Hybrid Beamforming Based Multi-user mmWave Communications via Opportunistic Scheduling

Kanchan Verma, L. Yashvanth, *Member, IEEE*, and Chandra R. Murthy, *Fellow, IEEE*

***Abstract*—We study the problem of excessive channel training and optimization overheads in hybrid analog-digital (A-D) precoded multi-user millimeter-wave (mmWave) systems, using opportunistic beamforming (OBF) methods based on randomized analog precoding and opportunistic scheduling of user equipments (UEs) in dense Internet-of-Things (IoT) deployments. Two OBF schemes are considered: a fully random precoder (Full-RP), in which the analog phase shifts are chosen independently and uniformly at random, and a channel-structure-aware random precoder (CSA-RP), where the random beams are sampled to match the array-response structure of geometric mmWave channels. We analytically show that, under the full-RP scheme, the probability of achieving a near-beamforming (BF) gain decays exponentially with the number of base-station antennas. In contrast, the CSA-RP scheme achieves successful beam alignment with high probability using a moderate number of UEs, enabling near-optimal throughput (comparable to coherent BF) with best-UE selection. Motivated by this, we analyze the achievable throughput of CSA-RP for both single- and multiple- radio-frequency (RF) chain architectures. First, for the single-RF-chain case, we derive analytical upper and lower bounds on the average throughput. Then, for the multiple-RF-chain case, we characterize the beam-alignment process using a Hypergeometric distribution and establish the asymptotic throughput scaling law under maximal ratio transmission-based digital precoding strategy. We validate our analytical findings using Monte-Carlo simulations and show that multiple RF chains provide significant OBF gains in sparse-user regime with low complexity.**

***Index Terms*—MmWave, Opportunistic beamforming, Multi-User diversity, Average throughput, Rate-based Scheduling.**

I. INTRODUCTION

Millimeter-wave (mmWave) communication is a key enabler for next-generation wireless systems due to the large bandwidths available in the 30–300 GHz spectrum, which can support multi-gigabit-per-second data rates [2]–[4]. To compensate for the severe propagation loss at mmWave frequencies, base stations (BSs) typically employ large antenna arrays to provide high beamforming (BF) gains to user equipments (UEs). However, large number of antennas substantially increases the overhead associated with channel state information (CSI) acquisition, feedback, and precoder optimization. These overheads are further exacerbated in hybrid analog–digital (A–D) architectures and become particularly prohibitive in dense-user scenarios such as Internet-of-Things (IoT) networks, where a large number of low-power devices must be served efficiently. In this paper, we address this challenge by developing opportunistic beamforming (OBF) schemes in hybrid A-D architectures based on randomized analog precoding and UE scheduling. This eliminates per-antenna CSI acquisition and joint precoder optimization while retaining most of the benefits of conventional BF.

Kanchan Verma is currently with Samsung, Bengaluru, India, and was with the Dept. of ECE, Indian Institute of Science (IISc), Bengaluru, India, 560012 during the development of this work. L. Yashvanth is currently with the Dept. of EEE, University College London (UCL), WC1E 6BT London, U.K. and was with the Dept. of ECE, IISc, Bengaluru, India, 560012 during the development of this work. Chandra R. Murthy is with the Dept. of ECE, IISc, Bengaluru, India, 560012. (e-mails: kanchanverma@iisc.ac.in, l.yashvanth@ucl.ac.uk, cmurthy@iisc.ac.in.). This work was presented in part at the National Conference on Communications (NCC), Feb. 2023 [1].

A. Challenges and Motivation

Next-generation IoT networks are envisioned to support massive numbers of heterogeneous devices, such as sensors, actuators, vehicles, appliances, and industrial machines, that communicate with a central controller or BS in a wireless mode. The large spectrum resources available in the mmWave bands make them a promising candidate for supporting such dense-device IoT scenarios. However, the use of mmWave frequencies introduces fundamental implementation challenges. Due to the high propagation loss, large antenna arrays are required at the BS to provide sufficient BF gain. Conventional mmWave transmission strategies therefore rely on a 3-step procedure to serve each UE: (i) acquisition of CSI between the BS antennas and the UE, (ii) feedback of the estimated CSI from the UE to the BS, and (iii) optimization of the precoding vectors based on the acquired CSI. Each of these steps incurs significant overhead. CSI acquisition requires training durations that scale at least linearly with the number of BS antennas, while CSI feedback demands reliable uplink signaling whose overhead also grows with antenna dimensions. Precoder optimization typically involves computationally intensive algorithms that must be executed frequently to adapt to time-varying channels. These overheads are further exacerbated in practical hybrid A–D architectures, where only low-dimensional observations of the channel are available. Importantly, all of these steps must be repeated for every UE and in every scheduling interval, making conventional CSI-driven mmWave BF unfavorable for large-scale IoT networks.

Addressing this challenge is critical for two main reasons: $a)$ excessive training and feedback overhead directly reduce the fraction of time available for data transmission, thereby limiting the achievable throughput away from the information-theoretic maximum, and $b)$ many IoT devices are designed to operate with minimal computational capability and limited energy budgets, making it impractical to require UEs to perform accurate channel estimation and frequent feedback. Unless these overheads are significantly reduced, the potential benefits of mmWave communication may be difficult to realize in IoT-centric deployments. This motivates the need for alternative transmission paradigms that can exploit the statistical and structural properties of large UE networks while minimizing signaling and computational overhead, yet still procuring the maximum benefits offered by mmWave communication.

