Toward Extra Large-Scale MIMO: New Channel Properties and Low-Cost Designs

Yu Han, *Member, IEEE*, Shi Jin, *Senior Member, IEEE*, Michail Matthaiou, *Fellow, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, and Chao-Kai Wen, *Senior Member, IEEE*

***Abstract*—Extra large-scale multiple-input–multiple-output (MIMO) has been recognized as one of the potential development directions of massive MIMO. By employing even more antennas than massive MIMO in the fifth-generation era, extra large-scale MIMO can further exploit the spatial domain resources and enable ultra-high data rates, low latency communications as well as emerging applications, such as sensing and localization, in sixth-generation mobile communication systems. However, with the increase of the size of the antenna array, and the decrease of the distance between a user and the array, new channel properties, that did not manifest in conventional massive MIMO, start to kick in. Most importantly, existing research strategies pertaining to massive MIMO cannot be directly applied or simply extended to fit the extra large-scale MIMO case. Moreover, increasing the number of antennas will inevitably boost the total cost, which refers to not only the high hardware cost, but also the burden of vast processing and computations as well as the substantial training overhead. In this article, we make a survey on the state-of-the-art on the new channel properties of and low-cost designs for extra large-scale MIMO systems. Particularly, we pursue a mathematical analysis to explain why the new features appear and illustrate how they affect the system model. Furthermore, we summarize and compare the low-cost designs from various perspectives and give our suggestions from a practical deployment point of view.**

***Index Terms*—Distributed processing, extra large-scale multiple-input–multiple-output (MIMO), low-cost architectures, low-overhead training, spatial nonstationarity, spherical wave, visibility region (VR).**

Manuscript received 7 January 2023; revised 20 April 2023; accepted 25 April 2023. Date of publication 5 May 2023; date of current version 8 August 2023. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62261160576, Grant 61921004, and Grant 62341107; in part by the National Research Foundation, Singapore; and in part by the Infocomm Media Development Authority under its Future Communications Research & Development Programme. The work of Michail Matthaiou was supported in part by the Research Grant from the Department for the Economy Northern Ireland through the US–Ireland R&D Partnership Programme and in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant 101001331. The work of Chao-Kai Wen was supported in part by the National Science and Technology Council of Taiwan under Grant NSTC 111-3114-E-110-001. *(Corresponding author: Shi Jin.)*

Yu Han and Shi Jin are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: hanyu@seu.edu.cn; jinshi@seu.edu.cn).

Michail Matthaiou is with the Centre for Wireless Innovation, Queen's University Belfast, BT3 9DT Belfast, U.K. (e-mail: m.matthaiou@qub.ac.uk).

Tony Q. S. Quek is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372, and also with the Department of Electronic Engineering, Kyung Hee University, Yongin 17104, South Korea (e-mail: tonyquek@sutd.edu.sg).

Chao-Kai Wen is with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (e-mail: chaokai.wen@mail.nsysu.edu.tw).

Digital Object Identifier 10.1109/JIOT.2023.3273328

I. Introduction

MASSIVE multiple-input–multiple-output (MIMO), also named as large-scale MIMO, has been a successful enabler to boost the data transmission rate in mobile communication systems in the fifth-generation (5G) era [1], and will keep serving as an important physical layer technology in future mobile communication systems. By employing tens or hundreds of antennas at the base station (BS), massive MIMO produces high spatial resolution and supports multiuser transmission on the same time–frequency resources. As the number of BS antennas grows unconventionally large, the multiuser interference and the uncorrelated noise diminish [2], providing preferable conditions for multiuser transmission. In order to further harness the gain caused by using more antennas, the concept of extra large-scale MIMO has been proposed for the sixth-generation mobile communications [3], [4], [5]. Extra large-scale MIMO employs hundreds or even thousands of antennas at the BS to simultaneously provide service to a certain set of users, and is an augmented version of massive MIMO.

In practical implementations, there are two deployment types of an extra large number of antennas, including the centralized type and the distributed type. The centralized type is a direct extension of 5G large-aperture arrays, where all the antennas are uniformly deployed in a co-located fashion, while we can sustain the half-wavelength distance between two adjacent antennas, forming an extra-large aperture array [6]. Alternatively, the antennas can be confined within a predetermined area, resulting in the concept of holographic MIMO [7]. If the practical environment does not allow the deployment of such a large array, then we can distribute the antennas across multiple sites, corresponding to the distributed type [8]. Each site is equipped with a small amount of antennas. These sites jointly serve a same set of users. A typical example is cell-free massive MIMO [9]. In this article, we focus on the centralized type with an extra large-aperture array.

Compared with distributed systems, synchronization among antennas is much easier in centralized extra large-scale MIMO systems. Moreover, extra large-aperture arrays can cover the external walls of buildings in populated city centres or be employed at stadiums/airports to provide wireless

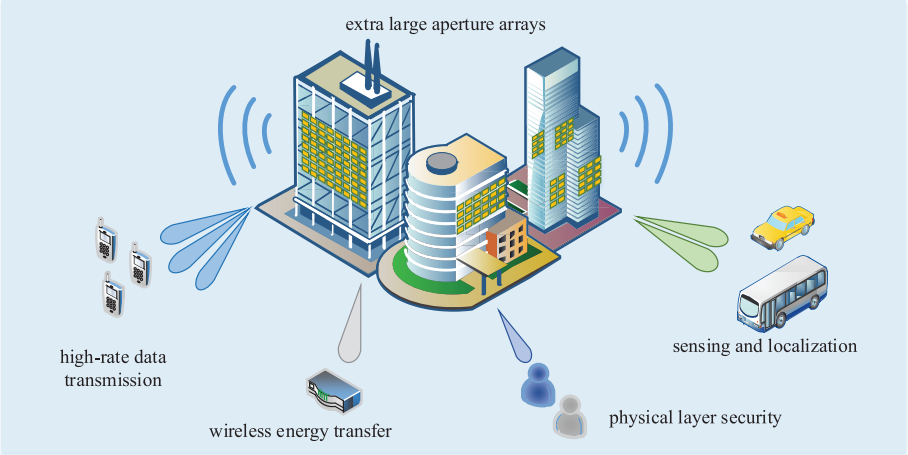

Fig. 1. Extra large-aperture arrays provide opportunities for high-rate data transmission, wireless energy transfer, physical-layer security, sensing, and localization.

communication services to a plethora of users. Therefore, in a centralized extra large-scale MIMO system, high beamforming gains can be harvested. Narrow beams with very low sidelobes can be generated by the extra large-aperture array and flexibly steered towards desired direction. Several orthogonal beams can be generated simultaneously, yielding an increase of the spatial–division multiplexing gain. Apart from satisfying the traditional requirements of high data rates, employing an extra large-aperture array enables new emerging applications. For instance, in indoor environments, such as in a factory, the autonomous driving of an electric car can be achieved by leveraging the high spatial resolution provided by such an array. There have been studies on extra large-aperture array-enabled new applications, including sensing and localization [10], [11], physical-layer security [12], [13], wireless energy transfer [14], [15], etc., as illustrated in Fig. 1.

Historically, the study of an emerging wireless architecture begins with the investigation of the propagation channel. Channel modeling of an extra large-aperture array system does not simply mean to expand the array size in a traditional MIMO channel model. With the increase of the array aperture, new channel properties kick in. First, the lower bound of the far field, known as the Rayleigh distance, is proportional to the array aperture. Considering that extra large-aperture arrays will generally be deployed in crowded urban or indoor factory environments, users will be close to the array. Different from traditional MIMO systems, where users are in the far field and signals experience plane wave propagation, in an extra large-aperture array system, there is a high probability that spherical waves will be created. Second, for users who are very close to the array, the pathloss between them fluctuates significantly across the array. If obstacles exist in the channel, then the channel power will be concentrated in a proportion of the array elements, known as the visibility region (VR). The spherical wave propagation and the existence of VR reflect the spatial nonstationarity of the channel. On the one hand, the new channel properties require new channel models for extra large-aperture array systems. On the other hand, these properties facilitate the above-mentioned new applications. Therefore, a deep and comprehensive study of the new channel properties is indispensable.

When translating a theoretical architecture into a commercial technology, the implementation and deployment costs are of pivotal importance. The employment of an extra large-aperture array entails the challenges of high hardware cost, high processing and computational complexity, and high training overhead. Regarding the hardware cost, a fully digital structure, where each active antenna is connected with a unique radio frequency (RF) chain, is unacceptably expensive when the number of active antennas grows large. Inspired by the low-cost designs in 5G millimeter wave systems, active antenna arrays with less RF chains can be adopted. Moreover, with the development of materials, extra large-aperture arrays can take the form of reconfigurable intelligent surfaces (RISs), which have the advantages of low cost and low power consumption. Therefore, the problem of high hardware cost can be tackled via different approaches.

In traditional MIMO systems with a limited number of antennas, signal processing, and computations are centralized at a common module, and the complexity is moderate. However, in an extra large-aperture array system, completely centralized processing and computations result in high complexity and are time consuming. In order to reduce the complexity and the processing latency, two approaches can be followed. One is to directly reduce the complexity of an algorithm in the centralized module. The other is to distribute the processing and computations to multiple local modules, thereby easing the burden in the centralized module. The distributed approach is more attractive, but the information exchange among the centralized module and the local modules affects the overall complexity and needs to be carefully assessed.

In a mobile communication system, an efficient transceiver design heavily depends on the precise knowledge of the wireless channel. The training overhead required to acquire the channel state information (CSI) usually increases with the number of antennas. Then, when an extra large-aperture array is deployed, the training overhead becomes substantial, which is evidently prohibitive for practical systems. Fortunately, the extra-large–dimensional channel shows directionality and sparsity in multiple domains. Traditional sparse channel estimation methods, such as compressed sensing, can be applied to reduce the training overhead. The directionality

of a spherical wave channel further supports localization and sensing. Further, the existence of the VR enables overhead reduction among multiple users. The feasibility of low-overhead communication and sensing, together with low-cost architectures and low-complexity processing and computations, guarantee the practical implementation of an extra large-aperture array.

This article makes a comprehensive survey on the new channel properties and the low-cost designs of extra large-scale MIMO systems. Section II investigates the spherical wave propagation by analyzing the channel responses on a point, an antenna, and an array step by step, and provides guidance to the selection of channel models in different fields/regions. With the analytical results on spherical waves, Section III explains why the VR appears and investigates the existing categories of VR and their definitions and models. The spatial nonstationarity is verified theoretically and further taken into account in the subsequent low-cost designs in Sections IV–VI. The low-cost architectures with active antenna arrays and RISs are illustrated in Section IV. A comparison of the hardware cost, implementation and synchronization difficulties, and scalability of different architectures is provided. Then, the low-complexity processing and computation designs are introduced in Section V. Existing methods to reduce the complexity in centralized and distributed processing structures are summarized. Finally, the low-overhead communication and sensing based on the directionality and channel sparsity in the transformation domains are studied in Section VI.

Notations: We use letters in normal fonts, lowercase, and uppercase letters in boldface for scalars, vectors, and matrices, respectively. The transpose, conjugate-transpose, and pseudo-inverse are indicated by the superscripts $(\cdot)^T$, $(\cdot)^H$, and $(\cdot)^\dagger$, respectively; $|\cdot|$ represents the absolute value of a scalar or the size of a set; $\|\cdot\|$ represents the modulus operation of a vector or a matrix; and $\mathbb{E}\{\cdot\}$ denotes expectation. For a matrix, $[\cdot]_{i,:}$, $[\cdot]_{:,j}$, and $[\cdot]_{i,j}$ return its ith row, the jth column, and the (i,j)th entry, respectively. The Hadamard and Kronecker products are denoted by $\odot$ and $\otimes$, respectively.

II. Spherical Wave

In a traditional MIMO system, the aperture of the BS antenna array is usually negligible when compared with the distance between it and a user served by the BS. The entire array can be regarded as one point. Thus, a signal sent from the user experiences an equal path loss and has a common angle-of-arrival (AoA) when arriving at different antennas of the BS array. Experiencing equal path loss and having a common AoA are two key features of a plane wave, which is typically modeled in the far-field region. However, when the aperture of the BS array grows large, the array cannot be regarded as one point any more. Then, spherical waves kick in and the plane wave model becomes irrelevant. In this section, we will make a comprehensive study on spherical waves.

A. Channel Response on Point

We start from the modeling of channel response. In a three-dimensional (3-D) free space, an isotropic point source

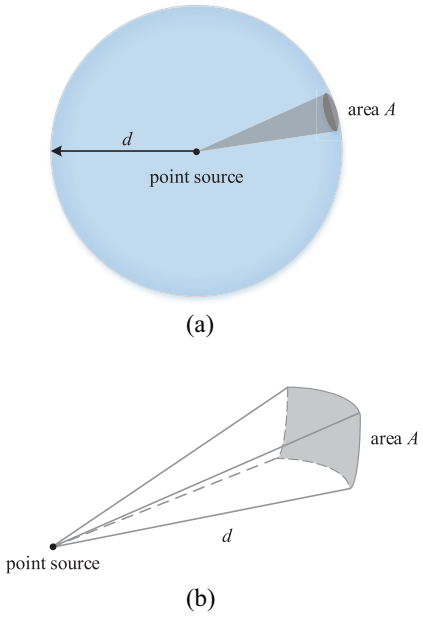

Fig. 2. (a) Isotropic point source radiates EM waves in all directions. (b) Receiver covers a continuous surface $\mathcal{A}$ on the sphere.

$\mathbf{s}$ is deployed at the origin of the coordinate system, i.e., $\mathbf{s} = [0,0,0]^T$, and radiates electromagnetic (EM) waves in all directions as the blue sphere shown in Fig. 2. For simplification, the transmit power of the point source is assumed uniform as 1. An antenna that covers a surface $\mathcal{A}$ with area A is located in the radiative field of $\mathbf{s}$, i.e., $\|\mathbf{p}-\mathbf{s}\| \gg \lambda$ holds for any point $\mathbf{p} = [x_p, y_p, z_p]^T \in \mathcal{A}$, where $\lambda = (c/f)$ is the wavelength of the EM wave with frequency f, while c is the speed of light. Based on the complexity of the models, we illustrate three channel response models that have been reported in the literature as follows.

1) Channel Response Model 1: The distance between the receiving point $\mathbf{p}$ and the source point $\mathbf{s}$ is $\|\mathbf{p}-\mathbf{s}\|$. At this distance, the power of the EM wave spreads uniformly on the sphere with radius $\|\mathbf{p}-\mathbf{s}\|$. Since the area of this sphere is $4\pi\|\mathbf{p}-\mathbf{s}\|^2$, the power on each point of this sphere equals [16]

$$\gamma(\mathbf{p},\mathbf{s}) = \frac{1}{4\pi\|\mathbf{p}-\mathbf{s}\|^2}. \tag{1}$$

Then, the channel response on point $\mathbf{p}$ can be expressed as follows:

$$\begin{aligned} h_{\mathrm{CR1}}(\mathbf{p},\mathbf{s}) &= \sqrt{\gamma(\mathbf{p},\mathbf{s})}e^{-j\frac{2\pi}{\lambda}\|\mathbf{p}-\mathbf{s}\|} \\ &= \frac{1}{\sqrt{4\pi}\|\mathbf{p}-\mathbf{s}\|}e^{-j\frac{2\pi}{\lambda}\|\mathbf{p}-\mathbf{s}\|} \end{aligned} \tag{2}$$

which is referred to as channel response model 1.

Model 1 describes an ideal case where the power on point $\mathbf{p}$ is perfectly and completely harvested. It requires that the normal direction of $\mathbf{p}$ with respect to surface $\mathcal{A}$, denoted as $\mathbf{v}_{\mathcal{A}}(\mathbf{p}) \in \mathbb{R}^{3\times 1}$ and satisfying $\|\mathbf{v}_{\mathcal{A}}(\mathbf{p})\| = 1$, exactly matches the radiation direction of the EM wave from source $\mathbf{s}$ [17]. As an example shown in Fig. 2(b), the surface $\mathcal{A}$ perfectly covers the sphere with radius $\|\mathbf{p}-\mathbf{s}\|$. Then, for any point $\mathbf{p} \in \mathcal{A}$, the normal line of $\mathbf{p}$ goes across the source $\mathbf{s}$, and the channel response model 1 is applicable.

2) Channel Response Model 2: In practice, patch antennas are widely utilized in mobile communication systems. Under this condition, the surface $\mathcal{A}$ of a patch antenna is a square. For a certain point $\mathbf{p} \in \mathcal{A}$, the normal direction does not always match the EM wave radiation direction. Then, the effective received power is a proportion of $\gamma(\mathbf{p}, \mathbf{s})$, and the proportionality factor is [18], [19], [20]

$$F(\mathbf{p}, \mathbf{s}) = \cos < \mathbf{p} - \mathbf{s}, \mathbf{v}_{\mathcal{A}}(\mathbf{p}) >= \frac{|(\mathbf{p} - \mathbf{s})^H \mathbf{v}_{\mathcal{A}}(\mathbf{p})|}{\|\mathbf{p} - \mathbf{s}\|} \quad (3)$$

satisfying $0 \le F(\mathbf{p}, \mathbf{s}) \le 1$. The expression of $F(\mathbf{p}, \mathbf{s})$ in (3) is a typical form of the *antenna pattern*.

The channel response on point $\mathbf{p}$ can be derived as follows:

$$\begin{aligned} h_{\mathrm{CR2}}(\mathbf{p}, \mathbf{s}) &= \sqrt{F(\mathbf{p}, \mathbf{s})} h_{\mathrm{CR1}}(\mathbf{p}, \mathbf{s}) \\ &= \sqrt{\frac{|(\mathbf{p} - \mathbf{s})^H \mathbf{v}_{\mathcal{A}}(\mathbf{p})|}{4\pi \|\mathbf{p} - \mathbf{s}\|^3}} e^{-j\frac{2\pi}{\lambda}\|\mathbf{p}-\mathbf{s}\|} \end{aligned} \quad (4)$$

which is referred to as channel response model 2. We see that when $F(\mathbf{p}, \mathbf{s}) = 1$ holds, model 2 is equivalent to model 1.

3) Channel Response Model 3: Papers [17], [21] considered the current density of the radiative EM waves from the source $\mathbf{s}$, which is written as follows:

$$\mathbf{J}(\mathbf{s}) = J_x(\mathbf{s})\mathbf{u}_x + J_y(\mathbf{s})\mathbf{u}_y + J_z(\mathbf{s})\mathbf{u}_z \quad (5)$$

where $\mathbf{u}_x = [1, 0, 0]^T$, $\mathbf{u}_y = [0, 1, 0]^T$, and $\mathbf{u}_z = [0, 0, 1]^T$ are the unit vectors along the x, y, and z directions, respectively, while $J_x(\mathbf{s})$, $J_y(\mathbf{s})$, and $J_y(\mathbf{s})$ represent the current density in the x, y, and z polarizations, respectively, satisfying the following normalization:

$$\|\mathbf{J}(\mathbf{s})\|^2 = |J_x(\mathbf{s})|^2 + |J_y(\mathbf{s})|^2 + |J_z(\mathbf{s})|^2 = 1. \quad (6)$$

Then, the effective received power at point $\mathbf{p}$ suffers further from the following proportionality factor:

$$\eta(\mathbf{p}, \mathbf{s}) = \left\| \left(\mathbf{I} - \frac{(\mathbf{p} - \mathbf{s})(\mathbf{p} - \mathbf{s})^H}{\|\mathbf{p} - \mathbf{s}\|^2} \right) \mathbf{J}(\mathbf{s}) \right\|^2. \quad (7)$$

As an example, [17] assumed that only the y direction is excited in $\mathbf{J}(\mathbf{s})$, which means $J_y(\mathbf{s}) = 1$ and $J_x(\mathbf{s}) = J_z(\mathbf{s}) = 0$. Under this condition

$$\eta(\mathbf{p}, \mathbf{s}) = 1 - \frac{[\mathbf{p} - \mathbf{s}]_2^2}{\|\mathbf{p} - \mathbf{s}\|^2}. \quad (8)$$

We see that $\eta(\mathbf{p}, \mathbf{s}) = 1$ happens when $[\mathbf{p} - \mathbf{s}]_2^2 = 0$, i.e., $y_p = 0$.

With $\eta(\mathbf{p}, \mathbf{s})$, the channel response on point $\mathbf{p}$ is written as follows:

$$\begin{aligned} h_{\mathrm{CR3}}(\mathbf{p}, \mathbf{s}) &= \sqrt{\eta(\mathbf{p}, \mathbf{s})} h_{\mathrm{CR2}}(\mathbf{p}, \mathbf{s}) \\ &= \left\| \left(\mathbf{I} - \frac{(\mathbf{p} - \mathbf{s})(\mathbf{p} - \mathbf{s})^H}{\|\mathbf{p} - \mathbf{s}\|^2} \right) \mathbf{J}(\mathbf{s}) \right\| \\ &\quad \times \sqrt{\frac{|(\mathbf{p} - \mathbf{s})^H \mathbf{v}_{\mathcal{A}}(\mathbf{p})|}{4\pi \|\mathbf{p} - \mathbf{s}\|^3}} e^{-j\frac{2\pi}{\lambda}\|\mathbf{p}-\mathbf{s}\|} \end{aligned} \quad (9)$$

which is referred to as channel response model 3.

