INVITED PAPER *Special Section on Multidisciplinary and Innovative Key technologies in Advanced communication systems (MIKA)*

User-Centric Design of Millimeter Wave Communications for Beyond 5G and 6G

Koji ISHIBASHI[†a)], ***Senior Member***, **Takanori HARA**[†], **Sota UCHIMURA**[†], ***Student Members***, **Tetsuya IYE**[††], **Yoshimi FUJII**[††], **Takahide MURAKAMI**[†††], ***and*** **Hiroyuki SHINBO**[†††], ***Members***

SUMMARY In this paper, we propose new radio access network (RAN) architecture for reliable millimeter-wave (mmWave) communications, which has the flexibility to meet users' diverse and fluctuating requirements in terms of communication quality. This architecture is composed of multiple radio units (RUs) connected to a common distributed unit (DU) via fronthaul links to virtually enlarge its coverage. We further present grant-free non-orthogonal multiple access (GF-NOMA) for low-latency uplink communications with a massive number of users and robust coordinated multi-point (CoMP) transmission using blockage prediction for uplink/downlink communications with a high data rate and a guaranteed minimum data rate as the technical pillars of the proposed RAN. The numerical results indicate that our proposed architecture can meet completely different user requirements and realize a *user-centric design* of the RAN for beyond 5G/6G.

key words: *beyond 5G, 6G, millimeter wave, radio access network (RAN), low-latency massive access, robust coordinated multi-point (CoMP), blockage prediction*

1. Introduction

Over the 130 years since Marconi's seminal work, mobile communications systems have evolved remarkably and have given rise to a wide variety of applications such as texting, voice/video communications, and streaming services. This paradigm shift has surely changed people's lifestyles. Fifth generation (5G) mobile communication systems were deployed worldwide in 2020. These systems can support three different quality requirements of communications: enhanced mobile broadband (eMBB) for high-speed communications, ultra-reliable low-latency communications (URLLC), and communications with a massive number of devices known as massive machine-type communications (mMTC), whereas early 5G systems mainly focused on eMBB applications [1]. Their even higher potential than former systems such as long-term evolution (LTE)-Advanced is expected to lead to the development of new applications such as the smart industry, automated driving, remote healthcare, augmented reality, and virtual reality referred to as the meta-verse.

However, the popularization of these emerging applications in all areas of society, including industry, education, public health, ocean, and space, is anticipated to render current 5G systems unable to handle the diverse requirements of communications from a large number of users (including all kinds of devices). This includes requirements such as ultra-high-speed access with guaranteed quality (for example, VR devices require a minimum data rate of 10 Gbps [2]) and low-latency communications with a massive number of users (namely a combination of URLLC and mMTC) [3]–[5]. Communication technologies beyond 5G and its successor, the sixth generation (6G), are already under discussion in many countries to accommodate these diverse requirements from users without compromising the quality [3], [6], [7]. To meet these requirements and realize highly reliable communications, it is essential to further exploit high-frequency bands, such as millimeter-wave (mmWave) bands, which enable reliable transmission at a high data rate with a massive number of users [2], [3], [5], [7].

However, these higher-frequency bands experience severe propagation losses, resulting in the inherent instability of wireless channels. In an attempt to solve the problem of path loss, beamforming has been used in current 5G systems to concentrate the radiated energy at specific angles, and has been intensively studied and applied over the last decade [8], [9]. However, as the coverage of base stations using mmWave bands is still not extensive, even with the aid of sophisticated beamforming compared to traditional microwave cellular systems, communication instability and large communication delays are inevitable, mainly because of the frequent handover among mmWave base stations. In addition, mmWave channels are susceptible to sudden and unpredictable obstruction by objects such as human bodies, buildings, and vehicles. In densely populated downtown areas or near roads carrying heavy traffic, the formed beams are either momentarily or continuously blocked, resulting in unstable communication with occurrence probabilities ranging between 20% and 60% [10]–[12].

In light of the above, research and development concerned with Beyond 5G and 6G mainly aims to develop communication technology that meets diverse requirements while solving the aforementioned problems associated with mmWave communications. In this paper, we propose new radio access network (RAN) architecture based on mmWave communications. The proposed architecture has the flexibil-

Manuscript received November 17, 2021.
Manuscript revised February 17, 2022.
Manuscript publicized July 13, 2022.

[†]The authors are with the Advanced Wireless & Communication Research Center (AWCC), The University of Electro-Communications, Chofu-shi, 182-8585 Japan.
[††]The authors are with the Kozo Keikaku Engineering, Inc., Tokyo, 164-0012 Japan.
[†††]The authors are with the KDDI Research, Inc., Fujimino-shi, 356-8502 Japan.
a) E-mail: koji@ieee.org

DOI: 10.1587/transcom.2021MEI0002

ity to meet users' requirements in terms of diversity and fluctuations with respect to communication quality. This architecture is composed of multiple radio units (RUs) connected to a common distributed unit (DU) via fronthaul links to virtually enlarge its coverage and has two technical pillars, grant-free non-orthogonal multiple access (GF-NOMA) for low-latency uplink communications with a massive number of users and robust coordinated multi-point (CoMP) transmission using blockage prediction for high-data-rate uplink/downlink communications with a guaranteed minimum data rate, to meet completely different user requirements and realize RAN architecture with a *user-centric design*.

The remainder of this paper is organized as follows. Section 2 provides an overview of our proposed RAN architecture including the transmission sequence. Section 3 and Sect. 4 elaborate on the basic ideas, technical details, and designs of the GF-NOMA and robust CoMP, respectively. Finally, we draw our conclusions in Sect. 5.

2. Flexible RAN for User-Centric Communications

Before describing the technical pillars of the proposed architecture, we provide an overview of the proposed RAN. As illustrated in Fig. 1, the system is composed of multiple distributed RUs connected to a common DU via fronthaul links. Every RU has a large number of antenna elements and is responsible for digital-to-analog (D/A) and analog-to-digital (A/D) conversion, fast Fourier transform (FFT), inverse FFT (IFFT), and high-power amplification of radio signals. The DU performs the tasks of modulation, demodulation, beamforming, scheduling, and hybrid automatic repeat request (ARQ). The coverage areas of RUs overlap with each other and virtually form a *service cell* using the same frequency band, thereby enabling the cell size of the system to be extended. Note that this centralized architecture can be found in the literature with different names such as cloud RAN, distributed multiple-input multiple-output (MIMO), network MIMO, and cell-free massive MIMO [13]–[16]. The main target of these conventional centralized architectures is to increase the network capacity by exploiting the inherently low spatial correlation of wireless channels realized by the distributed antennas and deterministically treating inter-cell interference. On the other hand, our target here is to enable the network to meet every user's demand for communications quality by overcoming the unreliable nature of mmWave channels through the proposed distributed architecture.

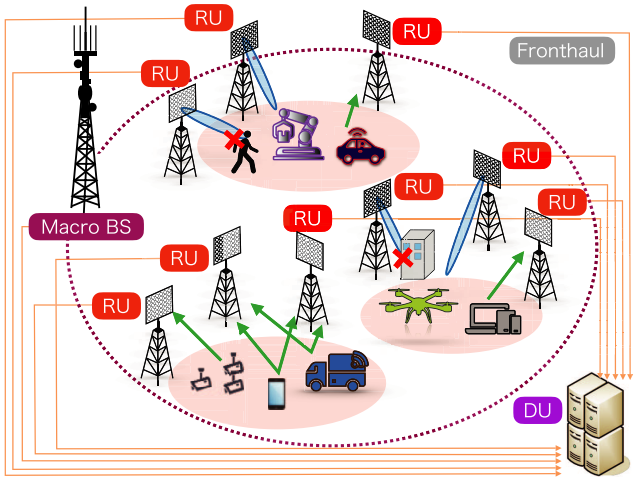

Fig. 1 Proposed RAN architecture composed of multiple radio units (RUs) connected to a common distributed unit (DU) via fronthaul links to virtually enlarge its coverage.

2.1 Initialization

To start the initial access process of the proposed RAN using mmWave bands, each UE requests access to the system through a control channel using microwave bands, namely a macrocell, to allocate a non-orthogonal pilot sequence as a user identifier. The DU is informed of this allocation, shared among all the RUs, and is exclusive to the system. Moreover, the information of resource blocks (RBs) allocated for GF access is also delivered to the UE for the initial access. This process is illustrated in (a) and (b) in Fig. 2, which is performed by the macro base station (BS), and details of the pilot sequences are provided in Sect. 3.

2.2 Initial Access

During the initial access process, every RU performs beam sweeping according to the pre-designed beam dictionary. This set of initial beams must be designed to guarantee a minimal signal-to-noise power ratio (SNR) even at the edge of the coverage area while maintaining the number of beams in the dictionary to reduce the unavoidable delay imposed by beam sweeping. To this end, we proposed the concept of the worst-case channel and an algorithm that generates the beamforming dictionary with the minimum number of beams while guaranteeing a minimal SNR in [17]. Beam sweeping is performed by every RU, as illustrated in (c) in Fig. 2, and the UE recognizes signals from multiple RUs through the appropriate initial beams. Then, the UE transmits the information about the available RUs, corresponding initial beam indexes, and a data transmission request with specific quality requirements such as minimum data rate and maximum delay to one of the RUs with the maximum SNR through the GF-NOMA using the pre-assigned user identifier, as shown in Fig. 2(d). Then, the connection between the UE and DU with the RUs is established.

