{\text{th}}\}$ with a threshold ξ , using Eq. (7), we get the minimum required QoS exponent θ as follows:

$$\theta = -\frac{\log(\xi)}{\bar{C}D_{\text{th}}}. \quad (8)$$

Consider a discrete-time arrival process with constant rate \bar{C} and the discrete-time time-varying departure process $R[k]$, where k is the time index. In order to guarantee the desired θ determined by Eq. (8), the statistical QoS theory [8]–[10] shows that the *effective capacity* $\mathcal{C}(\theta)$ of the service-rate process $R[k]$ needs to satisfy

$$\mathcal{C}(\theta) \geq \bar{C}, \quad (9)$$

under the given QoS exponent θ . The *effective capacity* function is defined in [9] as the maximum constant arrival rate which can be supported by the service rate to guarantee the specified QoS exponent θ . If the service-rate sequence $R[k]$ is stationary and time uncorrelated, the effective capacity can be written [13] as

$$\mathcal{C}(\theta) \triangleq -\frac{1}{\theta} \log \left(\mathbb{E} \left\{ e^{-\theta R[k]} \right\} \right), \quad (10)$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation.

In our distributed MIMO system, the BS selection result is designed as the function determined by the current CSI. Thus, the corresponding transmission rate (service rate) is time independent under the independent block-fading model (see Section II-A). Then, applying Eqs. (8)–(9), the delay QoS constraints given by Eq. (2) can be equivalently converted to:

$$\mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_n R_n} - e^{-\theta_n \bar{C}_n} \right\} \leq 0, \quad n = 1, 2, \dots, N_{\text{mu}}, \quad (11)$$

where $\theta_n = -\log(\xi_n) / (\bar{C}_n D_{\text{th}}^{(n)})$ and $\mathbb{E}_{\mathbf{H}}\{\cdot\}$ denotes the expectation over all \mathbf{H} .

IV. QOS-AWARE BS SELECTION FOR THE SINGLE-USER CASE

We focus on the scenario with a single mobile user in this section, where $K_{\text{mu}} = 1$. For presentation convenience, we use the term *transmission mode* L to denote the case where the cardinality of the selected BS subset is equal to L . Given transmission mode L , we denote by Ω_L the set of indices of the selected BS's, where $\Omega_L = \{i_{L,1}, i_{L,2}, \dots, i_{L,L}\}$ and $i_{L,\ell} \in \{1, 2, \dots, K_{\text{bs}}\}$ for $\ell = 1, 2, \dots, L$. Once a BS is selected, we use all its transmit antennas for data

transmissions. Then, we characterize the channel matrix for the selected BS subset by \mathbf{H}_{Ω_L} and write \mathbf{H}_{Ω_L} as $\mathbf{H}_{\Omega_L} \triangleq [\mathbf{H}_{1,i_{L,1}}, \mathbf{H}_{1,i_{L,2}}, \dots, \mathbf{H}_{1,i_{L,L}}]$, which is an $N_1 \times \mathcal{M}_L$ matrix with $\mathcal{M}_L \triangleq \sum_{\ell=1}^L M_{i_{L,\ell}}$. Accordingly, the physical-layer signal transmission is characterized by

$$\mathbf{y} = \mathbf{H}_{\Omega_L} \mathbf{s}_{\Omega_L} + \boldsymbol{\varsigma},$$

where \mathbf{y} is the $N_1 \times 1$ received signal vector and $\boldsymbol{\varsigma}$ denotes the $N_1 \times 1$ additive Gaussian noise vector whose elements are independent with unit power. The variable $\mathbf{s}_{\Omega_L} \triangleq [\mathbf{s}_{i_{L,1}}^\tau, \mathbf{s}_{i_{L,2}}^\tau, \dots, \mathbf{s}_{i_{L,L}}^\tau]^\tau$ is the input signal vector (transmitted signal vector) for the MIMO channel \mathbf{H}_{Ω_L} , where $\mathbf{s}_{i_{L,\ell}}$ is the $M_{i_{L,\ell}} \times 1$ signal vector transmitted from the $(i_{L,\ell})$ -th BS.

Clearly, for dynamic BS selections of distributed MIMO transmissions, we need to answer the following three questions: (i) For a specified transmission mode L , how do we determine the BS subset Ω_L ? (ii) How are the wireless resources shared if applying multiple transmission modes within a time frame? (iii) Which transmission modes will be used and how to quantitatively allocate the wireless resources? We first study associated issues for question (i) in Section IV-A. Then, we introduce the time-sharing transmission and probabilistic transmission to share the resources across different transmission modes in Section IV-B. Following the discussions in Sections IV-A and IV-B, we formulate two analytical optimization frameworks to answer question (iii). In particular, the first optimization framework is based on the incremental BS-selection algorithm and the time-sharing strategy; the second framework applies the ordered-gain based BS selection algorithm with probabilistic transmission, which will be detailed in Sections IV-C and IV-D, respectively. The BS-selection scheme derived under the first optimization framework can achieve better performance, while the second optimization framework is simpler to implement.

A. BS-Selection Strategy Given the Cardinality of the BS Subset

In this paper, we focus on the spatial multiplexing based MIMO transmissions. Given Ω_L in a fading state, the maximum achievable data rate (Shannon capacity), denoted by $R(\Omega_L)$ (nats/frame), is determined [5] by

$$R(\Omega_L) = \max_{\boldsymbol{\Xi}: \text{Tr}(\boldsymbol{\Xi}) = \mathcal{P}_L} \left\{ BT \log \left[\det \left(\mathbf{I} + \mathbf{H}_{\Omega_L} \boldsymbol{\Xi} \mathbf{H}_{\Omega_L}^\dagger \right) \right] \right\} \quad (12)$$

for $L = 1, 2, \dots, K_{\text{bs}}$ under given \mathbf{H} , where $(\cdot)^\dagger$ represents the conjugate transpose, $\det(\cdot)$ generates the determinant of a matrix, $\text{Tr}(\cdot)$ evaluates the trace of a matrix, and $\boldsymbol{\Xi}$ is the covariance matrix of \mathbf{s}_{Ω_L} . Also, we define $R(\Omega_L) \triangleq 0$ for $L = 0$, implying that no BS is selected and thus no data is transmitted in this fading state. For the details on how to achieve MIMO capacity under the given power budget and how to allocate power across multiple transmit antennas, please refer to [5].

Given L , the best selection strategy to optimize the achievable rate is the solution to the following optimization problem:

$$\max_{\Omega_L} \{ R(\Omega_L) \}. \quad (13)$$

```

00. Let  $\Psi := \{1, 2, \dots, K_{\text{bs}}\}$  and  $\overline{\Psi} := \emptyset$ , and  $Z = |\Psi|$ , where  $\emptyset$ 
    is the empty set and  $|\Psi|$  denotes the cardinality of the set  $\Psi$ ;
    ! Use variables  $\overline{\Psi}$  and  $\Psi$  to store all selected BS's and all other
    BS's, respectively.
01. For  $i := 1$  to  $L$  ! Add one BS to  $\overline{\Psi}$  in each step.
02. For  $z := 1$  to  $Z$  ! Examine  $Z$  BS's in  $\Psi$ , respectively
03.  $\Theta_z := \overline{\Psi} \cup \{\psi_z\}$ , where  $\psi_z$  is the  $z$ th element in  $\Psi$ ;
    ! Pick a BS of  $\Psi$  to form a new BS subset  $\Theta_z$  with  $\overline{\Psi}$ .
04.  $\tilde{R}_z := R(\Theta_z)$  based on Eq. (12) by setting  $\Omega_L := \Theta_z$ .
    ! Examine the achievable rate of  $\Theta_z$ .
05. End
06.  $z^* := \arg \max_{1 \leq z \leq Z} \{\tilde{R}_z\}$ ;
    ! Select the BS to maximize the achievable rate.
07.  $\overline{\Psi} := \Theta_{z^*}$ ,  $\Psi := \Psi \setminus \{\psi_{z^*}\}$ , and  $Z := |\Psi|$ ;
    ! Add the newly selected BS into the BS subset  $\overline{\Psi}$ .
08. End
09.  $\Omega_L := \overline{\Psi}$ . ! Complete the BS selection and get  $\Omega_L$ .

```

Fig. 3. The pseudo codes to determine Ω_L by using the incremental BS-selection algorithm for the single-user case.

To derive the optimal solution for this optimization problem, it is clear that we need to examine all $\binom{K_{\text{bs}}}{L}$ possible BS combinations, which leads to the prohibitively high computational complexity as K_{bs} gets large. Alternatively, we consider two suboptimal approaches with low complexities as follows.

1). *Incremental BS-Selection Algorithm*: In [7], the authors developed the fast antenna selection algorithm using the incremental-selection strategy. Although this incremental-selection strategy was developed for antenna selection without CSI feedback, it can be readily extended to the scenario for BS selection with CSI feedback to achieve the near optimal data rate. The pseudo codes of the incremental BS-selection algorithm are given in Fig. 3. In particular, the idea of this algorithm is to determine Ω_L through L steps, where in each step one BS is selected, as shown in lines 01-08 of Fig. 3. In each step, one selected BS is added to the BS subset denoted by $\overline{\Psi}$, where the selection criterion is to maximize the achievable rate of the updated BS subset $\overline{\Psi}$. Then, after L steps, we have totally added L BS's into $\overline{\Psi}$ and then assign $\Omega_L := \overline{\Psi}$. This algorithm only examines the achievable rates for $L(K_{\text{bs}} - (L - 1)/2)$ different BS combinations, which requires $L(K_{\text{bs}} - (L - 1)/2)$ times of singular value decomposition (SVD) to calculate the MIMO channel capacity, resulting in much less complexity than the optimal approach which examines all $\binom{K_{\text{bs}}}{L}$ BS combinations.

2). *Ordered-Gain Based BS-Selection Algorithm*: The ordered-gain (or ordered-SNR) based BS-selection algorithm selects L BS's with the largest aggregate power gain over all BS's, where the aggregate power gain is defined by Eq. (1). Since maximizing the aggregate power gain may not effectively optimize the achievable transmission rate for MIMO links, the incremental BS-selection algorithm usually dominates the ordered-gain based BS-selection algorithm. However, since the ordered-gain based BS-selection algorithm does not need to perform the SVD, its complexity is much lower than that of the incremental BS-selection algorithm.

B. Time Sharing and Probabilistic Transmissions

To get the more general framework for BS selection, we apply the time sharing and probabilistic transmission strate-

gies, respectively, over different transmission modes, which are described as follows.

1). *Time Sharing Transmissions*: Each time frame can be divided into $(K_{\text{bs}} + 1)$ time slots with the lengths equal to $\{T\alpha_m\}_{m=0}^{K_{\text{bs}}}$, where α_m is the normalized time-slot length and $\sum_{m=0}^{K_{\text{bs}}} \alpha_m = 1$. Within the m th time slot for $m > 1$, the transmission mode m will be used; for $m = 0$, no data is transmitted in the corresponding time slot. Then, the total service rate in a time frame is equal to $\sum_{L=0}^{K_{\text{bs}}} \alpha_L R(\Omega_L)$, where $R(\Omega_L)$ is given by Eq. (12) for $L \neq 0$ and $R(\Omega_L) = 0$ for $L = 0$. Furthermore, the total BS usage is given by $\sum_{L=0}^{K_{\text{bs}}} L\alpha_L$. The purpose of applying the time sharing based transmissions is to increase the system flexibility and to gain the continuous control on the BS usage within each time frame. Accordingly, we need to identify how to optimally adjust α with CSI and QoS constraints, where $\alpha \triangleq (\alpha_0, \alpha_1, \dots, \alpha_M)$.

2). *Probabilistic Transmissions*: Under this strategy, within each time frame only one transmission mode will be used for distributed MIMO transmissions. In particular, we will select transmission mode L with probability equal to ϕ_L and define $\phi \triangleq (\phi_0, \phi_1, \phi_2, \dots, \phi_{K_{\text{bs}}})^T$. Then, our target is to determine how to dynamically adjust ϕ according to the CSI and QoS requirements.