B. Channel of Antenna

By integrating the response across the entire surface $\mathcal{A}$, the channel between the source and the receiver antenna that covers the surface $\mathcal{A}$ is calculated by [17]

$$h_{\mathcal{A}} = \frac{1}{\sqrt{A}} \int_{\mathbf{p} \in \mathcal{A}} h(\mathbf{p}, \mathbf{s}) d\mathbf{p}. \quad (10)$$

Here, we provide the following three examples from the literature to illustrate the channel response in different cases.

1) Case 1: In this case, the receiver antenna is isotropic and located at $\mathbf{p} = [0, 0, z_p]^T$. The effective area of an isotropic antenna is [16]

$$A_{\mathrm{iso}} = \frac{\lambda^2}{4\pi}. \quad (11)$$

Under channel response model 1, the channel between the source and the isotropic receiver antenna is derived as follows:

$$h_{\mathcal{A},\mathrm{case\ 1}} = \sqrt{A_{\mathrm{iso}}} h_{\mathrm{CR1}}(\mathbf{p}, \mathbf{s}) = \frac{\lambda}{4\pi |z_p|} e^{-j\frac{2\pi}{\lambda}|z_p|}. \quad (12)$$

Then, the free space path loss seen by an isotropic receiver antenna at distance z_p can be expressed as follows:

$$\mathrm{PL}_{fs} = |h_{\mathcal{A},\mathrm{case\ 1}}|^2 = \frac{\lambda^2}{16\pi^2 z_p^2} \quad (13)$$

which is in accordance with the model in [22].

2) Case 2: Case 2 illustrates a patch antenna whose surface $\mathcal{A}$ is a square plane. For any point $\mathbf{p} \in \mathcal{A}$, the normal direction $\mathbf{v}_{\mathcal{A}}(\mathbf{p})$ is orthogonal to the surface $\mathcal{A}$. The area A_{pat} satisfies

$$A_{\mathrm{pat}} \le \frac{\lambda^2}{4} \quad (14)$$

because the length and the width of the patch antenna are less than or equal to the antenna spacing ($\lambda/2$). Let $\mathcal{A}$ be parallel with the xy plane. Then, we have $\mathbf{v}_{\mathcal{A}}(\mathbf{p}) = \mathbf{u}_z = [0, 0, 1]^T$.

In [20], the channel on the patch antenna under channel response model 2 was studied. By applying (10), the channel can be approximated by

$$h_{\mathcal{A},\mathrm{case\ 2}} \approx \sqrt{eA_{\mathrm{pat}}} h_{\mathrm{CR2}}(\mathbf{p}_c, \mathbf{s}) \quad (15)$$

where eA_{pat} is the effective area of the antenna [20], $0 < e \le 1$ is the proportionality factor, while $\mathbf{p}_c = [x_c, y_c, z_p]^T$ is the center point of $\mathcal{A}$. Notably, for an isotropic antenna in case 1, its area A_{iso} in (11) is its effective area. In [19], the proportionality factor e was not considered; that is to say, A_{pat} was regarded as the effective area of the antenna. By applying (4), we obtain

$$h_{\mathrm{CR2}}(\mathbf{p}_c, \mathbf{s}) = \frac{|z_p|^{\frac{1}{2}}}{\sqrt{4\pi}\left(x_c^2 + y_c^2 + z_p^2\right)^{\frac{3}{4}}} e^{-j\frac{2\pi}{\lambda}\sqrt{x_c^2+y_c^2+z_p^2}}. \quad (16)$$

Then, the channel between the source and a patch antenna is

$$h_{\mathcal{A},\mathrm{case\ 2}} \approx \frac{\sqrt{eA_{\mathrm{pat}}}|z_p|^{\frac{1}{2}}}{\sqrt{4\pi}\left(x_c^2 + y_c^2 + z_p^2\right)^{\frac{3}{4}}} e^{-j\frac{2\pi}{\lambda}\sqrt{x_c^2+y_c^2+z_p^2}}. \quad (17)$$

If $x_c = y_c = 0$, then $h_{\mathrm{CR2}}(\mathbf{p}_c)$ turns to be

$$h_{\mathrm{CR2}}(\mathbf{p}_c, \mathbf{s}) = \frac{1}{\sqrt{4\pi}|z_p|} e^{-j\frac{2\pi}{\lambda}|z_p|} \tag{18}$$

which is equivalent to $h_{\mathrm{CR1}}(\mathbf{p}, \mathbf{s})$ in (2). Notably, $F_{\mathrm{pat}}(\mathbf{p}, \mathbf{s}) \leq 1$, and the equation only holds for $\mathbf{p} = \mathbf{p}_c$. Furthermore, when $eA_{\mathrm{pat}} = A_{\mathrm{iso}} = (\lambda^2/4\pi)$, we have

$$|h_{\mathcal{A},\mathrm{case}\,2}|^2 < eA_{\mathrm{pat}}|h_{\mathrm{CR2}}(\mathbf{p}_c, \mathbf{s})|^2 = \frac{\lambda^2}{16\pi^2 z_p^2} \tag{19}$$

which is exactly $|h_{\mathcal{A},\mathrm{case}\,1}|^2$.

3) Case 3: Case 3 studies a more complicated modeling of the channel on a patch antenna under channel response model 3, which was considered in [17] and [21]. The patch antenna with area A_{pat} in case 2 is also considered; however, the proportionality factor e related to the effective area is not introduced in case 3. The center point is $\mathbf{p}_c = [0, 0, z_p]^T$. Then, for any point $\mathbf{p} \in \mathcal{A}$, its x and y coordinates satisfy

$$-\frac{\sqrt{A_{\mathrm{pat}}}}{2} \leq x_p, y_p \leq \frac{\sqrt{A_{\mathrm{pat}}}}{2}. \tag{20}$$

The current density $\mathbf{J}(\mathbf{s}) = [0, 1, 0]^T$. Thus, (8) holds, i.e.,

$$\eta(\mathbf{p}, \mathbf{s}) = \frac{x_p^2 + z_p^2}{x_p^2 + y_p^2 + z_p^2}. \tag{21}$$

By applying (16) and (21), we have

$$\begin{aligned} h_{\mathrm{CR3}}(\mathbf{p}, \mathbf{s}) &= \sqrt{\eta(\mathbf{p}, \mathbf{s})} h_{\mathrm{CR2}}(\mathbf{p}, \mathbf{s}) \\ &= \frac{|z_p|^{\frac{1}{2}} \left(x_p^2 + z_p^2\right)^{\frac{1}{2}}}{\sqrt{4\pi}\left(x_p^2 + y_p^2 + z_p^2\right)^{\frac{5}{4}}} e^{-j\frac{2\pi}{\lambda}\left(x_p^2 + y_p^2 + z_p^2\right)^{\frac{1}{2}}}. \end{aligned} \tag{22}$$

Given (20) and (22), according to (10), the channel between source $\mathbf{s}$ and the patch antenna is more difficult to derive in case 3 than in case 2. Therefore, [17] provided an upper bound of the channel gain as follows:

$$|h_{\mathcal{A},\mathrm{case}\,3}|^2 = \left|\int_{\mathcal{A}} h_{\mathrm{CR3}}(\mathbf{p}, \mathbf{s})\, d\mathbf{p}\right|^2 \leq \frac{1}{\pi}\left(\frac{1}{3}\alpha + \frac{2}{3}\beta\right) \tag{23}$$

where

$$\alpha = \frac{\frac{A_{\mathrm{pat}}}{4}|z_p|}{\left(\frac{A_{\mathrm{pat}}}{4} + z_p^2\right)\left(\frac{A_{\mathrm{pat}}}{2} + z_p^2\right)^{\frac{1}{2}}} \tag{24}$$

and

$$\beta = \arctan\left(\frac{\frac{A_{\mathrm{pat}}}{4}}{|z_p|\left(\frac{A_{\mathrm{pat}}}{2} + z_p^2\right)^{\frac{1}{2}}}\right). \tag{25}$$

If $A_{\mathrm{pat}} = (\lambda^2/4)$, then

$$\begin{aligned} \alpha &= \frac{\frac{\lambda^2}{16}|z_p|}{\left(\frac{\lambda^2}{16} + z_p^2\right)\left(\frac{\lambda^2}{8} + z_p^2\right)^{\frac{1}{2}}} < \frac{\lambda^2}{16 z_p^2} \\ \beta &= \arctan\left(\frac{\frac{\lambda^2}{16}}{|z_p|\left(\frac{\lambda^2}{8} + z_p^2\right)^{\frac{1}{2}}}\right) < \frac{\lambda^2}{16 z_p^2}. \end{aligned} \tag{26}$$

Under this condition

$$|h_{\mathcal{A},\mathrm{case}\,3}|^2 \leq \frac{\lambda^2}{16\pi z_p^2} \tag{27}$$

and this upper bound is π times $|h_{\mathcal{A},\mathrm{case}\,2}|^2$ in (19) because the proportionality factor e is not considered in case 3. Notably, recalling (21), we have $\eta(\mathbf{p}, \mathbf{s}) < 1$, and $|h_{\mathrm{CR3}}(\mathbf{p}, \mathbf{s})| \leq |h_{\mathrm{CR2}}(\mathbf{p}, \mathbf{s})|$ holds for arbitrary $\mathbf{p} \in \mathcal{A}$. Therefore, the upper bound in (27) is not tight.

C. Field Partition of Antenna

According to (2), (4), and (9), the channel response varies at different points on the surface spanned by an antenna. The variance of the channel response across the surface differs when the antenna is at different locations with respect to the source point $\mathbf{s}$. If the antenna is close to $\mathbf{s}$, then the channel response variance is significant across the surface. If the antenna is far from $\mathbf{s}$, then $\mathcal{A}$ can be viewed as a point from the perspective of $\mathbf{s}$ and the channel response variance is negligible. Based on the magnitude of variance of both the amplitude and phase, the entire field of the source $\mathbf{s}$ can be divided into three fields/regions [21], [23].

1) *Near field*, in which both the amplitude and the phase variations of the channel response are nonnegligible across the surface.
2) *Fresnel region*, in which the amplitude variance of the channel response is negligible but the phase variance of the channel response is nonnegligible across the surface.
3) *Fraunhofer region*, also known as *far field*, in which both the amplitude and the phase variations of the channel response are negligible across the surface.

Some research works have considered the two-region partition by focusing only on the phase variance. In [24] and [25], the two regions are the Fresnel and the Fraunhofer regions, where the phase of channel response is dependent on and independent from the distance between transmitter and receiver, respectively. Another two-region partition can be found in [26], [27], [28], and [29], where the two regions were named as near and far fields, respectively. In the near field, a plane wavefront is created, whilst in the far field, a spherical wavefront is created.

1) Rayleigh/Fraunhofer Distance: The Rayleigh or Fraunhofer distance is the boundary between the Fresnel and the Fraunhofer regions or that between the near and the far field [21], [26], [27]. It is defined by the maximum phase variance of the channel response. The maximum phase variance cannot exceed $(\pi/8)$ [23] in the Fraunhofer region or far field; otherwise, the receiver is in the Fresnel region or near field of the source. From (2), (4), and (9), we see that at point $\mathbf{p}$, and regardless of the channel response model, the phase of the channel response equals

$$\angle h_{\mathrm{CR}}(\mathbf{p}, \mathbf{s}) = -\frac{2\pi}{\lambda}\|\mathbf{p} - \mathbf{s}\|. \tag{28}$$

Consider the widely used patch antenna in cases 2 and 3 as the receiver, whose surface is parallel with the xy plane and the center $\mathbf{p}_c$ is on the z-axis. Then, the maximum phase variance can be computed by comparing the channel responses at the center and one vertex of the surface, respectively.

Given the area A_{pat}, the coordinate of one vertex of $\mathcal{A}$ can be $\mathbf{p}_v = [(\sqrt{A_{\text{pat}}}/2), (\sqrt{A_{\text{pat}}}/2), z_p]^T$. At the Rayleigh distance, the phase difference between $h_{\text{CR}}(\mathbf{p}_c, \mathbf{s})$ and $h_{\text{CR}}(\mathbf{p}_v, \mathbf{s})$ satisfies

$$|\angle h_{\text{CR}}(\mathbf{p}_c, \mathbf{s}) - \angle h_{\text{CR}}(\mathbf{p}_v, \mathbf{s})| = \frac{\pi}{8} \tag{29}$$

which can be further rewritten as follows:

$$\|\mathbf{p}_v - \mathbf{s}\| - \|\mathbf{p}_c - \mathbf{s}\| = \sqrt{\frac{A_{\text{pat}}}{2} + z_p^2} - |z_p| = \frac{\lambda}{2\pi} \cdot \frac{\pi}{8} = \frac{\lambda}{16}. \tag{30}$$

Given that

$$\sqrt{1+x} \approx 1 + \frac{x}{2} \tag{31}$$

we have

$$\sqrt{\frac{A_{\text{pat}}}{2} + z_p^2} - |z_p| \approx \frac{A_{\text{pat}}}{4|z_p|}. \tag{32}$$

By applying (30) and (32), we obtain

$$|z_p| = \frac{4A_{\text{pat}}}{\lambda}. \tag{33}$$

The patch antenna can also be described by its aperture D_{pat}, which satisfies $D_{\text{pat}}^2 = 2A_{\text{pat}}$. Under this condition

$$|z_p| = \frac{2D_{\text{pat}}^2}{\lambda}. \tag{34}$$

Therefore, the Rayleigh distance is calculated by

$$d_{\text{Rayleigh}} = \frac{4A_{\text{pat}}}{\lambda} = \frac{2D_{\text{pat}}^2}{\lambda}. \tag{35}$$

2) Lower Bound of Fresnel Region: Papers [20], [21], and [23] introduced a lower bound of the Fresnel region, which is defined by the maximum amplitude variance of the channel response across the surface. Unlike the variance of the phase, which is captured by the difference, the variance of the amplitude is described by the ratio

$$\Gamma = \frac{\min_{\mathbf{p}\in\mathcal{A}} |h_{\text{CR}}(\mathbf{p}, \mathbf{s})|}{\max_{\mathbf{p}\in\mathcal{A}} |h_{\text{CR}}(\mathbf{p}, \mathbf{s})|}. \tag{36}$$

Denote the lower bound of the Fresnel region as d_{Fresnel}. At distance d_{Fresnel}, the amplitude ratio is equal to a threshold, i.e., $\Gamma = \Gamma_{\text{th}} \in (0, 1)$. The value of Γ_{th} can be $\cos(\pi/8)$ [21], [23], or 0.9^2 [20]. Below this threshold, the variance of amplitude is nonnegligible across the surface. In [23], d_{Fresnel} was regarded as the boundary between the Fresnel region and near-field region. In [21], when $d_{\text{Fresnel}} < d_{\text{Rayleigh}}$ holds, the region between these two boundaries was named as the Fresnel region.

We still consider the patch antenna above. The amplitude of the channel response has different expressions when different models are applied. According to (2), (4), and (9)

$$\begin{aligned} \min_{\mathbf{p}\in\mathcal{A}} |h_{\text{CR}}(\mathbf{p}, \mathbf{s})| &= \left|h_{\text{CR}}(\mathbf{p}_v, \mathbf{s})\right| \\ \max_{\mathbf{p}\in\mathcal{A}} |h_{\text{CR}}(\mathbf{p}, \mathbf{s})| &= \left|h_{\text{CR}}(\mathbf{p}_c, \mathbf{s})\right|. \end{aligned} \tag{37}$$

Under channel response model 1

$$|h_{\text{CR1}}(\mathbf{p}, \mathbf{s})| = \frac{1}{\sqrt{4\pi}\|\mathbf{p} - \mathbf{s}\|} \tag{38}$$

and we have

$$\frac{\|\mathbf{p}_c - \mathbf{s}\|}{\|\mathbf{p}_v - \mathbf{s}\|} = \frac{|z_p|}{\sqrt{\frac{A_{\text{pat}}}{2} + z_p^2}} = \Gamma_{\text{th}}. \tag{39}$$

By further applying (36), we obtain

$$d_{\text{Fresnel, CR1}} = |z_p| = \sqrt{\frac{A_{\text{pat}}\Gamma_{\text{th}}^2}{2(1-\Gamma_{\text{th}}^2)}} = \frac{D_{\text{pat}}}{2}\sqrt{\frac{\Gamma_{\text{th}}^2}{1-\Gamma_{\text{th}}^2}}. \tag{40}$$

If $\Gamma_{\text{th}} = \cos(\pi/8)$, then $d_{\text{Fresnel, CR1}} \approx 1.2D_{\text{pat}}$ as given in [21], [23]. Under channel response model 2

$$h_{\text{CR2}}(\mathbf{p}, \mathbf{s}) = \sqrt{\frac{|(\mathbf{p}-\mathbf{s})^H \mathbf{v}_{\mathcal{A}}(\mathbf{p})|}{4\pi\|\mathbf{p}-\mathbf{s}\|^3}}. \tag{41}$$

Recalling $\mathbf{v}_{\mathcal{A}}(\mathbf{p}) = \mathbf{u}_z$, we derive that

$$\frac{\|\mathbf{p}_c - \mathbf{s}\|^{\frac{3}{2}}}{\|\mathbf{p}_v - \mathbf{s}\|^{\frac{3}{2}}} = \Gamma_{\text{th}}. \tag{42}$$

Compared with (39), d_{Fresnel} under channel model 2 satisfies

$$d_{\text{Fresnel, CR2}} = \sqrt{\frac{A_{\text{pat}}\Gamma_{\text{th}}^{\frac{4}{3}}}{2\left(1-\Gamma_{\text{th}}^{\frac{4}{3}}\right)}} = \frac{D_{\text{pat}}}{2}\sqrt{\frac{\Gamma_{\text{th}}^{\frac{4}{3}}}{1-\Gamma_{\text{th}}^{\frac{4}{3}}}} \tag{43}$$

as derived in [20]. Similarly, under channel response model 3, by directly applying (22), it can be obtained that

$$\frac{d_{\text{Fresnel, CR3}}^{\frac{3}{2}}\left(\frac{A_{\text{pat}}}{4} + d_{\text{Fresnel, CR3}}^2\right)^{\frac{1}{2}}}{\left(\frac{A_{\text{pat}}}{2} + d_{\text{Fresnel, CR3}}^2\right)^{\frac{5}{4}}} = \Gamma_{\text{th}}. \tag{44}$$

Generally, we have

$$d_{\text{Fresnel, CR3}} > d_{\text{Fresnel, CR2}} > d_{\text{Fresnel, CR1}}. \tag{45}$$

The lower bound of the Fresnel region can be alternatively calculated since the concept of near field is not unique. A *Fresnel distance* which equals $0.62\sqrt{(D_{\text{pat}}^3/\lambda)}$ is defined as the lower bound of the Fresnel region [24], [25], and this distance was also regarded as the upper bound of the reactive near field in [21].

D. Field Partition of Array

The field partition of a single antenna can be extended to that of a multiantenna array [20], [21]. Consider a widely applied uniform plane array (UPA) at the receiver. The UPA is composed of $N_h \times N_v$ antennas, where N_h and N_v are the numbers of columns and rows which are assumed to be even numbers. The distance between two horizontal or vertical adjacent antennas is $(\lambda/2)$. The UPA is parallel with the xy plane. The center of the UPA is $\mathbf{p}_c = [0, 0, d]^T$, where $d > 0$ is the distance between the source and the UPA. In an extreme case that the antennas are seamlessly deployed as shown in

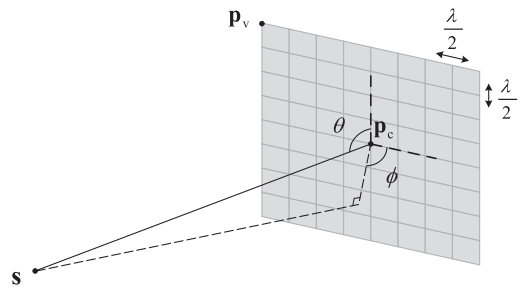

Fig. 3. Example of a UPA and its geometrical relation with the source.

Fig. 3, the entire antenna array can be regarded as a large patch antenna with area $A_{\text{UPA}} = N_h N_v A_{\text{pat}}$, where $A_{\text{pat}} = (\lambda^2/4)$. Then, one vertex of the UPA is at

$$\mathbf{p}_v = \left[\frac{N_h\lambda}{4}, \frac{N_v\lambda}{4}, d\right]^T. \tag{46}$$

The aperture of the UPA is

$$D_{\text{UPA}} = \frac{\sqrt{N_h^2 + N_v^2}\lambda}{2}. \tag{47}$$

1) Rayleigh/Fraunhofer Distance: The Rayleigh or Fraunhofer distance of the UPA is still defined by the maximum phase variance across the array, which equals $(\pi/8)$. Recalling (28) and (29), we can write that

$$\|\mathbf{p}_v - \mathbf{s}\| - \|\mathbf{p}_c - \mathbf{s}\| = \sqrt{\left(N_h^2 + N_v^2\right)\frac{\lambda^2}{16} + d^2} - d = \frac{\lambda}{2\pi}\cdot\frac{\pi}{8}. \tag{48}$$

Given (31) and after some derivations, we obtain the Rayleigh distance of the UPA as follows:

$$d_{\text{Rayleigh}} = \frac{\left(N_h^2 + N_v^2\right)\lambda}{2} = \frac{2D_{\text{UPA}}^2}{\lambda} \tag{49}$$

which is still determined by the aperture.