B. Related Work and Novelty

Hybrid A–D architectures have emerged as a practical solution for mmWave systems due to the limited number of radio-frequency (RF) chains [5], [6]. For single-user mmWave systems, sparse channel estimation and hybrid precoding based on orthogonal matching pursuit (OMP) were proposed in [7], leveraging channel sparsity to approach the performance of fully digital systems. Extensions to multi-user mmWave systems were considered in [8], and codebook-based precoding schemes were developed in [9], [10]. It has also been shown that, from the perspective of achieving array gain, purely analog beamforming can be sufficient under certain conditions [11], [12]. In the context of IoT networks, several recent works have explored mmWave system design to support dense device connectivity and low-overhead operation. These include beam-squint mitigation [13], modeling and optimization of reconfigurable intelligent surfaces (RIS)-aided mmWave systems [14], localization and mapping at mmWave frequencies [15], accelerated beam optimization strategies [16], cooperative transmission schemes [17], learning-based codebook selection and modulation-and-coding scheme (MCS) adaptation [18], and performance analysis of hybrid-precoded and extremely large-scale multiple-input multiple-output (MIMO) systems [19], [20]. However, these approaches generally rely on CSI acquisition, feedback, and precoder optimization, which induce substantial signaling and computational overhead and are challenging in dense IoT deployments. In particular, while RIS-assisted schemes improve performance through the joint optimization of the BS precoder and RIS phase shifts [21], [22], they exacerbate these overheads by requiring estimation of the cascaded BS-RIS-UE channels, whose dimensionality grows with the number of BS antennas, users, and RIS elements. The acquired cascaded CSI must then be conveyed to the BS for BF and phase-shift design; further, the joint optimization of the active BF and RIS phase shifts is computationally expensive and must be repeatedly performed as the wireless environment changes. Similarly, recent learning-based approaches seek to improve BF performance by learning optimized BF strategies in a data driven manner [23]–[27]. Although they reduce the online channel-estimation overhead by utilizing a limited number of pilot measurements or compressed channel observations, they still require channel-related measurements for beam prediction, large prior channel datasets, extensive offline training, and periodic retraining to adapt to changing propagation environments. Thus, despite their differences, both RIS- and learning-based approaches in the mmWave bands continue to incur significant signaling and computational overheads.

Motivated by this, in this paper, we develop opportunistic BF schemes based on randomized analog precoding and opportunistic UE scheduling. Our central premise is that, in IoT networks with a large number of UEs, a randomly chosen precoding configuration is likely to be near-optimal for at least one UE, which can then be scheduled for data transmission without explicit per-antenna CSI estimation or precoder optimization. Then, the system throughput approaches the throughput obtained via optimal BF to the scheduled UE, but without incurring the overheads, thanks to the *multi-user diversity* effect. We recognize that opportunistic BF has been studied in reasonable depth in sub-6 GHz systems in [28], [29]; however these ideas are relatively less explored in mmWave bands, which exhibit fundamentally different propagation characteristics compared to the sub-6 GHz bands. Similar ideas were explored in [30], [31] for mmWave bands; but these focus on asymptotic regimes with very large antenna arrays and fully digital architectures, which limits their generality and practicality. Thus, the broad novelty of our work can be summarized as follows:

- We provide a rigorous analytical treatment of randomized OBF schemes in *practical mmWave systems with hybrid analog-digital architectures and finite antenna arrays*, unlike prior works that focus on asymptotic or fully digital settings or sub-6 GHz environments.
- By exploiting the *resolvability of mmWave beams*, we derive a novel characterization for beam-alignment under randomized precoding; this is a first result of its kind.
- We derive *interpretable scaling laws*, which for the first time, characterize the interaction between array gain, multi-user diversity, and multi-path diversity under OBF.

We are now ready to explain our technical contributions.

C. Contributions of this Paper

The key contributions of this paper are as follows:

1) We propose low-complexity opportunistic OBF schemes for downlink multi-user mmWave systems with hybrid A-D architectures. First, focusing on a single-RF-chain configuration, we develop two schemes: a fully random precoder (Full-RP) and a channel-structure-aware random precoder (CSA-RP), both of which eliminate explicit channel estimation and analog precoder optimization. We also present a practical transmission protocol that enables seamless integration of the proposed schemes into existing wireless systems with minimal overhead.
2) For the Full-RP scheme, we analytically show that the probability of achieving a near-BF gain decays exponentially with the number of BS antennas. In contrast, for the CSA-RP scheme, we prove that the probability of successful beam alignment converges to 1 and, importantly, remains independent of the number of BS antennas. This difference arises from sampling the analog precoder in the latter according to the geometric structure of mmWave channels (see Theorems 1 and 2).
3) Leveraging the favorable alignment properties of CSA-RP, we derive tight analytical lower and upper bounds on the achievable average throughput in the single-RF-chain setting. Our analysis shows that CSA-RP achieves the full array gain while simultaneously harnessing the multi-user diversity gains at low complexity (see Theorems 3 and 4).

4) We extend the CSA-RP framework to systems with multiple RF chains and present a complete transmission protocol. Under randomized multi-beam transmission, we characterize the beam-alignment process using a discrete random variable and show that it follows a Hypergeometric distribution whose parameters depend on the number of RF chains, resolvable paths, and BS antennas (see Lemma 1).
5) Building on the above, we derive asymptotic throughput scaling laws under maximal ratio transmission (MRT)-based digital precoding for the multiple-RF-chain CSA-RP scheme. We show that the achievable throughput simultaneously benefits from (a) full array gain, (b) multi-user diversity, and (c) additional multipath diversity enabled by multiple RF chains, and we precisely quantify these gains using tools from extreme value theory (see Theorem 5).
6) Finally, we provide extensive Monte Carlo simulations to validate all analytical results. The simulations reveal distinct operating regimes in which single- or multiple-RF-chain architectures are preferable and provide clear insights into the performance-complexity trade-offs of OBF-based mmWave systems under realistic configurations. In particular, the results identify user-density-dependent regimes in which single-RF-chain architectures achieve near-optimal performance, while multiple RF chains offer significant gains primarily in sparse-user scenarios (see Sec. V).

Our results show that in an mmWave communication system, OBF schemes procure superior benefits even with moderate number of UEs compared to optimization-based methods, at low complexity, and without extensive CSI estimation/precoder optimization and associated feedback overheads.