2.3 Data Transmission

This flexible RAN mainly focuses on two different requirements: 1) low-latency uplink communications with a massive number of users and 2) high-data-rate uplink/downlink communications with a minimum rate guarantee. The former is suitable for applications that generate or receive sporadic traffic, such as vehicles/drone/robot controls and re-

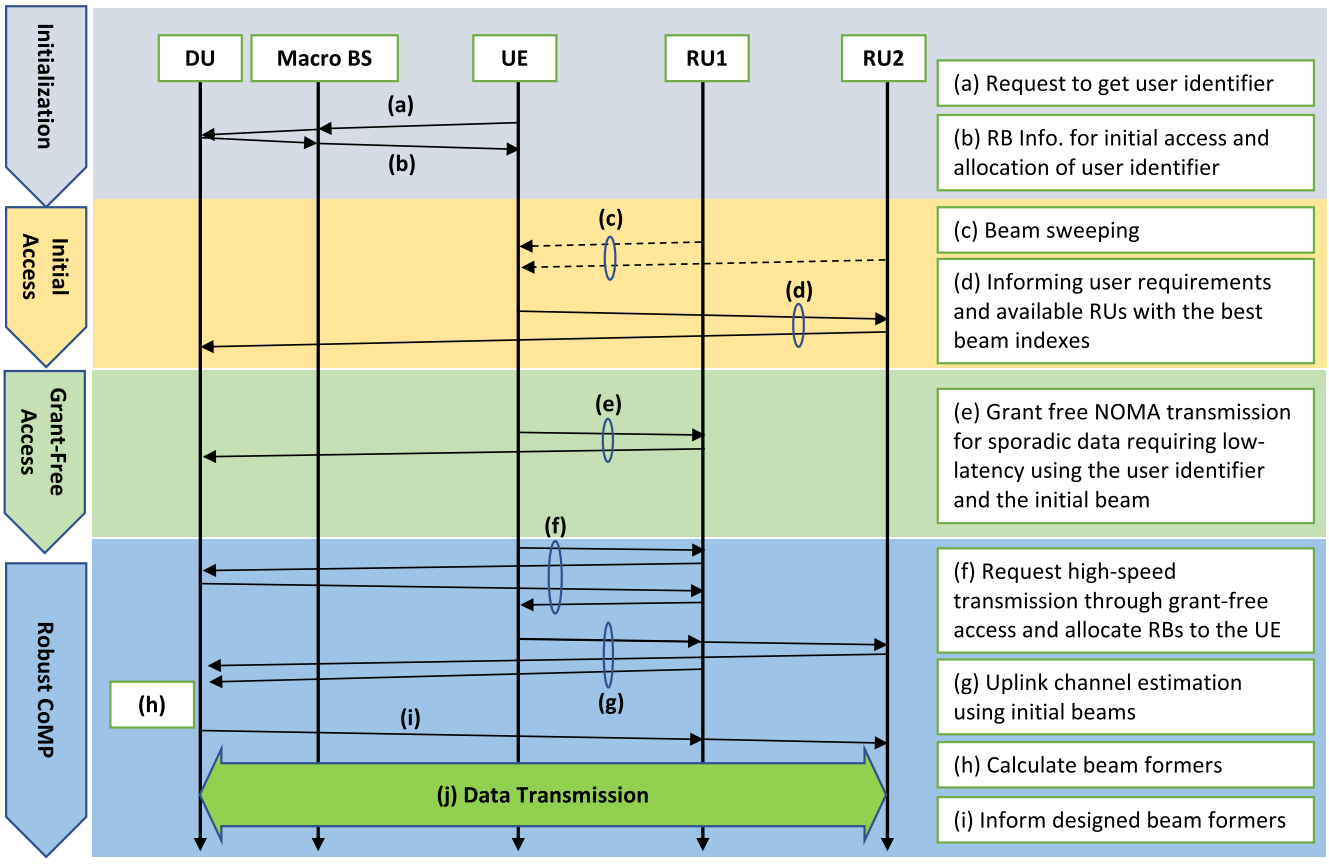

Fig. 2 Sequential procedure of proposed flexible RAN; initialization through the macro base station (BS) using microwave bands, initial access using grant-free NOMA, grant-free uplink transmission for low-latency applications, and robust CoMP for highly reliable uplink/downlink communications.

mote healthcare. The latter is suitable for remote driving with a 4K/8K high-resolution camera, real-time AR/VR systems, and so on.

Based on the requirements of every UE, the system optimally allocates RBs to UEs, such that the ratio of RBs for different requirements is optimally controlled by the central DU. For UEs with low-latency requirements, a sufficient number of RBs are reserved for GF-NOMA. The corresponding UE chooses one of the available RUs with the pre-determined initial beam to minimize the latency or maximize the resulting SNR and transmit data with their user identifiers, namely non-orthogonal pilot sequences, as shown in Fig. 2(e).

Every UE requiring high-data-rate transmission first sends the message to one of the RUs to request it with the required minimum rate via GF-NOMA. Then, the UE is informed of RB allocation. This process is illustrated in Fig. 2(f). Then, as shown in (g) in the figure, the UE transmits its identifier to all the connected RUs based on the initial beams at the given RBs, and the DU estimates the channel state information (CSI) between the UE and the RUs. The DU calculates the optimal beamformer for robust CoMP transmission with the aid of the *blockage prediction* [18], [19], which is depicted in (h). Note that this prediction is realized by the side information obtained by cameras installed on every RU [20]–[22] or sub-6 GHz signals [23], [24], and the algorithm, which is based on machine learning, predicts either the instantaneous blockage or the probabilities of blockage occurring for every UE, details of which are provided in Sect. 4. The obtained beamformers are then delivered to the corresponding RUs, as shown in (i), and uplink or downlink transmission is performed as in (j). Once the transmission is completed, the corresponding UE returns to the idle mode and restarts from process (f) when necessary.

Even if the UE was to exit the area of the corresponding RUs, the allocated user identifier is shared among all the RUs in the network; thus, the communications can be maintained. Moreover, even if the UE was to entirely leave the area of the DU, the DU can send the information of the assignment of the user identifier to the neighboring DUs and easily realize a smooth handover.

In the subsequent sections, we provide the technical particulars of GF-NOMA and robust CoMP with blockage prediction.

3. Grant-Free NOMA for Low-Latency Massive Access

As described in Sect. 2, grant-free access is a key enabler of low-latency initial access and low-latency data transmission for a large number of users. Uplink data transmission in current cellular networks is grant-based; every active user transmits its access request to the BS, and then the response is sent back from the BS as a *grant* for the data transmission. As described in [25], this granting procedure results in a latency of around 9.5 millisecond (ms), which hinders meeting the strict latency requirements of some IoT use cases [26]. Hence, grant-free access techniques such as K-repetition [27], variants of diversity slotted ALOHA [28]–[30] and code-domain GF-NOMA [25], [31]–[34] have been actively investigated to meet these low-latency requirements. Code-domain GF-NOMA can allow more users to transmit simultaneously without the BS's granting process and can theoretically achieve the most power-efficient communications [35]. This led us to propose a new code-domain GF-NOMA, which uses time and frequency domains that are compatible with the frame structure of 5G new radio (5G NR) [36]. Moreover, considering the requirements of Beyond 5G and 6G, we target 1 ms transmission over the air while maintaining a reasonable data rate[†].

3.1 Mathematical Model

Without loss of generality, we consider an uplink GF-NOMA system comprising K single-antenna UEs and a common RU equipped with an M-antenna uniform linear array[††]. The uplink transmission is organized into T orthogonal frequency-division multiplexing (OFDM) symbols with N subcarriers (samples) and a subcarrier spacing ΔB.

All UEs utilize radio resources following the frame structure illustrated in Fig. 3, in which $P \le N$ subcarriers are uniformly allocated as pilot subcarriers, and the others are used for data transmission. As shown in the figure, each user identifier (non-orthogonal sequence) is placed over time and frequency while each data symbol is repeated D times and placed in D subcarriers (namely only over frequency). The subsets of the indices of the pilot and data subcarriers are defined as $\mathcal{P}$ and $\mathcal{D}$, respectively.