2) Lower Bound of Fresnel Region: Following a similar approach as in the single-antenna case, we further study the lower bound of the Fresnel region of the UPA. Under channel response model 1, by applying (39), we get that

$$\frac{\|\mathbf{p}_c - \mathbf{s}\|}{\|\mathbf{p}_v - \mathbf{s}\|} = \frac{d}{\sqrt{\left(N_h^2 + N_v^2\right)\frac{\lambda^2}{16} + d^2}} = \Gamma_{\text{th}}. \tag{50}$$

Then, the lower bound of the Fresnel region is

$$d_{\text{Fresnel, CR1}} = \frac{\lambda}{4}\sqrt{\frac{\Gamma_{\text{th}}^2\left(N_h^2 + N_v^2\right)}{\left(1 - \Gamma_{\text{th}}^2\right)}} = \frac{D_{\text{UPA}}}{2}\sqrt{\frac{\Gamma_{\text{th}}^2}{1 - \Gamma_{\text{th}}^2}}. \tag{51}$$

Comparing (51) with (40), we see that d_{Fresnel} of the UPA can be easily calculated by applying the aperture D_{UPA} to the expression of d_{Fresnel} of an antenna. Thus, under channel response model 2, we have

$$d_{\text{Fresnel, CR2}} = \frac{D_{\text{UPA}}}{2}\sqrt{\frac{\Gamma_{\text{th}}^{\frac{4}{3}}}{1 - \Gamma_{\text{th}}^{\frac{4}{3}}}}. \tag{52}$$

TABLE I
EXAMPLES OF d_{RAYLEIGH} AND d_{FRESNEL}

	8×8 at 3.5 GHz	512×64 at 28 GHz
D_{UPA}	0.49 m	2.76 m
d_{Rayleigh}	5.49 m	1.426×10^3 m
$d_{\text{Fresnel,CR1}}$	0.59 m	3.34 m
$d_{\text{Fresnel,CR2}}$	0.73 m	4.14 m

Both d_{Rayleigh} and d_{Fresnel} increase with the aperture D_{UPA}. We take two typical examples in mobile communication systems to illustrate the partition of the radiative field under typical array apertures. The first example considers an 8×8 UPA working at 3.5 GHz. The second example involves a 512×64 UPA working at 28 GHz. Table I provides the values of D_{UPA}, d_{Rayleigh}, $d_{\text{Fresnel, CR1}}$, and $d_{\text{Fresnel, CR2}}$ in the two examples under the condition of $\Gamma_{\text{th}} = \cos(\pi/8)$.

For Example 1, the aperture of a small-scale array is limited. Then, the values of d_{Rayleigh} and d_{Fresnel} are small. When the array is employed at a 5G new radio (NR) BS, whose serving cell has a width of 200 m, it is very likely that users in the cell are beyond the Rayleigh distance of the array, being in the far-field region. However, for Example 2, although the wavelength of a millimeter wave is small, the aperture of a 512×64 UPA is much larger than the small-scale UPA in example 1. The Rayleigh distance of this extra-large array becomes 1.426×10^3 m, which is much larger than the size of the serving cell. Then, users are no longer in the far-field region of the array. For some users, their distances from the UPA may even be smaller than d_{Fresnel}.

E. Modeling of Channel Between Source and Array

Now, we study the channel model between the point source and the UPA. Denote by $\mathbf{H} \in \mathcal{C}^{N_h \times N_v}$ the channel matrix and $h(n_h, n_v)$ as the channel on the (n_h, n_v)th antenna, i.e.,

$$\mathbf{H} = \begin{bmatrix} h\left(-\frac{N_h}{2}, -\frac{N_v}{2}\right) & \cdots & h\left(\frac{N_h}{2} - 1, -\frac{N_v}{2}\right) \\ \vdots & \ddots & \vdots \\ h\left(-\frac{N_h}{2}, \frac{N_v}{2} - 1\right) & \cdots & h\left(\frac{N_h}{2} - 1, \frac{N_v}{2} - 1\right) \end{bmatrix}. \tag{53}$$

Based on the distance between the source and the UPA, which is denoted by d, three models of $h(n_h, n_v)$ can be derived [20], [30].

1) Channel Model 1: This model is for the region $d < d_{\text{Fresnel}}$. Considering that in this region, the channel's amplitude and phase variations are nonnegligible across the array, $h(n_h, n_v)$ in model 1 will have different amplitude and phase expressions for different (n_h, n_v). That is to say

$$h_{\text{CM1}}(n_h, n_v) = |h(n_h, n_v)|e^{j\angle h(n_h, n_v)} \tag{54}$$

which can be obtained by applying the geometrical information of antenna (n_h, n_v) in (10). Channel model 1 is referred to as the *spherical wave* channel model.

2) Channel Model 2: This model is for the region of $d_{\text{Fresnel}} \le d \le d_{\text{Rayleigh}}$, where the variance of amplitude is negligible across the array. A same $|h(n_h, n_v)|$ is shared by all (n_h, n_v) and is simplified by $|h|$, whose value can be assigned

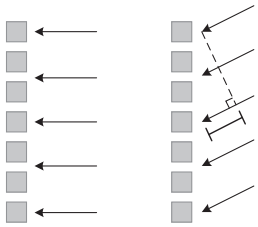

Fig. 4. Plane wave arrives at an array. (a) No phase difference exists across the array. (b) Phase difference is introduced across the array.

by $|h(n_h, n_v)|, \forall(n_h, n_v)$. Here, we select $|h| = |h(0,0)|$. Model 2 of $h(n_h, n_v)$ is then expressed as follows:

$$h_{\text{CM2}}(n_h, n_v) = |h| e^{j\angle h(n_h, n_v)}. \tag{55}$$

Channel model 2 is referred to as the *reduced spherical wave* channel model.

3) Channel Model 3: This model is for the region of $d > d_{\text{Rayleigh}}$, where the variations of amplitude and phase are both negligible across the array. A common $|h(n_h, n_v)|$ is still applied here. Moreover, a uniformed value $\angle h$ is shared by all (n_h, n_v). Similarly, we set $\angle h = \angle h(0,0)$. Model 3 of $h(n_h, n_v)$ is written as follows:

$$h_{\text{CM3}}(n_h, n_v) = |h| e^{j\angle h}. \tag{56}$$

We should note that according to (56), all the antennas of the UPA experience the same channel with no difference among them. This stems from the UPA orientation, which is parallel with the xy plane, while its center is $\mathbf{p}_c = [0, 0, d]^T$. That is to say, the source is exactly on the normal line of the UPA which goes across the UPA center. Then, no path difference exists when the wave arrives at different antennas, and thus no phase difference is introduced among the channels on these antennas, as illustrated in Fig. 4(a).

The model in (56) also means that the incident wave seen by each antenna comes from the same direction. That is, a plane wave instead of a spherical wave is experienced at the UPA. Hence, channel model 3 is referred to as the *plane wave* channel model. Consider a more general case that the UPA is parallel with the xy plane and $\mathbf{p}_c = [x_c, y_c, z_c]^T$. As shown in Fig. 3, the included angle between the incident wave and a column of the UPA is

$$\theta = \arccos \frac{y_c}{\sqrt{x_c^2 + y_c^2 + z_c^2}}. \tag{57}$$

The included angle between the projection of the incident wave on the UPA and a row of the UPA is

$$\phi = \arccos \frac{x_c}{\sqrt{x_c^2 + y_c^2}}. \tag{58}$$

The position of the (n_h, n_v)th antenna is

$$\mathbf{p}_{n_h,n_v} = \left[x_c + \frac{2n_h + 1}{4}\lambda, y_c + \frac{2n_v + 1}{4}\lambda, z_c \right]^T$$
$$n_h = -\frac{N_h}{2}, \ldots, \frac{N_h}{2} - 1, n_v = -\frac{N_v}{2}, \ldots, \frac{N_v}{2} - 1. \tag{59}$$

Model 3 of $h(n_h, n_v)$ becomes

$$h_{\text{CM3}}(n_h, n_v) = |h| e^{j(\angle h + \Delta\phi(n_h, n_v))} \tag{60}$$

under the plane incident wave from direction (θ, ϕ), where

$$\Delta\phi(n_h, n_v) = \pi (n_h \cos\theta + n_v \sin\theta \cos\phi) \tag{61}$$

is the difference between $\angle h(n_h, n_v)$ and $\angle h$ caused by the path different shown in Fig. 4(b). If $x_c = y_c = 0$, then $\theta = \phi = (\pi/2)$ and $\Delta\phi(n_h, n_v) = 0$, which corresponds to the case in (56).

Channel model 3 has been widely applied in the fourth-generation and 5G systems, since the aperture of antenna arrays is not large and users are in the far-field region of the array [31], [32], [33]. However, in 6G systems, extra large-aperture arrays, such as example 2 in Table I, will be employed. Then, users probably fall in the near-field region or the Fresnel region, and channel models 1 and 2 should be utilized. The presence of spherical waves instead of plane waves is one of the major unique characteristics of extra large-scale MIMO systems. Thus, far-field channel models will become inaccurate in the practical near or Fresnel field [11].

III. Visibility Region

When a user is very close to an extra large-aperture array, most of the channel power can be captured by only a part of the array. This part of the array is referred to as the VR of the user w.r.t. the array. The VR is another key characteristic in extra large-scale MIMO systems [3], [6], [34], [35]. In this section, we will make a comprehensive study on the origins, definition, and modeling of the VR.

A. Origins of the VR

The VR reflects the uneven distribution of the channel power over the array. There are two major manifestations behind the creation of the VR [6]. One is the unequal path loss across different antennas of the array. The other is the blockage stemming from the obstacles between the user and the array.

1) Unequal Path Loss: When the distance between a user and the array is below d_{Fresnel}, channel model 1 should be applied. Under this condition, $|h(n_h, n_v)|$, which reflects the path loss on the antenna array, varies significantly with (n_h, n_v). Let us revisit the UPA in Fig. 3, which is parallel with the xy plane while its center is at $\mathbf{p}_c = [0, 0, d]^T$, satisfying $d < d_{\text{Fresnel}}$. The source $\mathbf{s}$ is still located at the origin of the coordinate system. We analyze the value of $|h(n_h, n_v)|$ across the UPA under the three cases in Section II-B. For antenna (n_h, n_v) in case 1, by applying $\mathbf{p}_{n_h,n_v}$ in (12), we obtain

$$|h_{\text{case 1}}(n_h, n_v)| \propto \frac{1}{\|\mathbf{p}_{n_h,n_v} - \mathbf{s}\|}. \tag{62}$$

With d fixed and given (59), $\|\mathbf{p}_{n_h,n_v} - \mathbf{s}\|$ has the minimum value at $n_h = n_v = 0$ and the maximum value at $n_h = -(N_h/2)$, $n_v = -(N_v/2)$, which are the center and the vertex of the UPA, respectively. The minimum and the

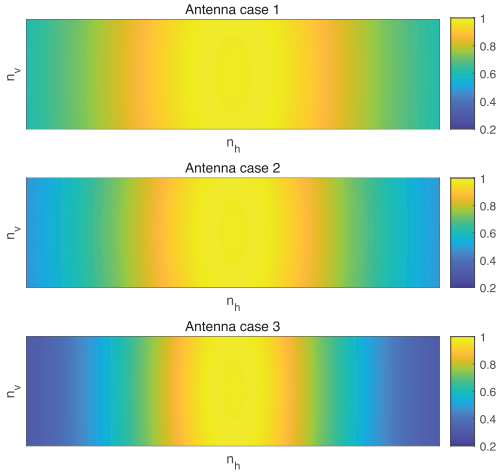

Fig. 5. Comparison of $(|h(n_h, n_v)|/[\max|h(n_h, n_v)|])$ in antenna cases 1, 2, and 3 when $N_h = 512$, $N_v = 64$, and $d = 100\lambda$.

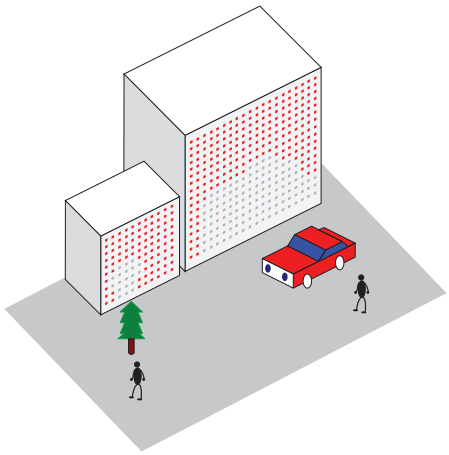

Fig. 6. Blockage of the channels on part of the array caused by obstacles, such as trees and cars. Red and gray squares represent the antennas whose channels are connected and blocked, respectively.

maximum values of $|h_{\text{case 1}}(n_h, n_v)|$ satisfy

$$\frac{\min|h_{\text{case 1}}(n_h, n_v)|}{\max|h_{\text{case 1}}(n_h, n_v)|} = \frac{\|\mathbf{p}_c - \mathbf{s}\|}{\|\mathbf{p}_v - \mathbf{s}\|} = \frac{d}{\sqrt{d^2 + \frac{D_{\text{UPA}}^2}{4}}}. \quad (63)$$

Recalling (51) and Table I, we know that the value of d_{Fresnel} is close to D_{UPA}. If $d = D_{\text{UPA}}$, then the ratio equals 0.89. With the decrease of d, the ratio drops. For an extra large-scale array which widely spreads over a long wall, it is possible that $d = D_{\text{UPA}}/10$. Then, the ratio approximates 0.2. If the source moves on the xy plane, then the ratio of the minimum and the maximum of $|h_{\text{case 1}}(n_h, n_v)|$ is further reduced.

A similar phenomenon can be observed when the UPA is customized as described in cases 2 and 3. Considering that

$$0 < F(\mathbf{p}_v, \mathbf{s}) < F(\mathbf{p}_c, \mathbf{s}) \le 1 \quad (64)$$
$$0 < \eta(\mathbf{p}_c, \mathbf{s}) < \eta(\mathbf{p}_c, \mathbf{s}) \le 1 \quad (65)$$

we have

$$\frac{\min|h_{\text{case 3}}(n_h, n_v)|}{\max|h_{\text{case 3}}(n_h, n_v)|} \le \frac{\min|h_{\text{case 2}}(n_h, n_v)|}{\max|h_{\text{case 2}}(n_h, n_v)|} \le \frac{\min|h_{\text{case 1}}(n_h, n_v)|}{\max|h_{\text{case 1}}(n_h, n_v)|}. \quad (66)$$

This phenomenon can be observed in Fig. 5, where the source is very close to the center of the array. We see that in case 3, the value of $|h(n_h, n_v)|$ is significantly larger for smaller $|n_h|$. In an extreme case that the ratio of the minimum and the maximum of $|h(n_h, n_v)|$ approaches zero, the channel power on a proportion of antennas in the array is negligible; for example, the first and the last columns of the array in case 3 of Fig. 5. The channel power can be captured by a proportion of antennas in the array. Particularly, the channel power is concentrated on the antennas that are close to the source. The channels on the antennas that are much farther to the source are significantly weaker.

2) Blockage Due to Obstacles: An extra large-aperture array can be widely spread on the wall of a building in an urban city. Then, D_{UPA} will be large. Normally, users are usually crowded in an urban environment and may be very close to the array. Trees, cars, and infrastructures can be seen everywhere and can all be possible obstacles in the channel between the array and a certain user.

Unlike in the far field, where the entire channel is blocked, in the near field or the Fresnel region, only a part of the array may be blocked. The blocked part of the array is determined by the geometry of the array, user, and obstacle, as shown in Fig. 6. Assume that only a line-of-sight (LoS) path exists. For antenna (n_h, n_v), if the line between $\mathbf{p}_{n_h,n_v}$ and $\mathbf{s}$ goes across the obstacle, then the channel on antenna (n_h, n_v) is blocked, i.e., $|h(n_h, n_v)| = 0$. The blocked part of the array also reflects the shape of the obstacle. As illustrated in Fig. 6, the blocked subarrays in gray color take identical patterns with the tree and the car, respectively. The uneven channel power distribution caused by blockage is independent of that resulting from the unequal path loss.

B. Definition of the VR

Uneven power distribution across the array is a new channel feature that appears when extra large-aperture arrays are employed in wireless systems. Then, the VR of a user w.r.t. the array is introduced to model the uneven channel power distribution [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45]. Actually, the VR is not a novel concept. In this section, we introduce different VR categories.

1) VR of User w.r.t. the Array: In the literature, the VR of a user w.r.t. the array is defined as the part of the array that captures the biggest proportion of the channel power over the entire array [6], [34], [35], [36]. It reflects the sparsity of a user channel in the antenna domain. Denote the VR of a user w.r.t. the array as Φ_{UA}. Then, Φ_{UA} is a set that contains the indices of antennas that the channel power of this user is concentrated on. The following property holds:

$$\frac{\sum_{(n_h,n_v)\in\Phi_{\text{UA}}}|h(n_h, n_v)|^2}{\|\mathbf{H}\|_F^2} \ge \zeta \quad (67)$$

where $0 < \zeta \le 1$ is a threshold with value close to 1. Note that Φ_{UA} contains the minimum number of antennas that satisfy the requirement of (67). The size of Φ_{UA} is denoted by $|\Phi_{\text{UA}}|$.

We first consider the channel under unequal path loss but without blockage. The VR caused simply by a spherical wavefront covers a continuous part of the array. Recall the example in Fig. 5, where Φ_{UA} covers the middle part of the UPA. Note that $\zeta = 1$ only holds when Φ_{UA} covers the entire array. However, when we set $\zeta < 1$, we can still find a proper Φ_{UA} to achieve (67). The Φ_{UA} obtained here is the size of a sliding window that covers the antennas in the array that captures ζ percentage of the channel power. Antennas out of the window still receive nonnegligible power and can be excluded from Φ_{UA}.

The VR can be obvious if a blockage occurs. At the first antenna in the blocked subarray, a sharp decrease of the channel power can be observed. Consider now the LoS channel case without any non-LoS (NLoS) paths. Then, the channel power on each antenna in the blocked subarray is zero. In (67), $\zeta = 1$ can be achieved even though $|\Phi_{\mathrm{UA}}| < N_h N_v$. Under this condition, Φ_{UA} contains the antennas that are not blocked by obstacles. If we set $\zeta < 1$, then the size of Φ_{UA} can be even smaller by discarding the antennas with the smallest power. Notably, a VR caused by blockage may not be continuous. One obstacle may block the channels on a continuous subarray. If the blocked subarray is at the array center, then the VR will not be continuous. If multiple obstacles exist, the VR may be composed of several discontinuous subarrays.

2) Two-Tier VRs: In the previous context, we focused on the case that only the LoS path exists in the channel. In practice, the wireless propagation environment is composed of various scatterers. Signals can be scattered and then arrive at the array along NLoS paths as well. Unlike the obstacles that block the signal propagation, scatterers provide new propagation paths and act as intermediate nodes. Then, the one-tier user–array channel becomes a two-tier user–scatterer and scatterer–array channel. Accordingly, the VR of a user w.r.t. the array is further partitioned by the VR of a scatterer w.r.t. the array and the VR of a user w.r.t. the scatterers [6], [39], [46], [47], [48], [49].

The scatterers are usually grouped into multiple clusters. Each cluster includes one or multiple neighboring scatterers. Scatterers in a cluster see the same antennas in the array and can be simultaneously observed by a user. The VR of a cluster w.r.t. the array, denoted by Φ_{CA}, contains the antennas that can be seen by the cluster. This definition is similar to that of the VR of a user w.r.t. the array. A cluster here corresponds to the user above; Φ_{CA} is also named as the cluster VR and is assumed to cover a continuous subarray [47]. The central antenna in Φ_{CA} has the highest channel power [47]. Furthermore, if Φ_{CA} includes all the array antennas, then the scatterers in this cluster are referred to as entirely visible scatterers; otherwise, they are referred to as partially visible scatterers [46].

The VR of a user w.r.t. the clusters, denoted by Φ_{UC}, contains the clusters that can be seen by the user. This is similar to the original concept of VR in COST channel models, which refers to a geometric region where a same set of scatterer clusters can be seen if the user is in this region [50]. If the user moves to another position, then the clusters that can be seen by this user vary. Note that, Φ_{UC} is also named as user VR in [47].

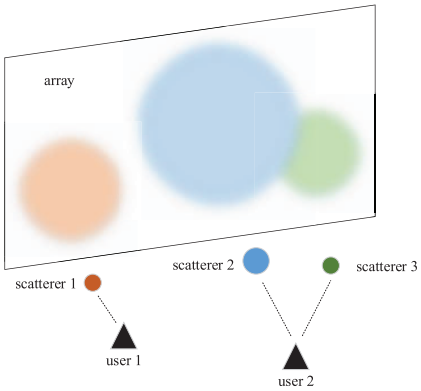

Fig. 7. Example of the two-tier VRs. A circular subarray in a particular color represents the VR of the cluster in the same color.