If setting $\phi = \alpha$ and using the same strategy to determine Ω_L over all fading states, we obtain the same BS usage. However, the effective capacities (see Eq. (10)) achieved under the time-sharing transmission and the probabilistic transmission, denoted by $\mathcal{C}_{\text{TS}}(\alpha, \theta_1)$ and $\mathcal{C}_{\text{PR}}(\phi, \theta_1)$, respectively, are different. Specifically, we derive

$$\begin{aligned}
 \mathcal{C}_{\text{TS}}(\alpha, \theta_1) &= -\frac{1}{\theta_1} \log \left(\mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \sum_{L=0}^{K_{\text{bs}}} \alpha_L R(\Omega_L)} \right\} \right) \\
 &\geq -\frac{1}{\theta_1} \log \left(\mathbb{E}_{\mathbf{H}} \left\{ \alpha_L e^{-\theta_1 \sum_{L=0}^{K_{\text{bs}}} R(\Omega_L)} \right\} \right) \\
 &= -\frac{1}{\theta_1} \log \left(\mathbb{E}_{\mathbf{H}} \left\{ \phi_L e^{-\theta_1 \sum_{L=0}^{K_{\text{bs}}} R(\Omega_L)} \right\} \right) \\
 &= \mathcal{C}_{\text{PR}}(\phi, \theta_1), \tag{14}
 \end{aligned}$$

where the inequality holds because $\mathbb{E}_{\mathbf{H}} \left\{ e^{-\sum_{L=0}^{K_{\text{bs}}} \alpha_L R(\Omega_L)} \right\}$ is convex over $(R(\Omega_0), R(\Omega_1), \dots, R(\Omega_{K_{\text{bs}}}))$. Eq. (14) suggests that the time-sharing transmission generally outperforms the probabilistic transmission. However, the probabilistic transmission is more realistic to implement than the time-sharing transmission due to the following reasons. On the one hand, for the optimized time-sharing transmission, the time-slot length $T\alpha_L$ may be quite small and thus very hard to implement. On the other hand, the multiple time slots (for the time-sharing transmission) within a time frame introduces more overhead than the single-slot case (for the probabilistic transmission).

C. Optimization Framework Using Time-Sharing Transmissions with Incremental BS Selection

As discussed in Section II-C, our major objective is to minimize the average BS usage. In this section, we focus on the framework which employs the incremental BS-selection algorithm for each transmission mode and apply the time-sharing transmission for different transmission modes. Then, we develop the efficient BS-selection scheme under the above

framework by solving the following optimization problem **A1**, which aims at minimizing the average BS usage while guaranteeing the delay-QoS requirement:

$$\begin{aligned} \mathbf{A1}: \min_{\alpha(\mathbf{H})} \{\bar{L}\} &= \min_{\alpha(\mathbf{H})} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\text{bs}}} \alpha_L(\mathbf{H})L \right\} \right\} \\ \text{s.t.}: 1). \sum_{L=0}^{K_{\text{bs}}} \alpha_L(\mathbf{H}) &= 1, \quad \forall \mathbf{H}; \\ 2). \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \sum_{L=0}^{K_{\text{bs}}} \alpha_L(\mathbf{H})R(\Omega_L)} - e^{-\theta_1 \bar{C}_1} \right\} &\leq 0, \end{aligned} \quad (15)$$

where Eq. (16) is the constraint to guarantee the delay-bound violation probability as derived in Eq. (2), we obtain Ω_L through the incremental selection algorithm listed in Fig. 2, which varies with \mathbf{H} , and we determine $R(\Omega_L)$ based on Eq. (12). Note that we denote the time-sharing vector (originally defined in Section IV-B) by $\alpha(\mathbf{H})$ in problem **A1** to emphasize that α is a function of \mathbf{H} . When the context is clear, we will then drop the symbol (\mathbf{H}) and only use α to simplify notations without causing confusions. We call the optimal solution to **A1**, denoted by α^* , as the *incremental BS-selection and time-sharing based* (IBS-TS) scheme.

Note that the optimization over α is an $(K_{\text{bs}} + 1)$ -dimensional problem. To reduce the dimension of optimization variables, for each \mathbf{H} we define

$$\begin{aligned} \tilde{R}(\mathcal{L}) &\triangleq \max_{\alpha} \left\{ \sum_{L=0}^{K_{\text{bs}}} \alpha_L R(\Omega_L) \right\}, \quad \text{for each } \mathbf{H} \quad (17) \\ \text{s.t.} \quad \sum_{L=0}^{K_{\text{bs}}} \alpha_L &= 1; \quad \sum_{L=0}^{K_{\text{bs}}} \alpha_L L = \mathcal{L}, \quad \text{for each } \mathbf{H}. \quad (18) \end{aligned}$$

Based on Eqs. (17)-(18), $\tilde{R}(\mathcal{L})$ is the maximum achievable rate over all α with the same BS usage. Since $\tilde{R}(\mathcal{L})$ is a convex combination [12] over $\{R(\Omega_L)\}_{L=1}^{K_{\text{bs}}}$, then the definition in Eqs. (17)-(18) suggests that $\tilde{R}(\mathcal{L})$ is a piece-wise linear and concave function, which can be written as

$$\tilde{R}(\mathcal{L}) = \begin{cases} \tilde{R}(m_{j-1}) + \nu_j (\mathcal{L} - m_{j-1}), & \text{if } \mathcal{L} \in (m_{j-1}, m_j], \\ & j=1, 2, \dots, \mathcal{K}; \\ 0, & \text{if } \mathcal{L} = 0; \end{cases} \quad (19)$$

for a certain integer \mathcal{K} , where $m_0 < m_1 < \dots < m_{\mathcal{K}}$, $m_0 = 0$, $m_{\mathcal{K}} = K_{\text{bs}}$, and $m_i \in \{0, 1, \dots, K_{\text{bs}}\}$. Moreover, using $(m_i, \tilde{R}(m_i))$ to represent the coordinates of a point in the two-dimensional plane, we can identify $\{(m_i, \tilde{R}(m_i))\}_{i=1}^{\mathcal{K}-1}$ through the following procedures: a). find vertices of the convex hull spanned by two-dimensional points $\{(L, \tilde{R}(L))\}_{i=0}^{K_{\text{bs}}}$; b). $\{(m_i, \tilde{R}(m_i))\}_{i=1}^{\mathcal{K}-1}$ are located above the line segment with end points $(0, 0)$ and $(K_{\text{bs}}, \tilde{R}(K_{\text{bs}}))$. Accordingly, ν_j is the slope of the line segment starting at the point $(m_i, \tilde{R}(m_i))$ and ending at the point $(m_{i-1}, \tilde{R}(m_{i-1}))$, which is determined by

$$\nu_i = \frac{\tilde{R}(m_i) - \tilde{R}(m_{i-1})}{m_i - m_{i-1}}, \quad i = 1, 2, \dots, \mathcal{K}. \quad (20)$$

For presentation convenience, we also define $\nu_0 \triangleq \infty$ and $\nu_{\mathcal{K}+1} \triangleq -\infty$. Note that $\tilde{R}(\mathcal{L})$ defined in Eq. (19) and the

associated variables, including Ω_L , \mathcal{K} , $\{m_j\}_{j=1}^{\mathcal{K}}$, and $\{\nu_i\}_{i=1}^{\mathcal{K}}$, are all functions of \mathbf{H} . Furthermore, given $\mathcal{L} \in [m_{j-1}, m_j]$, following the piece-wise linear property, we derive the corresponding α to achieve the service rate $\tilde{R}(\mathcal{L})$ as follows:

$$\alpha_L = \begin{cases} \frac{m_j - \mathcal{L}}{m_j - m_{j-1}}, & \text{if } L = m_{j-1}; \\ \frac{\mathcal{L} - m_{j-1}}{m_j - m_{j-1}}, & \text{if } L = m_j; \\ 0, & \text{otherwise.} \end{cases} \quad (21)$$

Applying Eqs. (17), (18), and (21) into problem **A1**, we can equivalently convert **A1** to the following problem **A1'**:

$$\begin{aligned} \mathbf{A1}': \min_{\mathcal{L}(\mathbf{H})} \{ \mathbb{E}_{\mathbf{H}} \{ \mathcal{L}(\mathbf{H}) \} \} \\ \text{s.t.}: \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \tilde{R}(\mathcal{L}(\mathbf{H}))} - e^{-\theta_1 \bar{C}_1} \right\} &\leq 0, \end{aligned} \quad (22)$$

where $\mathcal{L}(\mathbf{H})$ is a function of \mathbf{H} and we denote the optimal solution to problem **A1'** by $\mathcal{L}^*(\mathbf{H})$. To simplify notations, we will drop the symbol \mathbf{H} for problem **A1'**. After obtaining the optimal solution \mathcal{L}^* , we can then map \mathcal{L}^* to α through Eq. (21). Since $\tilde{R}(\mathcal{L})$ is an increasing and concave function, $e^{-\theta_1 \tilde{R}(\mathcal{L})}$ is convex over \mathcal{L} [14, pp. 84]. Thus, we can see that **A1'** satisfies: a) the objective function is convex; b) the inequality constraint function $\mathbb{E}_{\mathbf{H}} \{ e^{-\theta_1 \tilde{R}(\mathcal{L})} - e^{-\theta_1 \bar{C}_1} \}$ is convex. Therefore, **A1'** is a convex problem [14, pp. 137]. Then, using the Lagrangian method, we solve for the optimal solution of **A1'**, as summarized in the following Theorem 1.

Theorem 1: The optimal solution to **A1'**, if existing, is determined by

$$\mathcal{L}^* = \begin{cases} m_j, & \text{if } \nu_{j+1} \leq \frac{e^{\theta_1 \tilde{R}(m_j)}}{\theta_1 \lambda^*} \leq \nu_j; \\ \frac{\log(\theta_1 \lambda^* \nu_j)}{\nu_j \theta_1} - \frac{\tilde{R}(m_{j-1})}{\nu_j} + m_{j-1}, & \\ \text{if } \tilde{R}(m_{j-1}) < \frac{\log(\theta_1 \lambda^* \nu_j)}{\theta_1} < \tilde{R}(m_j), \end{cases} \quad (23)$$

where $\tilde{R}(\cdot)$ and ν_j are characterized by Eqs. (17) through (20). In Eq. (23), $\lambda^* \geq 0$ is a constant over all fading states, which needs to be selected such that equality in Eq. (22) holds.

Proof: The proof of Theorem 1 is provided in the Appendix A. ■

Remarks: (i) Having obtained the optimal \mathcal{L}^* for problem **A1'**, the optimal solution to **A1**, denoted by α^* , is obtained by setting $\mathcal{L} = \mathcal{L}^*$ in Eq. (21). (ii) Under the optimal solution, we do not allocate time slots for all transmission modes. As indicated by Eq. (21), within any time frames, we use at most two transmission modes. (iii) It is clear that by setting $\alpha_{K_{\text{bs}}} = 1$ for all time frames, we use the maximum transmit power and thus obtain the maximum achievable effective capacity, which is denoted by $\mathcal{C}_{\text{max}}^{(1)}$. If $\mathcal{C}_{\text{max}}^{(1)}$ is still smaller than \bar{C}_1 , the specified delay-QoS requirement cannot be satisfied since we have used up all power budget. As a result, no feasible solution exists for this case. In contrast, if $\mathcal{C}_{\text{max}}^{(1)} \geq \bar{C}_1$, we can always find the optimal solution. (iv) The analytical expressions for λ^* is usually intractable. Alternatively, we use the numerical-tracking method to derive the optimal λ^* . The details of the numerical-tracking method and the discussions on its complexity are provided in Appendix B.

D. Optimization Framework Using Probabilistic Transmissions with Ordered-Gain Based BS Selection

We in this section consider the framework using the ordered-gain based BS-selection algorithm and the probabilistic transmission strategy. Specifically, we formulate the optimization problem for this framework as follows:

$$\begin{aligned} \mathbf{A2} : \min_{\phi(\mathbf{H})} \{\bar{L}\} &= \min_{\phi(\mathbf{H})} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) L \right\} \right\} \\ \text{s.t.: 1). } & \sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) = 1, \quad \forall \mathbf{H}; \\ & 2). \mathbb{E}_{\mathbf{H}} \left\{ \left(\sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) e^{-\theta_1 R(\Omega_L)} \right) - e^{-\theta_1 \bar{C}_1} \right\} \leq 0, \end{aligned} \quad (24)$$

where the probabilistic transmission vector is denoted by a function $\phi(\mathbf{H})$ of \mathbf{H} . When the context is clear, we drop the symbol (\mathbf{H}) for problem **A2** to simplify notations. We call the optimal solution to the optimization problem **A2** as the ordered-gain and probability transmission based (OGBS-PT) scheme.

Theorem 2: The optimal solution to problem **A2**, denoted by ϕ^* , is given by

$$\phi_L^* = \begin{cases} 1, & \text{if } L = L^*; \\ 0, & \text{if } L \neq L^*, \end{cases} \quad (26)$$

for all \mathbf{H} , where

$$L^* = \arg \min_L \left\{ L + \lambda^* e^{-\theta_1 R(\Omega_L)} \right\}. \quad (27)$$

In Eq. (27), $\lambda^* \geq 0$ is a constant over all \mathbf{H} , which is selected such that the equality holds for Eq. (25) to guarantee the delay-QoS requirement.

Proof: The detailed proof of Theorem 2 is omitted due to lack of space, but is provided on-line in the technical-report version [19] of this paper. ■

Remarks: Theorem 2 suggests that under the optimal solution, the probabilistic transmission reduces to a deterministic strategy, where the only transmission mode L^* will be used for data transmission. Similar to problem **A1**, the parameter λ^* for **A2** also needs to be tracked through numerical searching.