Let $\mathbf{Y}_{\mathrm{p}}^{(t)} \in \mathbb{C}^{P\times M}$ denote the received signals in the subset $\mathcal{P}$ at the t-th OFDM symbol after cyclic prefix (CP) removal and discrete Fourier transform (DFT) modulation. The received signals are then given by

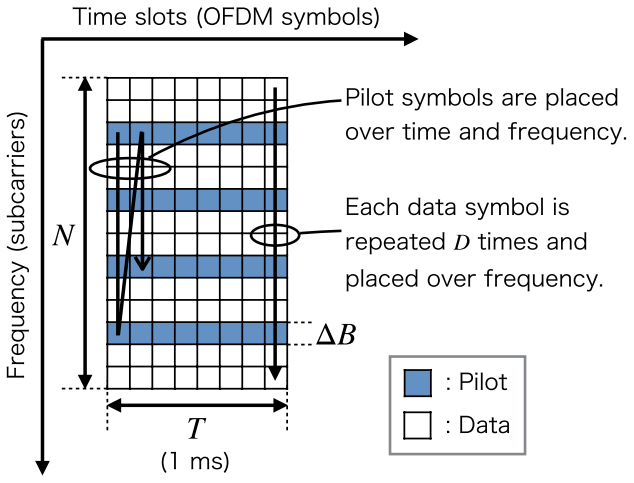

Fig. 3 Illustration of the uplink signal model.

$$\begin{aligned}\mathbf{Y}_{\mathrm{p}}^{(t)} &= \sum_{k\in\mathcal{A}} \mathrm{diag}(\mathbf{s}_k^{(t)})\mathbf{G}_k + \mathbf{Z}_{\mathrm{p}}^{(t)} \\ &= \sum_{k\in\mathcal{A}} \mathbf{S}_k^{(t)}\mathbf{G}_k + \mathbf{Z}_{\mathrm{p}}^{(t)}, \end{aligned} \tag{1}$$

where $\mathcal{A}$ denotes the set of active UEs and $\mathbf{S}_k^{(t)} = \mathrm{diag}(\mathbf{s}_k^{(t)})$ is a diagonal matrix based on the user identifier (namely, non-orthogonal pilot sequence) of the k-th UE, denoted by $\mathbf{s}_k^{(t)} \in \mathbb{C}^{P\times 1}$. The design of the user identifier is crucial to the exploitation of sparsity by the code-domain GF-NOMA owing to sporadic traffic. Hence, a few promising designs have been proposed, such as the frame-theoretical design [37] and deterministic design for odd lengths [38]. In this study, we assume that $\mathbf{s}_k^{(t)}$ is a random unimodular sequence with an amplitude of one element. In addition, $\mathbf{G}_k = [\mathbf{g}_{k,1}, \ldots, \mathbf{g}_{k,P}]^{\mathrm{T}} \in \mathbb{C}^{P\times M}$ is the channel frequency response (CFR) between the RU and the k-th UE, and the matrix $\mathbf{Z}_{\mathrm{p}}^{(t)} \in \mathbb{C}^{P\times M}$ represents the noise, in which the elements follow a complex Gaussian distribution with zero mean and variance σ_{n}^2.

We define $\mathbf{H}_k \in \mathbb{C}^{L\times M}$ as the channel impulse response (CIR) from the k-th UE to the RU with $L = \lceil \tau_{\max} W \rceil + 1$ taps, where W and $\tau_{\max}$ are the system bandwidth and maximum path delay, respectively. Then, the CFR can be expressed in terms of the CIR as follows:

$$\begin{aligned}\mathbf{G}_k &= \sqrt{N}\mathbf{F}_{P,L}\mathbf{H}_k \\ &= \sqrt{P}\bar{\mathbf{F}}_{P,L}\mathbf{H}_k, \end{aligned} \tag{2}$$

where $\mathbf{F}_{P,L} \in \mathbb{C}^{P\times L}$ is a matrix containing P rows according to $\mathcal{P}$ and the first L columns of the $N \times N$ DFT matrix $\mathbf{F}_N$ and $\bar{\mathbf{F}}_{P,L}$ is a column-normalized version of $\mathbf{F}_{P,L}$. Without loss of generality, we assume that $\tau_{\max}$ is smaller than the CP duration. Because the number of significant paths in the delay domain is limited [39], the CIR $\mathbf{H}_k$ is row sparse and the number of non-zero rows in $\mathbf{H}_k$ is less than L_{path}. Then, (1) can be rewritten as follows:

[†]Note that our GF-NOMA can be also used in microwave bands. However, a low-latency initial access and data transmission based on GF-NOMA are more important in mmWave bands since the unstable nature of mmWave channels frequently causes a large delay.

[††]In this section, we focus only on the case with a single RU because every UE chooses the best RU in terms of either the minimum latency or maximum SNR for grant-free transmission, as described above.

$$\begin{aligned}\mathbf{Y}_{\mathrm{p}}^{(t)} &= \sum_{k\in\mathcal{A}} \mathbf{S}_k^{(t)}\bar{\mathbf{F}}_{P,L} \times \sqrt{P}\mathbf{H}_k + \mathbf{Z}_{\mathrm{p}}^{(t)} \\ &= \sum_{k\in\mathcal{A}} \mathbf{A}_k^{(t)}\tilde{\mathbf{H}}_k + \mathbf{Z}_{\mathrm{p}}^{(t)} \\ &= \mathbf{A}^{(t)}\tilde{\mathbf{H}} + \mathbf{Z}_{\mathrm{p}}^{(t)}, \end{aligned} \tag{3}$$

where $\mathbf{A}_k^{(t)} \triangleq \mathbf{S}_k^{(t)}\bar{\mathbf{F}}_{P,L}$, $\tilde{\mathbf{H}}_k \triangleq \sqrt{P}\mathbf{H}_k$, $\mathbf{A}^{(t)} = [\mathbf{A}_1^{(t)}, \ldots, \mathbf{A}_K^{(t)}] \in \mathbb{C}^{P\times KL}$, and $\tilde{\mathbf{H}} = [\tilde{\mathbf{H}}_1^{\mathrm{T}}, \ldots, \tilde{\mathbf{H}}_K^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{KL\times M}$. Furthermore, based on the assumption that $\|\mathbf{s}_k^{(t)}\|_2 = \sqrt{P}$ and each column of $\bar{\mathbf{F}}_{P,L}$ is normalized, $\mathbf{A}^{(t)}$ is a unit-norm matrix.

To make use of the time domain, we consider the following signal model representing the signals received through T OFDM symbols:

$$\begin{aligned}\mathbf{Y}_{\mathrm{p}} &= [(\mathbf{Y}_{\mathrm{p}}^{(1)})^{\mathrm{T}}, \ldots, (\mathbf{Y}_{\mathrm{p}}^{(T)})^{\mathrm{T}}]^{\mathrm{T}} \\ &= \mathbf{A}\mathbf{X} + \mathbf{Z}_{\mathrm{p}} \in \mathbb{C}^{PT\times M}, \end{aligned} \tag{4}$$

where $\mathbf{A} = [(\mathbf{A}^{(1)})^{\mathrm{T}}, \ldots, (\mathbf{A}^{(T)})^{\mathrm{T}}]^{\mathrm{T}}/\sqrt{T} \in \mathbb{C}^{PT\times KL}$ and $\mathbf{X} = \sqrt{T}\tilde{\mathbf{H}}$. Then, all the columns of $\mathbf{A}$ are normalized. Note that the model presented in (4) implies that the proposed GF-NOMA system employs different sequences across different time slots (equivalently, OFDM symbols) to enlarge the dimension of the measurement, *that is*, the number of rows in $\mathbf{Y}_{\mathrm{p}}$, thereby enabling reliable sparse recovery based on compressed sensing (CS). Therefore, in this paper, a CS-based sparse recovery algorithm named *generalized multiple-measurement vector approximate message passing* (GMMV-AMP) is assumed to recover active UEs and CIRs, efficiently [40].

The received signal corresponding to the d-th data component at the t-th OFDM symbol and the m-th receiving antenna is given by:

$$\begin{aligned}\mathbf{y}_{m,\mathrm{d}}^{(t)} &= \sum_{k\in\mathcal{A}} \mathbf{g}_{k,m,\mathrm{d}} x_{k,\mathrm{d}}^{(t)} + \mathbf{z}_{m,\mathrm{d}}^{(t)} \\ &= \mathbf{G}_{m,\mathrm{d}}\mathbf{x}_{\mathrm{d}}^{(t)} + \mathbf{z}_{m,\mathrm{d}}^{(t)} \in \mathbb{C}^{D\times 1}, \end{aligned} \tag{5}$$

where $\mathbf{g}_{k,m,\mathrm{d}} \in \mathbb{C}^{D\times 1}$ and $\mathbf{z}_{m,\mathrm{d}}^{(t)} \sim \mathcal{CN}(\mathbf{0}_{\mathrm{d}}, \sigma_n^2\mathbf{I}_{\mathrm{d}})$ denote the CFRs between the RU and the k-th UE at the m-th receiving antenna and additive white Gaussian noise (AWGN), respectively. In addition, $x_{k,\mathrm{d}}^{(t)} \in \mathcal{X}$, where $\mathcal{X}$ is the set of Q-ary modulated symbols and represents the t-th data symbol transmitted by user k. The matrix $\mathbf{G}_{m,\mathrm{d}} \in \mathbb{C}^{D\times K_{\mathrm{a}}}$ and the vector $\mathbf{x}_{\mathrm{d}}^{(t)} \in \mathcal{X}^{K_{\mathrm{a}}\times 1}$ comprise the active users' CFRs and data symbols, respectively. For ease of data transmission, the transmitted data symbols are directly mapped onto subcarriers in subset $\mathcal{D}$, which comprises the indices that are uniformly selected from all available subcarriers with the exception of $\mathcal{P}$.