II. System Model and Problem Statement

A. System Model

We consider an mmWave system in which an N_t-antenna BS serves K single antenna UEs using a fully-connected hybrid A-D precoding architecture equipped with $N_{\rm RF} \le N_t$ RF chains as shown in Fig. 1. The digital and analog precoders at the BS are denoted by $\mathbf{F}_{\rm BB} \in \mathbb{C}^{N_{\rm RF}\times N_s}$ and $\mathbf{F}_{\rm RF} \in \mathbb{C}^{N_{\rm t}\times N_{\rm RF}}$, respectively, where $1 \le N_s \le N_{\rm RF}$ represents the number of independent data streams that the BS can support at any point in time.[1] Then, at time slot t, the transmitted signal from the BS can be written as

$$\mathbf{x}(t) = \mathbf{F}_{\rm RF}(t)\mathbf{F}_{\rm BB}(t)\mathbf{s}(t), \tag{1}$$

where $\mathbf{s}(t) \in \mathbb{C}^{N_s}$ is the transmit symbol vector at time t and has a total power constraint of $P_{\rm txt}$, i.e., $\mathbb{E}\left[\|\mathbf{s}(t)\|_2^2\right] = P_{\rm txt}$.

Due to the high attenuation losses and the directional propagation of signals in the mmWave bands, we model the small-scale downlink channel, $\mathbf{h}_k(t) \in \mathbb{C}^{1\times N_t}$ from the BS to UE-k using the well-known Saleh-Valenzuela model [7], given by

$$\mathbf{h}_k(t) = \sqrt{N_t\big/L}\sum\nolimits_{\ell=1}^{L} \alpha_{k,\ell}(t)\mathbf{a}_{\rm BS}^H(\psi_{k,\ell}(t)), \tag{2}$$

[1] In line with widely adopted practice, we consider a single-user (SU) MIMO setting in this paper [7]. So, every RF chain modulates the data to serve a single scheduled UE in a time slot. However, the core idea can be extended to multi-user (MU)-MIMO as well; we will pursue this in future.

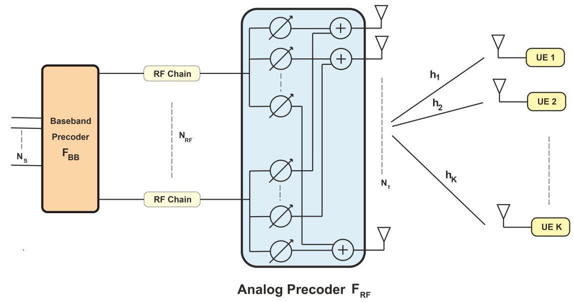


Fig. 1: Multi-user mmWave system with hybrid A-D BF.

where L is the number of paths between the BS and UE-k, $\alpha_{k,\ell} \sim \mathcal{CN}(0,1)$ is the complex gain of the ℓth path, and $\psi_{k,\ell}$ is the normalized angle of departure (AoD) of the ℓth path at the BS towards the kth UE, and is modeled as a uniformly distributed random variable with support $[-1,1)$. Further, we consider a uniform linear array (ULA) at the BS; so, its array steering response vector, $\mathbf{a}_{\rm BS}(\psi_{k,\ell}) \in \mathbb{C}^{N_t}$ is given by

$$\mathbf{a}_{\rm BS}\left(\psi_{k,\ell}\right) = \left(1\big/\sqrt{N_t}\right)\left[1, e^{j\pi\psi_{k,\ell}}, \ldots, e^{j\pi(N_t-1)\psi_{k,\ell}}\right]^T, \tag{3}$$

and the normalized AoD $\psi_{k,\ell}$ is related to its physical AoD $\phi_{k,\ell}$ as $\psi_{k,\ell} = \sin\phi_{k,\ell}$, where we considered that the inter-antenna spacing at the BS is set to half-signal wavelength. Thus, at time slot t, the received signal at UE-k is given by

$$\begin{aligned} y_k(t) &= \sqrt{\vartheta_k}\mathbf{h}_k(t)\mathbf{x}(t) + z_k(t) && (4)\\ &= \sqrt{\vartheta_k}\mathbf{h}_k(t)\mathbf{F}_{\rm RF}(t)\mathbf{F}_{\rm BB}\mathbf{s}(t) + z_k(t), && (5)\end{aligned}$$

where z_k is the additive white noise, distributed as $\mathcal{CN}(0,\sigma^2)$ and ϑ_k denotes the path loss at UE-k.

B. Problem Description

Consider an mmWave system design that aims to maximize the achievable system throughput under an SU-MIMO mode. From Sec. II, this translates to determining the optimal A-D precoders for each UE that maximizes the UE's throughput. Let $\mathbf{F}_{\rm opt}^*$ denote the optimal *composite precoder* across the A-D stages obtained by solving the corresponding rate maximization problem. Then, implementing $\mathbf{F}_{\rm opt}^*$ entails the following overheads that scales monotonically with N_t:

- Channel estimation at all UEs, which requires:
  1) Dedicated pilot resources in time, frequency, and power,
  2) Accurate CSI estimation algorithms.
- High computational complexity in determining $\mathbf{F}_{\rm RF}^*, \mathbf{F}_{\rm BB}^*$, for example, by first finding $\mathbf{F}_{\rm opt}^*$, and then decomposing it into A-D components by solving [7]:

$$\mathbf{F}_{\rm RF}^*, \mathbf{F}_{\rm BB}^* = \arg\min_{\mathbf{F}_{\rm RF},\mathbf{F}_{\rm BB}} \left\|\mathbf{F}_{\rm opt}^* - \mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\right\|_F^2. \tag{6}$$

Since these overheads are repeatedly incurred whenever a new UE is scheduled or the channel conditions change, the practical throughput achieved with an A-D architecture falls well below the information-theoretic maximum. To address the challenges posed by the significant overheads and complexity of optimal A-D precoding, we develop and analyze two low-complexity algorithms aimed at maximizing system throughput. Specifically, we exploit the inherent multi-user diversity gain, which does not require full channel estimation or precoder optimization, to enhance system throughput. To this end, we examine two OBF schemes: Full-RP and CSA-RP, both of which rely on random precoder configurations and opportunistic scheduling of UEs for data transmission. We then address the following:

1) How does the performance of the Full-RP scheme compare to that of the CSA-RP scheme?
2) How do the proposed schemes perform as the number of RF chains varies? In particular, how does the availability of multiple RF chains benefit OBF performance compared to the single-RF-chain case?