V. QOS-AWARE BS SELECTION FOR THE MULTI-USER CASE

We next consider the distributed MIMO transmissions for the case with *multiple mobile users*.¹ For efficient BS selection and distributed MIMO transmissions, the central server controls the selected distributed BS's and the mobile users to constitute the broadcast MIMO link, as mentioned in Section II. Specifically, given the transmission mode² L and BS-index subset $\Omega_L = \{i_{L,1}, i_{L,2}, \dots, i_{L,L}\}$ of the selected BS's, the channel matrix of the n th mobile user, modeled by $\mathbf{H}_{\Omega_L}^{(n)}$, is determined by

$$\mathbf{H}_{\Omega_L}^{(n)} \triangleq [\mathbf{H}_{n,i_{L,1}} \quad \mathbf{H}_{n,i_{L,2}} \quad \dots \quad \mathbf{H}_{n,i_{L,L}}], \quad n = 1, 2, \dots, K_{\text{mu}}$$

¹We use the terms of "mobile user" and "user" interchangeably in the rest of this paper.

²For the multi-user case, we also use the term of transmission mode L to denote the case where the BS-subset's cardinality is L .

where $\mathbf{H}_{\Omega_L}^{(n)}$ is an $N_n \times \left(\sum_{\ell=1}^L M_{i_{L,\ell}} \right)$ matrix. Then, the physical-layer signal transmissions can be characterized by

$$\mathbf{y}_{\Omega_L}^{(n)} = \mathbf{H}_{\Omega_L}^{(n)} \sum_{i=1}^{K_{\text{mu}}} \mathbf{s}_{\Omega_L}^{(i)} + \boldsymbol{\zeta}^{(n)}, \quad n = 1, 2, \dots, K_{\text{mu}}$$

where $\mathbf{s}_{\Omega_L}^{(i)}$ represents the i th user's input signal vector for the MIMO channel $\mathbf{H}_{\Omega_L}^{(n)}$, $\mathbf{y}_{\Omega_L}^{(n)}$ is the signal vector received at the n th user's receive antennas, and $\boldsymbol{\zeta}^{(n)}$ is the complex additive white Gaussian noise (AWGN) vector with unit power for each element of this vector.

It is well-known that the optimal capacity of the broadcast MIMO link can be achieved through the dirty-paper coding technique [16], which is, however, hard to implement due to its high complexity [17]. Alternatively, we apply the block-diagonalization precoding techniques [17] for distributed MIMO transmissions and then concentrate on developing efficient QoS-aware BS-selection schemes and associated resource allocation schemes, which are elaborated on in the following sections.

A. The Block Diagonalization Technique for Distributed MIMO Transmissions

The idea of block diagonalization (BD) [17] is to use a precoding matrix, denoted by $\mathbf{\Gamma}_{\Omega_L}^{(n)}$, for the n th user's transmitted signal vector, such that $\mathbf{H}_{\Omega_L}^{(i)} \mathbf{\Gamma}_{\Omega_L}^{(n)} = \mathbf{0}$ for all $i \neq n$. By setting $\mathbf{s}_{\Omega_L}^{(n)} = \mathbf{\Gamma}_{\Omega_L}^{(n)} \hat{\mathbf{s}}_{\Omega_L}^{(n)}$, where $\hat{\mathbf{s}}_{\Omega_L}^{(n)}$ is the n th user's data vector to be precoded by $\mathbf{\Gamma}_{\Omega_L}^{(n)}$, we can rewrite the received signal $\mathbf{y}_{\Omega_L}^{(n)}$ as

$$\begin{aligned} \mathbf{y}^{(n)} &= \mathbf{H}_{\Omega_L}^{(n)} \sum_{i=1}^{K_{\text{mu}}} \mathbf{\Gamma}_{\Omega_L}^{(i)} \hat{\mathbf{s}}_{\Omega_L}^{(i)} + \boldsymbol{\zeta}^{(n)} = \mathbf{H}_{\Omega_L}^{(n)} \mathbf{\Gamma}_{\Omega_L}^{(n)} \hat{\mathbf{s}}_{\Omega_L}^{(n)} + \boldsymbol{\zeta}^{(n)} \\ &= \hat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \hat{\mathbf{s}}_{\Omega_L}^{(n)} + \boldsymbol{\zeta}^{(n)}, \end{aligned} \quad (28)$$

where $\hat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \triangleq \mathbf{H}_{\Omega_L}^{(n)} \mathbf{\Gamma}_{\Omega_L}^{(n)}$. Under this strategy, the n th user's signal will not cause interferences to other users. Accordingly, the MIMO broadcast transmissions are virtually converted to K_{mu} orthogonal MIMO channels with channel matrices $\{\hat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}\}_{n=1}^{K_{\text{mu}}}$. Then, we can get the n th user's maximum achievable rate in a fading state, denoted by $R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})$, as follows:

$$\begin{aligned} R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)}) &\triangleq \max_{\boldsymbol{\Xi}^{(n)}} \left\{ BT \log \left[\det \left(\mathbf{I} + \hat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \boldsymbol{\Xi}^{(n)} \left(\hat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \right)^\dagger \right) \right] \right\}, \\ \text{s.t.: } & \text{Tr} \left(\boldsymbol{\Xi}^{(n)} \right) = \mathcal{P}_L^{(n)}, \end{aligned} \quad (29)$$

where $\boldsymbol{\Xi}^{(n)}$ is the covariance matrix of $\hat{\mathbf{s}}_{\Omega_L}^{(n)}$ and $\mathcal{P}_L^{(n)}$ is the power allocated for the n th user. Also, we define $R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)}) \triangleq 0$ for $L = 0$. Note that the aforementioned $\mathbf{\Gamma}_{\Omega_L}^{(n)}$ may not exist, implying that the total number of transmit antennas over the selected BS's cannot support enough freedom for block-diagonalization. In this case, no data will be transmitted to the n th user to avoid interferences caused to other users. For the procedures on how to determine $\mathbf{\Gamma}_{\Omega_L}^{(n)}$, please refer to [17].

B. Priority BS-Selection Strategy Given BS Subset Cardinality

When the transmission mode is specified, i.e., the cardinality of the BS subset is given, every user expects to select the BS subset that maximizes its own transmission rate. However, it is clear that this objective cannot be obtained for all users in the multi-user case. Moreover, the derivation of global optimal selection strategy in terms of minimizing the average BS usage is intractable due to the too high complexity, where we need to examine all $\binom{K_{\text{bs}}}{L}$ possible BS combinations. Therefore, we propose a simple yet efficient BS-selection algorithm, called priority BS-selection (PBS), which is detailed as follows.

For the n th user, the global maximum achievable transmission rate is attained when all BS's are used and all the other users do not transmit. Thus, the maximum achievable rate is given by

$$R^{(n)}(\Omega_{K_{\text{bs}}}, \mathcal{P}_{K_{\text{bs}}}) = \max_{\Xi^{(n)}: \text{Tr}(\Xi^{(n)}) = \mathcal{P}_{K_{\text{bs}}}} \left\{ BT \log \left[\det \left(\mathbf{I} + \mathbf{H}_n \Xi^{(n)} \mathbf{H}_n^\dagger \right) \right] \right\}. \quad (30)$$

Then, we get the maximum achievable effective capacity of the n th user, denoted by $\mathcal{C}_{\text{max}}^{(n)}$, as follows:

$$\mathcal{C}_{\text{max}}^{(n)} = -\frac{1}{\theta_n} \log \left(\mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_n R^{(n)}(\Omega_{K_{\text{bs}}}, \mathcal{P}_{K_{\text{bs}}})} \right\} \right) \quad (31)$$

for $n = 1, 2, \dots, K_{\text{mu}}$. We further define the effective-capacity fraction for the n th user as the ratio between the traffic loads and the maximum achievable effective capacity. Denoting the effective-capacity fraction by \widehat{C}_n , we have $\widehat{C}_n \triangleq \overline{C}_n / \mathcal{C}_{\text{max}}^{(n)}$. Clearly, for a higher \widehat{C}_n , the n th user needs more wireless resources to meet its QoS requirements. Thus, in order to satisfy the QoS requirements for all users, we assign higher BS-selection priority to the user with larger \widehat{C}_n . Following this principle, we design the priority BS-selection algorithm to determine Ω_L in each fading state and provide the pseudo code in Fig. 4. For presentation convenience, we sort $\{\widehat{C}_n\}_{n=1}^{K_{\text{mu}}}$ in the decreasing order and denote the sorted version by $\{\widehat{C}_{\pi(j)}\}_{h=1}^{K_{\text{mu}}}$, where $\widehat{C}_{\pi(1)} \geq \widehat{C}_{\pi(2)} \geq \dots \geq \widehat{C}_{\pi(K_{\text{mu}})}$ indicates the order from the higher priority to the lower priority. In the rest of this paper, we use the term of user $\pi(i)$ to denote the user with the i th largest effective-capacity fraction.

As shown in Fig. 4, in each fading state the BS-selection procedure starts with the selection for user $\pi(1)$, who has the highest priority. After picking one BS for user $\pi(1)$, we select one different BS for user $\pi(2)$. More generally, after selecting for user $\pi(j)$, we choose one BS for user $\pi(j+1)$ from the BS-subset Ψ , which consists of the BS's that have not been selected. This procedure repeats until L BS's are selected. For user- $\pi(j)$'s selection, we choose the BS with the maximum aggregate power gain (see Eq. (1) for its definition) over the subset Ψ . In addition, after user- $\pi(K_{\text{mu}})$'s selection, if the number of selected BS's is still smaller than L , we continue selecting one more BS for user $\pi(1)$, as shown in line 06 in Fig. 4, and repeat this iterative selection procedure until having selected L BS's.

C. The Optimization Framework for BS-Selection and Resource Allocation

C.1. Problem Formulation for Average BS-Usage Minimization

```

01. Let  $\Psi := \{1, 2, \dots, K_{\text{bs}}\}$  and  $\overline{\Psi} := \emptyset$ , and  $\ell = |\overline{\Psi}|$ ;
    ! Use variables  $\overline{\Psi}$  and  $\Psi$  to store selected BS's and all other BS's,
    respectively.
02.  $j := 1$ . ! User  $\pi(j)$  is selecting BS.
03. While ( $\ell < L$ ) ! Iterative selections until  $L$  BS's are selected.
04.  $m^* = \arg \min_{m \in \Psi} \{\gamma_{\pi(j), m}\}$ .
    !  $\gamma_{\pi(j), m}$  is the aggregate power gain associated with user  $\pi(j)$ .
    ! Select BS to maximum the aggregate power gain for user  $\pi(j)$ .
05.  $\overline{\Psi} := \overline{\Psi} \cup \{m^*\}$ ,  $\Psi := \Psi \setminus \{m^*\}$ , and  $\ell := \ell + 1$ .
    ! Update  $\overline{\Psi}$ ,  $\Psi$ , and  $\ell$ .
06. If  $j = K_{\text{mu}}$ , then  $j := 1$ ; else  $j := j + 1$ .
    ! Let next user with lower priority to select BS.
07. End
08.  $\Omega_L := \overline{\Psi}$ . ! Complete the BS selection and get  $\Omega_L$ .

```

Fig. 4. The pseudo codes to determine Ω_L in each fading state by using the priority BS-selection algorithm for the multi-user case.

We next study how to determine which transmission modes will be used, and how to derive the corresponding resource allocation strategy by integrating the block diagonalization and the priority BS selection. Similar to the OGBS-PT scheme for the single-user case, we also apply the probabilistic transmission (PT) for the multi-user case, where transmission mode L is used with a probability denoted by ϕ_L , $L = 1, 2, \dots, K_{\text{bs}}$. Note that for each transmission mode, there are K_{mu} coexisting links towards K_{mu} mobile users. Consequently, we also need to determine how to allocate the total power \mathcal{P}_L to these K_{mu} coexisting links. In particular, we describe the power allocation strategy in a fading state by $\mathcal{P} \triangleq (\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_{K_{\text{bs}}})$ with $\mathcal{P}_L \triangleq (\mathcal{P}_L^{(1)}, \mathcal{P}_L^{(2)}, \dots, \mathcal{P}_L^{(K_{\text{bs}})})$ for $L = 1, 2, \dots, K_{\text{bs}}$, where \mathcal{P} denotes the power-allocation policy for the entire system and the vector \mathcal{P}_L represents the power-allocation policy for transmission mode L . Then, we formulate the following optimization problem **A3** to derive the optimal QoS-aware probability-vector $\phi^* \triangleq (\phi_1^*, \phi_2^*, \dots, \phi_{K_{\text{bs}}}^*)$ and its corresponding power-allocation policy \mathcal{P}^* .

$$\mathbf{A3}: \min_{(\phi(\mathbf{H}), \mathcal{P}(\mathbf{H}))} \{\overline{L}\} = \min_{(\phi(\mathbf{H}), \mathcal{P}(\mathbf{H}))} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) L \right\} \right\}$$

$$\text{s.t.: } 1). \sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) = 1, \quad \forall \mathbf{H}; \quad (32)$$

$$2). \sum_{n=1}^{K_{\text{mu}}} \mathcal{P}_L^{(n)}(\mathbf{H}) = \mathcal{P}_L, \quad \forall L, \mathbf{H}; \quad (33)$$

$$3). \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) e^{-\theta_n R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)}(\mathbf{H}))} \right\} \leq e^{-\theta_n \overline{C}_n}, \quad \forall n, \quad (34)$$

where $\phi(\mathbf{H})$ and $\mathcal{P}(\mathbf{H})$ characterize the probabilistic transmission vector and the power allocation policy described in the beginning of Section V-C.1, both of which vary with \mathbf{H} , and $R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)}(\mathbf{H}))$ is determined through Eq. (29). After clearly defining $\phi(\mathbf{H})$ and $\mathcal{P}(\mathbf{H})$, we drop the symbol (\mathbf{H}) in problem **A3** to simplify notations. We call the optimal solution to **A3** as the PBS-BD-PT scheme.