The signals received by M antennas can be expressed as

$$\begin{aligned}\mathbf{y}_{\mathrm{d}}^{(t)} &= [(\mathbf{y}_{1,\mathrm{d}}^{(t)})^{\mathrm{T}}, \ldots, (\mathbf{y}_{M,\mathrm{d}}^{(t)})^{\mathrm{T}}]^{\mathrm{T}} \\ &= \begin{bmatrix}\mathbf{G}_{1,\mathrm{d}} \\ \vdots \\ \mathbf{G}_{M,\mathrm{d}}\end{bmatrix}\mathbf{x}_{\mathrm{d}}^{(t)} + \begin{bmatrix}\mathbf{z}_{1,\mathrm{d}}^{(t)} \\ \vdots \\ \mathbf{z}_{M,\mathrm{d}}^{(t)}\end{bmatrix} \\ &= \mathbf{G}_{\mathrm{d}}\mathbf{x}_{\mathrm{d}}^{(t)} + \mathbf{Z}_{\mathrm{d}}^{(t)} \in \mathbb{C}^{DM\times 1}. \end{aligned} \tag{6}$$

Note that, as is true for the pilot component, the CFRs in (6) can be obtained using the CIRs, *that is*

$$[\mathbf{g}_{k,1,\mathrm{d}}, \ldots, \mathbf{g}_{k,M,\mathrm{d}}] = \sqrt{N}\mathbf{F}_{D,L}\mathbf{H}_k,\ k \in \mathcal{A}, \tag{7}$$

where $\mathbf{F}_{D,L} \in \mathbb{C}^{D\times L}$ is a matrix comprising D rows according to $\mathcal{D}$ and the first L columns of $\mathbf{F}_N$. The proposed GF-NOMA efficiently estimates the transmitted data using symbol-level Gaussian belief propagation (GaBP) [41] by considering the relationship between the CFR and CIR.

3.2 Design Guideline in Asymptotic Resume

The proposed GF-NOMA utilizes GMMV-AMP to estimate the active UEs and CIRs. The performance of this algorithm determines the overall performance of the proposed GF-NOMA; thus, the corresponding system parameters must be chosen appropriately to guarantee the recovery quality of GMMV-AMP. To this end, we propose a system design method based on the phase transition of the algorithm in an asymptotic resume [42], [43].

The theoretical phase transition can be obtained by evaluating the minimum mean squared error (MSE), which is defined as $\mathcal{M}(\epsilon|\eta)$, where $\epsilon = \rho\delta$ and η denote, respectively, a sparsity ratio and a denoiser for approximate message passing (AMP) [42]. Then, AMP succeeds with a high probability when the following condition is satisfied [42]:

$$\delta > \mathcal{M}(\epsilon|\eta). \tag{8}$$

As a convincingly successful region, we contemplated using the phase transition derived in [43], in which a single-measurement vector recovery (SMVR) problem in the complex domain is considered. Because this is a special case of a multiple-measurement vector recovery (MMVR) problem, its estimation performance can serve as the lower bound of that of an MMVR problem. According to [43, Theorem III.5], ρ and δ for complex AMP (CAMP) satisfy the following relation for $\tau \in [0, \infty)$:

$$\rho = \frac{\chi_1(\tau)}{(1+\tau^2)\chi_1(\tau) - \tau\chi_2(\tau)}, \tag{9}$$

$$\delta = \frac{4(1+\tau^2)\chi_1(\tau) - 4\tau\chi_2(\tau)}{-2\tau + 4\chi_2(\tau)}, \tag{10}$$

where $\chi_1(\tau) \triangleq \int_{\omega\geq\tau} \omega(\tau-\omega)\mathrm{e}^{-\omega^2}d\omega$ and $\chi_2(\tau) \triangleq \int_{\omega\geq\tau} \omega(\tau-\omega)^2\mathrm{e}^{-\omega^2}d\omega$, respectively. The largest phase transition for CAMP can then be obtained by exploiting τ that maximizes the value of ρ in (9).

However, contrary to this, we decided to exploit a theoretical approach to recover block-sparse signals [42] to obtain the largest achievable phase transition. This approach utilizes the fact that MMVR problems can be expressed using SMVR problems with block-sparse signals. In addition, this approach considers the performance of AMP using a block soft-thresholding denoiser. To clarify the ultimate

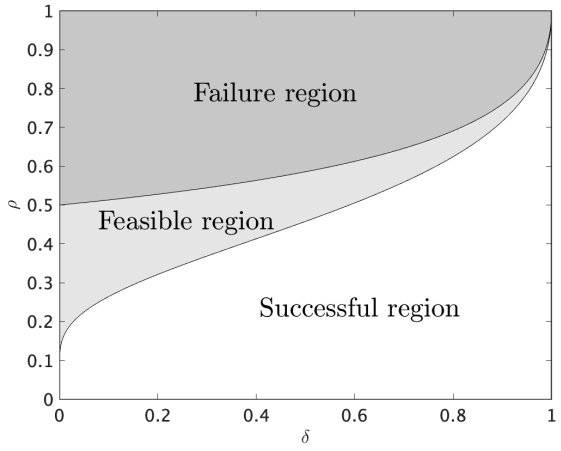

Fig. 4 Illustration of regions classified by phase transitions in [42] and [43] as guidelines for appropriate parameter design.

boundary, we focus on the case of an infinite block size, *that is*, $M \to \infty$. Following [42, Lemma 3.3], we obtain the minimum MSE for a large block size:

$$\mathcal{M}(\epsilon|\eta) = 2\epsilon - \epsilon^2. \tag{11}$$

According to (8) and (11), the largest phase transition for a large block size is given by $\delta = 2\epsilon - \epsilon^2$.

These theoretically derived phase transitions serve as a guideline for appropriate parameter design, as indicated by the three regions in Fig. 4, which are divided by boundaries obtained from the largest phase transition for CAMP and using (11) with $M \to \infty$. In Fig. 4, the "Failure region" represents the region in which the accurate estimation of $\mathbf{X}$ using (4) is absolutely impossible even if $M \to \infty$. In contrast, the "Successful region" is where the highest degree of accuracy of estimation is surely achievable and the "Feasible region" is where accurate estimation can be performed because it has a boundary for $M > 1$. Thus, the figure indicates that the values of ρ and δ in the proposed GF-NOMA system should exist in at least the "Feasible region."

Hereafter, the variables ρ and δ in the proposed GF-NOMA system are given for the sake of simplicity by

$$\rho = \frac{K_{\mathrm{a}} L_{\mathrm{path}}}{PT}, \quad \delta = \frac{PT}{KL}, \tag{12}$$

where $K_{\mathrm{a}} L_{\mathrm{path}}$ corresponds to the maximum number of non-zero rows in $\mathbf{X}$, which varies as a result of the randomness of path delays. The values of T that satisfy the latency requirement are strictly limited by the subcarrier spacing ΔB, and the value of L depends on both ΔB and the system bandwidth W. Accordingly, we can use the number of pilot subcarriers and OFDM symbols for active user detection (AUD) and channel estimation (CE) to determine the successful (or feasible) region characterized by phase transitions by taking into account ΔB and W.

Table 1 Simulation parameters.

Meaning	Character	Value
Num. of UEs	K	500
Num. of active UEs	K_{a}	50
Num. of antennas at the RU	M	8
Num. of subcarriers	N	4096
Num. of pilot subcarriers	P	36 (3RBs)
Num. of data subcarriers	D	16
Num. of used OFDM symbols	T	28
Num. of (visible) taps in CIRs	L	25
Num. of significant paths	L_{path}	6
Subcarrier spacing	ΔB	30 [kHz]
System bandwidth	W	10 [MHz]
Modulation order	Q	4

3.3 Numerical Examples

We evaluated the performance of the proposed GF-NOMA system using computer simulation. The values of the simulation parameters are listed in Table 1. Note that $T = 28$ is equivalent to the number of transmissible OFDM symbols of length 1 ms within a 5G NR subframe with a subcarrier spacing of 30 kHz. As each RB in the 5G NR comprises 12 subcarriers, we based the number of pilot subcarriers P on the assumption that RB is one unit. Moreover, the value of P in Table 1 is determined according to the design guidelines described in Sect. 3.2. Specifically, for a given K, K_{a}, L, L_{path}, and T, we searched for the value of P, where ρ and δ in (12) are slightly above the boundary between the feasible and successful regions in Fig. 4. In all simulations, the SNR was defined as $\mathrm{SNR} \triangleq 1/\sigma_{\mathrm{n}}^2$, and Gray-coded QPSK was employed as a modulation scheme. Furthermore, we set the maximum iterations of GMMV-AMP and GaBP to 200 and 16, and the damping factors to 0.3, and 0.5, respectively.

Figure 5 shows the activity error rate (AER) performance with $\mathrm{SNR} = 10\,[\mathrm{dB}]$, $K_{\mathrm{a}} = 50$, and the number of OFDM symbols varying from 14 to 56. The performance results of the conventional scheme with $P = 60$ (5RBs) are also shown. These results reveal that, although the performance of the proposed scheme is inferior to that of the conventional scheme when $T < 24$ owing to the shrinkage of the dimensionality of the received signals, the former significantly outperforms the latter otherwise. Note that the performance of the proposed scheme can be improved by using more pilot subcarriers, even when T is small, as the dimensionality of the received signals is determined by the product of P and T. In contrast, increasing P in the conventional scheme does not yield any performance gain. These results indicate that the proposed GF-NOMA can take advantage of both the time and frequency domains to support a massive number of users more efficiently than the conventional approach.

In addition, we evaluate the achievable throughput per active user of the proposed GF-NOMA in terms of the following effective throughput metric:

$$R_{\mathrm{eff}} \triangleq (1 - \mathrm{FER}) N_{\mathrm{sym}} T \log_2 Q, \tag{13}$$

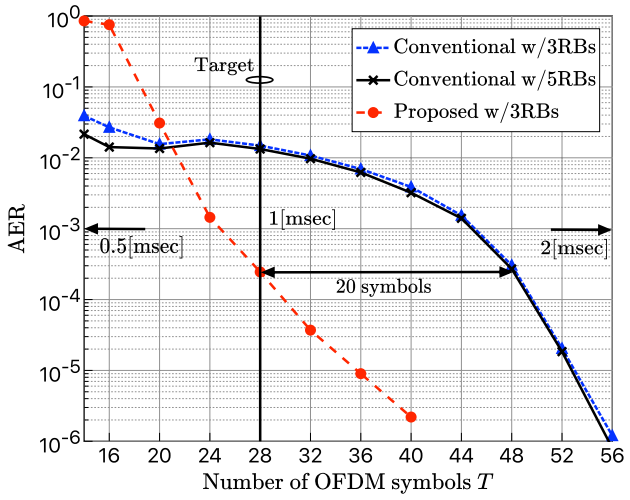

Fig. 5 AER as a function of T with SNR = 10 dB and K_a = 50. The value of T varies from 14 (0.5 ms) to 56 (2 ms).