By cascading the two-tier VRs, the VR of a user w.r.t. the array can be obtained. For user k, its one-tier VR and two-tier VRs have the following relation:

$$\Phi_{\mathrm{UA},k} = \bigcup_{c \in \Phi_{\mathrm{UC},k}} \Phi_{\mathrm{CA},c} \tag{68}$$

where the one-tier VR $\Phi_{\mathrm{UA},k}$ denotes the VR of user k w.r.t. the array, while the two-tier VRs $\Phi_{\mathrm{UC},k}$ and $\Phi_{\mathrm{CA},c}$ represent the VR of user k w.r.t. the clusters and the VR of cluster c w.r.t. the array, respectively. As an example in Fig. 7, $\Phi_{\mathrm{UC},1} = \{1\}$ and $\Phi_{\mathrm{UC},2} = \{2, 3\}$. Thus, we have $\Phi_{\mathrm{UA},1} = \Phi_{\mathrm{CA},1}$ and $\Phi_{\mathrm{UA},2} = \Phi_{\mathrm{CA},2} \bigcup \Phi_{\mathrm{CA},3}$.

C. Channel Modeling With VR

Now, we investigate the modeling of a channel with VR. According to (67), the VR almost harvests the total power of channel. To simplify the expression, only the channel in VR is modeled to be nonzero, and the channel out of the VR is assumed to be zero. There have been channel models that capture the new feature of VR [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [49], where the VR is described in different ways.

1) Channel Covariance Matrix With VR: A channel covariance matrix reflects the statistical covariance of channels across different antennas. It has been widely applied in the modeling of multiantenna channels. When the channel experiences correlated Rayleigh fading, the channel between the single-antenna user k and the N-dimensional array can be modeled as [51], [52], [53]

$$\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_k) \tag{69}$$

where $\mathbf{h}_k \in \mathbb{C}^{N \times 1}$ is the multiantenna complex channel with zero mean, and $\mathbf{R}_k \in \mathbb{C}^{N \times N}$ is the channel covariance matrix satisfying

$$\mathbf{R}_k = \mathbb{E}\{\mathbf{h}_k \mathbf{h}_k^H\}. \tag{70}$$

This model is equivalent to

$$\mathbf{h}_k = \mathbf{R}_k^{\frac{1}{2}} \mathbf{h}_{w,k} \tag{71}$$

where $\mathbf{h}_{w,k} \in \mathbb{C}^{N\times 1}$ is the small-scale fading coefficient vector, whose entries are independent, identically distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance. In traditional multiantenna systems, the diagonal entries of $\mathbf{R}_k$ are nonzero. However, if the VR is introduced, then only the diagonal entries in the VR are nonzero [35], [36], [37], [38], [40], [41], [42], [44]. Moreover, $\mathbf{R}_k$ shows block sparsity. For $n_1, n_2 \in [1, N]$, if $n_1 \notin \Phi_{\mathrm{UA},k}$ or $n_2 \notin \Phi_{\mathrm{UA},k}$, then $\mathbf{R}(n_1, n_2) = 0$. In a typical case that $\Phi_{\mathrm{UA},k}$ covers a continuous region, $\mathbf{R}_k$ has the following structure:

$$\mathbf{R}_k = \begin{bmatrix} \mathbf{0} & & \\ & \mathbf{R}_{\mathrm{UA},k} & \\ & & \mathbf{0} \end{bmatrix} \tag{72}$$

where $\mathbf{R}_{\mathrm{UA},k} \in \mathbb{C}^{|\Phi_{\mathrm{UA},k}|\times|\Phi_{\mathrm{UA},k}|}$ is the covariance submatrix with nonzero entries. Given the block sparsity of $\mathbf{R}_k$, the channel model (71) can be further rewritten as follows:

$$\mathbf{h} = \mathbf{D}_{\mathrm{UA},k}\mathbf{R}_{\mathrm{UA},k}^{\frac{1}{2}}\mathbf{h}_{\mathrm{UA},w,k} \tag{73}$$

where $\mathbf{D}_{\mathrm{UA},k} = \{0, 1\}^{N\times|\Phi_{\mathrm{UA},k}|}$ is a selection matrix that selects the antennas in $\Phi_{\mathrm{UA},k}$, while the dimension of $\mathbf{h}_{\mathrm{UA},w,k}$ is $|\Phi_{\mathrm{UA},k}| \times 1$.

When scatterer clusters are further considered, the scatterers can be regarded as a virtual antenna array [46]. In traditional multiantenna systems, the covariance matrix-based scattering channel model is [52]

$$\mathbf{h}_k = \mathbf{R}_A^{\frac{1}{2}}\mathbf{H}_w\mathbf{R}_S^{\frac{1}{2}}\mathbf{h}_{w,k} \tag{74}$$

where $\mathbf{R}_A \in \mathbb{C}^{N\times N}$ and $\mathbf{R}_S \in \mathbb{C}^{S\times S}$ are the covariance matrices at the array and the scatterer side, respectively, S is the number of scatterers, and $\mathbf{H}_w \in \mathbb{C}^{N\times S}$ and $\mathbf{h}_{w,k} \in \mathbb{C}^{S\times 1}$ are small-scale fading matrices (vectors). In extra large-aperture array systems, since different clusters of scatterers have different VRs, (74) needs to be rewritten. Assume that the number of clusters is C. In cluster c, there are S_c scatterers, satisfying $\sum_{c=1}^{C} S_c = S$. The total number of scatterers that can be seen by user k is $\tilde{S}_k = \sum_{c\in\Phi_{\mathrm{UC},k}} S_c$. By cascading the array–scatterer channel and the scatterer-user channel together, the channel between the array and user k is expressed as follows:

$$\mathbf{h} = [\mathbf{G}_1, \ldots, \mathbf{G}_C]\mathbf{R}_S^{\frac{1}{2}}\mathbf{D}_{\mathrm{UC},k}\mathbf{h}_{w,k} \tag{75}$$

where $\mathbf{D}_{\mathrm{UC},k} = \{0, 1\}^{S\times\tilde{S}_k}$ is the selection matrix that selects the scatterers that can be seen by user k, $\mathbf{h}_{w,k}\mathbb{C}^{\tilde{S}_k\times 1}$ is the small-scale fading vector, while

$$\mathbf{G}_c = \mathbf{D}_{\mathrm{CA},c}\mathbf{R}_{\mathrm{CA},c}^{\frac{1}{2}}\mathbf{H}_{w,c} \in \mathbb{C}^{N\times S_c} \tag{76}$$

is the separate channel between the array and cluster c, $\mathbf{D}_{\mathrm{CA},c} = \{0, 1\}^{N\times|\Phi_{\mathrm{CA},c}|}$ selects the antennas that can be seen by cluster c, $\mathbf{R}_{\mathrm{CA},c} \in \mathbb{C}^{|\Phi_{\mathrm{CA},c}|\times|\Phi_{\mathrm{CA},c}|}$ is the covariance matrix across antennas within $\Phi_{\mathrm{CA},c}$, while $\mathbf{H}_{w,c} \in \mathbb{C}^{|\Phi_{\mathrm{CA},c}|\times S_c}$ models the small-scale fading. This model has been applied in [46], [47], and [48].

Channel covariance matrix-based channel models pave the way for the analysis of key performance indicators, such as the signal-to-interference and noise (SINR) [35] and the ergodic capacity [46], which further helps the design of transceivers.

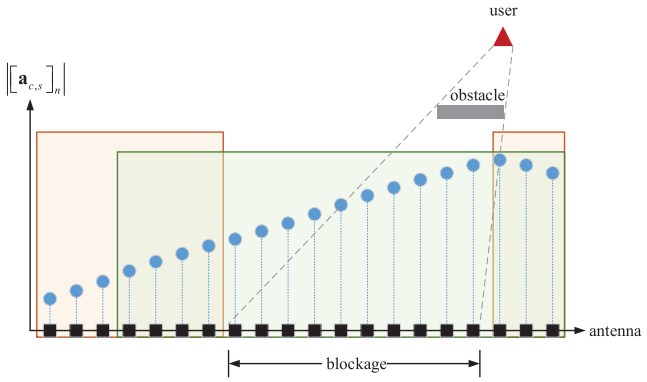

Fig. 8. Example of different VR masks $\mathbf{p}_c$ with and without blockage.

2) Steering Vectors With VR: The discrete physical model is another widely used multiantenna channel model [54], [55]. It focuses on the distinguished paths in the environment. The discrete physical channel model is expressed as follows:

$$\mathbf{h}_k = \sum_{c\in\Phi_{\mathrm{UC},k}} \sum_{s=1}^{S_c} \beta_{c,s}\mathbf{a}_{c,s} \tag{77}$$

where $\beta_{c,s}$ is the complex coefficient of the path resulting from scatterer s in cluster c, which also represents the response of this path on the reference antenna, and $\mathbf{a}_{c,s} \in \mathbb{C}^{N\times 1}$ is the steering vector of the path that involves the difference of response on each antenna w.r.t. the reference antenna. In traditional multiantenna systems, the plane wave channel model 3 in (60) is utilized to construct the steering vector $\mathbf{a}_{c,s}$, satisfying $[\mathbf{a}_{c,s}]_n = e^{j\Delta\phi_n}$, where $\Delta\phi_n$ is expressed in (61). Each element of $\mathbf{a}_{c,s}$ has amplitude equal to 1.

In extra large-aperture array systems, when introducing the concept of VR, the limited dimensional channel model becomes [39], [49]

$$\mathbf{h}_k = \sum_{c\in\Phi_{\mathrm{UC},k}} \sum_{s=1}^{S_c} \beta_{c,s}\mathbf{a}_{c,s} \odot \mathbf{p}_c \tag{78}$$

where $\mathbf{p}_c = \{0, 1\}^{N\times 1}$ is the VR mask vector of cluster c with the following entries:

$$[\mathbf{p}_c]_n = \begin{cases} 1, & \text{if } n \in \Phi_{\mathrm{CA},c} \\ 0, & \text{else.} \end{cases} \tag{79}$$

The steering vector with VR mask, i.e., $\mathbf{a}_{c,s} \odot \mathbf{p}_c$, can be regarded as the effective steering vector. Notably, when an extra large-aperture array is deployed, $\mathbf{a}_{c,s}$ has the forms of the spherical wave channel models 1 and 2 in (54) and (55). In fact, when applying channel model 1, the entries in the steering vector $\mathbf{a}_{c,s}$ have different amplitudes, directly reflecting the VR caused by unequal path loss.

Depending on whether blockage happens or not, the VR mask $\mathbf{p}_c$ should be set in two different ways. Take an example in Fig. 8. The amplitude of $\mathbf{a}_{c,s}$ varies significantly across the array as shown by the blue circles. If an obstacle exists, part of the array is blocked; then, $\mathbf{p}_c$ covers the red windows where the blockage does not take effect. However, if there is no obstacle, then $\mathbf{p}_c$ selects the green window that captures the majority of power with the minimum window size. Notably, in the case

of no blockage, $\mathbf{p}_c$ can be an all-one vector, contributing to a precise extra large-aperture array channel model. Introducing a zero-one mask vector will result in an approximated channel model with a reduced dimension, which further helps to reduce the complexity of transceiver design.

D. Spatial Nonstationarity

The spherical wave propagation as well as the VR caused by blockages contribute to spatial nonstationarity, which is the new channel property that appears in extra large-aperture array systems. The concept of spatial stationarity of a multiantenna channel is derived from the wide sense stationarity of a stochastic process [56], where the stochastic process becomes the multiantenna channel $\mathbf{h}$ here. Note that the spatial stationarity of a multiantenna channel is different from the stationarity of a time-varying channel [57]. The multiantenna channel $\mathbf{h}$ is spatially stationary if the correlation of any two distinct elements of $\mathbf{h}$ only depends on the difference of the two-element indices. That is to say

$$\mathbb{E}\left\{[\mathbf{h}]^*_{l+m}[\mathbf{h}]_{l+n}\right\} = \mathbb{E}\left\{[\mathbf{h}]^*_{m}[\mathbf{h}]_{n}\right\} \tag{80}$$

holds for arbitrary $l \in [0, N-1]$. Otherwise, the multiantenna channel $\mathbf{h}$ is spatially nonstationary.

If the VR of the user w.r.t. the array does not cover the entire array, then the channel is definitely spatially nonstationary. This is because

$$\mathbb{E}\{[\mathbf{h}]_n\}\begin{cases} > 0, & \text{if } n \in \Phi_{\text{UA}} \\ = 0, & \text{else} \end{cases} \tag{81}$$

holds regardless of which channel model from (73), (76), and (78) is applied. Thereafter, we have

$$\mathbb{E}\left\{[\mathbf{h}]^*_{m}[\mathbf{h}]_{n}\right\}\begin{cases} > 0, & \text{if } m, n \in \Phi_{\text{UA}} \\ = 0, & \text{else.} \end{cases} \tag{82}$$

The requirement for spatial stationarity cannot be satisfied when $\Phi_{\text{UA}} \subsetneq [1, N]$.

If the VR of the user w.r.t. the array covers the entire array, but the user is in the near field or Fresnel region of the array, then the multiantenna channel $\mathbf{h}$ is still spatially nonstationary [58]. Under this condition, the channel models (54) and (55) should be utilized. More specifically, when applying (54), $\mathbb{E}\{[\mathbf{h}]_n\}$ has unequal amplitudes for different n due to the unequal path loss. Furthermore, the phase of $[\mathbf{h}]_n$ is dependent on the index n whenever (54) or (55) is applied. An equal phase difference between two adjacent channel entries cannot be supported. Thus, $\mathbb{E}\{[\mathbf{h}]^*_{l+m}[\mathbf{h}]_{l+n}\}$ is dependent on the particular l, m and n, instead of m-n. Spatial nonstationarity of a near-field channel has been mathematically verified in [59] and experimentally observed through measurements in [60] and [61].

IV. Low-Cost Extra Large-Aperture Array Architectures

The new channel properties brought by an extra large-aperture array will inform the hardware and transceiver design. The multiantenna arrays used in traditional systems do not have a large size, and a fully digital architecture is widely employed to connect each active antenna with a unique RF chain. However, with the increase of the antenna array size, the fully digital architecture with high resolution will be expensive and not suitable for practical applications. Low-cost architecture designs are of great importance for the commercial deployment of extra large-aperture arrays. Moreover, for an active antenna array, each antenna is driven by a power amplifier (PA) or a low noice amplifier (LNA) and has the ability to transmit and receive wireless signals. Thereafter, the power consumption of an active antenna array is usually large as well. Fortunately, the new channel features provide room for cost reduction. By jointly considering the hardware and power cost as well as the new channel properties, in this section, we will introduce the potential low-cost extra large-aperture array architectures.

A. Active Arrays With Less RF Chains

Research in this type of architectures originates in the beginning of the 5G era [62], [63], [64], [65], [66], [67], [68], [69], [70], [71]. A large array with massive active antennas is controlled by a small amount of RF chains. The numbers of active antennas and RF chains are denoted as N and N_{RF}, respectively, satisfying $N \gg N_{\text{RF}}$. One RF chain is connected to one or multiple antennas and controls them through RF devices, such as phase shifters (PSs) and/or switches. After analog processing, such as analog beamforming, combining, and selection in the RF module, a base band (BB) processing is further applied among the signals on these RF chains. Therefore, a hybrid RF and BB structure is modeled as follows:

$$\mathbf{r} = \mathbf{F}_{\text{BB}}\mathbf{F}_{\text{RF}}\mathbf{y} \tag{83}$$

where $\mathbf{y} \in \mathbb{C}^{N\times 1}$ is the signal received at the antennas, $\mathbf{F}_{\text{RF}} \in \mathbb{C}^{N_{\text{RF}}\times N}$ is the RF processing matrix, $\mathbf{F}_{\text{BB}} \in \mathbb{C}^{K\times N_{\text{RF}}}$ is the BB processing matrix, K is the number of data streams, and $\mathbf{r} \in \mathbb{C}^{K\times 1}$ is used for signal detection. The format of $\mathbf{F}_{\text{RF}}$ is determined by the type of connections among the RF chains and antennas as well as the type of RF devices on each connection.

1) Connection Type: The connection type directly determines the hardware cost, transceiver design, transmission performance, as well as the scalability of the architecture. Generally, there are two main types. One is the *single-RF chain single-antenna* type, and the other is the *single-RF chain multiple-antenna* type [69], [70].

1) *Single-RF Chain Single Antenna:* When this connection type is adopted, a single RF chain can be only connected to a single antenna. A switch is required at each RF chain to enable *antenna selection*, that is, to determine whether this RF chain is activated and which antenna it is connected with. If the RF chain is activated, then only one antenna will be connected with it. A total of N_{RF} switches are deployed. No PSs are needed because beamforming is solely implemented at the BB module. Antenna selection can be further categorized into two types, including *full array selection* and *partial array selection*. Full array selection enables an RF chain to connect with any antenna in the array. Partial array

TABLE II
ARCHITECTURES OF ACTIVE ARRAYS WITH LESS RF CHAINS THAT HAVE APPEARED IN EXISTING STUDIES

single-RF chain single-antenna		single-RF chain multiple-antennas				
full array selection	partial array selection	full array connection		partial array connection		
		PSs	ON/OFF switches	fixed subarray		dynamic subarray
				PSs	ON/OFF switches	
i	*ii*	*iii*	*iv*	*v*	*vi*	*vii*

selection means that each RF chain can select from a subarray which is physically closest to it. For a certain RF chain, the partial array for antenna selection is usually *fixed*, and the size of the partial array is determined by the sweeping space of the switch at the RF chain side. Partial arrays corresponding to different RF chains can be disjoint or overlapped. If two RF chains select antennas from a same partial array, then their selection strategy needs to be different.

2) *Single-RF Chain Multiple Antennas:* When applying this connection type, a single RF chain can be connected with multiple antennas. Signal combination or beamforming is achieved at the RF module, and then the array gain can be harvested. Most studies focus on this connection type.

Similar to antenna selection, one RF chain can be connected with the full array or a partial array close to it, corresponding to the *full array connection* structure and *the partial array connection* structure, respectively. In the full array connection structure, each antenna can be connected with all the RF chains and vice versa. A unique physical link is established between each RF chain and each antenna. In each link, a PS can be deployed at the antenna side to enable analog beamforming, or an ON/OFF switch can be deployed at the antenna side to reduce the cost and achieve a simple signal combination. A total of $N_{\mathrm{RF}}N$ PSs or ON/OFF switches are required in the full array connection structure.

In the partial array connection structure, one RF chain can be connected with a proportion of antennas, but one antenna can be connected with only one RF chain. For a certain RF chain, the partial array that can be connected with is *fixed* or *dynamic*. In the former case, a physical link exists between the RF chain and each antenna in the partial array. In each link, a PS or an ON/OFF switch can be deployed at the antenna side as well. The size of each partial array or subarray is fixed, and a total of N PSs or ON/OFF switches are required in the fixed subarray structure. In the latter case, apart from these PSs or ON/OFF switches, an extra switch is employed at each antenna to determine which RF chain it will be connected with. Notably, unlike the ON/OFF switch, this switch is used for RF chain selection and it sweeping space covers all the RF chains. No switch is further needed at the RF chain side for antenna selection. The size of each subarray can be adjusted in a real-time manner. This structure is more suitable for extra large-aperture array systems under spatial nonstationarity.

2) Component Type: Now, we turn our attention on the three component mentioned above, including the PS, the ON/OFF switch, and the switch for selection.

1) *PS:* A PS can adjust the phase of an RF signal. It is a key enabler of analog beamforming in multiantenna systems. When PSs are deployed, the RF matrix $\mathbf{F}_{\mathrm{RF}}$ is called the analog beamforming matrix, contributing to the hybrid beamforming structure together with the BB precoding. However, the cost of a PS is analogous to its operating frequency, as well as its resolution.
2) *ON/OFF Switch:* An ON/OFF switch can be turned ON or OFF to determine whether the signal can pass through the connection. When a switch is in the physical link between one RF chain and one antenna, this connection can be activated or inactivated by choosing the ON and OFF status, respectively. The cost of an ON/OFF switch is significantly lower than that of a PS, but the insertion loss is a major problem.
3) *Switch for selection:* A switch for selection has a sweeping space and can be connected to one of the physical links in this sweeping space. It can be deployed at the RF chain side to achieve antenna selection, or be deployed at the antenna side to make RF chain selection. A switch for selection is more expansive than an ON/OFF switch.