C.2. The Properties of $R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})$

Before solving **A3**, we need to study the properties of

$R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})$. Let us consider the n th user with $\mathbf{\Gamma}_{\Omega_L}^{(n)}$ not equal to zero. Similar to the results summarized in Section IV-A, the n th user's MIMO channel $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}$ (after the block diagonalization) can be converted to $Z_L^{(n)}$ parallel Gaussian sub-channels, where $Z_L^{(n)}$ is the rank of $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}$, the z th sub-channel's SNR is equal to $\varepsilon_{L,z}^{(n)}$ and $\sqrt{\varepsilon_{L,z}^{(n)}}$ is the z th largest nonzero singular value of $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}$. The optimal power $\rho_{L,z}^{(n)}$ allocated to the z th sub-channel follows the water-filling allocation, which is equal to $\rho_{L,z}^{(n)} = [\mu_L^{(n)} - 1/\varepsilon_{L,z}^{(n)}]^+$, where $\mu_L^{(n)} \geq 1/\varepsilon_{L,1}^{(n)}$ and $\mu_L^{(n)}$ is selected such that $\sum_{z=1}^{Z_L^{(n)}} \rho_{L,z}^{(n)} = \mathcal{P}_L^{(n)}$. Since $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}$ has only $Z_L^{(n)}$ non-zero singular values, for presentation convenience, we define $1/\varepsilon_{L,i}^{(n)} \triangleq \infty$ for $i = Z_L^{(n)} + 1$. Accordingly, we can show that

$$\frac{dR^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})}{d\mathcal{P}_L^{(n)}} = \frac{BT}{\mu_L^{(n)}} \quad (35)$$

holds and that $R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})$ is strictly concave over $\mathcal{P}_L^{(n)}$. Moreover, if $\mu_L^{(n)} \in [1/\varepsilon_{L,i}^{(n)}, 1/\varepsilon_{L,i+1}^{(n)})$ for $i = 1, 2, \dots, Z_L^{(n)}$, we get:

$$\begin{cases} \text{(a). } \mathcal{P}_L^{(n)} = i\mu_L^{(n)} - \sum_{j=1}^i \frac{1}{\varepsilon_{L,j}^{(n)}} \\ \text{(b). } R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)}) = BT \log \left(\prod_{j=1}^i \varepsilon_{L,j}^{(n)} \right) + BTi \log \mu_L^{(n)}. \end{cases} \quad (36)$$

C.3. The Optimal Solution to **A3**

Theorem 3: The optimal power-allocation policy \mathcal{P}^* for optimization problem **A3**, if existing, is given as follows:

$$\left(\mathcal{P}_L^{(n)}\right)^* = i^* \left(\prod_{j=1}^{i^*} \varepsilon_{L,j}^{(n)} \right)^{-\frac{BT\theta_n}{1+i^*BT\theta_n}} \left(\frac{\zeta_{\mathbf{H},L}^*}{BT\theta_n \lambda_n^*} \right)^{-\frac{1}{1+i^*BT\theta_n}} - \sum_{j=1}^{i^*} \frac{1}{\varepsilon_{L,j}^{(n)}}, \quad (37)$$

for all n, L , and \mathbf{H} , where $\varepsilon_{L,j}^{(n)}$ is the square of $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}$'s j th largest singular value, and i^* is the unique solution satisfying the following condition:

$$\mu_L^{(n)} \in \left[\frac{1}{\varepsilon_{L,i^*}^{(n)}}, \frac{1}{\varepsilon_{L,i^*+1}^{(n)}} \right), \quad \forall n, L, \mathbf{H}, \quad (38)$$

where

$$\mu_L^{(n)} = \max \left\{ \frac{1}{\varepsilon_{L,1}^{(n)}}, \left(\prod_{j=1}^{i^*} \varepsilon_{L,j}^{(n)} \right)^{-\frac{BT\theta_n}{1+i^*BT\theta_n}} \left(\frac{\zeta_{\mathbf{H},L}^*}{BT\theta_n \lambda_n^*} \right)^{-\frac{1}{1+i^*BT\theta_n}} \right\}. \quad (39)$$

The corresponding optimal probability-transmission policy is determined by

$$\phi_L^* = \begin{cases} 1, & \text{if } L = L^*; \\ 0, & \text{otherwise} \end{cases} \quad (40)$$

with

$$L^* = \arg \min_{0 \leq L \leq K_{\text{bs}}} \left\{ L + \sum_{n=1}^{K_{\text{mu}}} \lambda_n^* e^{-\theta_n BT (\log(\prod_{j=1}^{i^*} \varepsilon_{L,j}^{(n)}) + i^* \log \mu_L^{(n)})} \right\}, \quad \forall \mathbf{H}, \quad (41)$$

where given $\{\lambda_n^*\}_{n=1}^{K_{\text{mu}}}$ for Eqs. (37)-(39), the optimal $\zeta_{\mathbf{H},L}^*$ is selected to satisfy the equation $\sum_{n=1}^{K_{\text{mu}}} (\mathcal{P}_L^{(n)})^* = \mathcal{P}_L$ for all L and \mathbf{H} ; $\{\lambda_n^*\}_{n=1}^{K_{\text{mu}}}$ are constants, which are selected such that the equality of Eq. (34) holds.

Proof: We construct **A3**'s Lagrangian function, denoted by $\mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H}})$, as $\mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H}}) = \mathbb{E}_{\mathbf{H}} \{ \mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H}}) \}$ with

$$\begin{aligned} \mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H}}) &\triangleq \sum_{n=1}^{K_{\text{mu}}} \lambda_n \left[\sum_{L=0}^{K_{\text{bs}}} \phi_L e^{-\theta_n R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})} \right. \\ &\quad \left. - e^{-\theta_n \bar{C}_n} \right] + \sum_{L=1}^{K_{\text{bs}}} \zeta_{\mathbf{H},L} \left(\sum_{n=1}^{K_{\text{mu}}} \mathcal{P}_L^{(n)} - \mathcal{P}_L \right) + \sum_{L=0}^{K_{\text{bs}}} \phi_L L \end{aligned} \quad (42)$$

under $\sum_{L=1}^{K_{\text{bs}}} \phi_L = 1$, where $\lambda_n \geq 0$ for $n = 1, 2, \dots, K_{\text{mu}}$ are the Lagrangian multipliers associated with Eq. (34), which are constants over all fading states, and $\lambda \triangleq (\lambda_1, \lambda_2, \dots, \lambda_{K_{\text{mu}}})$; $\{\zeta_{\mathbf{H},L}\}_{L=1}^{K_{\text{bs}}}$ are the Lagrangian multipliers associated with Eq. (33), which vary with \mathbf{H} and L , and $\zeta_{\mathbf{H},L} \triangleq (\zeta_{\mathbf{H},1}, \zeta_{\mathbf{H},2}, \dots, \zeta_{\mathbf{H},K_{\text{bs}}})$.

Problem **A3**'s Lagrangian dual function [12], [14], denoted by $\tilde{\mathcal{J}}_{A3}(\lambda, \zeta_{\mathbf{H}})$, is determined by

$$\begin{aligned} \tilde{\mathcal{J}}_{A3}(\lambda, \zeta_{\mathbf{H}}) &\triangleq \min_{(\phi, \mathcal{P})} \left\{ \mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H}}) \right\} \\ &= \mathbb{E}_{\mathbf{H}} \left\{ \min_{(\phi, \mathcal{P})} \left\{ \mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H}}) \right\} \right\}. \end{aligned} \quad (43)$$

We denote the minimizer pair in Eq. (43) by $(\tilde{\phi}, \tilde{\mathcal{P}})$. Then, we can derive

$$\tilde{\phi} = \arg \min_{\phi: \sum_{L=1}^{K_{\text{bs}}} \phi_L = 1} \left\{ \sum_{L=1}^{K_{\text{bs}}} \phi_L \left(L + \sum_{n=1}^{K_{\text{mu}}} \lambda_n e^{-\theta_n R^{(n)}(\Omega_L, \tilde{\mathcal{P}}_L^{(n)})} \right) \right\}, \quad \forall \mathbf{H}, \quad (44)$$

where Eq. (44) holds by plugging Eq. (42) into Eq. (43) and removing the terms independent of ϕ . Solving Eq. (44), we obtain

$$\tilde{\phi}_L = 1, \quad \text{if } L = \arg \min_{1 \leq \ell \leq K_{\text{bs}}} \left\{ \ell + \sum_{n=1}^{K_{\text{mu}}} \lambda_n e^{-\theta_n R^{(n)}(\Omega_\ell, \tilde{\mathcal{P}}_\ell^{(n)})} \right\}; \quad (45)$$

otherwise, $\tilde{\phi}_L = 0$. Following the above derivations, $\tilde{\mathcal{P}}_L$ needs to minimize the function $\mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H}})$ given $\phi_L = 1$ and $\phi_j = 0$ for all $j \neq L$. Then, we define a set of functions $\mathcal{J}_{A3,L}(\mathcal{P}; \lambda, \zeta_{\mathbf{H},L})$ for $L = 1, 2, \dots, K_{\text{bs}}$, where $\mathcal{J}_{A3,L}(\mathcal{P}; \lambda, \zeta_{\mathbf{H},L}) \triangleq \mathcal{J}_{A3}(\phi, \mathcal{P}; \lambda, \zeta_{\mathbf{H},L})|_{\phi_L=1; \phi_j=0, j \neq L}$. Taking the derivative of $\mathcal{J}_{A3,L}(\mathcal{P}; \lambda, \zeta_{\mathbf{H},L})$ w.r.t. $\mathcal{P}_L^{(n)}$ and letting the derivative equal to zero, we get

$$\zeta_{\mathbf{H},L} - BT \lambda_n \theta_n \mu_L^{(n)} e^{-\theta_n R^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})} = 0 \quad (46)$$

for all n , L , and \mathbf{H} , where $\mu_L^{(n)} = dR^{(n)}(\Omega_L, \mathcal{P}_L^{(n)})/d\mathcal{P}_L^{(n)}$ as given in Eq. (35). Plugging Eq. (36)-(b) into Eq. (46) and solving for the optimal $\mu_L^{(n)}$ under the boundary condition of $\mu_L^{(n)} \geq 1/\varepsilon_{L,1}^{(n)}$, we obtain Eq. (39) with $i = i^*$. Since Eq. (36) is obtained under the condition of $\mu_L^{(n)} \in [1/\varepsilon_{L,i}^{(n)}, 1/\varepsilon_{L,i+1}^{(n)}]$, the variable i^* in Eq. (39) must satisfy the condition of $\mu_L^{(n)} \in [1/\varepsilon_{L,i^*}^{(n)}, 1/\varepsilon_{L,i^*+1}^{(n)}]$, as shown in Eq. (38). Moreover, we can show that $J_{A3,L}(\mathcal{P}; \lambda, \zeta_{\mathbf{H},L})$ is a strictly convex function, and thus i^* for Eq. (39) is unique.