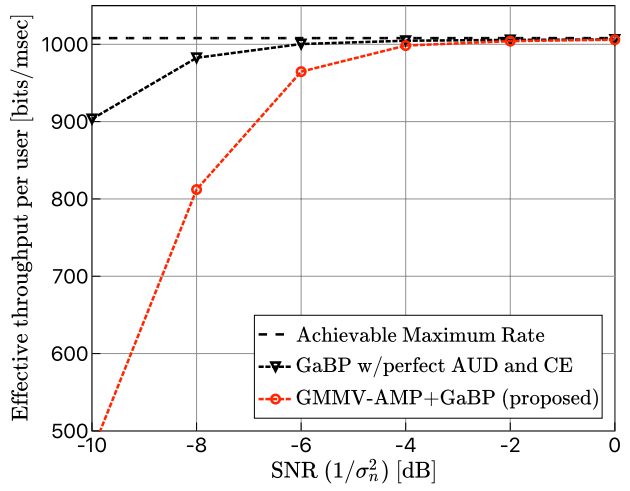

Fig. 6 Effective throughput of proposed scheme with K_a = 50, T = 28, P = 36, and D = 16.

where N_{sym} denotes the number of transmissible data symbols in a single OFDM symbol. Note that a system using the setup defined in Table 1 can use 24RBs for data transmission while supporting K_a = 50 active users. Thus, when D = 16, each active user can transmit N_{sym} = 18 data symbols in a single OFDM symbol.

Figure 6 shows the effective throughput of the proposed scheme with K_a = 50, T = 28, P = 36, and D = 16. As benchmarks, the achievable maximum rate is plotted in the figure to indicate the performance limit of the proposed scheme. The GaBP with ideal AUD and CE values is also plotted. The results demonstrate that, while performing AUD, CE, and MUD, the proposed scheme enables active UEs to attain the maximum rate when the SNR is sufficiently high. The high achievable throughput beyond 32 bytes/ms is worth noting and indicates that the proposed GF-NOMA has the potential to satisfy the general requirement for URLLC defined under the 3GPP standard [44]†. These results indicate that the proposed scheme is suitable for massive low-latency communication at moderate data rates.

4. Robust CoMP with Blockage Prediction for Reliable High-Speed Access

In this section, we present the technical details of robust CoMP transmission to realize uplink/downlink transmission at a high data rate with guaranteed quality. Here, two different blockage prediction methods are considered, and different beam designs that employ these prediction methods are studied. Note that although we describe the design of downlink communications here, the extension to uplink communications is straightforward.

4.1 Mathematical Model

We consider the downlink CoMP transmission over mmWave channels, in which multiple RUs equipped with a uniform plane array (UPA) consisting of N_t antenna elements cooperatively serve a single-antenna UE subjected to unpredictable blockages. Let $b \in \mathcal{B} \triangleq \{1, 2, \ldots, B\}$ and $u \in \mathcal{U} \triangleq \{1, 2, \ldots, U\}$ denote the RU and UE indices, respectively, where B and U denote the total number of RUs and UEs, respectively.

Following [45], we assume that the communication channel between the b-th RU and the u-th UE consists of $K_{b,u}$ clusters, where $K_{b,u}$ is modeled as $K_{b,u} \sim \max(1, \text{Poisson}(\lambda))$ with the intensity parameter λ [46]. One of the clusters corresponds to the line-of-sight (LOS) path, and the other corresponds to the non-line-of-sight (NLOS) paths. The CSI between the b-th RU and the u-th UE can be estimated by assuming a time division duplex (TDD) and the reciprocity between the uplink and downlink as described in Sect. 2. However, in practice, unpredictable sudden blockages of wireless paths are inevitable because of human bodies, buildings, vehicles, and so on, which results in system outage. In this study, every path blockage is modeled by the Bernoulli random variable $\omega_{b,u}^k \in \{0, 1\}$ with mean $p_{b,u}^k$ depicting the corresponding blockage probability. The actual channel between the bth RU and the uth UE can then be modeled as

$$\mathbf{h}_{b,u} = \sqrt{\frac{1}{K_{b,u}}} \sum_{k=1}^{K_{b,u}} \omega_{b,u}^k g_{b,u}^k \mathbf{a}_{N_t}\left(\theta_{b,u}^k, \phi_{b,u}^k\right), \tag{14}$$

where $\theta_{b,u}^k$ and $\phi_{b,u}^k$ are the elevation and azimuth angle of departure (AoD) of the k-th cluster from the b-th RU toward the u-th UE, respectively, and $\mathbf{a}_{N_t}\left(\theta_{b,u}^k, \phi_{b,u}^k\right)$ is the array response vector of the UPA on the transmitter side. In addition, $g_{b,u}^k \sim \mathcal{CN}\left(0, 10^{-\text{PL}_{b,u}^k/10}\right)$, and the path loss $\text{PL}_{b,u}^k$ is defined by [46]. Without loss of generality, k = 1 corresponds

†Concretely, the requirement is defined as $1 - 10^{-5}$ reliability within 1 ms user plane latency for 32 bytes.

to the LOS component.

Let $\mathbf{f}_{b,u} \in \mathbb{C}^{N_\mathrm{t}\times 1}$ denote the beamforming vector from the b-th RU toward the u-th UE such that the received signal y_k at the u-th UE is written as

$$\begin{aligned} y_u &= \sum_{b\in\mathcal{B}} \mathbf{h}_{b,u}^{\mathrm{H}}\mathbf{f}_{b,u}x_u + \sum_{u'\in\mathcal{U}\setminus u}\sum_{b\in\mathcal{B}} \mathbf{h}_{b,u}^{\mathrm{H}}\mathbf{f}_{b,u'}x_{u'} + n_u \\ &= \mathbf{h}_u^{\mathrm{H}}\mathbf{f}_u x_u + \sum_{u'\in\mathcal{U}\setminus u} \mathbf{h}_u^{\mathrm{H}}\mathbf{f}_{u'}x_{u'} + n_u, \end{aligned} \tag{15}$$

where $x_u \sim \mathcal{CN}(0,1)$ is the transmitted signal intended for the u-th UE, and $n_u \sim \mathcal{CN}\left(0,\sigma_u^2\right)$ is AWGN with power density σ_u^2. The vectors $\mathbf{h}_u$ and $\mathbf{f}_u$ are defined as $\mathbf{h}_u \triangleq \left[\mathbf{h}_{1,u}^{\mathrm{T}},\ldots,\mathbf{h}_{B,u}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{BN_\mathrm{t}\times 1}$ and $\mathbf{f}_u \triangleq \left[\mathbf{f}_{1,u}^{\mathrm{T}},\ldots,\mathbf{f}_{B,u}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{BN_\mathrm{t}\times 1}$. Moreover, the signal-to-noise interference ratio (SINR) is given by:

$$\Gamma_u(\mathbf{h}_u,\mathbf{f}) = \frac{\left|\mathbf{h}_u^{\mathrm{H}}\mathbf{f}_u\right|^2}{\sum_{u'\in\mathcal{U}\setminus u}\left|\mathbf{h}_u^{\mathrm{H}}\mathbf{f}_{u'}\right|^2 + \sigma_u^2}, \tag{16}$$

where $\mathbf{f} \triangleq \left[\mathbf{f}_1^{\mathrm{T}},\ldots,\mathbf{f}_U^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{UBN_\mathrm{t}\times 1}$.

4.2 Beamforming Design with Different Blockage Prediction Methods

Based on the model defined above, we continue our discussion of the beamforming design of the robust CoMP with two different blockage prediction methods: instantaneous blockage prediction and blockage probability prediction. Note that, in the following, we do not consider the specific algorithm for the prediction because it is beyond the scope of the work presented in this paper; instead, we only consider the side information that can be used in the design of beamformers.

4.2.1 Instantaneous Blockage Prediction

We first consider the case in which the system predicts the occurrence of instantaneous path blockages, typically using cameras and an algorithm based on machine learning. Based on the output of the blockage prediction, the estimated CSI can be written as:

$$\hat{\mathbf{h}}_{b,u} = \sqrt{\frac{1}{K_{b,u}}}\sum_{k=1}^{K_{b,u}} \hat{\omega}_{b,u}^k g_{b,u}^k \mathbf{a}_{N_\mathrm{t}}\left(\theta_{b,u}^k,\phi_{b,u}^k\right), \tag{17}$$

where $\hat{\omega}_{b,u}^k \in \{0,1\}$ is an estimate of the blockage corresponding to the kth path. If the prediction incorrectly estimates the blockage, an error event, $\hat{\omega}_{b,u}^k \neq \omega_{b,u}^k, \forall b,u,k$, occurs, and the estimated CSI does not match the actual channel response. In practice, it is significantly difficult to predict the blockage events of NLOS paths; therefore, most existing algorithms only estimate the blockage corresponding to the LOS path. Therefore, in the following, we assume that the system predicts only $\hat{\omega}_{b,u}^1$ and sets $\hat{\omega}_{b,u}^k = 1$ for $k \neq 1$.