3) State-of-the-Art Architectures: The various connection types and device types can jointly form many different combinations, each corresponding to a particular architecture. Here, we introduce the architectures that have appeared in existing studies, which are listed in Table II, and make an analysis on their signal model, advantages, and drawbacks.

a) Single-RF chain single antenna in full array selection: This is the traditional antenna selection architecture as shown in Fig. 9 (i). In this architecture, we have $[\mathbf{F}_{\mathrm{RF}}]_{i,j} \in \{0, 1\}$ for $i = 1, \ldots, N_{\mathrm{RF}}, j = 1, \ldots, N$ and

$$0 \leq \sum_{j=1}^{N} [\mathbf{F}_{\mathrm{RF}}]_{i,j} \leq 1 \ \forall i$$
$$0 \leq \sum_{i=1}^{N_{\mathrm{RF}}} [\mathbf{F}_{\mathrm{RF}}]_{i,j} \leq 1 \ \forall j. \tag{84}$$

Since the sweeping space of a switch is confined, this architecture is widely adopted in traditional multiantenna systems due to the limited array size. However, when employing an extra large-aperture array, it may be impractical to find a switch that could be connected to all antennas in a massive array.

b) Single-RF chain single antenna in partial array selection: This architecture is more easily implemented in an extra large-aperture array system. Considering the scalability issue as well, a subarray-based antenna selection architecture is naturally considered. As shown in Fig. 9(ii), the entire array

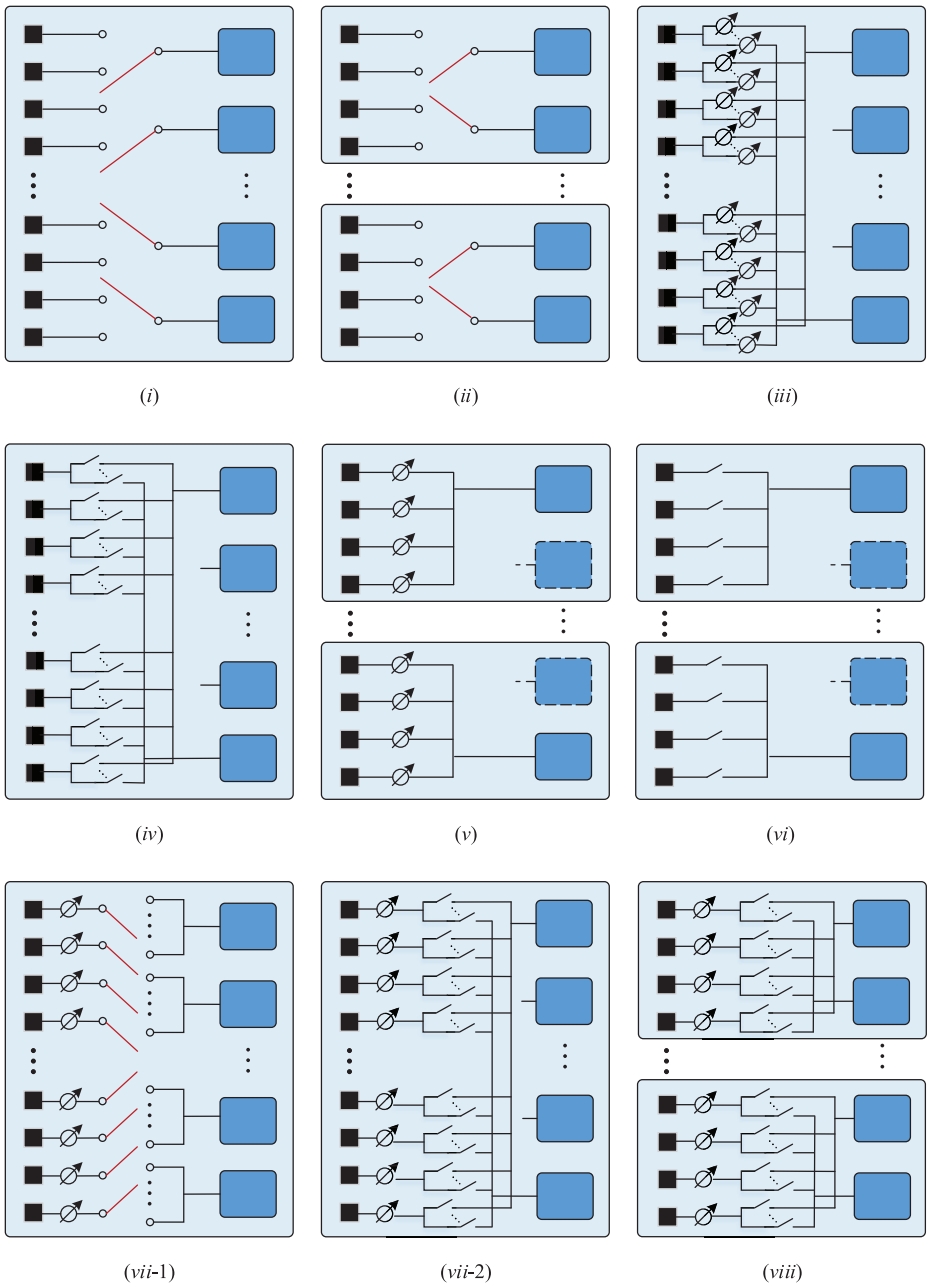

Fig. 9. Architectures of active arrays with less RF chains that have appeared in existing studies (i)–(vii) versus the proposed double layer architecture (viii).

is composed of multiple subarrays. Each subarray has completely the same topology, including the number of antennas and the number of RF chains. Denote the number of subarrays as B. Then, each subarray has (N/B) antennas and (N_{RF}/B) RF chains. One RF chain can select no more than one antenna within the same subarray, and one antenna can be selected by no more than one RF chain within the same subarray. Thus, $\mathbf{F}_{\mathrm{RF}}$ has a block diagonal structure

$$\mathbf{F}_{\mathrm{RF}} = \begin{bmatrix} \mathbf{F}_{\mathrm{RF},1} & & \\ & \ddots & \\ & & \mathbf{F}_{\mathrm{RF},B} \end{bmatrix} \tag{85}$$

where $\mathbf{F}_{\mathrm{RF},b} \in \{0,1\}^{(N_{\mathrm{RF}}/B)\times(N/B)}$ is the RF submatrix of the bth subarray, satisfying

$$0 \le \sum_{j=1}^{\frac{N}{B}} \left[\mathbf{F}_{\mathrm{RF},b}\right]_{i,j} \le 1 \ \forall i$$
$$0 \le \sum_{i=1}^{\frac{N_{\mathrm{RF}}}{B}} \left[\mathbf{F}_{\mathrm{RF},b}\right]_{i,j} \le 1 \ \forall j. \tag{86}$$

When applying this architecture, multiple identical subarrays can be directly combined together to construct an extra large-aperture array. This scalability facilitates the design, fabrication and production of the array. Moreover, local antenna selection within a single subarray is supported, thereby giving room for complexity reduction. However, the selected antennas are usually discontinuous and cannot cover a continuous VR. Thus, the array gain will be compromised.

c) Single-RF chain multiple antennas in full array connection with PSs: This is the widely studied full-connection hybrid beamforming architecture in 5G millimeter wave systems [65], [66], [67], [68], [69], [70] and has been considered in the extra large-aperture array system as in [29]. As shown in Fig. 9(iii), each RF chain is connected with all antennas through PSs. The RF matrix $\mathbf{F}_{\mathrm{RF}}$ has the following format:

$$\mathbf{F}_{\mathrm{RF}} = \begin{bmatrix} \mathbf{f}_{\mathrm{RF},1}^{T} \\ \vdots \\ \mathbf{f}_{\mathrm{RF},N_{\mathrm{RF}}}^{T} \end{bmatrix} \tag{87}$$

where $\mathbf{f}_{\mathrm{RF},i} \in \mathbb{C}^{N\times 1}$ is the analog beamforming vector in the ith RF chain with $[\mathbf{f}_{\mathrm{RF},i}]_j = e^{j\phi_{i,j}}$, and $\phi_{i,j} \in [0, 2\pi]$ is the phase shift introduced by the PS in the physical link between RF chain i and antenna j. In sparse channel conditions, the performance of this architecture can be very close to that of the fully digital architecture. However, this architecture has the following drawbacks. First of all, both the cost and the energy consumption of $N_{\mathrm{RF}}N$ PSs are high. Second, with the increase of the array size, the length of the transmission line that connects the antenna array edges grows, and the transmission latency differs significantly across the array. Then, the synchronization across antennas in an RF chain becomes problematic. Third, this architecture lacks scalability. If the array is expanded and more antennas are added to the array, then an equal amount of components need to be added to each RF chain as well, and hence, the structure of each RF chain will change. Finally, integrating such a large number of PSs in an RF module is difficult. Therefore, this architecture is not recommended for extra large-aperture array systems.

d) Single-RF chain multiple- antennas in full array connection with ON/OFF switches: This architecture is a reduced version of architecture *iii* by replacing the expensive PSs with low-cost ON/OFF switches as illustrated in Fig. 9(iv). The RF matrix $\mathbf{F}_{\mathrm{RF}}$ sustains the format in (87). The difference is that $[\mathbf{f}_{\mathrm{RF},i}]_j \in \{0, 1\}$ for $i = 1, \ldots, N_{\mathrm{RF}}$ and $j = 1, \ldots, N$. Note that this architecture is not subject to the antenna selection constraints in (84). It can simultaneously achieve antenna selection and dynamic partial array connection. However, it also entails integration, synchronization, and scalability challenges.

e) Single-RF chain multiple antennas in fixed partial array connection with PSs: This is the well known subarray hybrid beamforming architecture [63], [64], [68], [69], [70]. In Fig. 9(v), each subarray has equal size with only one RF chain and (N/N_{RF}) antennas. The RF matrix $\mathbf{F}_{\mathrm{RF}}$ also has a block diagonal structure

$$\mathbf{F}_{\mathrm{RF}} = \begin{bmatrix} \mathbf{f}_{\mathrm{RF},1}^{T} & & \\ & \ddots & \\ & & \mathbf{f}_{\mathrm{RF},N_{\mathrm{RF}}}^{T} \end{bmatrix} \tag{88}$$

where $\mathbf{f}_{\mathrm{RF},i} \in \mathbb{C}^{(N/N_{\mathrm{RF}})\times 1}$ is the RF vector in the ith subarray with $[\mathbf{f}_{\mathrm{RF},i}]_j = e^{j\phi_{i,j}}$ for $i = 1, \ldots, N_{\mathrm{RF}}$ and $j = 1, \ldots, (N/N_{\mathrm{RF}})$. Given an equal number of RF chains, the performance of this architecture is inferior to that of architecture *iii*. However, the number of PSs in this architecture is much smaller, thereby the cost is greatly reduced. Moreover, the synchronization problem ceases to exist, since antennas connected with the same RF chain are within a single subarray, whose size is usually limited. Besides, the array size can be easily scaled up by using more subarrays. For all these reasons, this architecture finally managed to earn a commercial deployment opportunity in 5G.

f) Single-RF chain multiple antennas in fixed partial array connection with ON/OFF switches: This architecture is deduced from architecture *v* by replacing PSs with ON/OFF switches as shown in Fig. 9(vi). The RF matrix $\mathbf{F}_{\mathrm{RF}}$ still follows the format (88). The difference is that $[\mathbf{f}_{\mathrm{RF},i}]_j \in \{0, 1\}$ for $i = 1, \ldots, N_{\mathrm{RF}}$ and $j = 1, \ldots, (N/N_{\mathrm{RF}})$. The constraints in (86) do not need to be considered. Notably, when the array size is large, the channel power will be concentrated on the VR of the user w.r.t. the extra large-aperture array. Note that the VR caused by an unequal path loss usually covers a continuous part of the array. To increase the energy efficiency, only the continuous part of the array in VR can be turned ON. That is, the effective size of each subarray is dynamic as well. This architecture also has the advantages of easy synchronization and scalability as well as the lowest hardware cost (only N ON/OFF switches), thereby becoming suitable for extra large-aperture array systems [44].

g) Single-RF chain multiple antennas in dynamic partial array connection with PSs: The concept of a dynamic partial array or dynamic subarray appeared in [71]. It is an improved version of architecture *v*. As the name suggests, segmentation of the subarrays can be flexibly adjusted instead of being fixed. That is, even though $\mathbf{F}_{\mathrm{RF}}$ follows the format in (88), the size of $\mathbf{f}_{\mathrm{RF},i}$ varies with i. Each antenna can be flexibly connected to an arbitrary RF chain or be deactivated. This architecture not only harvests the array gain but adjusts the effective subarray sizes based on the real-time channel condition. Equally importantly, when a VR exists, dynamic subarrays are verified to achieve better performance than fixed subarrays [6].

However, this architecture has the following drawbacks. First, it is hard to implement. There are two solutions denoted as architectures *vii*-1 and *vii*-2, respectively. The first solution is to deploy a switch for selection at each antenna to select one of the N_{RF} RF chains.[1] An extra combiner is actually needed at each RF chain to enable the connection with multiple antennas, as shown in Fig. 9(vii-1). However, it is difficult to connect such a combiner with N switches for selection at the antenna side. The second solution is to modify architecture *iv* by adding a PS before each antenna in Fig. 9(vii-2). However, the integration of NN_{RF} ON/OFF switches and N PSs and the synchronization among them are challenging. Moreover, the lack of scalability further makes this architecture hard to be deployed in practical extra large-aperture array systems.

4) Proposed Double-Layer Architecture: Considering the advantage of dynamic subarrays as well as the practical implementation and scalability, in this article, we integrate the full-connection and subarray structures and propose a double-layer architecture, which is refered to as architecture *viii*.

[1]This solution is inspired from [72], but moves the switch for selection from the RF chain to the antenna.

TABLE III
COMPARISON OF ARCHITECTURES OF ACTIVE ARRAYS WITH LESS RF CHAINS

architectures	*i*	*ii*	*iii*	*iv*	*v*	*vi*	*vii*-1	*vii*-2	*viii*
number of PSs	0	0	$N_{\mathrm{RF}}N$	0	N	0	N	N	N
number of ON/OFF switches	0	0	0	$N_{\mathrm{RF}}N$	0	N	0	$N_{\mathrm{RF}}N$	$\frac{N_{\mathrm{RF}}N}{B^2}$
number of switches for selection	N_{RF}	N_{RF}	0	0	0	0	N	0	0
implementation difficulty	high	low	high	high	low	low	high	high	low
synchronization difficulty	high	low	high	high	low	low	high	high	low
scalability	×	√	×	×	√	√	×	×	√

As shown in Fig. 9(viii), the outer layer follows the fixed subarray structure, and the inner layer follows the dynamic subarray structure. The extra large-aperture array is composed of B physical subarrays. Each physical subarray has the same hardware topology, including (N_{RF}/B) RF chains and (N/B) antennas. The values of (N_{RF}/B) and (N/B) are not large. For example, we can let $(N_{\mathrm{RF}}/B) = 4$ and $(N/B) = 64$.

Architecture *vii* is adopted in each physical subarray. For convenient implementation, a physical link is established between each RF chain and each antenna in the common physical subarray. An ON/OFF switch is deployed in each physical link. For a certain antenna, only one RF chain can be selected, and thus no more than one physical link connected with this antenna is finally turned ON. To enable analog beamforming, each antenna is further equipped with a PS. A total of $(N_{\mathrm{RF}}N/B^2)$ ON/OFF switches and (N/B) PSs are integrated in a physical subarray.

In the proposed double-layer architecture, the RF matrix $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{RF}}\times N}$ follows the block diagonal structure in (85). The submatrix $\mathbf{F}_{\mathrm{RF},b} \in \mathbb{C}^{(N_{\mathrm{RF}}/B)\times(N/B)}$ has the following format:

$$\mathbf{F}_{\mathrm{RF},b} = \mathbf{S}_b \odot \left(\mathbf{1}_{\frac{N_{\mathrm{RF}}}{B}} \otimes \mathbf{f}_b^T\right) \tag{89}$$

where $\mathbf{S}_b \in \{0,1\}^{(N_{\mathrm{RF}}/B)\times(N/B)}$ is the ON/OFF matrix satisfying

$$0 \le \sum_{i=1}^{\frac{N_{\mathrm{RF}}}{B}} [\mathbf{S}_b]_{i,j} \le 1 \ \forall j. \tag{90}$$

The column vector $\mathbf{1}_{(N_{\mathrm{RF}}/B)}$ has (N_{RF}/B) ones, while $\mathbf{f}_b \in \mathbb{C}^{(N/B)\times 1}$ is the phase shifting vector. If the ith RF chain is activated, then it controls an effective subsubarray. The effective subsubarray has a dynamic size, which depends on the number of antennas whose physical links with the ith RF chain are turned ON. Analog beamforming is also supported within the effective subsubarray, and thus the array gain can be harvested.

The proposed double-layer architecture sustains the advantage of easy synchronization and scalability of the subarray structure. Equally importantly, the hardware cost is greatly reduced compared with architecture *vii*. The insertion loss is substantially mitigated by using much less switches. This architecture also can harvest the full array gain by activating all antennas in a subarray simultaneously. Alternatively, in spatial nonstationary channel conditions, we can only activate the antennas where the biggest proportion of channel power is concentrated in. For the above-mentioned reasons, this is a potential architecture for extra large-aperture arrays.

Table III summarizes the hardware cost, advantages, and disadvantages of the nine architectures, including the two solutions of architecture *vii* and the proposed architecture *viii*. Considering the scalability, architectures *ii*, *v*, *vi*, and *viii* are promising in the deployment of an extra large-aperture active array with less RF chains.

B. Reconfigurable Intelligent Surfaces

Another low-cost extra large-aperture array is the RIS [73], [74], [75], [76], [77], which is also known as metasurface [75], [78], [79], or intelligent reflecting surface (IRS) [80]. An RIS is composed of low-cost near passive unit cells, each with independently tunable EM responses controlled by external signals. An incident EM wave can be reflected or refracted by the RIS, or the reflection and the refraction happen simultaneously [81], [82], [83]. An RIS flexibly adjusts the amplitude, phase, or polarization of the incident EM wave in real time. Then, a preferable EM propagation environment can be customized by properly controlling the RIS.

The widely studied category of RISs reflect the EM waves toward the desired directions by adjusting their phases. An RIS works as a controllable reflector in the wireless environment, providing an additional controllable link between the transmitter and the receiver to assist the wireless communication. Suppose the transmitter and the receiver are equipped with a single antenna, respectively. The number of unit cells in the RIS is N. Then, the signal at the receiver can be modeled as follows:

$$r = gs + \mathbf{h}_2^T \Lambda \mathbf{h}_1 s + z \tag{91}$$

where s is the transmitted signal, $g \in \mathbb{C}$ is the direct channel between the transmitter and receiver, $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{C}^{N\times 1}$ are the channel between the transmitter and RIS and the channel between RIS and the receiver, respectively, while

$$\Lambda = \mathrm{diag}\{\mathbf{v}\}, \quad \mathbf{v} = \left[e^{j\phi_1}, \ldots, e^{j\phi_N}\right]^T \tag{92}$$

include the phase shift of signal introduced by the RIS, ϕ_n is the phase shift on the nth unit cell, and z is the complex Gaussian noise. Apart from the direct link g, an RIS link $\mathbf{h}_2^T \Lambda \mathbf{h}_1$ is added in. If the direct link is blocked by obstacles, then the RIS can reconstruct the wireless link and recover the communication service. The effective channel in an RIS-assisted wireless communication system is

$$g_{\mathrm{eff}} = g + \mathbf{h}_2^T \mathbf{\Lambda} \mathbf{h}_1. \tag{93}$$

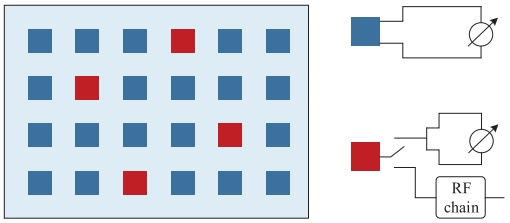

Fig. 10. Semi-passive RIS. Unit cells in blue are passive and only have the reflection mode with phase shift capability. Unit cells in red are active and have the reflection and receiving modes.

1) Fully Passive RIS: Most existing RISs that work in the reflection mode are fully passive regardless of the low external control voltage. No signal processing module exists at the RIS, and, thus, the RIS is not able to transmit or receive wireless signals. Since the individual channels $\mathbf{h}_1$ and $\mathbf{h}_2$ are cascaded together, channel estimation can only be applied at the receiver side. Under this condition, it is convenient to directly estimate the effective channel g_{eff}. Alternatively, by rewriting

$$\mathbf{h}_2^T \mathbf{\Lambda} \mathbf{h}_1 = \mathbf{h}_2^T \text{diag}\{\mathbf{v}\}\mathbf{h}_1 = \mathbf{h}_2^T \text{diag}\{\mathbf{h}_1\}\mathbf{v}$$

it is feasible to estimate the cascaded channel $\mathbf{h}_2^T \text{diag}\{\mathbf{h}_1\}$. The estimate of $\mathbf{h}_2^T \text{diag}\{\mathbf{h}_1\}$ can further guide the design of $\mathbf{v}$. However, the training overhead required to estimate $\mathbf{h}_2^T \text{diag}\{\mathbf{h}_1\} \in \mathbb{C}^{1\times N}$ at the single-antenna receiver is large. Therefore, the fully passive RIS faces intrinsic difficulties in channel estimation.

2) Semi-Passive RISs: To tackle the channel estimation problem, semi-passive RISs were proposed in [84], [85], [86], and [87]. As shown in Fig. 10, a semi-passive RIS introduces a few active sensors that can receive signals to enable channel estimation at the RIS. These active sensors are connected with RF chains and have two modes. One is the reflection mode, same as a common RIS unit cell. The other is the reception mode, in which the incident signals are received and conveyed to the signal processing module through RF chains. Suppose $\bar{N}$ unit cells are active sensors, satisfying $1 \le \bar{N} \le N$. Under this condition, the two individual channels from the transmitter and from the receiver to RIS, denoted by $\bar{\mathbf{h}}_1 \in \mathbb{C}^{\bar{N}\times 1}$, and $\bar{\mathbf{h}}_2 \in \mathbb{C}^{\bar{N}\times 1}$, respectively, can be estimated at the RIS. By leveraging the sparsity of channels and the correlation among different unit cells, the large-dimensional channels $\mathbf{h}_1$ and $\mathbf{h}_2$ can be extrapolated from their reduced dimensional versions $\bar{\mathbf{h}}_1$ and $\bar{\mathbf{h}}_2$ when the channel experiences spatial stationarity. In practice, one of the transmitter and receiver can be the BS or access point. Considering that the locations of BS/AP and RIS are fixed, the channel between them (denoted as $\mathbf{h}_1$), remains unchanged within a long time period. Therefore, $\mathbf{h}_1$ does not need to be frequently estimated, saving a great amount of training overhead. However, the channel extrapolation method may not work well in frequency-division duplexing systems, where reciprocity does not hold between the uplink and downlink channels.