The Lagrangian duality principle [12] shows that $\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}}) = \mathcal{J}(\tilde{\phi}, \tilde{\mathcal{P}}; \lambda, \zeta_{\mathbf{H}})$ is concave over λ and $\zeta_{\mathbf{H}}$. Moreover, the original problem (also called the primal problem) **A3**'s dual problem is defined by $\max_{(\lambda, \zeta_{\mathbf{H}})} \{\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})\}$. We denote the optimal objective of **A3** by \bar{L}^* . The equation $\bar{L}^* \geq \max_{(\lambda, \zeta_{\mathbf{H}})} \{\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})\}$ always holds [12], where the difference between \bar{L}^* and $\max_{(\lambda, \zeta_{\mathbf{H}})} \{\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})\}$ is known as the duality gap [12]. We can further show that $\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})$ is differentiable w.r.t. λ and $\zeta_{\mathbf{H}}$. Then, based on the Lagrangian duality principle [12], we can obtain

$$\begin{cases} \frac{\partial \tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})}{\partial \lambda_n} = \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\text{bs}}} \tilde{\phi}_L e^{-\theta_n R^{(n)}(\Omega_L, \tilde{\mathcal{P}}_L^{(n)})} - e^{-\theta_n \bar{C}_n} \right\}, & \forall n; \\ \frac{\partial \tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})}{\partial \zeta_{\mathbf{H},L}} = \left(\sum_{n=1}^{K_{\text{mu}}} \tilde{\mathcal{P}}_L^{(n)} - \mathcal{P}_L \right) g(\mathbf{H}) d\mathbf{H}, & \forall L, \mathbf{H} \end{cases} \quad (47)$$

where $g(\mathbf{H})$ is the probability density function of \mathbf{H} and $d\mathbf{H}$ denotes the integration variable. As a result, the maximizer $\zeta_{\mathbf{H},L}^*$ must be selected such that $\sum_{n=1}^{K_{\text{mu}}} \tilde{\mathcal{P}}_L^{(n)} - \mathcal{P}_L = 0$. Such a $\zeta_{\mathbf{H},L}^*$ exists because $\zeta_{\mathbf{H},L} \rightarrow 0$ and $\zeta_{\mathbf{H},L} \rightarrow \infty$ leads to $\tilde{\mathcal{P}}_L^{(n)} \rightarrow \infty$ and $\tilde{\mathcal{P}}_L^{(n)} \rightarrow 0$, respectively, as indicated by Eqs. (36)-(a) and (39).

Having obtained $\zeta_{\mathbf{H}}^*$, we next focus on the optimal λ^* to maximize $\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}}^*)$. Due to the concavity of $\tilde{\mathcal{J}}_{A3}(\lambda, \zeta_{\mathbf{H}})$, $\partial \tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}}^*)/\partial \lambda_n$ is a decreasing function of λ_n . Also, we can readily show that $\partial \tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}}^*)/\partial \lambda_n|_{\lambda_n=0} > 0$. Then, if there does not exist λ such that $\partial \tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}}^*)/\partial \lambda_n = 0$ for all n , we have $\lambda_n^* \rightarrow \infty$ for some n th user and $\partial \tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}}^*)/\partial \lambda_n > 0$ always holds. For this case, we get $\bar{L}^* \geq \tilde{\mathcal{J}}(\lambda^*, \zeta_{\mathbf{H}}^*) \rightarrow \infty$, implying that there is no feasible solution for **A3**.

In contrast, if there exists λ^* such that $\partial \tilde{\mathcal{J}}(\lambda^*, \zeta_{\mathbf{H}}^*)/\partial \lambda_n = 0$ for all n , the pair of $(\lambda^*, \zeta_{\mathbf{H}}^*)$ is the optimal solution to the dual problem given by Eq. (47). Plugging $\partial \tilde{\mathcal{J}}(\lambda^*, \zeta_{\mathbf{H}}^*)/\partial \lambda_n = 0$ into Eq. (47), we can see that the effective-capacity constraint required by Eq. (34) is satisfied for every user. Also note that the optimum for $\max_{(\lambda, \zeta_{\mathbf{H}})} \{\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})\}$ is achieved by using $(\tilde{\mathcal{P}}, \tilde{\phi})$ under λ^* and $\zeta_{\mathbf{H}}^*$, which satisfy the constraints imposed by Eqs. (32)-(33). Therefore, all constraints for problem **A3** are satisfied, implying that this policy is feasible to problem **A3**. As a result, $\bar{L}^* = \max_{(\lambda, \zeta_{\mathbf{H}})} \{\tilde{\mathcal{J}}(\lambda, \zeta_{\mathbf{H}})\}$ holds with zero duality gap. Thus, this policy is the optimal solution to **A3**. Then, setting $\mathcal{P}^* = \tilde{\mathcal{P}}$ and $\phi^* = \tilde{\phi}$ with λ^* and $\zeta_{\mathbf{H}}^*$ in Eq. (45), we obtain Eqs. (40)-(41). Further plugging Eq. (39) into Eq. (36)-(a), we prove that Eq. (37) holds. Finally, comparing $\partial \tilde{\mathcal{J}}(\lambda^*, \zeta_{\mathbf{H}}^*)/\partial \lambda_n = 0$ with Eq. (47), we show that the equality of Eq. (34) holds, which completes the proof of Theorem 3. ■

Note that there are no closed-form solutions for the optimal Lagrangian multipliers $\zeta_{\mathbf{H},L}^*$ and λ^* . However, we can determine the values of $\zeta_{\mathbf{H},L}^*$ and λ^* by using the numerical searching method similar to the approach given in Appendix B. The detailed searching techniques are omitted due to lack of space, but are provided on-line in the technical-report version [19] of this paper.

D. The TDMA Based BS-Selection Scheme

We next propose the TDMA based QoS-aware BS-selection scheme for the comparative analyses. In the TDMA based BS-selection, we also apply the priority BS-selection algorithm given by Fig. 4 when transmission mode L is specified. For transmission mode L , we further divide each time frame into K_{mu} time slots for data transmissions to K_{mu} users, respectively. The n th user's time-slot length is set equal to $T \times t_{L,n}$ for $n = 1, 2, \dots, K_{\text{mu}}$, where $t_{L,n}$ is the normalized time-slot length. Moreover, we still use the probabilistic transmission strategy across different transmission modes, where the probability of using transmission mode L to transmit data is equal to ϕ_L . Then, we derive the TDMA and probabilistic transmission policies through solving the following optimization problem **A4**.

$$\begin{aligned} \mathbf{A4} : \quad & \min_{(\mathbf{t}(\mathbf{H}), \phi(\mathbf{H}))} \{\bar{L}\} = \min_{(\mathbf{t}(\mathbf{H}), \phi(\mathbf{H}))} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\text{bs}}} L \phi_L(\mathbf{H}) \right\} \right\} \\ \text{s.t.: } & 1). \sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) = 1, \quad \forall \mathbf{H}, \quad (48) \\ & 2). \sum_{n=1}^{K_{\text{mu}}} t_{L,n}(\mathbf{H}) = 1, \quad \forall \mathbf{H}, L = 1, 2, \dots, K_{\text{bs}}, \quad (49) \\ & 3). \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\text{bs}}} \phi_L(\mathbf{H}) e^{-\theta_n t_{L,n}(\mathbf{H}) R^{(n)}(\Omega_L, \mathcal{P}_L)} \right\} \\ & \leq e^{-\theta_n \bar{C}_n}, \quad \forall n, \quad (50) \end{aligned}$$

where $\phi(\mathbf{H})$ and $\mathbf{t}(\mathbf{H})$ are the probabilistic transmission vector and the time-division policy. In particular, in problem **A4** we have $\phi(\mathbf{H}) \triangleq (\phi_0(\mathbf{H}), \phi_1(\mathbf{H}), \dots, \phi_{K_{\text{mu}}}(\mathbf{H}))$ and $\mathbf{t}(\mathbf{H}) \triangleq (t_{1,1}(\mathbf{H}), t_{1,2}(\mathbf{H}), \dots, t_{K_{\text{bs}}}(\mathbf{H}))$ with $t_L \triangleq (t_{L,1}(\mathbf{H}), t_{L,2}(\mathbf{H}), \dots, t_{L,K_{\text{bs}}}(\mathbf{H}))$. We use $\phi(\mathbf{H})$ and $\mathbf{t}(\mathbf{H})$ to denote the probabilistic transmission vector and the time-division policy to address that they are function of \mathbf{H} . When the context is clear, we drop the symbol (\mathbf{H}) in problem **A4** to simplify notations. We call the optimal solution of problem **A4** as the PBS-TDMA-PT scheme.

Theorem 4: Problem **A4**'s optimal solution pair (\mathbf{t}^*, ϕ^*) , if existing, is determined by

$$t_{L,n}^* = \left[\frac{1}{\theta_n R^{(n)}(\Omega_L, \mathcal{P}_L)} \log \left(\frac{\lambda_n^* \theta_n R^{(n)}(\Omega_L, \mathcal{P}_L)}{\delta_{\mathbf{H},L}^*} \right) \right]^+, \quad (51)$$

for all L , n , and \mathbf{H} , and

$$\phi_L^* = \begin{cases} 1, & \text{if } L = \arg \min_{\ell=0,1,\dots,K_{\text{bs}}} \left\{ \ell + \sum_{n=1}^{K_{\text{mu}}} \lambda_n^* e^{-\theta_n t_{L,n}^* R^{(n)}(\Omega_L, \mathcal{P}_L)} \right\}; \\ 0, & \text{otherwise} \end{cases} \quad (52)$$

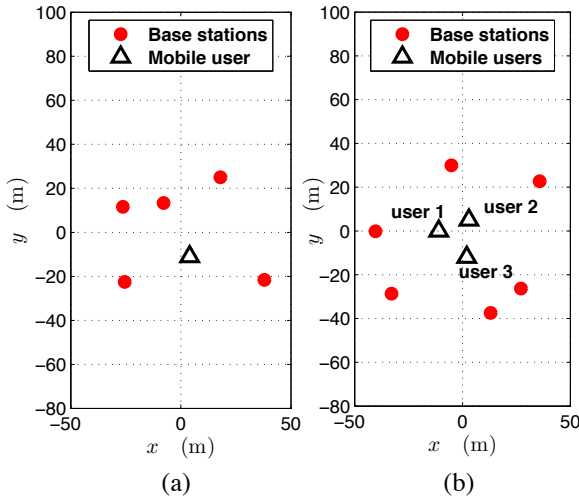


Fig. 5. The deployment of BS's and the positions of mobile users. (a) Single-user case: $K_{\text{bs}} = 5$ BS's, whose coordinates are $(37.96, -21.56)$, $(-7.83, 13.33)$, $(25.50, -22.49)$, $(17.98, 25.00)$, and $(-26.34, 11.62)$; the mobile station's coordinates are $(4, -11)$. (b) Multi-user case: $K_{\text{bs}} = 6$, whose coordinates are $(35.77, 22.69)$, $(13.06, -37.45)$, $(27.15, -26.33)$, $(-40.28, -0.14)$, $(-32.86, -28.65)$, and $(-5.10, 29.98)$; $K_{\text{mu}} = 3$, whose coordinates are $(-11, 0)$, $(3, 5)$, and $(2, -12)$.

for all L and \mathbf{H} , where $\delta_{\mathbf{H},L}^*$ under given $\{\lambda_n^*\}_{n=1}^{K_{\text{mu}}}$ is determined by satisfying $\sum_{n=1}^{K_{\text{mu}}} t_{L,n}^* = 1$, and $\{\lambda_n^*\}_{n=1}^{K_{\text{mu}}}$ needs to be selected such that the equality of Eq. (50) holds.

Proof: The detailed proof of Theorem 4 is omitted due to lack of space, but is provided on-line in the technical-report version [19] of this paper. ■

VI. SIMULATION EVALUATIONS

We use simulations to evaluate the performances of our proposed QoS-aware BS-selection schemes for distributed MIMO links. In particular, for the single-user case, the BS's deployment and the mobile user's position are shown in Fig. 5(a), where $K_{\text{bs}} = 5$. For the multi-user case, the BS's deployment and the mobile users' positions are given by Fig. 5(b), where $K_{\text{bs}} = 6$ and $K_{\text{mu}} = 3$. Moreover, we set $T = 10$ ms and $B = 10^5$ Hz throughout the simulations. In the simulations, we focus on the case that the positions of users do not change. This can be mapped to the scenarios where users move slowly (e.g., pedestrians) or do not move. In such scenarios, the users' statistical channel conditions, depending on the distances between users and distributed BS's, often change very slightly in tens of seconds, during which many (thousands of) packets will be transmitted before the users' statistical channel conditions vary significantly. Then, our developed statistical delay-QoS aware BS-selection schemes, which are based on the current statistical channel information, can effectively guarantee the corresponding delay QoS during this period. For the scenarios that the users move quickly, the delay-QoS performances high depend on the users' moving speeds and trajectories, which are unlike the evaluations for the average transmission error rate in cellular networks, where only the distribution information of users' positions is needed. Thus the delay-QoS-aware BS-selection schemes which need to be re-designed, which will be studied in our future work.