Using this estimated CSI, standard deterministic robust optimization approaches are applicable to the design of beamformers that prevent degradation of the resulting data rate due to path blockages [47], [48]. In particular, the beamforming design based on the worst-case optimization framework can be formulated as the following sum-rate maximization (SRM) problem:

$$\underset{\mathbf{f},\alpha_u,\forall u}{\text{maximize}} \sum_{u\in\mathcal{U}} \log_2(1+\alpha_u), \tag{18a}$$

$$\text{subject to } \Gamma_u\left(\hat{\mathbf{h}}_u,\mathbf{f}\right) \geq \alpha_u, \forall u \in \mathcal{U}, \tag{18b}$$

$$\sum_{u\in\mathcal{U}} \left\|\mathbf{f}_{b,u}\right\|_2^2 \leq P_{\max,b}, \forall b \in \mathcal{B}, \tag{18c}$$

where α_u is an auxiliary variable. The SINR constraint given by (18b) is non-convex, and thus we resort to successive convex approximation (SCA) to approximate the SINR constraints as a convex function, which enables the solution to be obtained efficiently [48].

On the other hand, Charan et al. [49] pointed out the difficulty of synchronization between the RUs in the CoMP transmission. They proposed an alternative that entailed switching the RUs based on the prediction of the occurrence of LOS component blockage to avoid system outage due to path blockages. The best RU with the highest received power, namely $b_u^{\mathrm{opt}} = \underset{b\in\mathcal{B}}{\mathrm{argmin}} \left\|\hat{\mathbf{h}}_{b,u}\right\|_2$, is assumed to transmit the information to the u-th UE. Then, the DU solves (18) while limiting the terms corresponding to b_u^{opt} in (16).

4.2.2 Blockage Probability Prediction

Precise prediction requires instantaneous blockage prediction to frequently update the prediction result because the update interval must be even shorter than the movements of the objects around the UE. Moreover, the prediction error directly affects the overall performance and stability of communications. Instead of predicting the instantaneous blockages, estimating the probability of blockage occurrence along every path is a judicious alternative that relies on received-signal-power prediction based on machine learning [20]. Based on the output of the predictor, the estimated CSI available for the beamforming design is given by:

$$\tilde{\mathbf{h}}_{b,u}^m = \sqrt{\frac{1}{K_{b,u}}}\sum_{k=1}^{K_{b,u}} \Omega_{b,u}^{k,m} g_{b,u}^k \mathbf{a}_{N_\mathrm{t}}\left(\theta_{b,u}^k,\phi_{b,u}^k\right), \tag{19}$$

where $\Omega_{b,u}^{k,m} \in \{0,1\}$ denotes a Bernoulli random variable that takes zero with the probability given by the blockage prediction. Because this estimated CSI is a random variable, the estimated SINR also becomes a random variable. Therefore, to guarantee the minimum rate, it is necessary to design robust beamforming by solving the sum outage probability minimization (OutMin) problem:

$$\underset{\mathbf{f}}{\text{minimize}} \sum_{u\in\mathcal{U}} \Pr\left\{\Gamma_u\left(\tilde{\mathbf{h}}_u^m,\mathbf{f}\right) < \gamma_u\right\}, \tag{20a}$$

Table 2 Simulation parameters.

Meaning	Character	Value
Num. of RUs	B	4
Num. of transmit antennas	N_{t}	16
Num. of UEs	U	2
Carrier frequency	f_{c}	28 [GHz]
Bandwidth	W	100 [MHz]
Noise power	σ_u^2	−88.93 [dBm]
Transmit power constraint	$P_{\max,b}$	30 [dBm]

$$\text{subject to} \sum_{u\in\mathcal{U}} \left\|\mathbf{f}_{b,u}\right\|_2^2 \le P_{\max,b}, \tag{20b}$$

where γ_u is the target SINR calculated from the target rate, and $P_{\max,b}$ is the maximum transmit power for each RU. In [18], the OutMin beamforming design was reformulated as an empirical loss minimization (ERM) problem [50] by introducing a generalized smooth hinge surrogate function ν:

$$\nu\left(\Gamma_u\left(\mathbf{h}_u,\mathbf{f}\right)\right) = \begin{cases} 0 & \text{if } 1-\frac{\Gamma_u(\mathbf{h}_u,\mathbf{f})}{\gamma_u} < 0 \\ 1-\frac{\Gamma_u(\mathbf{h}_u,\mathbf{f})}{\gamma_u} & \text{otherwise} \end{cases} \tag{21}$$

From (20) and (21), we obtain

$$\underset{\mathbf{f}}{\text{minimize}} \sum_{u\in\mathcal{U}} \mathbb{E}_\Omega\left[\nu\left(\tilde{\mathbf{h}}_u^m,\mathbf{f}\right)\right], \tag{22a}$$

$$\text{subject to} \sum_{u\in\mathcal{U}} \left\|\mathbf{f}_{b,u}\right\|_2^2 \le P_{\max,b}, \tag{22b}$$

where $\mathbb{E}_\Omega[\cdot]$ denotes the expected value operation for the blockage patterns. Because the hinge function ν is a convex function concerning the SINR Γ_u, the optimal solution of (22) can be obtained efficiently using the mini-batch gradient descent method by replacing the expected value in (22a) with the ensemble mean calculated by multiple channel realizations with different blockage patterns $\tilde{\mathbf{h}}_u^m$ $(m = 1, 2, \ldots, M_{\min})$.

4.3 Numerical Examples

In this section, the performance of the robust CoMP transmissions with different blockage prediction methods are quantitatively evaluated via computer simulations, which reveal the relationship between the accuracy of the blockage prediction and the achievable throughput.

The simulation parameters used in this section are listed in Table 2. We assume that the RUs are placed at each corner of a square area with sides of 100 m, and the number of antenna elements of each RU is $N_{\mathrm{t}} = 4 \times 4 = 16$. The probability of blockage occurrence of each path $p_{b,u}^k$ follows a uniform distribution in the interval $[0.2, 0.6]$, as in [11], [12]. The outage probability and effective throughput are defined as $\Pr\{\log_2(1+\Gamma_u) < \log_2(1+\gamma_u)\}$ and $\mathbb{E}[a_u \log_2(1+\Gamma_u)]$, respectively, where a_u is a variable, which equals 0 when an outage occurs, and 1 otherwise. The initial value required for each beamforming design is given by the minimum mean square error (MMSE) approach. The parameters in the gradient descent are taken from [18], and the SRM problem (18) is solved using SDPT3, which is a convex optimization solver for CVX [51].

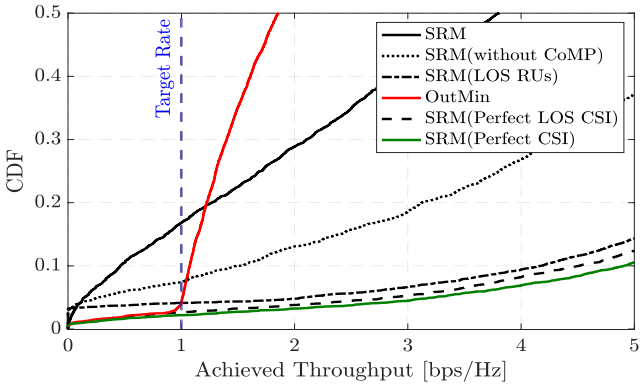

Fig. 7 CDF of achievable throughput with the given target rate $\log_2(1+\gamma_u) = 1.0$ [bps/Hz].

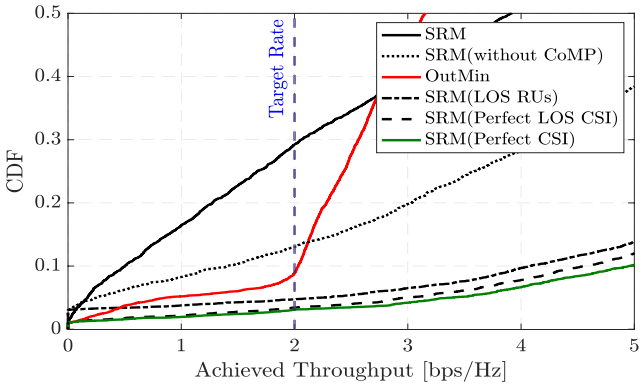

Fig. 8 CDF of achievable throughput with the given target rate $\log_2(1+\gamma_u) = 2.0$ [bps/Hz].