The above low-cost architectures enable the deployment of extra large-aperture arrays. Active antenna arrays and RISs can be jointly applied to satisfy specific service requirements in different application scenarios.

V. Low-Complexity Processing and Computation

Apart from the problem of high cost, the implementation of an extra large-aperture array also requires high-complexity processing and computations. In multiantenna systems, the computational complexity of the widely used linear signal processing algorithms usually has an order of $\mathcal{O}(N)$, where N is the number of antennas. If a matrix multiplication or inversion is further involved, then the order of computational complexity grows. When N grows large, the complexity of these algorithms that jointly process signals across all antennas will grow explosively. The high-complexity processing and computations usually result in unacceptably high latency. The centralized control over the entire extra large-aperture array requires an extremely powerful central process unit (CPU) as shown in Fig. 11(a). Therefore, in extra large-aperture array systems, low-complexity processing and computation design is also a key objective.

A. Complexity Reduction at CPU

One method is to directly reduce the complexity of some high-complexity algorithms for their simplified or scalable implementation in the CPU. Complexity reduction in massive MIMO systems is not a novel concept [88], [89], [90], [91], [92]. Some of these methods can be extended to fit in extra large-aperture array systems.

There have been studies focusing on the complexity reduction in the CPU of extra large-aperture array systems [36], [40], [93], [94], [95]. Most of these studies focus on zero-forcing (ZF), which is a widely used linear signal processing method in multiuser multiantenna systems. The ZF precoder and combiner can be applied at the transmitter and the receiver, respectively, to cancel out the interuser interference. For example, let us denote the downlink channel between the extra large-aperture array at the BS and the single-antenna user k as $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$. The channels of K users are stacked together as $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{C}^{N\times K}$. Then, the ZF precoder is calculated as follows:

$$\mathbf{W}_{\text{ZF}} = \mathbf{H}\left(\mathbf{H}^H\mathbf{H}\right)^{-1} \in \mathbb{C}^{N\times K}. \tag{94}$$

Since matrix multiplication and inversion are involved, the computational complexity of calculating $\mathbf{W}_{\text{ZF}}$ reaches $\mathcal{O}(NK^2)+\mathcal{O}(K^3)$. To reduce the complexity, Ribeiro et al. [93] proposed a double-layer precoding method. The inner-layer precoder is applied to a group of users that share a similar elevation angle. The outer-layer precoder decreases the interference among different user groups. For each user group, channels on a column of antennas are summed up for the calculation of the inner and outer layer precoders. Suppose $\bar{\mathbf{h}}_k \in \mathbb{C}^{N_h\times 1}$, whose ith entry represents the sum of channels on the ith column of antennas. Then, complexity reduction is achieved by utilizing the low-dimensional $\bar{\mathbf{h}}_k$ instead of the extra large-dimensional $\mathbf{h}_k$. Another example in [40] focused on the acceleration of the calculation of the ZF combiner at the receiver. The algorithm acceleration problem was addressed from the perspective of linear equation systems and addressed by the randomized Kaczmarz (RK) algorithm.

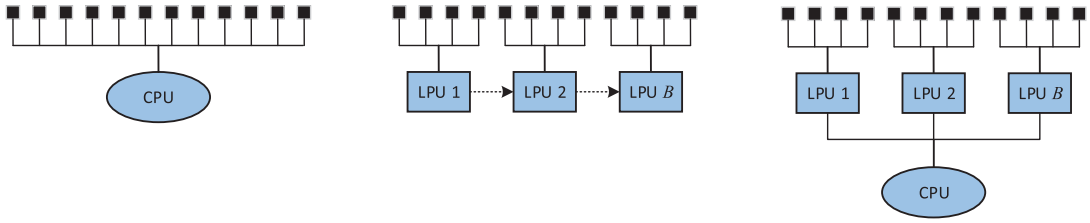

Fig. 11. Extra large-scale array is controlled by (a) single CPU, (b) multiple LPUs, or (c) CPU and multiple LPUs.

In addition to the ZF receiver, variational message passing (VMP) is another widely used multiuser MIMO detector, which has lower complexity than ZF because no matrix inversion is involved. In this context, Amiri et al. [36] applied VMP in the extra large-aperture array system under spatial nonstationary channel conditions, and further utilized a maximal ratio combiner (MRC) for initialization. The complexity of VMP and MRC is linear with N and K, which is much smaller than that of ZF.

Some other works focused on the complexity reduction of antenna selection [94] and user scheduling [95] in extra large-aperture array systems. Given the number of antennas N and the number of RF chains N_{RF}, the exhaustive searching-based antenna selection method requires a search over $\mathcal{O}(N^{N_{\mathrm{RF}}})$ combinations of antennas and RF chains, which is unacceptably high in extra large-aperture array systems. To reduce the complexity, [94] proposed a suboptimal method, which initially sets a coarse antenna selection result and then iteratively refines it based on a closed-form analytical expression of the energy efficiency, effectively avoiding the exhaustive search over a huge combination set. Moreover, when the EM wave experiences spherical propagation, then the channel is reconstructed by both the distance and the angle of the incident signal. Based on this channel feature, [95] introduced an effective distance between a user and the extra large-aperture array and then proposed a low-complexity user scheduling scheme that simply compares the effective distances of different users, making the scheduling problem simple and easy.

B. Distributed Processing and Computation

Assigning all the processing and computation tasks to a single CPU is not a reasonable choice in the extra large-aperture array system. An alternative is to partition the entire array into multiple subarrays and distribute the tasks to the subarrays [6], [34], [37], [38], [39], [41], [42], [43], [96]. This is a logical concept of subarray different from the physical subarray above. A logical subarray may have a fully digital physical architecture, but it has its own local processing unit (LPU) as shown in Fig. 11(b) and (c). Some processing and computation tasks of an individual logical subarray, such as channel estimation, antenna selection, etc., can be handled by its own LPU. When LPUs exist, there can be arranged via two logical architectures.

1) Single Layer With LPUs: This logical architecture is illustrated in Fig. 11(b) and solely composed of LPUs. That is to say, all the processing and computation tasks are distributed and performed at the LPUs, without a centralized control over the LPUs. Since no CPU exists, this architecture can be easily scaled up.

Notably, some tasks are local tasks and can be handled by a single LPU. A typical example of a local task is channel estimation. The channel across the entire array can be uniformly partitioned into B subchannels, i.e., $\mathbf{h} = [\mathbf{h}_1^T, \ldots, \mathbf{h}_B^T]^T$, where $\mathbf{h}_b \in \mathbb{C}^{(N/B)\times 1}$ is the subchannel on subarray b. The estimation of $\mathbf{h}_b$ can be independently performed by LPU b based on the pilots received by subarray b, without the cooperation with other LPUs [39]. When linear channel estimation is performed, the complexity of estimating $\mathbf{h}_b$ is $\mathcal{O}(N/B)$, significantly lower than $\mathcal{O}(N)$ of estimating $\mathbf{h}$. Denote the estimation result as $\hat{\mathbf{h}}_b$. The final channel estimation result across the entire array can be obtained by simply stacking the subchannel estimates together, that is, $\hat{\mathbf{h}} = [\hat{\mathbf{h}}_1^T, \ldots, \hat{\mathbf{h}}_B^T]^T$.

Most of the tasks are global tasks that require the cooperation among LPUs. A typical example is signal detection. We write the uplink signal model in a time-division duplexing system as follows:

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{h}_k s_k + \mathbf{n} \tag{95}$$

where s_k is the transmit signal from user k. The task of signal detection is to estimate $\mathbf{s} = [s_1, \ldots, s_K]^T$ from the $\mathbf{y} \in \mathbb{C}^{N\times 1}$, which is the signal received by the entire array. Denote $\mathbf{y} = [\mathbf{y}_1^T, \ldots, \mathbf{y}_B^T]^T$, where $\mathbf{y}_b \in \mathbb{C}^{(N/B)\times 1}$ is the signal received by subarray b. If LPU b independently performs signal detection based on $\mathbf{y}_b$, then there is a high probability that different LPUs provide different estimates of $\mathbf{s}$. This is because the channel vector $\mathbf{h}_b$ and the random noise $\mathbf{n}_b$ vary across different b, especially in multipath propagation scenarios and when spatial nonstationarity exists. Considering that only one final detection result is required, while the CPU that can make the final decision is absent, a serial detection method was proposed in [41]. VMP is normally combined with belief propagation for multiuser data detection. The output of LPU $b < B$ is the soft information of $\mathbf{s}$ and serves as an input of LPU $b+1$. The outputs of LPU B are the estimates of $\mathbf{s}$ and serve as the final detection result. The serial cooperation among the LPUs brings the benefit of easy scalability, but still suffers from the high processing latency. Moreover, the working procedure among the LPUs is fixed and cannot be flexibly adjusted according to practical channel conditions.

2) Double Layers With CPU and LPUs: A more reasonable and widely studied logical architecture is the double-layer architecture with LPUs in the lower layer and CPU in the upper layer as shown in Fig. 11(c). When spatial nonstationarity holds, different users have different VRs w.r.t. the array. If subarray b is not in the VR of user k, then the CPU can inform LPU b to deactivate the processing and computation related to user k. Therefore, a more efficient transceiver design can be deployed at the CPU, thereby enabling complexity reduction.

In this architecture, each LPU is connected with the CPU. Having completed the distributed processing and calculation, each LPU feeds its local result back to the CPU. Then, the CPU integrates the local results from all the LPUs and obtains the final global result by means of hard decision or data fusion [6]. At the receiver, [37] decentralizes the RK-ZF algorithm and applies it in multiuser signal detection in extra large-scale MIMO systems. LPU b calculates its local linear combiner matrix $\mathbf{V}_b \in \mathbb{C}^{K\times(N/B)}$, applies it on the received signal on subarray b, and computes the estimate of $\mathbf{s}$ at LPU b as follows:

$$\hat{\mathbf{s}}_b = \mathbf{V}_b \mathbf{y}_b. \tag{96}$$

If the VR of user k does not cover subarray b, then the entries in the kth row of $\mathbf{V}_b$ are zero. Thereafter, $\hat{\mathbf{s}}_b$ is sent to the CPU. Having received all the estimates from B LPUs, the CPU integrates $\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_B$ and makes the final decision through data fusion. Similarly, [42] and [47] applied VMP for multiuser signal detection, and LPU b outputs the symbol probability $q_b(\mathbf{s})$ instead of the estimate $\hat{\mathbf{s}}_b$. The estimates of multiuser signals are only obtained at the CPU.

The concept of LPUs of subarrays can be extended to LPUs of users. In [38], transmit antenna selection and user mapping were studied. Considering that different users have unequal VRs, parallel user mapping convolutional neural networks (CNNs) were proposed to learn the selected antennas for each user independently. The kth CNN outputs $N_{\max}$ antennas for user k. The CPU further makes antenna selection from the $N_{\max}$ antennas for user k by jointly considering the sum-rate of all K users. In the above studies [6], [37], [38], [42], [47], the LPUs work independently in parallel, and information exchange between one LPU and the CPU occurs only once. Therefore, the working procedure has relatively low latency.

Some recent works proposed the information exchange among LPUs or iterations between CPU and LPUs to gradually improve the performance. Information exchange between two distinct LPUs can be achieved with the assistance of CPU, or, a direct connection can be further established between the two LPUs. At the receiver, the LPUs in [34] performed ZF-based signal detection on a per user basis, while the detection results of a certain user were shared by the LPUs for the detection of signal from the next user. This serial interference cancelation method was also applied in [47], where VMP is employed in each LPU. Notably, given the VR of each user, the operation order of LPUs as well as the detection order of user signals can be initially determined by CPU [34], which further improves the detection performance.

Apart from ZF and VMP, expectation propagation (EP) is another effective algorithm that has been utilized at the receiver for multiuser signal detection in extra large-aperture array systems [97], [98]. EP in a centralized processing strategy that has excellent performance and moderate complexity. In this context, [97] initially implemented EP in a decentralized manner and made efforts on the reduction of computational complexity and information exchange amount, while [98] further refined the decentralized EP by approximating the matrix inversion at the CPU, whose complexity is $\mathcal{O}(K^3)$, with a polynomial expansion. Given that EP is an iterative algorithm, the decentralized EP method also requires information exchange among the CPU and the LPUs.

In [96], antenna selection and resource allocation were considered at the downlink transmitter. Even though in this work the LPUs operate in parallel, back-and-force information exchange between CPU and LPUs occurs since a genetic algorithm was adopted. Successive operation of LPUs and iterative optimization between two layers inevitably increase the latency.

Multilayer processing can be further applied in extra large-aperture RIS-assisted mobile communication systems [43]. The RIS can be uniformly partitioned into B logical subarrays, corresponding to the lowest processing layer. In the design of the RIS reflection codebook, a reduced dimensional local subcodebook can be first designed for each subarray. Then, subcodebooks in the second lowest layer is obtained from the ones in the lowest layer. Through this sequential design, the fully dimensional codebook can be finally derived in the higher layer. The multilayer processing reduces not only the complexity, but the huge training overhead caused by the extra large-aperture RIS.

VI. Low-Overhead Communication and Sensing

In this section, we focus on low-overhead design in extra large-aperture array systems. Training is an effective and reliable approach to acquire CSI. With the increase of user equipments and the diversification of device types that are connected to the extra large-aperture array system, the amount of pilots required will grow prohibitively high if independent training is performed across them. Furthermore, for an extra large-aperture array with massive active antennas but less RF chains, estimation of the huge dimensional channel on each antenna inevitably involves a beam sweeping or antenna switching process, which will be time consuming if the number of RF chains is much smaller than the number of active antennas.

Fortunately, the directionality and sparsity of propagation channels create room for overhead reduction, which will be explained in detail in the following part of this section. Furthermore, the extra large-aperture array has an extremely high spatial resolution, and the high-dimensional channel contains the environment information, such as knowledge about the user location and surrounding obstacles. Therefore, sensing can be achieved together with communication during the training process [99]. In this section, we study the low-overhead communication and sensing paradigm.

A. Directionality and Channel Sparsity

In a traditional multiantenna system, the serving area of a BS is large, and users are in the far-field region of the array. The plane wave channel model (60) is then applied, and the plane wave is expressed by its AoA/AoD as shown in (61). Due to the high spatial resolution of the large-aperture array, and the much smaller number of propagation paths than the number of antennas, the channel shows significant sparsity and directionality in the angular domain. In an extra large-aperture array system, there is a high probability that the distance between a user and the BS is smaller than the Rayleigh distance. Under these conditions, the spherical wave channel model (54) should be introduced, and the spherical wave is expressed by the position of the source (28). Moreover, the VR kicks in when blockage exists, which means that the effective array size is reduced. Then, whether the channel sparsity and directionality hold becomes a question.

Assume the BS is equipped with an extra large-aperture uniform linear array (ULA) with N elements lying on the x-axis, where N is even for the simplification of analysis. Considering that the horizontal ULA has flexible control over only the xz plane, and we describe the positions through (x, z) coordinates. The center of the ULA is at the origin of the coordinate system, and the position of antenna n is $(-([2n+1]/2)d, 0)$, where $n = -(N/2), \ldots, (N/2)-1$. User k is located at $\mathbf{s}_k = (x_k, z_k)$. By applying the limited dimensional channel model (78) and assume that only the LoS path exists, the channel between the BS and user k can be simplified as follows:

$$\mathbf{h}_k = \beta_k \mathbf{a}(\mathbf{s}_k) \odot \mathbf{p}(\Phi_k) \tag{97}$$

where $\mathbf{a}(\mathbf{s}) \in \mathbb{C}^{N\times 1}$ is the steering vector, satisfying

$$[\mathbf{a}(\mathbf{s})]_n = \frac{\lambda}{4\pi d_{k,n}} e^{-j\frac{2\pi}{\lambda}d_{k,n}}. \tag{98}$$

When applying (12)

$$d_{k,n} = \sqrt{\left(x_k + \frac{2n+1}{2}d\right)^2 + z_k^2} \tag{99}$$

is the distance between the source and antenna n, Φ_k is the VR of the user w.r.t. the array, and $\mathbf{p}(\Phi)$ follows the structure in (79).

1) Angular Domain: We start by investigating whether the directionality and sparsity hold for $\mathbf{h}_k$ in the angular domain when the VR covers the entire array. The angular domain transformation is derived from the plane wave model where equal phase deviation is experienced by each pair of adjacent antennas as shown in (61). Therefore, the discrete Fourier transformation (DFT) matrix is usually adopted as the angular domain transformation matrix. Denote the N-dimensional DFT matrix as $\mathbf{U}_A \in \mathbb{C}^{N\times N}$, where $[\mathbf{U}_A]_{n_1,n_2} = e^{-j2\pi(n_1/N)n_2}$, $n_1, n_2 = 0, \ldots, N-1$. The nth row corresponds to the direction with an angle of $\theta = \arccos(n/N)$. The rows of $\mathbf{U}_A$ are orthogonal with each other. Then, the angular domain channel of user k is written as $\tilde{\mathbf{h}}_{A,k} = \mathbf{U}_A \mathbf{h}_k$, where $\tilde{\mathbf{h}}_{A,k} \in \mathbb{C}^{N\times 1}$ has the same dimension with $\mathbf{h}_k$. Under the plane wave model, the amplitude of $[\tilde{\mathbf{h}}_{A,k}]_n$ will be large if the angle of the LoS path is close to $\arccos(n/N)$, and, thus, $\tilde{\mathbf{h}}_{A,k}$ would have a sparse pattern. However, under the spherical wave model, the entire array does not experience a common angle, and a significant angular spread appears. As shown in Fig. 12(a), $\tilde{\mathbf{h}}_{A,k}$ shows directionality around $\cos\theta = 0$ when $z_k = 5000\lambda$. With the decrease of z_k, and wherever user k moves toward the array, the angular spread increases, and $\tilde{\mathbf{h}}_{A,1}$ has more continuous nonzero entries than $\tilde{\mathbf{h}}_{A,2}$ and $\tilde{\mathbf{h}}_{A,3}$. In an extreme but unpractical case that $z_k = 0$, the angular spread will cover the entire angle value region, and then directionality and sparsity no longer exist in $\tilde{\mathbf{h}}_{A,k}$.

2) Cartesian Domain: From (98), we see that $[\mathbf{a}(\mathbf{s})]_n$ is determined by the 2-D Cartesian coordinate (x_k, z_k), instead of a 1-D angle θ. Therefore, under the spherical wave model, it is more reasonable to transform $\mathbf{h}_k$ to a 2-D domain than to a 1-D domain. Paper [39] proposed to transform the radio channel to the Cartesian domain. The transformation matrix $\mathbf{U}_c \in \mathbb{C}^{N_c\times N}$ is composed of N_c row vectors of $([\mathbf{a}^H(\bar{x}, \bar{z})]/\|\mathbf{a}(\bar{x}, \bar{z})\|)$, where $\bar{x}$ and $\bar{z}$ are the samples of x and z, respectively, and N_c is the number of sample pairs $(\bar{x}, \bar{z})$.

Let $N_c = N_x N_z$, where N_x and N_z are the numbers of x and z samples, respectively, by uniformly and separately sampling x and z as follows:

$$\begin{aligned}\{&\bar{x} = x_{\min}, x_{\min} + \Delta x, \ldots, x_{\max}\\ &\bar{z} = z_{\min}, z_{\min} + \Delta z, \ldots, z_{\max}\}\end{aligned} \tag{100}$$

where $x_{\min}, x_{\max}, z_{\min}$, and $z_{\max}$ jointly define the rectangular region that users may appear in, while Δx and Δz are the sampling steps in the x and z axis, respectively. Different from $\mathbf{U}_A$, the orthogonality among the rows of $\mathbf{U}_c$ cannot be guaranteed. The channel in the Cartesian domain is obtained by $\tilde{\mathbf{h}}_{C,k} = \mathbf{U}_c \mathbf{h}_k$, whose dimension is $N_x N_z$, i.e., not equal to that of $\mathbf{h}_k$. The $N_x N_z$-dimensional vector $\tilde{\mathbf{h}}_{C,k}$ can be rearranged to an $N_x \times N_z$-dimensional matrix $\tilde{\mathbf{H}}_{C,k}$. Fig. 12(b) illustrates the normalized amplitudes of the 2-D matrices $\tilde{\mathbf{H}}_{C,k}, k = 1, 2, 3$. For the sample pair which satisfies $(\bar{x}, \bar{z}) = (x_k, z_k)$, the corresponding entry of $\tilde{\mathbf{H}}_{C,k}$ has the largest amplitude as expected, demonstrating the directionality in the Cartesian domain. When $z_k = 50\lambda$, even though most entries of $\tilde{\mathbf{H}}_{C,k}$ are nonzero, their amplitudes are still obviously lower than the maximal one. With the increase of z_k, the number of nonzero entries decreases. The sparsity of $\tilde{\mathbf{H}}_{C,k}$ gradually becomes significant and can be found solely in the x-domain.