III. Opportunistic BF with a Single RF chain

We first begin by considering a single RF chain scenario, i.e., $N_{\mathrm{RF}} = 1$, in order to demonstrate the key idea and the ability of the proposed scheme to achieve the optimal throughout without incurring any of the precoding optimization overheads. Since only one UE can be served using a single RF chain, it follows that $N_s = 1$. For simplicity of exposition, we let $\mathbf{F}_{\mathrm{BB}} = 1$, which simplifies the overall received signal at UE-k in (5) as

$$y_k(t) = \sqrt{\vartheta_k}\mathbf{h}_k(t)\mathbf{F}_{\mathrm{RF}}(t)s_k(t) + z_k(t),\ \ k = 1, 2, \ldots, K. \tag{7}$$

Then, the overall signal-to-noise ratio (SNR) at UE-k in slot t is given by

$$\gamma_k(t) = \rho_k|\mathbf{h}_k(t)\mathbf{F}_{\mathrm{RF}}(t)|^2, \tag{8}$$

where $\rho_k \triangleq P_{\mathrm{txt}}\vartheta_k/\sigma^2$ denotes the average received SNR of UE-k. As a consequence, the achievable data rate (normalized by the bandwidth) of UE-k in time slot t is

$$R_k(t) = \log_2(1 + \gamma_k(t)). \tag{9}$$

Next, we describe the mechanism of the OBF schemes.

A. Transmission Mechanism

The proposed OBF schemes entails the following 4-steps:

- *Step 1:* In each time slot, the BS randomly samples the phase shifts of the RF precoder $\mathbf{F}_{\mathrm{RF}}(t)$ and broadcasts a pilot signal to all UEs.
- *Step 2:* Each UE measures its received SNR $\gamma_k(t)$, based on the effective channel using (8).
- *Step 3:* The BS identifies the *best user* $\hat{k}(t)$, i.e., the UE experiencing the highest SNR according to

$$\hat{k}(t) \triangleq \arg\max_{k\in[K]\triangleq\{1,\ldots,K\}} \gamma_k(t), \tag{10}$$

  using efficient, low-complexity feedback mechanisms such as timer-based schemes [32] (elaborated below).
- *Step 4:* The BS schedules the selected UE $\hat{k}(t)$ for data transmission in the current time slot.

This procedure is repeated in every time slot, with the precoding vector $\mathbf{F}_{\mathrm{RF}}(t)$ independently drawn from a suitably designed distribution. Then, the resulting average system throughput is given by [29]

$$R^{(K)} = \mathbb{E}\left[\log_2\left(1 + \max_{k\in[K]}\gamma_k(t)\right)\right], \tag{11}$$

where the expectation is taken over the joint distribution of $\mathbf{F}_{\mathrm{RF}}(t)$ and the channels $\mathbf{h}_1(t), \ldots, \mathbf{h}_K(t)$.

Feedback Mechanism: A key component of the above scheme is to enable the BS identify the UE with the highest SNR in each time slot. In IoT systems with a large number of UEs, K, conventional feedback approaches incur substantial overhead that scales with K, making them impractical. To address this, we can adopt the timer-based feedback scheme proposed in [32]. In this scheme, after measuring its SNR, each UE starts a timer whose expiry time is inversely proportional to the measured SNR. The UE whose timer expires first (i.e., the UE with the highest SNR) transmits a short feedback message containing its identity. Upon receiving this feedback packet, the BS designates the corresponding UE as the best UE and schedules it for data transmission in the current slot [32].[2]

Thus, the proposed random analog precoding-based OBF schemes incur minimal feedback/optimization overhead and eliminate the need for high-dimensional channel estimation or precoder configuration. In the sequel, we analyze two OBF schemes: full RP and CSA-RP, which differ in the distribution used to randomly configure the analog precoder.

B. Full RP Scheme

In this scheme, we construct $\mathbf{F}_{\mathrm{RF}}(t)$ at time t as

$$\mathbf{F}_{\mathrm{RF}}(t) = 1\Big/\sqrt{N_t}\left[e^{j\theta_1(t)}, \ldots,\ e^{j\theta_{N_t}(t)}\right]^T, \tag{12}$$

where $\theta_1(t), \ldots, \theta_{N_t}(t)$ are independent and identically distributed (i.i.d.) variables which are uniformly drawn at random from $[0, 2\pi)$. Going forward, we drop the index t for brevity. Clearly, the choice of precoder as in (12) is simple to implement, as it requires no knowledge of the channel states from the BS to the K UEs. However, we next demonstrate that the full RP scheme performs worse than schemes with optimized precoders for a given UE, primarily due to a mismatch between the subspace spanned by the UEs' channel vectors and the *search* subspace induced by the random precoding vectors at the BS. To formalize this, we present the following result assuming many BS antennas for analytical tractability (rest of the results in this paper do not require this assumption); however, we numerically show that it remains valid even with a modest number of antennas (as few as 4 antennas).

[2]Note that practical timer-based feedback schemes can be efficiently implemented using a finite set of discrete timer levels, which have been shown to maximize the probability of successful UE selection [32], [33]. Such timer mappings can improve the reliability of the UE selection process by reducing timer collisions (and hence feedback errors) while preserving the low-feedback nature of the protocol. Moreover, since timer values are fixed based on the channel quality metric, even if occasional feedback imperfections or timer collisions prevent the selection of the optimal UE, the selected UE is still likely to deliver large gains, thereby limiting the degradation of performance. Thus, for simplicity, and in line with [34], we assume error-free feedback in our analysis. A rigorous characterization of imperfect feedback within the OBF framework is beyond the scope of this work and is left for future research.