Figures 7 and 8 compare the achievable throughputs of different CoMP approaches described in Sect. 4.2 in terms of the cumulative distribution function (CDF), where their target throughputs are given as 1.0 and 2.0 [bps/Hz], respectively. Note that perfect prediction of the path blockages is assumed here. The solid black curve represents the performance of the SRM beamforming without the blockage prediction, and the dotted black curve denoted "SRM (without CoMP)" represents the performance of the methods that choose the best RU with the highest received power and solving the SRM with the RU. Moreover, "SRM (LOS RUs)" with the dotted/dashed black curve indicates the performance of the SRM-based CoMP transmission only with the RUs, of which the LOS path is predicted as having no blockage, and "OutMin" with the solid red curve indicates the performance of the CoMP using the blockage probability prediction. For reference, the SRM-based CoMP with perfect knowledge of the blockage of the LOS path is plotted as a black dashed curve denoted "SMR (Perfect LOS CSI)." In this case, all the RUs transmit signals even if the LOS paths are blocked; namely, the RUs assume that NLOS components are never blocked. Moreover, the SRM-based

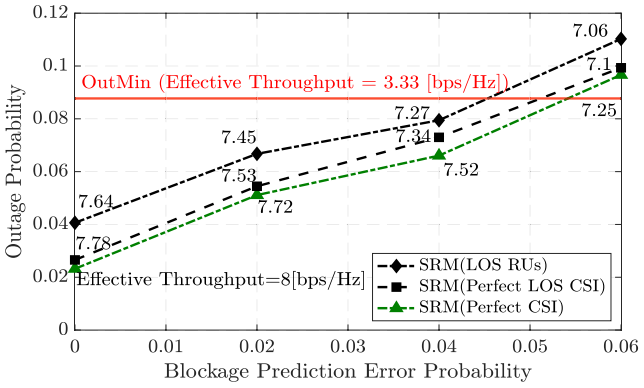

Fig. 9 Outage probabilities and effective throughputs with the given target throughput $\log_2(1+\gamma_u) = 2.0$ [bps/Hz] for different blockage prediction accuracy.

CoMP with perfect CSI, including the effects of LOS/NLOS blockages, is additionally plotted as a green solid curve to indicate the achievable lower bound.

As shown in the figures, the performance of the SRM without prediction is even worse than that of the other approaches. The system cannot meet the required data rate of approximately 20% of users when the target rate is 1 [bps/Hz] and approximately 30% when the target rate is 2 [bps/Hz]. Using the prediction, "SRM (without CoMP)" improves the performance. Although this approach can prevent the effect of sudden path blockage by choosing one RU without blockage from among multiple RUs, this limits the spatial degrees of freedom and results in the difference from "SRM (LOS RUs)." At the same time, the discrepancy between the performance of "SRM (LOS RUs)" and "SRM (Perfect LOS CSI)" still remains because the "SRM (Perfect LOS CSI)" can exploit the NLOS components using all the RUs. The gap between "SRM (Perfect LOS CSI)" and "SRM (Perfect CSI)" indicates a mismatch between the actual channels and the estimated channels in terms of the NLOS components. Surprisingly, the performance of "OutMin" is superior to that of "SRM (LOS RUs)" even though "OutMin" does not track the instantaneous realization of the blockages and approaches the performance of "SRM (Perfect CSI)" at the target rate when the target rate is 1 [bps/Hz]. However, when the target rate is 2 [bps/Hz], "SRM (LOS RUs)" outperforms "OutMin" which clearly indicates the advantage of tracking the instantaneous channel fluctuation.

Finally, Fig. 9 evaluates the performance of the outage probability with different blockage-prediction accuracies. The horizontal axis represents the probability that $\hat{\omega}^k_{b,u} \neq \omega^k_{b,u}, \forall b, u, k$. In addition, the numbers attached to the markers indicate the resulting effective throughput. As shown in the figure, when the error probability of the blockage prediction exceeds approximately 5%, the performance of the CoMP approaches with the instantaneous blockage prediction becomes worse than that with the blockage probability prediction. Therefore, considering the simplicity of the blockage probability prediction, "OutMin" can be a judicious option as an access scheme with a high and guaranteed data rate. However, because the effective throughput of "OutMin" is lower than the others, the highly accurate prediction of blockages is also promising for realizing user-centric communications, particularly for high-data-rate access.

5. Conclusion

In this paper, we presented a flexible RAN to realize user-centric communications capable of meeting users' diverse requirements in terms of communication quality. We described the two technical pillars of the proposed RAN, GF-NOMA, and robust CoMP, which use blockage prediction. As is evident from the numerical examples, these technical pillars support low-latency access with a massive number of users and high-data-rate access with a guaranteed data rate.

We conclude this paper with several future tasks that would need to be accomplished to realize the flexible RAN.

- Time and frequency synchronization among RUs and UEs: In GF-NOMA, the carrier frequency offset (CFO) and the timing offset cause performance degradation [52]. In addition, precise time and frequency synchronization are necessary for robust CoMP [49]. These synchronization issues have already been reported in the literature [52], [53] but have not been fully addressed.
- Efficient resource allocation for different requests: The proposed flexible RAN has to accommodate a massive number of completely different types of user requirements, such that highly efficient resource allocation is essential. This design is not considered in this study, but would have to be addressed.
- Core network design: To guarantee end-to-end performance, the latency and the resources in the core network must be considered. Hence, integrating the design of the RAN and core network is a future task.

Acknowledgments

This research was supported by the Ministry of Internal Affairs and Communications in Japan (JPJ000254).

References

[1] ITU-R, "IMT Vision — "Framework and overall objectives of the future development of IMT for 2020 and beyond," ITU-R Recommendation M.2083, Sept. 2015.
[2] "Terahertz communication for vehicular networks," IEEE Trans. Veh. Technol., vol.66, no.7, pp.5617–5625, 2017.
[3] M.Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y.M. Jang, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," IEEE Open J. Commun. Soc., vol.1, pp.957–975, 2020.
[4] T. Murakami, Y. Kishi, K. Ishibashi, K. Kasai, H. Shinbo, M. Tamai, K. Tsuda, M. Nakazawa, Y. Tsukamoto, H. Yokoyama, Y. Fujii, Y. Seki, S. Nanba, T. Hara, F. Adachi, and T. Sotoyama, "Research project to realize various high-reliability communications in