In the simulations, we employ the following average power degradation propagation model [15]. For the given reference distance d_{ref} , if the transmission distance, denoted by d , is smaller than or equal to d_{ref} , the free-space propagation model is used; if $d > d_{\text{ref}}$, the power degradation is proportional to $(d/d_{\text{ref}})^\eta$, where η is the path loss exponent and typically varies from 2 to 6 indoor environments without LOS [15]. Accordingly, the variance $\bar{h}_{n,m}$ of $H_{n,m}$'s elements can be determined by

$$\bar{h}_{n,m} = \begin{cases} Gd_{n,m}^{-2}, & \text{if } d_{n,m} \in (0, d_{\text{ref}}]; \\ G\left(\frac{d_{\text{ref}}}{d_{n,m}}\right)^\eta, & \text{if } d_{n,m} \in (d_{\text{ref}}, \infty), \end{cases}$$

where $d_{n,m}$ is the distance between the mobile user and the m th BS, and G is the aggregate power gain generated by the antenna and other factors. In simulations, we set $d_{\text{ref}} = 1$ m [15] and $\eta = 3$. Furthermore, we select G such that $\bar{h}_{n,m} = 0$ dB at $d_{n,m} = 50$ m. Also, we set $\sigma_{\text{th}}^2 = 0$ dB for the evaluation of the average interfering range (see Section II-C for its definition).

We also consider the following baseline BS-selection schemes for comparative analyses:

1). *Base-Station Selection with Fixed Cardinality for Single-User Cases*

This scheme does not change the selected BS-subset's cardinality over all fading states. The fixed cardinality, denoted by L_{fix} , is determined by

$$L_{\text{fix}} = \min \left\{ L \left| \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 R(\Omega_L)} \right\} \leq \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \bar{C}_1} \right\}, L \text{ is fixed} \right\}, \quad (53)$$

where Ω_L for each \mathbf{H} is determined through the ordered-gain based BS-selection algorithm given by Section IV-A, as also used by our derived OGBS-PT scheme.

2). *Optimal Time-Sharing Based Scheme for Single-User Cases*

The optimal time-sharing (TS) based scheme is the same to our proposed IBS-TS scheme except that it uses exhaustive search to find the optimal Ω_L for Eq. (13) in each fading state. It is clear that the optimal TS based scheme provides the lower-bound of the average BS usage for our proposed IBS-TS scheme.

3). *Semi-Random BS-Selection Scheme for Multi-user Cases*

In this scheme, given the cardinality L , the central controller randomly selects Ω_L with the uniform distribution across all BS's. Furthermore, the cardinality L of the selected BS subset is determined by using the algorithm derived through Theorem 3. We call this scheme as the *semi-random BD-PT* scheme in that the selection of L is not in a complete random manner. We evaluate the semi-random BD-PT scheme to provide an upper-bound of the average BS-usage for our derived priority BS-selection algorithm given in Fig. 4. This semi-random strategy can be also applied to the TDMA based scheme, where Ω_L is randomly selected and L is determined by using the algorithm obtained through Theorem 4.

Figures 6(a) and 6(b) plot the average BS usage and the average interfering range, respectively, versus the incoming traffic loads for the single-user case. As shown in Fig. 6(a), our proposed IBS-TS and OGBS-PT schemes both effectively

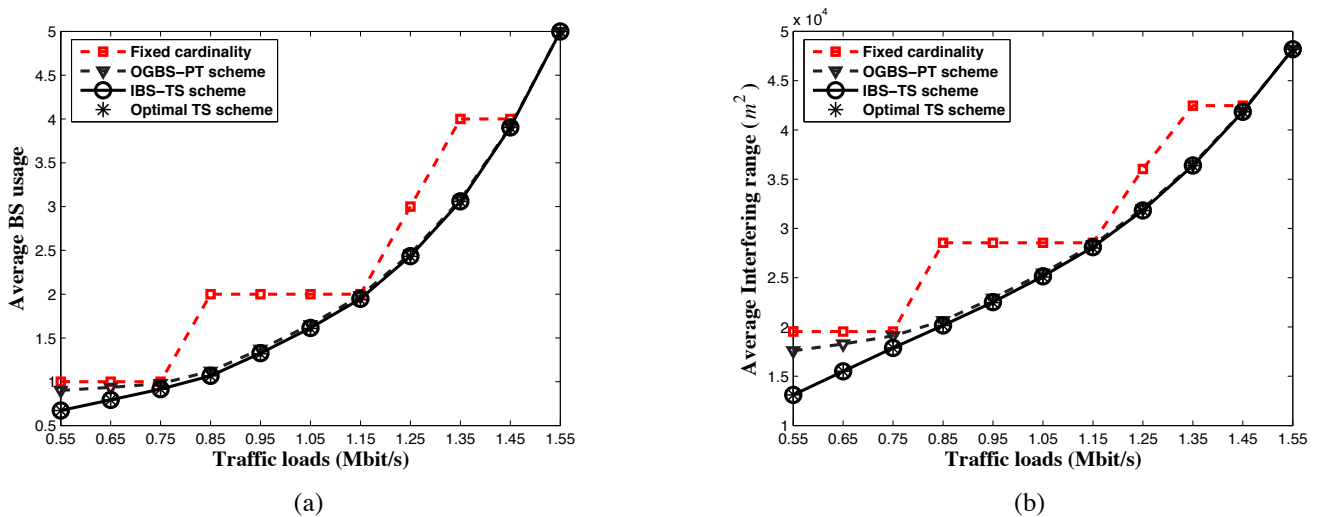


Fig. 6. Simulations of the single-user case for the specified delay-QoS requirements given by $D_{th}^{(1)} = 50$ ms and $\xi_1 = 10^{-4}$, where $K_{bs} = 5$, each BS has two transmit antennas, the mobile user has two receive antennas, $\mathcal{P}_{ref} = 4$, and $\kappa = 2.4$. (a) Average BS usage versus traffic load. (b) Average interfering range versus traffic load.

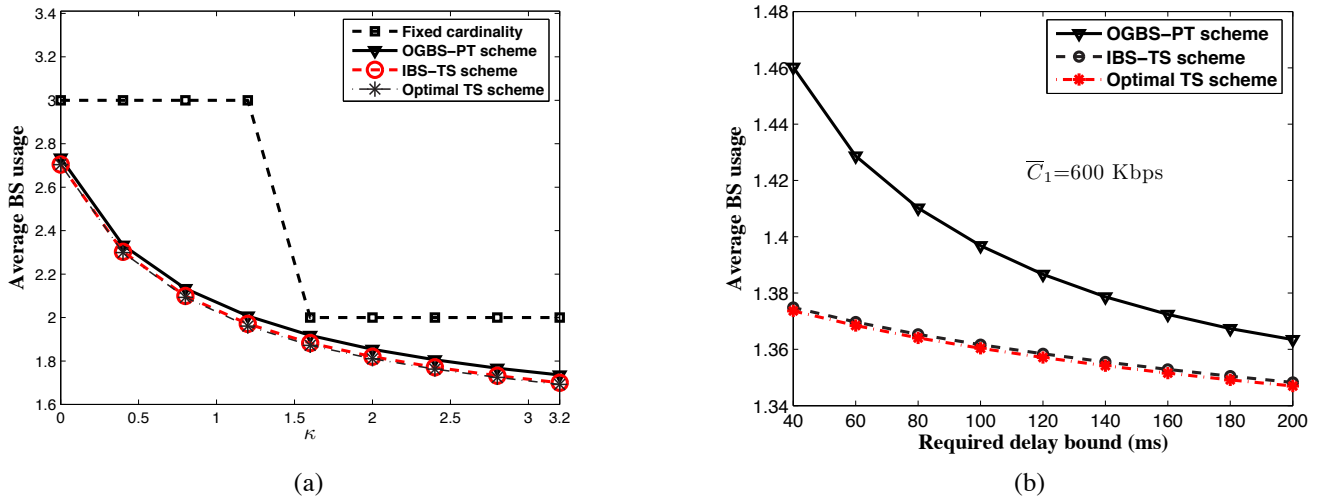


Fig. 7. Simulation results for single-user case varying the specified delay-QoS requirements and system parameters, where $K_{bs} = 5$, each BS has two transmit antennas, the mobile user has two receive antennas. (a) Average BS usage versus κ , where $D_{th}^{(1)} = 50$ ms, $\xi_1 = 10^{-4}$, $\bar{C}_1 = 1.1$ Mbps, and $\mathcal{P}_{ref} = 1$. (b) Average BS usage versus delay-bound QoS requirements, where $\xi_1 = 10^{-6}$, $\bar{C}_1 = 600$ Kbps, $\kappa = 0.4$, and $\mathcal{P}_{ref} = 1$.

decrease the average BS usage and the interfering range as compared to the fixed BS-selection scheme. This is expected because our proposed BS-selections can adaptively adjust BS selection in each fading state based on the CSI, traffic loads, and the QoS requirements. In contrast, the resulted average BS usage and the interfering range by applying the fixed BS-selection scheme cannot smoothly vary with traffic loads, which may cause unnecessary BS usage with more power consumption and thus larger interferences to the entire wireless networks. We also observe from Fig. 6 that the IBS-TS scheme needs less BS usage than the OGBS-PT scheme to support the incoming traffic loads under the specified QoS requirements and therefore generates lower interferences accordingly, which verifies our discussions in Sections IV-A and IV-B. However, we can see that the performance differences between the IBS-TS and OGBS-PT schemes are slight, especially when the incoming traffic loads are relatively high. In addition, we can

observe from Fig. 6 that the BS-usage performance and the interfering range of our proposed IBS-TS scheme are almost the same as compared to the optimal TS scheme, which verifies the effectiveness of our proposed IBS-TS scheme.

Figure 7(a) depicts the dynamics of the average BS usage and the interfering range as functions of κ , where κ is the power increasing rate against the number of selected BS's as defined in Eq. (3). When κ gets larger, the power budget used for distributed MIMO transmission in each fading state increases. As a result, the average BS usage is reduced correspondingly for all schemes, as illustrated in Fig. 7(a). Again, Fig. 7(a) demonstrates the inflexibility of the fixed BS-selection, and the significant BS-usage reduction and the interfering-range reduction as a result of using our proposed IBS-TS and OGBS-PT schemes.

Figure 7(b) illustrates the impact of delay-QoS requirements on the average BS-usage among different BS-selection

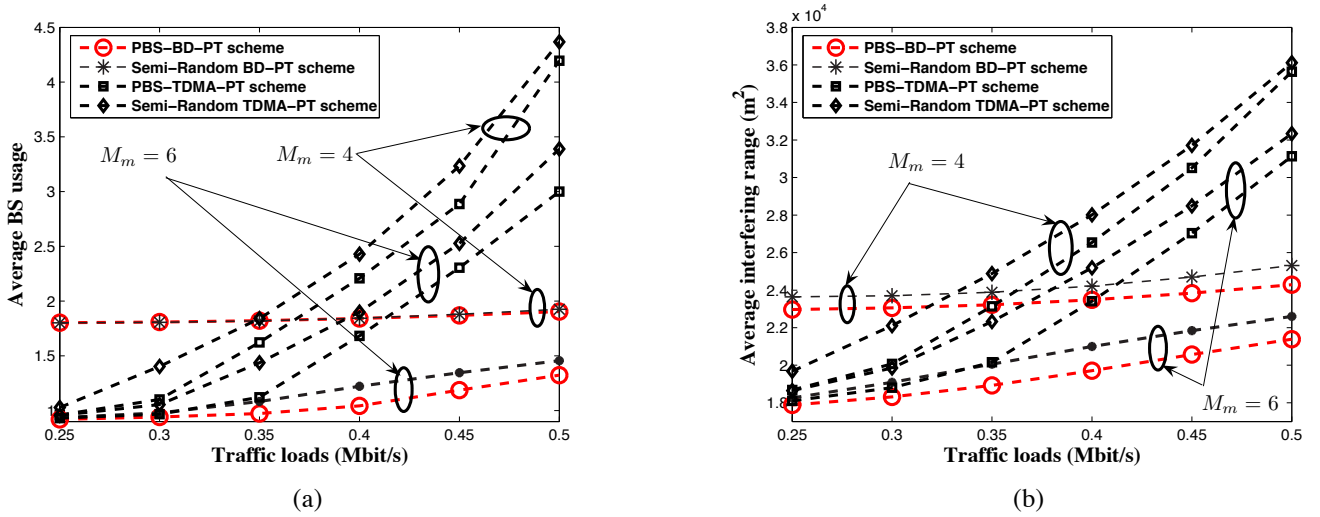


Fig. 8. Simulation results of the multi-user case for the specified delay-QoS requirements given by $\xi_n = 10^{-4}$ for all n and $D_{th}^{(1)} = D_{th}^{(2)} = 50$ ms and $D_{th}^{(3)} = 40$ ms, where $K_{bs} = 6$, $K_{mu} = 3$, $N_n = 2$ for all $n = 1, 2, \dots, K_{mu}$, $\mathcal{P}_{ref} = 4$, $\kappa = 1.2$, and M_m is the same for all users. (a) Average BS usage versus traffic load. (b) Average interfering range versus traffic load.