Theorem 1. *Let $\eta \in (0,1)$ be a small number. The probability that a randomly chosen BS precoder configuration under the full-RP scheme procures at least a $(1-\eta)$ factor of the beamforming gain at UE-k, when N_t is large, is given by*

$$P_{succ,k}^{(\eta)} = e^{-(1-\eta)N_t}, \tag{13}$$

where N_t is the number of antennas at the BS.

Proof. Define the $(1-\eta)$-success event of the full RP OBF scheme at UE-k for some $\eta > 0$ as

$$\mathcal{E}_k^{(\eta)} \triangleq \left\{\mathbf{h}_k, \mathbf{F}_{\mathrm{RF}} : |\mathbf{h}_k \mathbf{F}_{\mathrm{RF}}|^2 \geq (1-\eta)\|\mathbf{h}_k\|_2^2\right\}, \tag{14}$$

which quantifies that the channel gain seen by UE-k using the randomly chosen BS precoder is at least a $(1-\eta)$ fraction of the channel gain obtained under beamforming configuration. Now, using the law of total probability, we can write

$$\mathsf{Pr}\left(\mathcal{E}_k^{(\eta)}\right) = \mathbb{E}_{\mathbf{h}_k}\left[\mathsf{Pr}\left(\mathcal{E}_k^{(\eta)} \middle| \sigma(\mathbf{h}_k)\right)\right], \tag{15}$$

where $\sigma(\mathbf{h}_k)$ is the sigma-algebra generated by the random vector $\mathbf{h}_k$. Using (12), we can obtain $\mathsf{Pr}\left(\mathcal{E}_k^{(\eta)} \middle| \sigma(\mathbf{h}_k)\right) =$

$$\mathsf{Pr}\left(\left|\sum_{n=1}^{N_t} [\mathbf{h}_k]_n e^{j\theta_n}\right|^2 \geq (1-\eta)N_t \|\mathbf{h}_k\|_2^2 \middle| \sigma(\mathbf{h}_k)\right). \tag{16}$$

For large N_t, by the Central Limit Theorem, $\sum_{n=1}^{N_t} [\mathbf{h}_k]_n e^{j\theta_n} \to X \sim \mathcal{CN}(\mu_X, \sigma_X^2)$, where for a fixed $\mathbf{h}_k$,

$$\mu_X = \sum_{n=1}^{N_t} [\mathbf{h}_k]_n \mathbb{E}\left[e^{j\theta_n}\right] = \sum_{n=1}^{N_t} \frac{[\mathbf{h}_k]_n}{2\pi}\int_0^{2\pi} e^{j\theta_n}\mathrm{d}\theta_n = 0, \text{ and}$$
$$\sigma_X^2 \overset{(a)}{=} \sum_{n=1}^{N_t} \mathsf{Var}([\mathbf{h}_k]_n e^{j\theta_n}) = \sum_{n=1}^{N_t} |[\mathbf{h}_k]_n|^2 \mathbb{E}\left[\left|e^{j\theta_n}\right|^2\right] = \|\mathbf{h}_k\|_2^2, \tag{17}$$

where in (a), we used the fact that conditioned on $\mathbf{h}_k$, $\left\{[\mathbf{h}_k]_n e^{j\theta_n}\right\}_{n=1}^{N_t}$ are independent. So, $|X|^2 \sim \exp(1/\|\mathbf{h}_k\|_2^2)$. Now, using (16), we evaluate the conditional probability as

$$\mathsf{Pr}\left(\mathcal{E}_k^{(\eta)} \middle| \sigma(\mathbf{h}_k)\right) = e^{-(1-\eta)N_t}. \tag{18}$$

Noticing that (18) is independent of $\mathbf{h}_k$ and using this value in (15) completes the proof of the theorem. ☐

Theorem 1 shows that $P_{\text{succ},k}^{(\eta)} \sim \mathcal{O}(e^{-N_t})$, indicating that the probability of achieving a near-beamforming gain under the full-RP scheme decays exponentially with the number of BS antennas. This scaling behavior can be understood as follows. Recall that in the full-RP scheme, the analog precoder is sampled in a i.i.d. fashion from a uniform distribution over $[0, 2\pi)$. Then, achieving a near-beamforming event requires the randomly chosen precoder to align with the UE channel across all antenna dimensions simultaneously. However, since this alignment must occur independently in each dimension, the overall probability decreases exponentially as the number of antennas N_t increases. In effect, the result highlights a fundamental mismatch between the search subspace spanned by random precoding vectors and the channel subspace at the UEs. We empirically validate the accuracy of the derived result in Theorem 1 in Fig. 2.

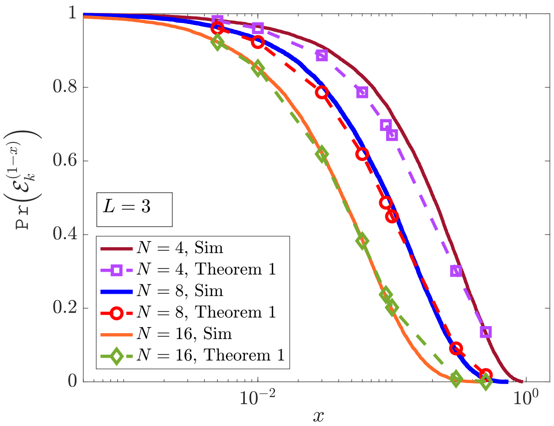


Fig. 2: Tightness in the CCDF of $|\mathbf{h}_k^H \mathbf{F}_{\mathrm{RF}}|^2 / \|\mathbf{h}_k\|_2^2$.