advanced 5G network," Proc. IEEE Wireless Communications and Networking Conference (WCNC), pp.1–8, 2020.
[5] H. Shinbo, T. Murakami, Y. Tsukamoto, and Y. Kishi, "R&D of technology for high reliability management in advanced 5G network to meet the various requirements of different communication services," Proc. 2021 IEEE VTS 17th APWCS, pp.1–5, 2021.
[6] K. David and H. Berndt, "6G vision and requirements: Is there any need for beyond 5G?," IEEE Veh. Technol. Mag., vol.13, no.3, pp.72–80, 2018.
[7] U. Gustavsson, P. Frenger, C. Fager, T. Eriksson, H. Zirath, F. Dielacher, C. Studer, A. Pärssinen, R. Correia, J.N. Matos, D. Belo, and N.B. Carvalho, "Implementation challenges and opportunities in beyond-5G and 6G communication," IEEE J. Microw., vol.1, no.1, pp.86–100, 2021.
[8] O.E. 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, 2014.
[9] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays," IEEE J. Sel. Areas Commun., vol.35, no.7, pp.1432–1443, 2017.
[10] G.R. MacCartney, T.S. Rappaport, and S. Rangan, "Rapid fading due to human blockage in pedestrian crowds at 5G millimeter-wave frequencies," Proc. IEEE GLOBECOM'17, pp.1–7, 2017.
[11] S. Ju, O. Kanhere, Y. Xing, and T.S. Rappaport, "A millimeter-wave channel simulator NYUSIM with spatial consistency and human blockage," Proc. IEEE GLBECOM'19, pp.1–6, 2019.
[12] V. Raghavan, L. Akhoondzadeh-Asl, V. Podshivalov, J. Hulten, M.A. Tassoudji, O.H. Koymen, A. Sampath, and J. Li, "Statistical blockage modeling and robustness of beamforming in millimeter-wave systems," IEEE Trans. Microw. Theory Tech., vol.67, no.7, pp.3010–3024, 2019.
[13] A. Checko, H.L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M.S. Berger, and L. Dittmann, "Cloud RAN for mobile networks — A technology overview," IEEE Commun. Surveys Tuts., vol.17, no.1, pp.405–426, 2015.
[14] M. Alonzo, S. Buzzi, A. Zappone, and C. D'Elia, "Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave," IEEE Trans. Green Commun. Netw., vol.3, no.3, pp.651–663, 2019.
[15] G. Kwon and H. Park, "Joint user association and beamforming design for millimeter wave UDN with wireless backhaul," IEEE J. Sel. Areas Commun., vol.37, no.12, pp.2653–2668, 2019.
[16] J. Kim, H.W. Lee, and S. Chong, "Virtual cell beamforming in cooperative networks," IEEE J. Sel. Areas Commun., vol.32, no.6, pp.1126–1138, 2014.
[17] A.S. Guerreiro, H. Iimori, and K. Ishibashi, "Low latency beam-sweeping for millimeter wave systems via pessimistic optimization," IEEE Wireless Commun. Lett., vol.10, no.12, pp.2742–2746, 2021.
[18] H. Iimori, G.T.F. de Abreu, O. Taghizadeh, R.A. Stoica, T. Hara, and K. Ishibashi, "Stochastic learning robust beamforming for millimeter-wave systems with path blockage," IEEE Wireless Commun. Lett., vol.9, no.9, pp.1557–1561, 2020.
[19] H. Iimori, G.T.F. De Abreu, O. Taghizadeh, R.A. Stoica, T. Hara, and K. Ishibashi, "A stochastic gradient descent approach for hybrid mmWave beamforming with blockage and CSI-error robustness," IEEE Access, vol.9, pp.74471–74487, 2021.
[20] T. Nishio, H. Okamoto, K. Nakashima, Y. Koda, K. Yamamoto, M. Morikura, Y. Asai, and R. Miyatake, "Proactive received power prediction using machine learning and depth images for mmWave networks," IEEE J. Sel. Areas Commun., vol.37, no.11, pp.2413–2427, 2019.
[21] Y. Koda, J. Park, M. Bennis, K. Yamamoto, T. Nishio, M. Morikura, and K. Nakashima, "Communication-efficient multimodal split learning for mmwave received power prediction," IEEE Commun. Lett., vol.24, no.6, pp.1284–1288, 2020.
[22] S. Mihara, S. Ito, T. Murakami, and H. Shinbo, "Positioning for user equipment of a mmWave system using RSSI and stereo camera images," Proc. IEEE WCNC'21, pp.1–7, 2021.
[23] A. Ali, N. González-Prelcic, and R.W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," IEEE Trans. Wireless Commun., vol.17, no.2, pp.1038–1052, 2018.
[24] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," IEEE Trans. Commun., vol.68, no.9, pp.5504–5518, 2020.
[25] L. Liu, E.G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," IEEE Signal Process. Mag., vol.35, no.5, pp.88–99, 2018.
[26] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," IEEE Commun. Mag., vol.56, no.12, pp.119–125, 2018.
[27] 3rd Generation Partnership Project. 3GPP, TR 38.214 v15.9.0, "5G; NR; Physical layer procedures for data," April 2020.
[28] S. Ogata, K. Ishibashi, and G.T.F. de Abreu, "Optimized frameless ALOHA for cooperative base stations with overlapped coverage areas," IEEE Trans. Wireless Commun., vol.17, no.11, pp.7486–7499, 2018.
[29] M. Oinaga, S. Ogata, and K. Ishibashi, "Design of coded ALOHA with ZigZag decoder," IEEE Access, vol.7, pp.168527–168535, 2019.
[30] S. Ogata and K. Ishibashi, "Application of ZigZag decoding in frameless ALOHA," IEEE Access, vol.7, pp.39528–39538, 2019.
[31] T. Hara and K. Ishibashi, "Grant-free non-orthogonal multiple access with multiple-antenna base station and its efficient receiver design," IEEE Access, vol.7, pp.175717–175726, 2019.
[32] T. Hara, H. Iimori, and K. Ishibashi, "Hyperparameter-free receiver for grant-free NOMA systems with MIMO-OFDM," IEEE Wireless Commun. Lett., vol.10, no.4, pp.810–814, 2021.
[33] M.B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S.J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," IEEE Commun. Surveys Tuts., vol.22, no.3, pp.1805–1838, 2020.
[34] H. Iimori, T. Takahashi, K. Ishibashi, G.T.F. de Abreu, and W. Yu, "Grant-free access via bilinear inference for cell-free MIMO with low-coherence pilots," IEEE Trans. Wireless Commun., vol.20, no.11, pp.7694–7710, 2021.
[35] Y. Polyanskiy, "A perspective on massive random-access," Proc. 2017 IEEE International Symposium on Information Theory (ISIT), pp.2523–2527, 2017.
[36] 3rd Generation Partnership Project. 3GPP, TR 25.996 Ver. 14.2.0, "Study on new radio (NR) access technology; Physical layer aspects," Sep. 2017.
[37] R.A. Stoica, G.T.F. de Abreu, T. Hara, and K. Ishibashi, "Massively concurrent non-orthogonal multiple access for 5G networks and beyond," IEEE Access, vol.7, pp.82080–82100, 2019.
[38] N.Y. Yu, K. Lee, and J. Choi, "Pilot signal design for compressive sensing based random access in machine-type communications," Proc. 2017 IEEE Wireless Commun. Netw. Conf., San Francisco, CA, USA, pp.1–6, March 2017.
[39] C.R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," IEEE Commun. Mag., vol.48, no.11, pp.164–174, Nov. 2010.
[40] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," IEEE Trans. Signal Process, vol.68, pp.764–779, Jan. 2020.
[41] T. Takahashi, S. Ibi, and S. Sampei, "Design of adaptively scaled belief in multi-dimensional signal detection for higher-order modulation," IEEE Trans. Commun., vol.67, no.3, pp.1986–2001, March 2019.
[42] D.L. Donoho, I. Jhonstone, and A. Montanari, "Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising," IEEE Trans. Inf. Theory, vol.59, no.6, pp.3396–

3433, June 2013.
[43] A. Maleki, L. Anitori, Z. Yang, and R.G. Baraniuk, "Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)," IEEE Trans. Inf. Theory, vol.59, no.7, pp.4290–4308, July 2013.
[44] 3rd Generation Partnership Project. 3GPP, TR 38.913 v16.0.0, "Study on scenarios and requirements for next generation access technologies," July 2020.
[45] R.W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A.M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," IEEE J. Sel. Topics Signal Process., vol.10, no.3, pp.436–453, Feb. 2016.
[46] M.R. Akdeniz, Y. Liu, M.K. Samimi, S. Sun, S. Rangan, T.S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," IEEE J. Sel. Areas Commun., vol.32, no.6, pp.1164–1179, June 2014.
[47] D. Kumar, J. Kaleva, and A. Tolli, "Rate and reliability trade-off for mmWave communication via multi-point connectivity," Proc. IEEE GLOBECOM'19, Waikoloa, HI, USA, pp.1–6, 2019.
[48] D. Kumar, J. Kaleva, and A. Tölli, "Blockage-aware reliable mmWave access via coordinated multi-point connectivity," IEEE Trans. Wireless Commun., vol.20, no.7, pp.4238–4252, Feb. 2021.
[49] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-Aided 6G Wireless Communications: Blockage Prediction and Proactive Handoff," IEEE Trans. Veh. Technol., vol.70, no.10, pp.10193–10208, Aug. 2021.
[50] V. Vapnik, Principles of Risk Minimization for Learning Theory, Morgan Kaufmann Publishers Inc., San Francisco, CA, Systems, Dec. 1991.
[51] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, March 2014.
[52] G. Sun, Y. Li, X. Yi, W. Wang, X. Gao, L. Wang, F. Wei, and Y. Chen, "Massive grant-free OFDMA with timing and frequency offsets," IEEE Trans. Wireless Commun., vol.21, no.5, pp.3365–3380, 2022.
[53] T. Hara, H. Iimori, and K. Ishibashi, "Activity detection for uplink grant-free NOMA in the presence of carrier frequency offsets," Proc. ICC'20 Workshops, pp.1–6, 2020.

Koji Ishibashi received the B.E. and M.E. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2002 and 2004, respectively, and the Ph.D. degree in engineering from Yokohama National University, Yokohama, Japan, in 2007. From 2007 to 2012, he was an Assistant Professor with the Department of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu, Japan. Since April 2012, he has been with the Advanced Wireless and Communication Research Center (AWCC), The University of Electro-Communications, where he is currently a Professor. From 2010 to 2012, he was a visiting scholar at the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. His current research interests include grant-free access, cell-free architecture, millimeter-wave communications, energy harvesting communications, wireless power transfer, channel codes, signal processing, and information theory. He served as an associate editor of the IEICE Transactions on Communications from 2015 to 2020.

Takanori Hara received the B.E. and M.E. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree at the University of Electro-Communications. His current research interests include communication theory, channel coding, GF access, and information theory.

Sota Uchimura received the B.E. degree in engineering from The University of Electro-Communications, Tokyo, Japan, in 2020. His current research interests include millimeter-wave communications and signal processing.

Tetsuya Iye was born in Tokyo, Japan, in 1986. He received the B.S. degree in physics from Waseda University, Tokyo, Japan, in 2009 and M.S. and Ph.D. degrees in physics from Kyoto University, Kyoto, Japan, in 2011 and 2014, respectively. From 2013 to 2014, he was a Research Fellow of the Japan Society for the Promotion of Science. He joined Kozo Keikaku Engineering, Inc., Tokyo, Japan, as a Systems Engineer in 2014. His current research interests include millimeter-wave communications antennas and propagation, beamforming, and software-defined radio (SDR).

Yoshimi Fujii received B.S. and M.S. degrees in computer science and communication engineering from Kyushu University, Fukuoka, Japan, in 1989 and 1991, respectively. Since 1991, he has been working as a software engineer in the telecommunications field with Kozo Keikaku Engineering, Inc., Tokyo, Japan. His research interests include the various layers of wireless communication technology, especially the lower layers, such as the physical layer and media access layer. Currently, he is involved in a series of projects related to the implementation of the wireless communication PHY layer using a software-defined radio (SDR) approach. In addition to telecommunication, he is interested in the implementation of GNSS signal generators and receivers with an SDR.

Takahide Murakami received the B.E., M.E., and Ph.D. degrees all in communication engineering from Tohoku University, Sendai, Japan, in 2002, 2004, and 2007, respectively. He joined KDDI Corp. in 2007 and has been with KDDI Research, Inc. (formerly KDDI R&D Laboratories). His research interests include wireless communications and radio access networks.

Hiroyuki Shinbo received the B.S. degree in electro information communication in 1987 and the M.S. degree from the Graduate School of Information and Systems in 1990, both degrees from the University of Electro-Communications, Tokyo, Japan. He joined KDD Corp. (now KDDI Corp.) in 1999 and is employed in the KDD R&D Laboratory (now KDDI Research, Inc.). He was seconded to work at the Advanced Telecommunications Research Institute International from 2013 to 2016. He is currently a senior manager at the Advanced Radio Application Laboratory at KDDI Research, Inc. His research interests include beyond 5G/6G systems (especially radio access networks), TCP/IP, flying base stations, network operation systems, and lunar communications.