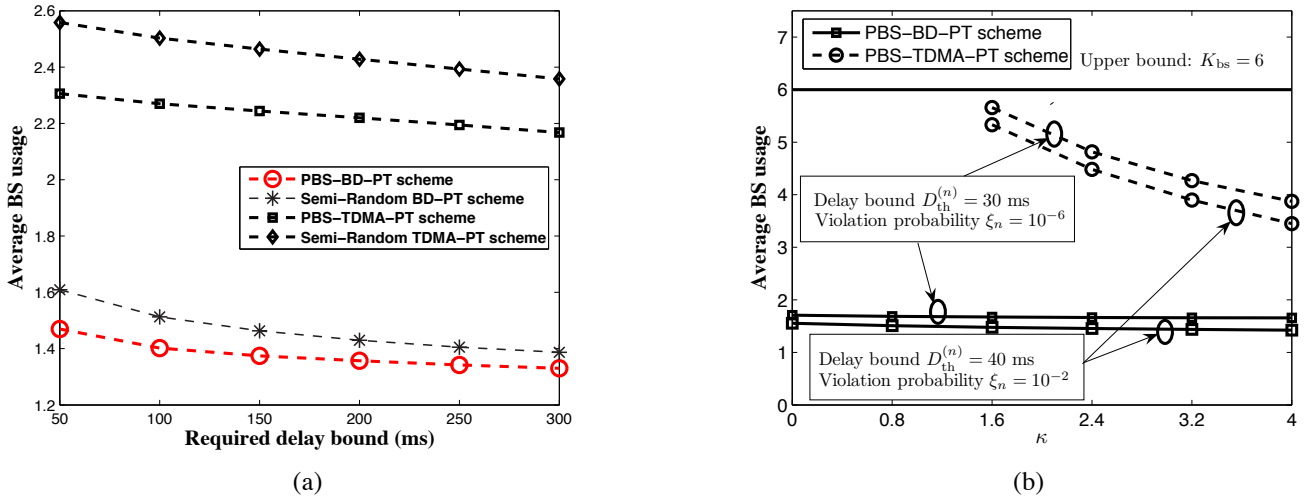


Fig. 9. Simulation results for impact of QoS requirement on the multi-user case, where $K_{bs} = 6$, $K_{mu} = 3$, and $N_n = 2$ for all $n = 1, 2, \dots, K_{mu}$ and $M_m = 6$ for all $m = 1, 2, \dots, K_{bs}$; $\{\xi_n\}_{n=1}^{K_{mu}}$ are equal; $\{D_{th}^{(n)}\}_{n=1}^{K_{mu}}$ are equal; $\{\bar{C}_n\}_{n=1}^{K_{mu}}$ are equal. (a) Average BS usage versus delay-bound QoS requirements, where $\bar{C} = 300$ Kbps, $\xi_n = 10^{-4}$, $\kappa = 0.1$, and $\mathcal{P}_{ref} = 1$. (b) Average BS usage versus κ , where $\bar{C} = 600$ Kbps and $\mathcal{P}_{ref} = 4$.

algorithms. As shown in Fig. 7(b), the average BS-usage decreases as the required delay-bound increases, which is expected because of the following reasons. When the delay QoS constraint becomes looser, suggesting a larger delay bound, the central controller does not have to use more BS's to improve the transmission rate when the wireless channel experiences deep fading. Correspondingly, the efficiency of using the distributed BS's can be improved, thus reducing the average BS usage. The above observations verify that our proposed BS-selection algorithms can effectively adapt to the variation of delay QoS requirements. Also, from Fig. 7(b) we can observe again that the IBS-TS results in less BS usage than the OGBS-PT scheme, and the IBS-TS scheme achieves almost the same performance as the optimal TS scheme.

Figures 8(a) and 8(b) compare the average BS usage and interfering range among our proposed BS-selection scheme for multi-user case under various system parameters. Figures 8(a)

and 8(b) show that as the traffic loads increase, all schemes' average BS usages and interfering ranges increase to satisfy the specified QoS requirements for the incoming traffic loads. Moreover, the PBS-TDMA-PT scheme's BS usage increases much more rapidly than the PBS-BD-PT scheme. As shown in Fig. 8(a), the PBS-BD-PT scheme results in less average BS usage than the PBS-TDMA-PT scheme when the traffic load is high. The superiority of the PBS-BD-PT over the PBS-TDMA-PT scheme gradually vanishes as the traffic loads become low. Furthermore, we can observe that given $M_m = 6$, the PBS-BD-PT scheme achieves lower BS usage than the PBS-TDMA-PT scheme under various traffic loads. This is expected since it has been demonstrated that the BD technique usually outperforms the TDMA technique in achieving high throughput. However, for $M_m = 4$, the PBS-TDMA-PT scheme leads to less BS usage when the traffic load is relatively low. The above result is caused by the following reasons.

In order to support simultaneous transmissions across multiple users by using the BD technique, a necessary condition is that the total number of transmit antennas over all distributed BS's can provide enough freedoms for spatial multiplexing. Thus, if the number M_m of transmit antenna per BS is smaller, more BS's have to involve the transmission to gain enough spatial-multiplexing freedoms, even when the traffic load is low. In contrast, when the traffic load is very low, TDMA does not have such requirement on spatial multiplexing because it achieves cross-interference-free transmissions through time-slot division. The above observations will motivate us to develop hybrid TDMA-BD BS-selection schemes in future studies.

Figure 8(a) further illustrates that our proposed PBS-BD-PT and PBS-TDMA-PT schemes in general yields less BS usages than the semi-random BD-PT and the semi-random TDMA-PT schemes, respectively. Note that when $M_m = 4$, the BS-usages of the semi-random BD-PT scheme and our proposed PBS-BD-PT scheme are the same, which also increase very slowly as the traffic loads increase. This is because for small number of transmit antennas, more BS's are needed to support the space-multiplexing, as aforementioned. When more BS's are used, the achievable rate for each dimension/user often exceed the traffic load for the corresponding user when the traffic load is low. In such a case, even randomly selecting BS can satisfy the specified delay-QoS constraint under the given traffic loads. However, even when the BS usages of the semi-random BD-PT scheme and our proposed PBS-BD-PT scheme are the same, we can see from Fig. 8(b) that the PBS-BD-PT scheme results in a smaller interfering range by selecting the BS based CSI.

Figure 9(a) plots the dynamics of the BS usage versus the required delay bound. As depicted in Fig. 9(a), the larger delay bound leads to less BS usage for all schemes, which demonstrates that our proposed schemes can effectively adjust the BS selections to meet diverse QoS requirements across multiple co-existing users. Also, we can see that the semi-random based schemes cause significantly higher BS-usages than our derived PBS based schemes. Fig. 9(b) illustrates our proposed PBS-BD-PT scheme and PBS-TDMA-PT based scheme's performances versus κ (see Eq. (3)) under diverse QoS requirements. Similar to the single-user case, the average BS usage decreases as κ gets larger. Moreover, as shown in Fig. 9, the lower delay bound and the smaller violation probability threshold, implying more stringent delay-QoS requirements, cause more BS usages and thus larger interfering ranges.

VII. CONCLUSIONS

We proposed the QoS-aware BS-selection schemes for the distributed wireless MIMO downlink to minimize the BS usages and to reduce the interfering range affected by the distributed MIMO system, while satisfying diverse statistical delay-QoS constraints characterized by the delay-bound violation probability and the effective capacity technique. For the single-user scenario, we developed the scheme using the incremental BS selection and time-sharing strategy and also developed the scheme employing the ordered-gain based BS-selection and probabilistic transmission strategy. The former scheme archives better performance, while the latter scheme

is easier to implement. For the multi-user scenario, we developed the joint priority BS-selection, block-diagonalization precoding, and probabilistic transmission schemes. We also studied the TDMA based BS-selection scheme for multi-user link. The extensive simulation results show that our proposed schemes can effectively support the incoming traffic loads for the specified QoS requirements and significantly outperform baseline schemes in terms of minimizing the average BS usage and decreasing the interfering range.

APPENDIX A PROOF OF THEOREM 1

Proof: We construct the Lagrangian function of $\mathbf{A1}$, denoted by $\mathcal{J}_{A1}(\mathcal{L}, \lambda)$, given as $\mathcal{J}_{A1}(\mathcal{L}, \lambda) = \mathbb{E}_{\mathbf{H}}\{J_{A1}(\mathcal{L}, \lambda)\}$ with

$$J_{A1}(\mathcal{L}, \lambda) = \mathcal{L} + \lambda \left(e^{-\theta_1 \tilde{R}(\mathcal{L})} - e^{-\theta_1 \bar{C}_1} \right), \quad (54)$$

where λ is the Lagrangian multiplier associated with the constraint of $\mathbf{A1}'$. Then, the optimal \mathcal{L}^* and the optimal Lagrangian multiplier λ^* are solutions to the following equations [12]:

$$\begin{cases} 0 \in \partial_{\mathcal{L}} J_{A1}(\mathcal{L}, \lambda), & \forall \mathbf{H}; \\ 0 = \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \tilde{R}(\mathcal{L})} - e^{-\theta_1 \bar{C}_1} \right\}. \end{cases} \quad (55)$$

where $\partial_{\mathcal{L}} J_{A1}(\mathcal{L}, \lambda)$ denotes the *subdifferential* [12] of the function $J_{A1}(\mathcal{L}, \lambda)$ with respect to \mathcal{L} . Note that the subdifferential is defined for nondifferentiable convex functions (e.g., piece-wise linear functions), which is the counterpart concept for the gradient of differentiable convex functions. Based on [12], for a convex function $h(\mathbf{b})$ defined on $\mathbf{b} \in \mathcal{B} \subset \mathbb{R}^n$, where \mathbb{R} is the real-number set and \mathcal{B} is a convex set, an $n \times 1$ real-valued vector ϖ is a *subgradient* of $h(\mathbf{b})$ if $h(\mathbf{b}') \geq h(\mathbf{b}) + \varpi^T(\mathbf{b}' - \mathbf{b})$ for all $\mathbf{b}' \in \mathcal{B}$. Accordingly, the collection of subgradients at \mathbf{b} is the *subdifferential* of $h(\mathbf{b})$. For more details and properties of subdifferential, please refer to [12].

Applying the piece-wise linear property and the concavity of $\tilde{R}(\mathcal{L})$, we derive

$$\begin{aligned} \partial_{\mathcal{L}} J_{A1}(\mathcal{L}, \lambda) &= \begin{cases} \left[1 - \theta_1 \lambda \nu_j e^{-\theta_1 \tilde{R}(m_j)}, 1 - \theta_1 \lambda \nu_{j+1} e^{-\theta_1 \tilde{R}(m_j)} \right), \\ \quad \text{if } \mathcal{L} = m_j, j = 0, 1, \dots, \mathcal{K}; \\ \left\{ 1 - \theta_1 \lambda \nu_j e^{-\theta_1 \tilde{R}(\mathcal{L})} \right\}, \\ \quad \text{if } \mathcal{L} \in (m_{j-1}, m_j), j = 1, \dots, \mathcal{K}. \end{cases} \end{aligned} \quad (56)$$

Plugging Eq. (56) into Eq. (55) and solving for the optimal solution, we get Eq. (23). Also, the equality of Eq. (22) needs to hold as required by Eq. (55). which completes the proof of Theorem 1. \blacksquare

APPENDIX B THE NUMERICAL-TRACKING METHOD TO DERIVE λ^* AND DISCUSSIONS ON ITS COMPLEXITY

As defined in Appendix A, the Lagrangian function under the optimal solution derived in Theorem 1 is given by

REFERENCES

$$\mathcal{J}_{A1}(\mathcal{L}^*, \lambda) = \mathbb{E}_{\mathbf{H}}\{\mathcal{L}^*\} + \lambda \mathbb{E}_{\mathbf{H}}\left\{e^{-\theta_1 \tilde{R}(\mathcal{L}^*)} - e^{-\theta_1 \bar{C}_1}\right\}.$$

Note that in this equation we remove the superscript $(\cdot)^*$ of λ^* because we have not obtained the actual value of λ^* . Based on the convex optimization theory, $\mathcal{J}_{A1}(\mathcal{L}^*, \lambda)$, known as the Lagrangian dual function [12], [14], is concave over λ . We can show that the derivative $d\mathcal{J}_{A1}(\mathcal{L}^*, \lambda)/d\lambda = \mathbb{E}_{\mathbf{H}}\{e^{-\theta_1 \tilde{R}(\mathcal{L}^*)} - e^{-\theta_1 \bar{C}_1}\}$. Moreover, if the optimal solution exists, $\mathcal{J}_{A1}(\mathcal{L}^*, \lambda)$ reaches its maximum at $\lambda = \lambda^*$ with $\mathbb{E}_{\mathbf{H}}\{e^{-\theta_1 \tilde{R}(\mathcal{L}^*)} - e^{-\theta_1 \bar{C}_1}\} = 0$. Therefore, we can use the gradient descent method [14] to search for λ^* numerically. The gradient descent method is implemented through the following iterative equation:

$$\lambda := \lambda + \epsilon \mathbb{E}_{\mathbf{H}}\left\{e^{-\theta_1 \tilde{R}(\mathcal{L}^*)} - e^{-\theta_1 \bar{C}_1}\right\},$$