In hindsight, the full-RP OBF scheme did not exploit the subspace structure of UE channels, and this resulted in a lower success rate of the OBF scheme. To address this limitation, we next introduce the CSA-RP scheme, in which, we employ a channel-structure aware sampling distribution for selecting the RF precoder, which radiates its energy only within the subspace spanned by the UEs' channels. This, in turn, can enable at least one UE to experience a near-beamforming configuration with much higher probability.

C. CSA-RP Scheme

To address the dimensional mismatch in the distribution of the channel phase and the precoder in the full-RP scheme, we now propose the CSA-RP scheme, in which the structure of the random analog precoder is chosen as per the array response of the geometric channel model given in (3):

$$\mathbf{F}_{\mathrm{RF}}(t) = \mathbf{a}_{\mathrm{BS}}(\beta(t)) \tag{19}$$

where $\beta(t) \sim \mathcal{U}[-1, 1)$ is a random normalized angle associated with the CSA-RP scheme and is chosen so that the precoder matches the channel phase distribution. The rest of the scheme works as described in Sec. III-A.

Under the CSA-RP scheme, we say that a randomly chosen precoder as in (19) is η-*aligned* if

$$\min_{k\in[K],\ell\in[L]} |\beta(t) - \psi_{k,\ell}(t)| < \eta, \tag{20}$$

for some $\eta \in [0, 1)$, where $\beta(t)$, $\psi_{k,\ell}(t)$ denote the normalized angles of the precoder (see (19)) and AoD of the of the ℓth path of the channel to UE-k (see (2)), respectively. Specifically, (20) characterizes the event that the BS precoder is aligned to at least one of the paths for at least one UE. Dropping the index t for brevity, we have the following result.

Theorem 2. *In a system with K UEs each containing L channel paths from the BS, the probability that the CSA-RP scheme yields an η-aligned precoder for $\eta \in [0, 1)$ is given by*

$$P_{succ}^{(\eta)} = 1 - (1-\eta)^{KL}. \tag{21}$$

Proof. Using (20), the probability that a CSA-RP scheme yields an η-aligned precoder, denoted by $P_{\text{succ}}^{(\eta)}$, is given by

$$P_{\text{succ}}^{(\eta)} = \Pr\left(\min_{k\in[K],\ell\in[L]} |\beta - \psi_{k,\ell}| < \eta\right) \tag{22}$$

$$= 1 - \Pr\left(\min_{k\in[K],\ell\in[L]} |\beta - \psi_{k,\ell}| \geq \eta\right) \tag{23}$$

$$= 1 - \Pr\left(\bigcap_{k=1}^{K}\bigcap_{\ell=1}^{L} \{|\beta - \psi_{k,\ell}| \geq \eta\}\right) \tag{24}$$

$$\stackrel{(a)}{=} 1 - \prod_{k=1}^{K}\prod_{\ell=1}^{L} \Pr\left(\{|\beta - \psi_{k,\ell}| \geq \eta\}\right), \tag{25}$$

where in (a), we used the fact that $\psi_{k,\ell}$ are independent random variables across $k \in [K]$, and $\ell \in [L]$ which in turn implies that the events $\mathcal{G}_{k,\ell}^{(\eta)} \triangleq \{|\beta - \psi_{k,\ell}| < \eta\}$ are independent. Next, it is easy to verify that

$$\Pr\left(\mathcal{G}_{k,\ell}^{(\eta)}\right) = \Pr\left(\beta \in (\psi_{k,\ell} - \eta, \psi_{k,\ell} + \eta)\right) = \eta. \tag{26}$$

Using the above in (25), the proof of (21) is completed. □

From Theorem 2, for any $\eta > 0$, we have $\lim_{K\to\infty} P_{\text{succ}}^{(\eta)} = 1$, implying that, with a large number of UEs, the CSA-RP scheme indeed yields a beam that is aligned with at least one path of one of the UEs in the system, thereby procuring *multi-user diversity* benefits. Importantly, the success probability is independent of the number of antennas, unlike the full-RP scheme. We next evaluate the throughput of CSA-RP.

Let $\gamma_{\hat{k}} = \max_{k\in[K]}\gamma_k$ be the SNR at UE-$\hat{k}$ determined as per (10). When the UEs' SNRs, $\{\gamma_k\}_{k=1}^{K}$ are i.i.d., the probability density function (PDF) of $\gamma_{\hat{k}}$ can be obtained as

$$f_{\gamma_{\hat{k}}}(x) \triangleq K f_{\gamma}(x)[F_{\gamma}(x)]^{K-1},$$

where $f_{\gamma}(x)$ is the PDF of $\gamma \stackrel{d}{=} \gamma_k$ and $F_{\gamma}(x)$ is the cumulative distribution function (CDF) of γ. As a consequence, the achievable average throughput of the transmission protocol given in Sec. III-A under the CSA-RP scheme is given by

$$R^{(K)} = \int_0^{\infty} \log_2(1+x) f_{\gamma_{\hat{k}}}(x)dx. \tag{27}$$

However, evaluating the integral in (27) is analytically intractable. To address this, we exploit a key characteristic of next-generation wireless systems to simplify the analysis. Specifically, primary use cases of 5G and beyond include IoT and massive machine-type communications (mMTC), where the network is expected to support up to 10^6 devices per square kilometer. Even when a modest user activity rate of 10% is considered, it would require the BS to serve several hundred active UEs simultaneously. Leveraging the high-UE-density scenario of next-generation wireless systems and assuming that all UEs experience the same path loss[3] (i.e., $\rho_k = \rho,\ \forall k$), we now derive upper and lower bounds on the achievable rate, which are shown to be numerically tight in Sec. V.