3) Polar Domain: The spherical wave channel is more frequently expressed by the polar coordinates (D_k, θ_k), where D_k and θ_k represent the distance and angle between the ULA's center and user k, respectively, satisfying

$$\begin{aligned} D_k &= \sqrt{x_k^2 + z_k^2}, \quad \theta_k = \arcsin\frac{x_k}{D_k}\\ x_k &= D_k \sin\theta_k, \quad z_k = D_k \cos\theta_k. \end{aligned} \tag{101}$$

Then, $d_{k,n}$ in (99) is calculated by

$$d_{k,n} = \sqrt{D_k^2 + \frac{(2n+1)^2}{4}d^2 + (2n+1)dD_k \sin\theta_k}. \tag{102}$$

The polar transformation matrix can be defined as $\mathbf{U}_P \in \mathbb{C}^{N_P\times N}$ with row vectors of $([\mathbf{a}^H(\bar{D}\sin\bar{\theta}, \bar{D}\cos\bar{\theta})]/\|\mathbf{a}(\bar{D}\sin\bar{\theta}, \bar{D}\cos\bar{\theta})\|)$, where $\bar{D}$ and $\bar{\theta}$ are samples of D

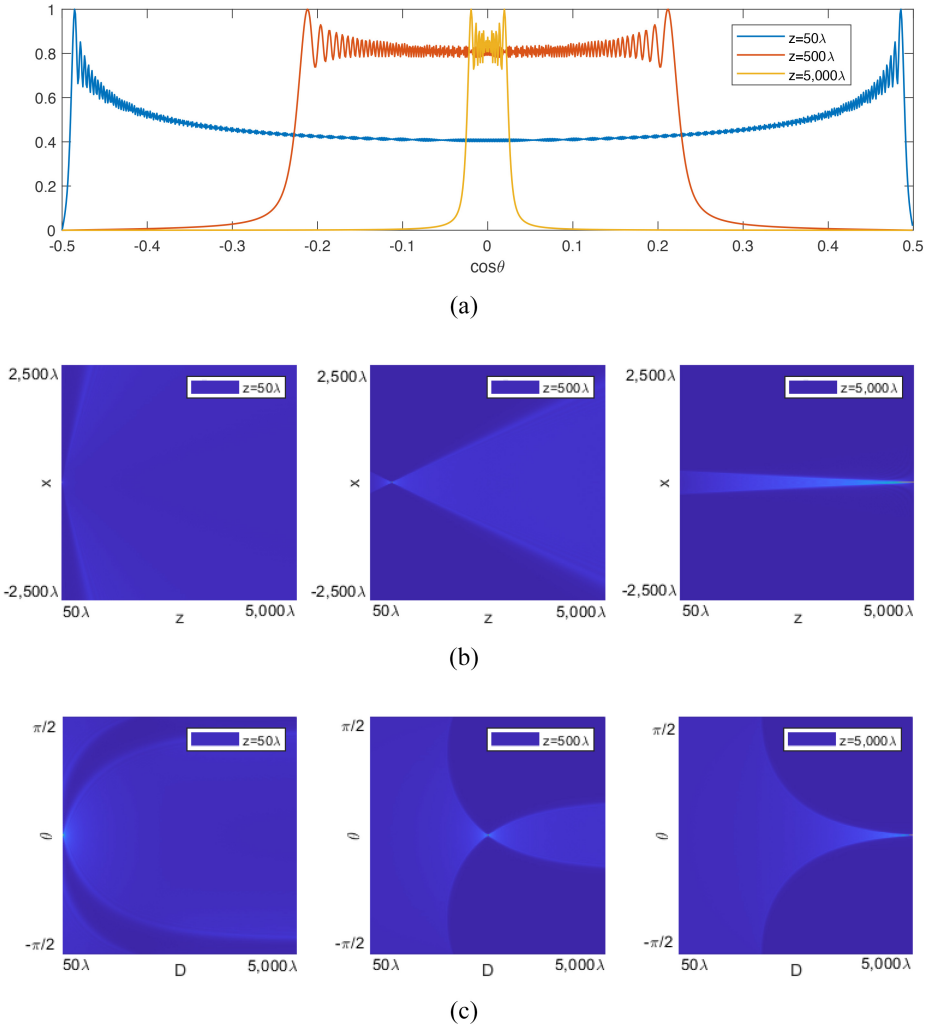

Fig. 12. Directionality and sparsity of channels in (a) angular, (b) Cartersian, and (c) polar domains, respectively, when $N = 1024$ and $\lambda = 0.01$ m. Users 1, 2, and 3 are located at $(0, 50\lambda)$, $(0, 500\lambda)$, and $(0, 5,000\lambda)$, respectively, and their VRs cover the entire array. (a) Illustrates the normalized amplitudes of vectors $\tilde{\mathbf{h}}_{A,k}, k = 1, 2, 3$. (b) and (c) Show the normalized amplitudes of matrices $\tilde{\mathbf{H}}_{C,k}$ and $\tilde{\mathbf{H}}_{P,k}, k = 1, 2, 3$.

and θ, respectively, and N_P is the number of sample pairs $(\bar{D}, \bar{\theta})$.

Similar to the Cartesian domain samples, we can let $N_P = N_D N_\theta$ by taking N_D samples of D and N_θ samples of θ independently, but choose

$$\begin{aligned}\big\{&\lg \bar{D} = \lg D_{\min}, \lg D_{\min} + \Delta D, \ldots, \lg D_{\max}\\ &\bar{\theta} = \theta_{\min}, \theta_{\min} + \Delta\theta, \ldots, \theta_{\max}\big\}\end{aligned} \tag{103}$$

where $D_{\min}, D_{\max}, \theta_{\min}$, and $\theta_{\max}$ define the fan-shaped region that users may appear in, and ΔD and $\Delta\theta$ are the sampling steps of $\lg D$ and θ, respectively. Here, $\lg D$ instead of D is uniformly sampled. This is because with the increase of D, the spherical wave channel becomes less sensitive to D, and, thus, the sampling interval of D can grow with D. Similar to $\mathbf{U}_c$, the orthogonality among different rows of $\mathbf{U}_P$ cannot be guaranteed as well. Thereafter, we obtain the polar domain channel as $\tilde{\mathbf{h}}_{P,k} = \mathbf{U}_P \mathbf{h}_k$, whose dimension is $N_D N_\theta$. Similarly, the $N_D N_\theta$-dimensional vector $\tilde{\mathbf{h}}_{P,k}$ can be rearranged to an $N_D \times N_\theta$-dimensional matrix $\tilde{\mathbf{H}}_{P,k}$ As shown in Fig. 12(c), the entry corresponding to $(\bar{D}, \bar{\theta}) = (D_k, \theta_k)$ has the maximal amplitude, verifying the directionality in the polar domain. Moreover, even though the sparsity of the channel in polar domain is not obvious when D is small, the amplitudes of nonzero entries are definitely much lower than the maximum one. The sparsity gets apparent with the increase of D, and is shown only in the angular domain when $D = 5000\lambda$.

To decrease the correlation among rows of $\mathbf{U}_P$, a joint angle and distance sampling grid was proposed in [29], where θ is uniformly sampled with $N_\theta = N$ and $\Delta\theta = (\pi/N)$. Specifically, the sampling of the distance depends on that of the angle. For a particular sample of angle $\bar{\theta}$, we acquire a unique sample set of $\bar{D}$, where the obtained vectors of $([\mathbf{a}^H(\bar{D}\sin\bar{\theta}, \bar{D}\cos\bar{\theta})]/\|\mathbf{a}(\bar{D}\sin\bar{\theta}, \bar{D}\cos\bar{\theta})\|)$ are nearly orthogonal to each other. To achieve this near orthogonality, the size of the distance sample set varies with the value of $\bar{\theta}$. When $\cos\bar{\theta}$ approaches 0, the sample set of the distance is expanded. Otherwise, the size of the sample set of distance decreases, resulting in an insufficient sampling grid of the entire space. Despite this drawback, the row vectors of $\mathbf{U}_P$ are approximately orthogonal to each other under this setting, and the polar domain channel $\tilde{\mathbf{h}}_{P,k}$ shows sparsity.

4) Antenna Domain: When the user is very close to the array as the example in Fig. 5, or severe blockage happens as illustrated in Fig. 6, the VR of the user w.r.t. the array is a small-scale subset of antennas in the array. Then, the channel shows sparsity in the antenna domain. In the simplest case that the VR of user k is a continuous subarray, the channel can be approximated as follows:

$$\mathbf{h}_k \approx \begin{bmatrix} \mathbf{0} \\ \mathbf{h}_{k,VR} \\ \mathbf{0} \end{bmatrix} \tag{104}$$

where $\mathbf{h}_{k,VR}$ is the subvector of $\mathbf{h}_k$ corresponding to the entries within the VR.

B. Low-Overhead Design

Channel directionality and sparsity in the transformation domains provide room for overhead reduction. More particularly, channel directionality guarantees the accuracy of user localization, which further supports channel reconstruction and sensing. Channel sparsity enables the application of compressed sensing techniques in the estimation of channels and the orthogonal transceiver design among multiple users. Details are given as follows.

Consider an extra large-aperture array system with less RF chains than active antennas at the BS. The spatially nonstationary channel $\mathbf{h}_k$ follows the limited dimensional model in (77) and can be rewritten as follows:

$$\mathbf{h}_k = \sum_{l=1}^{L_k} \beta_{k,l} \mathbf{a}(\mathbf{s}_{k,l}) \odot \mathbf{p}(\Phi_{k,l}) \tag{105}$$

where L_k is the number of paths in the channel of user k, while $\mathbf{s}_{k,l}$ and $\Phi_{k,l}$ are determined by the scatterers, reflectors, and obstacles in the environment. In the uplink training phase, user k transmits a pilot sequence to the BS for channel estimation and sensing. The received pilot sequence at the BS at time instance t is expressed as follows:

$$\mathbf{Y}_t = \sqrt{P}\mathbf{F}_{\mathrm{RF},t} \sum_{k=1}^{K} \mathbf{h}_k \mathbf{x}_k^H + \mathbf{F}_{\mathrm{RF},t}\mathbf{N}_t \tag{106}$$

where $\mathbf{Y}_t \in \mathbb{C}^{N_{\mathrm{RF}} \times Q}$ is the received pilot sequence with length Q on N_{RF} RF chains at time instance t, P is the transmit power of each user, $\mathbf{F}_{\mathrm{RF},t} \in \mathbb{C}^{N_{\mathrm{RF}} \times N}$ is the RF matrix at time instance t, $\mathbf{x}_k \in \mathbb{C}^{Q \times 1}$ is the pilot sequence of user k satisfying $\mathbf{x}_k^H \mathbf{x}_k = 1$ and $\mathbf{x}_k^H \mathbf{x}_j = 0, j \neq k$, $\mathbf{N}_t \in \mathbb{C}^{N \times Q}$ is the noise matrix with i.i.d. entries, with entry following a complex Gaussian distribution with zero mean and unit variance. A total of T time instances are used for uplink pilot transmission. By stacking $\mathbf{Y}_t, t = 1, \ldots, T$ together and multiplying them with $\mathbf{x}_k$, we have

$$\mathbf{y}_k = \sqrt{P}\mathbf{F}\mathbf{h}_k + \mathbf{n}_k \tag{107}$$

where $\mathbf{y}_k \in \mathbb{C}^{N_{\mathrm{RF}}T \times 1}$, $\mathbf{F} = [\mathbf{F}_{\mathrm{RF},1}^H, \ldots, \mathbf{F}_{\mathrm{RF},T}^H]^H$, and $\mathbf{n}_k = [(\mathbf{F}_{\mathrm{RF},1}\mathbf{N}_1\mathbf{x}_k)^H, \ldots, (\mathbf{F}_{\mathrm{RF},T}\mathbf{N}_T\mathbf{x}_k)^H]^H$. The independent linear estimation of the channel on each antenna requires $T = (N/N_{\mathrm{RF}})$. Then, the value of T will be large if $N \gg N_{\mathrm{RF}}$, resulting in a huge amount of training overhead.

1) Localization Based on Directionality: When an LoS path exists between user k and the BS, it is usually set as $l = 1$ in (105), and then $\mathbf{s}_{k,1}$ is the position of user k. The LoS path has stronger power than other NLoS components due to the smallest pathloss. Given the directionality of the near-field channel in Cartesian and polar domains, the matching method of [39] can be applied to find the position $\mathbf{s}_{k,1}$ from $\mathbf{y}_k$. Applying (105) in (107), the received pilot can be rewritten as follows:

$$\mathbf{y}_k = \sum_{l=1}^{L_k} \sqrt{P}\beta_{k,l}\mathbf{F}\big(\mathbf{a}\big(\mathbf{s}_{k,l}\big) \odot \mathbf{p}\big(\Phi_{k,l}\big)\big) + \mathbf{n}_k. \tag{108}$$

We now assume that $\Phi_{k,1}$ has been successfully identified. Then, the codebook for matching can be defined as $\bar{\mathbf{c}}(\bar{x}, \bar{z}) = \mathbf{F}(\mathbf{a}(\bar{x}, \bar{z}) \odot \mathbf{p}(\Phi_{k,1}))$, where $(\bar{x}, \bar{z})$ are in (100), or $\bar{\mathbf{c}}(\bar{D}, \bar{\theta}) = \mathbf{F}(\mathbf{a}(\bar{D}\sin\bar{\theta}, \bar{D}\cos\bar{\theta}) \odot \mathbf{p}(\Phi_{k,1}))$, where $(\bar{D}, \bar{\theta})$ are in (103). Utilizing the directionality, we obtain

$$\big(\hat{x}_{k,1}, \hat{z}_{k,1}\big) = \arg \max_{(\bar{x},\bar{z}) \in (100)} \frac{\bar{\mathbf{c}}(\bar{x}, \bar{z})^H \mathbf{y}_k}{\|\bar{\mathbf{c}}(\bar{x}, \bar{z})\|} \tag{109}$$

or

$$\Big(\hat{D}_{k,1}, \hat{\theta}_{k,1}\Big) = \arg \max_{(\bar{D},\bar{\theta}) \in (103)} \frac{\bar{\mathbf{c}}\Big(\big(\bar{D}, \bar{\theta}\big)^H\Big)\mathbf{y}_k}{\|\bar{\mathbf{c}}\big(\bar{D}, \bar{\theta}\big)\|} \tag{110}$$

and the localization result is $\hat{\mathbf{s}}_{k,l} = (\hat{x}_{k,1}, \hat{z}_{k,1})$ or $(\hat{D}_{k,1}\sin\hat{\theta}_{k,1}, \hat{D}_{k,1}\cos\hat{\theta}_{k,1})$. This localization method can work well when $T \ll (N/N_{\mathrm{RF}})$.

For sensing, given the estimates of position and VR, we can generally decide where the obstacle is. With more paths interacting with a common obstacle, the localization, size, and even shape of the obstacle can be more accurately determined from the positions and VRs of these paths. Then, the environment can be identified.

2) Channel Estimation Based on Sparsity: In practical environments, when the system works in higher frequency bands, the NLoS paths becomes fewer due to the severe pathloss and blockage. In an extra large-aperture array system, we usually have $L_k \ll N$. Therefore, the large dimensional channel $\mathbf{h}_k$ can be expressed by a limited amount of paths. In the Cartesian or polar domain, most of the channel power is concentrated on nearly L_k entries. Based on whether the orthogonality holds among the rows of transformation matrix, there are two categories of low-cost channel estimation methods. One is channel reconstruction, and the other is compressed sensing. Channel reconstruction focuses on the estimation of the limited amount of path parameters instead of the large-dimensional channel [39]. The parameters to be estimated include $\beta_{k,l}$, $\mathbf{s}_{k,l}$, and $\Phi_{k,l}$. When an LoS path exists, the user position can be obtained by the above matching method in (109) or (110). If the LoS component $\sqrt{P}\beta_{k,1}\mathbf{F}(\mathbf{a}(\mathbf{s}_{k,1}) \odot \mathbf{p}(\Phi_{k,1}))$ is extracted from $\mathbf{y}_k$ in (108), then the second largest path component can be extracted from the residual of $\mathbf{y}_k$ through the same matching method. The L_k paths can be iteratively extracted from their mixture. Finally, the large-dimensional channel $\mathbf{h}_k$ can be reconstructed by applying the estimates of $\beta_{k,l}$, $\mathbf{s}_{k,l}$ and $\Phi_{k,l}$ into (105). The training

overhead of channel reconstruction is comparable to that of localization based on directionality.

Compressed sensing aims to estimate the reduced dimensional sparse channel in a transformation domain. The precondition is that the row vectors of the transformation matrix maintain the orthogonality between them, which can be achieved by the polar domain transformation in [29]. For $\tilde{\mathbf{h}}_{P,k} \in \mathbb{C}^{N_P \times 1}$, we denote the indices of its nonzero entries as $\Upsilon_k = \{n_{k,1}, \ldots, n_{k,\tilde{N}_k}\}, \tilde{N}_k \ll N_P$. Then, the reduced dimension subchannel $[\tilde{\mathbf{h}}_{P,k}]_{\Upsilon_k}$ contains almost all the information in $\mathbf{h}_k$. When $\mathbf{U}_P$ and $[\mathbf{U}_P]_{\Upsilon_k,:}$ have full ranks, (107) can be further written as follows:

$$\mathbf{y}_k = \sqrt{P}\mathbf{F}\mathbf{U}_P^{\dagger}\tilde{\mathbf{h}}_{P,k} + \mathbf{n}_k \approx \sqrt{P}\mathbf{F}[\mathbf{U}_P]^{\dagger}_{\Upsilon_k,:}\left[\tilde{\mathbf{h}}_{P,k}\right]_{\Upsilon_k} + \mathbf{n}_k. \quad (111)$$

Then, the objective becomes to estimate the reduced dimensional channel $[\tilde{\mathbf{h}}_{P,k}]_{\Upsilon_k}$, which can be realized through compressed sensing. The key point lies in the identification of Υ_k from $\{1, \ldots, N_P\}$. Following the compressed sensing-based channel estimation methods in millimeter wave hybrid beamforming systems, the estimates $\hat{\Upsilon}_k$ and $[\hat{\tilde{\mathbf{h}}}_{P,k}]_{\hat{\Upsilon}_k}$ can be estimated through the orthogonal matching pursuing (OMP) algorithm, where the matching step is the same as (109). Then, the large-dimensional channel can be obtained by

$$\hat{\mathbf{h}}_k = [\mathbf{U}_P]_{\hat{\Upsilon}_k,:}\left[\hat{\tilde{\mathbf{h}}}_{P,k}\right]_{\hat{\Upsilon}_k}. \quad (112)$$

Notably, since the sampling grid cannot cover the entire space, there is a high probability that the positions estimated by OMP are not the real positions, and a further refinement of the estimated positions toward the real positions is required [29] if localization needs to be achieved simultaneously.

3) Multiuser Pilot Transmission Based on Sparsity: The nonoverlapping sparsity of different users' antenna-domain channels enables the simultaneous transmission of pilots from or to these users. A common pilot sequence can be shared among users that have nonoverlapping VRs, and the orthogonal pilot sequences are assigned to users with overlapping VRs. Due to the limited amount of orthogonal pilot sequences, the nonoverlapping sparsity among users creates potential for the reduction of the overall training time. By knowing the VR of user k, i.e., $\Phi_{\mathrm{UA},k}$, the BS directly transmits or receives the pilot of user k through $\Phi_{\mathrm{UA},k}$. For instance, suppose $\Phi_{\mathrm{UA},1}, \ldots, \Phi_{\mathrm{UA},B}$ cover subarrays $1, \ldots, B$, respectively, and they are nonoverlapped with each other. While $\Phi_{\mathrm{UA},B+1}$ covers the entire array. Then, pilot sequences $\mathbf{x}_1$ and $\mathbf{x}_2$ are assigned to users $1, \ldots, B$ and user $B+1$, respectively. In the uplink, the received pilots at the BS from all users can be expressed as follows:

$$\mathbf{Y} = \sqrt{P}\mathbf{F}_{\mathrm{RF}}\left(\sum_{b=1}^{B}\mathbf{h}_b\mathbf{x}_1^H + \mathbf{h}_{B+1}\mathbf{x}_2^H\right) + \mathbf{F}_{\mathrm{RF}}\mathbf{N}. \quad (113)$$

By multiplying $\mathbf{Y}$ with $\mathbf{x}_1$, the pilots from users $1, \ldots, B$ are extracted:

$$\mathbf{y} = \sqrt{P}\mathbf{F}_{\mathrm{RF}}\sum_{b=1}^{B}\mathbf{h}_b + \mathbf{n} \quad (114)$$

where $\mathbf{y} = [y_1, \ldots, y_B]^T$ contains the received pilot on each subarray, and $\mathbf{n} = [n_1, \ldots, n_B]^T = \mathbf{F}_{\mathrm{RF}}\mathbf{N}\mathbf{x}_1$. By further recalling (88) and (104), we can rewrite y_b as follows:

$$y_b = \sqrt{P}\mathbf{f}_{\mathrm{RF},b}^T\mathbf{h}_{b,VR} + n_b \quad (115)$$

which involves only the channel of user b. That is to say, only two instead of $B+1$ orthogonal pilot sequences are required for multiuser training without introducing interference among them. The nonoverlapping sparsity in the antenna domain has been utilized in [45], where the overhead for random access was greatly reduced and the efficiency was enhanced.