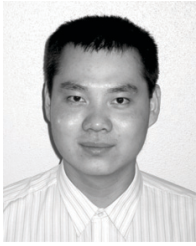
where ϵ is a positive real number much smaller than 1. For the fast tracking of λ^* , we can use the time average via the first-order autoregressive low-pass filter to estimate and replace the expectation $\mathbb{E}_{\mathbf{H}}\{e^{-\theta_1 \tilde{R}(\mathcal{L}^*)} - e^{-\theta_1 \bar{C}_1}\}$ for the above gradient descent method. In particular, we generate a large number (more than 10^4) of channel realizations, representing different fading states, which are indexed by time sequence $k = 1, 2, \dots$. Denoting the estimate of $\mathbb{E}_{\mathbf{H}}\{e^{-\theta_1 \tilde{R}(\mathcal{L}^*)} - e^{-\theta_1 \bar{C}_1}\}$ at the k -th time by $\mathcal{J}'_{\lambda}[k]$, we use the following first-order autoregressive low-pass filter to update $\mathcal{J}'_{\lambda}[k]$:

$$\begin{aligned} \mathcal{J}'_{\lambda}[k] &= \vartheta \mathcal{J}'_{\lambda}[k-1] \\ &+ (1-\vartheta) \left(e^{-\theta_1 \tilde{R}(\mathcal{L}^*[k])} - e^{-\theta_1 \bar{C}_1} \right) \Big|_{\mathbf{H}=\mathbf{H}[k]}, \end{aligned}$$

where $\vartheta \in (0, 1)$ is a real number close to 1, $\mathcal{L}^*[k]$ is determined by using Eq. (23) under the current λ of the k th time frame, and $\mathbf{H}[k]$ denotes the CSI at the k th time frame. With appropriately selected ϵ and ϑ , the above algorithm can converge and oscillate around the optimal λ^* within a small range. The above tracking method has been used and validated in many optimization problems for resource allocation over fading channels. By using this algorithm, we can conveniently find the optimal solution without deriving the closed-form expressions for the expectation and λ^* , which are usually hard to obtain for general fading channel models.

Under the above strategy, the computational complexity for our proposed BS-selection scheme in each fading state is dominated by: 1). the incremental BS-selection algorithm given in Fig. 2 to determine Ω_L ; 2). singular value decomposition (SVD) of the channel matrix to calculate the MIMO capacity [5]; 3). the procedures to find the convex hull of $\{(L, \tilde{R}(L))\}_{i=0}^{K_{\text{bs}}}$ for deriving Eq. (19). It is not difficult to see that the complexity of the incremental algorithm is upper-bounded by $O(K_{\text{bs}}^2)$; the complexity for SVD is known upper-bounded by $O(K_{\Sigma}^3)$, where $K_{\Sigma} \triangleq \sum_{m=1}^{K_{\text{bs}}} M_m$; the complexity to identify the convex hull is upper-bounded by $O(K_{\text{bs}} \log(K_{\text{bs}}))$. Therefore, the total complexity for each fading state is upper-bounded by $O(K_{\Sigma}^3)$, implying the major computational complexity is caused by the calculation for the MIMO capacity.

- [1] A. Sanderovich, S. Shamai (Shitz), and Y. Steinberg, "Distributed MIMO receiver-Achievable rates and upper bounds," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4419-4438, Oct. 2009.
- [2] D. Richard Brown III and H. Vincent Poor, "Time-slotted round-trip carrier synchronization for distributed beamforming," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5630-5643, Nov. 2008.
- [3] R-Mudumbai, D.R. Brown III, U. Madhow, and H.V. Poor, "Distributed transmit beamforming: Challenges and recent progress," *IEEE Commun. Mag.*, vol. 47, no. 2, pp. 102-110, Feb 2009.
- [4] P. Shang, G. Zhu, L. Tan, G. Su, and T. Li, "Transmit antenna selection for the distributed MIMO systems," *International Conference on Networks Security, Wireless Communications and Trusted Computing*, 2009, pp. 449-453.
- [5] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecommun.*, vol. 10, no. 6, pp. 585C596, Nov. 1999.
- [6] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, no. 10, pp. 68-73, October 2004.
- [7] M. Gharavi-Alkhangari, A. B. Gershman, "Fast antenna subset selection in MIMO systems," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 339-347, Feb. 2004.
- [8] C.-S. Chang, *Performance Guarantees in Communication Networks*, Springer-Verlag London, 2000.
- [9] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. on Wireless Commun.*, vol. 2, no. 4, July 2003, pp. 630-643.
- [10] X. Zhang, J. Tang, H.-H. Chen, S. Ci, and M. Guizni, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," *IEEE Commun. Mag.*, pp. 100-106, Jan. 2006.
- [11] Q. Du and X. Zhang, "Cross-layer design based rate control for mobile multicast in cellular networks," in *Proc. IEEE GLOBECOM 2007*, Washington DC, USA, Nov. 26-30, 2007, pp. 5180-5184.
- [12] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed., John Wiley & Sons, Inc., 2006.
- [13] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058-3068, Aug. 2007.
- [14] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
- [15] T. S. Rappaport, *Wireless Communications: Principles & Practice*, Prentice Hall, 1996.
- [16] H. Weingarten, Y. Steinberg and S. Shamai "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936-3964, Sep. 2006.
- [17] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-Forcing methods for downlink spatial multiplexing in multisuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, Feb. 2004.
- [18] C. Meyer, *Matrix Analysis and Applied Linear Algebra*, Society for Industrial and Applied Mathematics, 2000.
- [19] Q. Du and X. Zhang, "QoS-Aware Base-Station Selections for Distributed MIMO Links in Broadband Wireless Networks," *Networking and Information Systems Labs., Dept. Electr. and Comput. Eng., Texas A&M Univ., College Station, Tech. Rep.* [Online.] Available: http://www.ece.tamu.edu/~xizhang/papers/distributed_mimo.pdf.
- [20] Z.-L. Zhang, *End-to-end support for statistical quality-of-service guarantees in multimedia networks*, Ph.D. dissertation, Department of Computer Science, University of Massachusetts Amherst, 1997.
- [21] J. Tang and X. Zhang, "Cross-Layer Modeling for Quality of Service Guarantees Over Wireless Links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4504-4512, Dec. 2007.
- [22] J. Wang and L.B. Milstein, "CDMA overlay situations for microcellular mobile communications," *IEEE Trans. Commun.*, vol. 43, no. 234, pp. 603-614, Feb./Mar./Apr. 1995.



Qinghe Du [S'09] received the B.S. and M.S. degrees from Xi'an Jiaotong University, China, in 2001 and 2004, respectively, both in electrical engineering, and the Ph.D. degree in computer engineering from Texas A&M University, College Station, Texas, USA, in 2010.

He joined Department of Information and Communications Engineering, Xian Jiaotong University, China, in 2011. His research interests include mobile wireless communications and networks with emphasis on mobile multicast, statistical QoS provisioning,

QoS-driven resource allocations, cognitive radio techniques, and cross-layer design over wireless networks. His work co-authored with his Ph.D. advisor Prof. Xi Zhang received the Best Paper Award in the IEEE GLOBECOM 2007 for the paper "Cross-Layer Design Based Rate Control for Mobile Multicast in Cellular Networks". He has published multiple papers in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE *Communications Magazine*, IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, etc. He is serving as the Publicity Chairs for IEEE GLOBECOM 2011 and ICST WICON 2011, respectively, and has served as a Publicity Co-Chair for the ICST QShine 2010. He has served as the Technical Program Committee (TPC) member for IEEE GLOBECOM, IEEE ICC, IEEE WCNC, IEEE PIMRC, etc.



Xi Zhang (S'89-SM'98) received the B.S. and M.S. degrees from Xidian University, Xi'an, China, the M.S. degree from Lehigh University, Bethlehem, PA, all in electrical engineering and computer science, and the Ph.D. degree in electrical engineering and computer science (Electrical Engineering-Systems) from The University of Michigan, Ann Arbor.

He is currently an Associate Professor and the Founding Director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station. He was an Assistant Professor and the Founding Director of the Division of Computer Systems Engineering, Department of Electrical Engineering and Computer Science, Beijing Information Technology Engineering Institute, China, from 1984 to 1989. He was a Research Fellow with the School of Electrical Engineering, University of Technology, Sydney, Australia, and the Department of Electrical and Computer Engineering, James Cook University, Australia, under a Fellowship from the Chinese National Commission of Education. He was with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hills, NJ, and with AT&T Laboratories Research, Florham Park, NJ, in 1997. He has published more than 190 research papers in the areas of wireless networks and communications systems, mobile computing, network protocol design and modeling, statistical communications, random signal processing, information theory, and control theory and systems.

Prof. Zhang received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He is an IEEE Communications Society Distinguished Lecturer. He received the Best Paper Awards in the IEEE WCNC 2010, IEEE GLOBECOM 2009, and the IEEE GLOBECOM 2007, respectively. He also received the TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering at Texas A&M University, College Station, in 2006. He is currently serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the special issue on "Broadband Wireless Communications for High Speed Vehicles", a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the special issue on "Wireless Video Transmissions", an Associate Editor for the IEEE COMMUNICATIONS LETTERS, a Guest Editor for the *IEEE Communications Magazine* for the special issue on "Advances in Cooperative Wireless Networking", a Guest Editor for the *IEEE Wireless Communications Magazine* for the special issue on "next generation of CDMA versus OFDMA for 4G wireless applications", an Editor for the JOHN WILEY'S JOURNAL ON WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, an Editor for the JOURNAL OF COMPUTER SYSTEMS, NETWORKING, AND COMMUNICATIONS, an Associate Editor for the JOHN WILEY'S JOURNAL ON SECURITY AND COMMUNICATIONS NETWORKS, an Area Editor for the ELSEVIER JOURNAL ON COMPUTER COMMUNICATIONS, and a Guest Editor for JOHN WILEY'S JOURNAL ON WIRELESS COMMUNICATIONS AND MOBILE COMPUTING for the special issue on "next generation wireless communications and mobile computing". He has frequently served as the Panelist on the U.S. National Science Foundation Research-Proposal Review Panels.

Prof. Zhang is serving or has served as the Technical Program Committee (TPC) Co-Chair for the IEEE INFOCOM 2013, the TPC Chair for the IEEE GLOBECOM 2011, Area TPC Chair for the IEEE INFOCOM 2012, the General Chair for IEEE ICC 2011 - Workshop on Advanced Networking Technologies for Smart-Services Based Clouding Computing, TPC Co-Chair for the IEEE ICDCS 2011 - Workshop on Data Center Performance, Panel/Demo/Poster Chairs for the ACM MobiCom 2011, TPC Vice-Chair for IEEE INFOCOM 2010, General Chair for the ACM QShine 2010, TPC Co-Chair for IEEE INFOCOM 2009 Mini-Conference, TPC Co-Chair for IEEE GLOBECOM 2008 - Wireless Communications Symposium, TPC Co-Chair for the IEEE ICC 2008 - Information and Network Security Symposium, Symposium Chair for IEEE/ACM International Cross-Layer Optimized Wireless Networks Symposium 2006, 2007, and 2008, respectively, the TPC Chair for IEEE/ACM IWCMC 2006, 2007, and 2008, respectively, the Demo/Poster Chair for IEEE INFOCOM 2008, the Student Travel Grants Co-Chair for IEEE INFOCOM 2007, the General Chair for ACM QShine 2010, the Panel Co-Chair for IEEE ICCCN 2007, the Poster Chair for IEEE/ACM MSWiM 2007 and IEEE QShine 2006, Executive Committee Co-Chair for QShine, the Publicity Chair for IEEE/ACM QShine 2007 and IEEE WirelessCom 2005, and the Panelist on the Cross-Layer Optimized Wireless Networks and Multimedia Communications at IEEE ICCCN 2007 and WiFi-Hotspots/WLAN and QoS Panel at IEEE QShine 2004. He has served as the TPC members for more than 90 IEEE/ACM conferences, including IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, IEEE WCNC, IEEE VTC, IEEE/ACM QShine, IEEE WoWMoM, IEEE ICCCN, etc.

Prof. Zhang is a Senior Member of the IEEE and a Member of the Association for Computing Machinery (ACM).