Theorem 3. *Consider an mmWave system where an N_t-antenna BS equipped with a single RF chain serves K UEs each containing L channel paths from the BS using the CSA-RP OBF scheme. Then, as $K \to \infty$, the achievable average throughput, $R^{(K)}$ defined in (27) is lower bounded by*

$$R_{LB}^{(K)} \triangleq \max_{\epsilon_0\in[0,1]} \sum_{k_s=1}^{K} P_{k_s} \log_2\left(1 + \rho\left(\frac{N_t}{L} - \delta\right)\ln(k_s L)\right) \times \left(1 - \left(1 - \frac{1}{Lk_s}\right)^{LK}\right), \tag{28}$$

where $P_{k_s} \triangleq \binom{K}{k_s}(p_{\epsilon_0})^{k_s}(1-p_{\epsilon_0})^{K-k_s}$ with $p_{\epsilon_0} \triangleq \epsilon_0 - \frac{\epsilon_0^2}{4}$, and $\delta = \left(\pi^2 N_t(N_t-1)\epsilon_0^2\right)/(12L)$.

Proof. Using the expression for channel in (2), the SNR at UE-k via the ith path under the CSP-RP scheme is given by

$$\gamma_{k,i} = \rho|\alpha_{k,i}|^2 \left|\sum_{n=0}^{N_t-1} \frac{1}{\sqrt{N_t L}} e^{-j\pi n(\sin\phi_{k,i} - \sin\theta)}\right|^2, \tag{29}$$

where $\theta = \sin^{-1}(\beta)$, and β is the normalized angle of the precoder randomly selected by the CSA-RP. Let $\gamma_{\hat{k},\hat{i}} \triangleq \max_{k\in[K],\ i\in[L]} \gamma_{k,i}$ be the largest SNR among all the K UEs over L paths. To characterize $\gamma_{\hat{k},\hat{i}}$, we note that $\gamma_{k,i}$ is the product of two independent random variables. Then, a lower bound can be obtained by considering an appropriate lower bound for each of the terms given in (29).

The second term can take a maximum value of N_t/L when the terms in the summation add up to N_t. For this to happen, we need $\sin\phi_{k,i}$ to be close to $\sin\theta$ for some (k,i). Specifically, suppose that, for k_s of the UEs (with $0 \leq k_s \leq K$), $|\sin\phi_{k,i} - \sin\theta| < \epsilon_0$ holds for some $i \in [L]$ and a chosen ϵ_0. Then, due to the i.i.d. nature of $\phi_{k,i}$, k_s follows a Binomial distribution with parameters K and $p_{\epsilon_0} \triangleq \Pr\left(|\sin\phi_{k,i} - \sin\theta| < \epsilon_0\right)$. It is easy to see that

$$p_{\epsilon_0} = \Pr\left(|\psi_{k,i} - \beta| < \epsilon_0\right) \stackrel{(a)}{=} \epsilon_0 - \frac{\epsilon_0^2}{4}, \tag{30}$$

where $\psi_{k,i} = \sin\phi_{k,i}$, and in (a), we used the fact the random variable $\psi_{k,i} - \beta$ follows a triangle distribution obtained from the linear convolution of the PDFs of the uniformly distributed random variables $\psi_{k,i}$ and β. Now, for a chosen ϵ_0, note that whenever $|\sin\phi_{k,i} - \sin\theta| < \epsilon_0$, we have

$$\left|\sum_{n=0}^{N_t-1} \frac{1}{\sqrt{N_t L}} e^{-j\pi n(\sin\phi_{k,i} - \sin\theta)}\right|^2 > \frac{1}{N_t L}\left|\sum_{n=0}^{N_t-1} e^{-j\pi n\epsilon_0}\right|^2$$

$$\stackrel{(b)}{=} \frac{1}{N_t L}\left|\sum_{n=0}^{N_t-1}\left(1 - j\pi n\epsilon_0 - \frac{\pi^2 n^2\epsilon_0^2}{12}\right)\right|^2 + \mathcal{O}\left(\frac{\pi^4 n^4 \epsilon^4}{144 N_t}\right)$$

$$\stackrel{(c)}{=} N_t/L - \left(\pi^2 N_t(N_t-1)\epsilon_0^2\right)/(12L) = N_t/L - \delta, \tag{31}$$

where in (b), we used the Taylor's series expansion of e^x around $x = 0$, and in (c), we performed a few algebraic simplifications and then neglected the higher-order terms; this

[3] We adopt the equal path-loss assumption only for analytical tractability in characterizing the order statistics, similar to [35]. The scheme itself applies uniformly under unequal path losses also. In particular, this assumption is used to isolate the multi-user diversity gain arising from small-scale channel fluctuations, which in the unequal path-loss case, lead to intractable and less interpretable expressions. While path-loss-aware opportunistic scheduling policies can be employed to address this, we adopt the well-known max-rate scheduling policy for clarity and simplicity of exposition.

is justified for small ϵ_0, which is of interest in this analysis. Under the above condition, $\gamma_{k,i} \geq \rho |\alpha_{k,i}|^2 \left(\frac{N_t}{L} - \delta\right)$. Now, the term $z_{k,i} \triangleq |\alpha_{k,i}|^2$ follows an exponential distribution with mean 1, and are i.i.d. under the equal path loss assumption. Then, by using the results from extreme value theory [36, theorem 10.5.2], [29, lemma 2], we can show that $z_{\hat{k},\hat{i}} \triangleq \max_{k\in[K],\ i\in[L]} z_{k,i}$ has a CDF $F(\cdot)$ that belongs to the domain of attraction of the degenerate Gumbel distribution and that $z_{\hat{k},\hat{i}}$ grows as $\ln(Lk_s)$ as K becomes large.