In extra large-aperture RIS-assisted systems, directionality, and channel sparsity still hold in the angular, Cartesian, polar, and RIS unit domains at the RIS side. Therefore, the low-cost designs are also applicable in RIS-assisted systems. Notably, when applying the multiuser pilot transmission scheme, the RIS should be equipped with signal reception capabilities.

VII. Conclusion

We investigated the new channel properties of spatial nonstationarity, including the spherical wave propagation and the VR, and made a survey about existing works in the context of hardware cost, processing and computation complexity, and training overhead for extra large-scale MIMO systems. We also studied the origins of spatial nonstationarity and illustrated the modifications of channel modeling when spatial nonstationarity was considered. This new property paves the way for low-cost hardware architectures. Through a detailed comparison, we proposed a double-layer architecture and the RIS as the most promising implementation architecture of an extra large-aperture array. Then, the complexity reduction problem was investigated and the distributed solution with one CPU and multiple LPUs demonstrated the most promising potential. Finally, the low-overhead communication and sensing strategies were investigated, which can be realized given the directionality and sparsity of the channel in the Cartesian, polar, and antenna domains. Summarizing, this article reviewed the early stage research efforts of extra large-scale MIMO, and highlighted the importance of low-cost designs in future practical implementations.

References

[1] F. Rusek et al., "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[2] M. Matthaiou, O. Yurduseven, H. Q. Ngo, D. Morales-Jimenez, S. L. Cotton, and V. F. Fusco, "The road to 6G: Ten physical layer challenges for communications engineers," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 64–69, Jan. 2021.

[3] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next?: Five promising research directions for antenna arrays," *Digit. Signal Process.*, vol. 94, pp. 3–20, Nov. 2019.

[4] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[5] Z. Wang et al., "Extremely large-scale MIMO: Fundamentals, challenges, solutions, and future directions," Sep. 2022, *arXiv:2209.12131*.

[6] E. De Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, "Non-stationarities in extra-large-scale massive MIMO," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 74–80, Aug. 2020.

[7] A. Pizzo, T. L. Marzetta, and L. Sanguinetti, "Spatially-stationary model for holographic MIMO small-scale fading," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 1964–1979, Sep. 2020.

[8] E. Björnson, M. Matthaiou, and M. Debbah, "Massive MIMO with non-ideal arbitrary arrays: Hardware scaling laws and circuit-aware design," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4353–4368, Aug. 2015.

[9] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99878–99888, 2019.

[10] A. Guerra, F. Guidi, D. Dardari, and P. M. Djurić, "Near-field tracking with large antenna arrays: Fundamental limits and practical algorithms," *IEEE Trans. Signal Process.*, vol. 69, pp. 5723–5738, 2021.

[11] H. Chen, A. Elzanaty, R. Ghazalian, M. F. Keskin, R. Jäntti, and H. Wymeersch, "Channel model mismatch analysis for XL-MIMO systems from a localization perspective," in *Proc. IEEE GLOBECOM*, Dec. 2022, pp. 1588–1593.

[12] L. Mucchi et al., "Physical-layer security in 6G networks," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1901–1914, 2021.

[13] G. J. Anaya-López, J. P. González-Coma, and F. J. López-Martínez, "Spatial degrees of freedom for physical layer security in XL-MIMO," in *Proc. IEEE VTC*, Jun. 2022, pp. 1–5.

[14] L. Zhao, Z. Wang, and X. Wang, "Wireless power transfer empowered by reconfigurable intelligent surfaces," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2121–2124, Jun. 2021.

[15] J. Wang, Y. Li, Y. Jia, J. Zhang, S. Jin, and T. Q. Quek, "Wireless energy transfer in extra-large massive MIMO Rician channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5628–5641, Sep. 2021.

[16] H. Friis, "A note on a simple transmission formula," *Proc. IRE*, vol. 34, no. 5, pp. 254–256, May 1946.

[17] E. Björnson and L. Sanguinetti, "Power scaling laws and near-field Behaviors of massive MIMO and intelligent reflecting surfaces," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1306–1324, 2020.

[18] H. Lu and Y. Zeng, "Multi-user communication with extremely large-scale MIMO," Jun. 2021, *arXiv:2106.06901*.

[19] H. Lu and Y. Zeng, "Near-field modeling and performance analysis for multi-user extremely large-scale MIMO communication," *IEEE Commun. Lett.*, vol. 26, no. 2, pp. 277–281, Feb. 2022.

[20] H. Lu and Y. Zeng, "Communicating with extremely large-scale array/surface: Unified modeling and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4039–4053, Jun. 2022.

[21] E. Björnson, Ö. T. Demir, and L. Sanguinetti, "A primer on near-field beamforming for arrays and reconfigurable intelligent surfaces," in *Proc. IEEE ASILOMAR*, Nov. 2021, pp. 105–112.

[22] *Calculation of Free-Space Attenuation*, Rec. P.525-4, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[23] J. Sherman, "Properties of focused apertures in the fresnel region," *IRE Trans. Antennas Propag.*, vol. 10, no. 4, pp. 399–408, Jul. 1962.

[24] K. T. Selvan and R. Janaswamy, "Fraunhofer and fresnel distances: Unified derivation for aperture antennas," *IEEE Antennas Propag. Mag.*, vol. 59, no. 4, pp. 12–15, Aug. 2017.

[25] Y. Krivosheev, A. Shishlov, A. Tobolev, and I. Vilenko, "Fresnel field to far field transformation using sparse field samples," in *Proc. IEEE MMET*, Aug. 2012, pp. 237–242.

[26] X. Yin, S. Wang, N. Zhang, and B. Ai, "Scatterer localization using large-scale antenna arrays based on a spherical wave-front parametric model," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6543–6556, Oct. 2017.

[27] X. Wei and L. Dai, "Channel estimation for extremely large-scale massive MIMO: Far-field, near-field, or hybrid-field?" *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 177–181, Jan. 2022.

[28] M. Cui, L. Dai, R. Schober, and L. Hanzo, "Near-field wide-band beamforming for extremely large antenna array," Dec. 2021, *arXiv:2109.10054v2*.

[29] M. Cui and L. Dai, "Channel estimation for extremely large-scale MIMO: Far-field or near-field?" *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2663–2677, Apr. 2022.

[30] Z. Abu-Shaban, K. Keykhosravi, M. F. Keskin, G. C. Alexandropoulos, G. Seco-Granados, and H. Wymeersch, "Near-field localization with a reconfigurable intelligent surface acting as lens," in *Proc. IEEE ICC*, Jun. 2021, pp. 1–6.

[31] Y.-H. Nam et al., "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, Jun. 2013.

[32] Y. Han, H. Zhang, S. Jin, X. Li, R. Yu, and Y. Zhang, "Investigation of transmission schemes for millimeter-wave massive MU-MIMO systems," *IEEE Syst. J.*, vol. 11, no. 1, pp. 72–83, Mar. 2017.

[33] Y. Han, Q. Liu, C.-K. Wen, S. Jin, and K.-K. Wong, "FDD massive MIMO based on efficient downlink channel reconstruction," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4020–4034, Jun. 2019.

[34] A. Amiri, M. Angjelichinoski, E. De Carvalho, and R. W. Heath, "Extremely large aperture massive MIMO: Low complexity receiver architectures," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2018, pp. 1–6.

[35] A. Ali, E. De Carvalho, and R. W. Heath, "Linear receivers in non-stationary massive MIMO channels with visibility regions," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 885–888, Jun. 2019.

[36] A. Amiri, C. N. Manchón, and E. De Carvalho, "A message passing based receiver for extra-large scale MIMO," in *Proc. IEEE CAMSAP*, Dec. 2019, pp. 564–568.

[37] V. C. Rodrigues, A. Amiri, T. Abrao, E. De Carvalho, and P. Popovski, "Low-complexity distributed XL-MIMO for multiuser detection," in *Proc. IEEE ICC Workshops*, Jun. 2020, pp. 1–6.

[38] A. Amiri, C. N. Manchón, and E. de Carvalho, "Deep learning based spatial user mapping on extra large MIMO arrays," Feb. 2020, *arXiv:2002.00474*.

[39] Y. Han, S. Jin, C.-K. Wen, and X. Ma, "Channel estimation for extremely large-scale massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 633–637, May 2020.

[40] V. Croisfelt, A. Amiri, T. Abrão, E. De Carvalho, and P. Popovski, "Accelerated randomized methods for receiver design in extra-large scale MIMO arrays," *IEEE Trans. Veh. Technol.*, vol. 70, no. 7, pp. 6788–6799, Jul. 2021.

[41] A. Amiri, C. N. Manchón, and E. De Carvalho, "Uncoordinated and decentralized processing in extra-large MIMO arrays," *IEEE Wireless Commun. Lett.*, vol. 11, no. 1, pp. 81–85, Jan. 2022.

[42] V. Croisfelt, T. Abrão, A. Amiri, E. de Carvalho, and P. Popovski, "Decentralized design of fast iterative receivers for massive MIMO with spatial non-stationarities," in *Proc. IEEE ACSSC*, Nov. 2021, pp. 1242–1249.

[43] Y. Zhang and A. Alkhateeb, "Learning reflection beamforming codebooks for arbitrary RIS and non-stationary channels," Oct. 2021, *arXiv:2109.14909v2*.

[44] X. Yang, F. Cao, M. Matthaiou, and S. Jin, "On the uplink transmission of extra-large scale massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15229–15243, Dec. 2020.

[45] J. C. M. Filho, G. Brante, R. D. Souza, and T. Abrão, "Exploring the non-overlapping visibility regions in XL-MIMO random access and scheduling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6597–6610, Aug. 2022.

[46] X. Li, S. Zhou, E. Björnson, and J. Wang, "Capacity analysis for spatially non-wide sense stationary uplink massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 7044–7056, Dec. 2015.

[47] A. Amiri, S. Rezaie, C. N. Manchón, and E. De Carvalho, "Distributed receiver processing for extra-large MIMO arrays: A message passing approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2654–2667, Apr. 2022.

[48] D. W. M. Guerra and T. Abrão, "Clustered double-scattering channel modeling for XL-MIMO with uniform arrays," *IEEE Access*, vol. 10, pp. 20173–20186, 2022.

[49] Y. Zhu, H. Guo, and V. K. Lau, "Bayesian channel estimation in multi-user massive MIMO with extremely large antenna array," *IEEE Trans. Signal Process.*, vol. 69, pp. 5463–5478, 2021.

[50] L. Liu et al., "The COST 2100 MIMO channel model," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, Dec. 2012.

[51] C.-N. Chuah, D. N. C. Tse, J. M. Kahn, and R. A. Valenzuela, "Capacity scaling in MIMO wireless systems under correlated fading," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 637–650, Mar. 2002.

[52] H. Shin and J. H. Lee, "Capacity of multiple-antenna fading channels: Spatial fading correlation, double scattering, and keyhole," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2636–2647, Oct. 2003.

[53] M. R. McKay, A. J. Grant, and I. B. Collings, "Performance analysis of MIMO-MRC in double-correlated rayleigh environments," *IEEE Trans. Commun.*, vol. 55, no. 3, pp. 497–507, Mar. 2007.

[54] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[55] A. S. Poon, R. W. Brodersen, and D. N. Tse, "Degrees of freedom in multiple-antenna channels: A signal space approach," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 523–536, Feb. 2005.

[56] D. Tjostheim and J. Thomas, "Some properties and examples of random processes that are almost wide sense stationary," *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 257–262, May 1975.

[57] K.-W. Yip and T.-S. Ng, "Karhunen–Loeve expansion of the WSSUS channel output and its application to efficient simulation," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 4, pp. 640–646, May 1997.
[58] J. Liu et al., "A non-stationary channel model with correlated NLoS/LoS states for ELAA-mMIMO," in *Proc. IEEE GLOBECOM*, Dec. 2021, pp. 1–6.
[59] Z. Dong and Y. Zeng, "Near-field spatial correlation for extremely large-scale array communications," *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1574–1578, Jul. 2022.
[60] C. Wang et al., "Characteristics of 5.3 GHz MIMO channels with an extremely large antenna array in urban macro scenarios," in *Proc. IEEE VTC*, Jun. 2022, pp. 1–5.
[61] R. Feng, C.-X. Wang, J. Huang, Y. Zheng, F. Lai, and W. Zhou, "Mutual coupling analysis of 6G ultra-massive MIMO channel measurements and models," in *Proc. IEEE ICC*, May 2022, pp. 956–961.
[62] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.
[63] O. El Ayach, R. W. Heath, S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 3476–3480.
[64] W. Roh et al., "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.
[65] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
[66] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[67] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.
[68] S. Han, I. Chih-Lin, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag*, vol. 53, no. 1, pp. 186–194, Jan. 2015.
[69] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.
[70] M. Xiao et al., "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.
[71] S. Park, A. Alkhateeb, and R. W. Heath, "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2907–2920, May 2017.
[72] Z. Wu, M. Cui, Z. Zhang, and L. Dai, "Distance-aware precoding for near-field capacity improvement in XL-MIMO," in *Proc. IEEE VTC*, Jun. 2022, pp. 1–5.
[73] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.
[74] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238–8242, Aug. 2019.
[75] W. Tang et al., "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 421–439, Jan. 2021.
[76] M. D. Renzo et al., "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.
[77] Y. Han, X. Li, W. Tang, S. Jin, Q. Cheng, and T. J. Cui, "Dual-polarized RIS-assisted mobile communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 591–606, Jan. 2022.
[78] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao, and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light Sci. Appl.*, vol. 3, no. 10, pp. e218–e218, Oct. 2014.
[79] W. Tang et al., "Wireless communications with programmable metasurface: Transceiver design and experimental results," *China Commun.*, vol. 16, no. 5, pp. 46–61, May 2019.
[80] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.
[81] Y. Liu et al., "STAR: Simultaneous transmission and reflection for 360° coverage by intelligent surfaces," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 102–109, Dec. 2021.
[82] J. Xu, Y. Liu, X. Mu, and O. A. Dobre, "STAR-RISs: Simultaneous transmitting and reflecting reconfigurable intelligent surfaces," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 3134–3138, Sep. 2021.
[83] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3083–3098, May 2022.
[84] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Deep learning for large intelligent surfaces in millimeter wave and massive MIMO systems," in *Proc. IEEE GLOBECOM*, Dec. 2019, pp. 1–6.
[85] S. Liu, Z. Gao, J. Zhang, M. D. Renzo, and M.-S. Alouini, "Deep denoising neural network assisted compressive channel estimation for mmWave intelligent reflecting surfaces," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9223–9228, Aug. 2020.
[86] M. Jian and Y. Zhao, "A modified off-grid SBL channel estimation and transmission strategy for RIS-assisted wireless communication systems," in *Proc. IWCMC*, Jun. 2020, pp. 1848–1853.
[87] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1035–1071, 2nd Quart., 2022.
[88] S. Zarei, W. Gerstacker, R. R. Müller, and R. Schober, "Low-complexity linear precoding for downlink large-scale MIMO systems," in *Proc. IEEE PIMRC*, Sep. 2013, pp. 1119–1124.
[89] A. Kammoun, A. Müller, E. Björnson, and M. Debbah, "Linear precoding based on polynomial expansion: Large-scale multi-cell MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 861–875, Oct. 2014.
[90] M. N. Boroujerdi, S. Haghighatshoar, and G. Caire, "Low-complexity statistically robust precoder/detector computation for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6516–6530, Oct. 2018.
[91] J. V. Alegría, F. Rusek, and O. Edfors, "Trade-offs in decentralized multi-antenna architectures: The WAX decomposition," *IEEE Trans. Signal Process.*, vol. 69, pp. 3627–3641, 2021.
[92] J. V. Alegría and F. Rusek, "Enabling Decentralized computation of the WAX decomposition," in *Proc. IEEE ICC*, May 2022, pp. 1–6.
[93] L. N. Ribeiro, S. Schwarz, and M. Haardt, "Low-complexity zero-forcing precoding for XL-MIMO transmissions," in *Proc. EUSIPCO*, Aug. 2021, pp. 1621–1625.
[94] J. C. Marinello, T. Abrão, A. Amiri, E. De Carvalho, and P. Popovski, "Antenna selection for improving energy efficiency in XL-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13305–13318, Nov. 2020.
[95] J. P. González-Coma, F. J. López-Martínez, and L. Castedo, "Low-complexity distance-based scheduling for multi-user XL-MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2407–2411, Nov. 2021.
[96] J. H. I. de Souza, A. Amiri, T. Abrao, E. De Carvalho, and P. Popovski, "Quasi-distributed antenna selection for spectral efficiency maximization in subarray switching XL-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 7, pp. 6713–6725, Jul. 2021.
[97] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, "Expectation propagation detector for extra-large scale massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2036–2051, Mar. 2020.
[98] Z. Sun, X. Pu, S. Shao, S. Jin, and Q. Chen, "A low complexity expectation propagation detector for extra-large scale massive MIMO," in *Proc. ICCC*, Jul. 2021, pp. 746–751.
[99] J. Yang, Y. Zeng, S. Jin, C.-K. Wen, and P. Xu, "Communication and localization with extremely large lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3031–3048, May 2021.

Yu Han (Member, IEEE) received the B.S. degree in communications engineering from Hangzhou Dianzi University, Hangzhou, China, in 2012, and the M.S. and Ph.D. degrees in information and communications engineering from Southeast University, Nanjing, China, in 2015 and 2020, respectively.

She was a Postdoctoral Fellow with Singapore University of Technology and Design, Singapore, till 2022. She is currently an Associate Professor with Southeast University, Nanjing, China. Her research interests include extra large-scale MIMO and reconfigurable intelligent surface.

Shi Jin (Senior Member, IEEE) received the B.S. degree in communications engineering from Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2007.

From June 2007 to October 2009, he was a Research Fellow with the University College London (Adastral Park Research Campus), London, U.K. He is currently the faculty with the National Mobile Communications Research Laboratory, Southeast University. His research interests include wireless communications, random matrix theory, and information theory.

Dr. Jin and his coauthors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory, the 2022 Best Paper Award, and the 2010 Young Author Best Paper Award by the IEEE Signal Processing Society. He was an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and *IET Communications*. He is serving as an Area Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and *Electronics Letters* (IET).

Michail Matthaiou (Fellow, IEEE) was born in Thessaloniki, Greece, in 1981. He received the Diploma degree (five years) in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2004, the M.Sc. degree (with Distinction) in communication systems and signal processing from the University of Bristol, Bristol, U.K., in 2005, and the Ph.D. degree from The University of Edinburgh, Edinburgh, U.K., in 2008.

From September 2008 to May 2010, he was with the Institute for Circuit Theory and Signal Processing, Munich University of Technology, Munich, Germany, working as a Postdoctoral Research Associate. He is currently a Professor of Communications Engineering and Signal Processing and the Deputy Director of the Centre for Wireless Innovation, Queen's University Belfast, Belfast, U.K., after holding an Assistant Professor position with Chalmers University of Technology, Gothenburg, Sweden. His research interests span signal processing for wireless communications, beyond massive MIMO, intelligent reflecting surfaces, mm-wave/THz systems, and deep learning for communications.

Dr. Matthaiou and his coauthors received the IEEE Communications Society (ComSoc) Leonard G. Abraham Prize in 2017. He was awarded the prestigious 2018/2019 Royal Academy of Engineering/The Leverhulme Trust Senior Research Fellowship and also received the 2019 EURASIP Early Career Award. His team was also the Grand Winner of the 2019 Mobile World Congress Challenge. He was the recipient of the 2011 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East, and Africa Region, and a co-recipient of the 2006 IEEE Communications Chapter Project Prize for the best M.Sc. dissertation in the area of communications. He has coauthored papers that received Best Paper Awards at the 2018 IEEE WCSP and 2014 IEEE ICC. In 2014, he received the Research Fund for International Young Scientists from the National Natural Science Foundation of China. He currently holds the ERC Consolidator Grant BEATRICE (2021–2026) focused on the interface between information and electromagnetic theories. He is currently the Editor-in-Chief of *Physical Communication* (Elsevier), a Senior Editor for IEEE WIRELESS COMMUNICATIONS LETTERS and *IEEE Signal Processing Magazine*, and an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS.

Tony Q. S. Quek (Fellow, IEEE) received the B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008.

He is currently the Cheng Tsang Man Chair Professor with Singapore University of Technology and Design (SUTD), Singapore, and the ST Engineering Distinguished Professor. He also serves as the Director of the Future Communications Research and Development Programme, the Head of ISTD Pillar, and the Deputy Director of the SUTD-ZJU IDEA. His current research topics include wireless communications and networking, network intelligence, nonterrestrial networks, open radio access network, and 6G.

Dr. Quek is currently serving as an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards—Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, the 2020 IEEE Communications Society Young Author Best Paper Award, the 2020 IEEE Stephen O. Rice Prize, the 2020 Nokia Visiting Professor, and the 2022 IEEE Signal Processing Society Best Paper Award. He has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is a Fellow of the Academy of Engineering Singapore.

Chao-Kai Wen (Senior Member, IEEE) received the Ph.D. degree from the Institute of Communications Engineering, National Tsing Hua University, Hsinchu, Taiwan, in 2004.

He was with Industrial Technology Research Institute, Hsinchu, and MediaTek Inc., Hsinchu, from 2004 to 2009, where he was engaged in broadband digital transceiver design. In 2009, he joined the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, where he is currently a Professor. His research interests center around the optimization of wireless multimedia networks.