Combining the order statistic from both the random variables, we see that the order statistic, $\gamma_{\hat{k},\hat{i}}$, is lower bounded by the maximum of the channel gains among the k_s UEs for whom the precoder is already in near beamforming configuration. Thus, the rate obtained by selecting the UE among the k_s UEs with the highest SNR grows at least as fast as

$$R_{k_s} \triangleq \log_2\left(1 + \rho\left(\frac{N_t}{L} - \delta\right)\ln(Lk_s)\right), \tag{32}$$

when k_s is large. But, for finite and small values of k_s, by using the law of total probability, a valid lower bound can be obtained by multiplying (32) by $\Pr(z_{\hat{k},\hat{i}} \geq \ln(Lk_s))$. Now, using the statistics of $z_{k,i}$, we can show that

$$\Pr(z_{\hat{k},\hat{i}} \geq \ln(Lk_s)) = 1 - (1 - e^{-\ln(Lk_s)})^{LK} \tag{33}$$

$$= 1 - \left(1 - \frac{1}{Lk_s}\right)^{LK} \overset{K\to\infty}{\longrightarrow} 1. \tag{34}$$

Subsequently, multiplying (33) with R_{k_s}, the overall rate can be lower bounded by

$$R_{k_s,\mathrm{LB}} \triangleq \log_2\left(1 + \rho\left(\frac{N_t}{L} - \delta\right)\ln(Lk_s)\right)\Pr(z_{\hat{k},\hat{i}} \geq \ln(Lk_s)). \tag{35}$$

Finally, let κ be a random variable that counts the number of UEs for whom the precoder offers near-BF gain. Using the law of iterated expectations, the achievable throughput under the CSA-RP scheme is lower bounded as

$$R_{\mathrm{LB}}^{(K)} \geq \sum_{k_s=1}^{K} \Pr(\kappa = k_s)\, R_{k_s,\mathrm{LB}} \tag{36}$$

$$= \sum_{k_s=1}^{K} \binom{K}{k_s} (p_{\epsilon_0})^{k_s} (1 - p_{\epsilon_0})^{K-k_s} R_{k_s,\mathrm{LB}}. \tag{37}$$

Substituting (30) and (35) in the above and taking the maximum of such lower-bounds over all $\epsilon_0 \in [0,1]$ yields (28), as desired. This completes the proof. $\square$

Next, we provide an asymptotic upper bound to $R^{(K)}$.

Theorem 4. *Consider an mmWave system where an N_t-antenna BS equipped with a single RF chain serves K UEs, with L channel paths from the BS to each UE, using the CSA-RP OBF scheme. Then, as $K \to \infty$, the achievable average throughput, $R^{(K)}$ in (27), satisfies an upper bound that behaves as*

$$\lim_{K\to\infty}\left(R^{(K)} - \mathcal{O}\left\{\log_2\left(1 + \rho\frac{N_t}{L}\ln(LK)\right)\right\}\right) = 0. \tag{38}$$

Proof. A valid upper bound can be obtained by deducing the individual upper bounds of the two terms in (29). For the first term, using extreme value theory [36], $z_{\hat{k},\hat{i}} \triangleq \max_{k\in[K],i\in[L]} z_{k,i}$ has the following property:

$$\max_{k\in[K],i\in[L]} z_{k,i} \overset{K\to\infty}{\longrightarrow} \ln(LK), \tag{39}$$

i.e., the order statistic asymptotically concentrates to $\ln(LK)$. The second term is at most N_t/L. Combining the two terms, $\gamma_{\hat{k},\hat{i}} \leq (\rho N_t/L)\ln(LK)$. Using the monotonicity of $\log(\cdot)$ function and substituting this in (11), the result follows. $\square$

Theorems 3 and 4 reveal that the achievable SNR under the CSA-RP scheme scales as $\mathcal{O}(N_t \ln(K))$ in each time slot. In contrast, a scheme that selects a UE in a round-robin fashion and optimizes the precoder via (6) achieves only an array gain of N_t. The CSA-RP scheme, thus, yields an additional multiplicative gain of $\ln(K)$, attributed to multi-user diversity arising from opportunistic scheduling, on top of the array gain. Therefore, the proposed OBF scheme not only eliminates the computational and time overheads associated with channel estimation and precoder optimization, but also achieves a higher throughput than conventional A-D architecture-based mmWave systems with round-robin scheduling.

Having established the advantages of the proposed OBF scheme, we now examine how its performance scales with the number of RF chains available at the BS.

IV. Opportunistic BF with Multiple RF Chains

In this section, we consider $N_{\mathrm{RF}} > 1$ RF chains adopt the CSA-RP OBF scheme to schedule a UE with the best SNR in every time resource. Then, by letting $N_s = 1$, the received signal in (5) becomes

$$y_k(t) = \sqrt{\vartheta_k}\mathbf{h}_k(t)\mathbf{F}_{\mathrm{RF}}(t)\mathbf{f}_{\mathrm{BB}}(t)s(t) + z_k(t), \tag{40}$$

$$= \sqrt{\vartheta_k}\mathbf{h}_k(t)\sum_{n=1}^{N_{\mathrm{RF}}}\mathbf{f}_{\mathrm{RF},n}(t)\nu_n(t)s(t) + z_k(t), \tag{41}$$

where the digital precoder is given by $\mathbf{f}_{\mathrm{BB}}(t) = [\nu_1(t), \ldots, \nu_{N_{\mathrm{RF}}}(t)]^T$ and the analog precoder is given by $\mathbf{F}_{\mathrm{RF}}(t) = [\mathbf{f}_{\mathrm{RF},1}(t), \ldots, \mathbf{f}_{\mathrm{RF},N_{\mathrm{RF}}}(t)]$, where $\mathbf{f}_{\mathrm{RF},n}$ denotes the beam formed by the BS on the nth RF chain. Substituting the expression for the channel at UE-k given in (2) into (41) and dropping the index t, the received signal at UE-k is

$$y_k = \sqrt{\vartheta_k}\sqrt{\frac{N_t}{L}}\sum_{\ell=1}^{L}\sum_{n=1}^{N_{\mathrm{RF}}}\alpha_{k,\ell}\nu_n\mathbf{a}_{\mathrm{BS}}^H(\psi_{k,\ell})\mathbf{f}_{\mathrm{RF},n}s + z_k. \tag{42}$$

We next describe the procedure for generating the N_{RF} beams and the transmission mechanism of the proposed OBF mechanism with multiple RF chains.

A. Generation of N_{RF} Beams

Since the UEs may be located arbitrarily, it is essential to form beams that collectively cover a wide area. At the same time, minimizing the number of beams is desirable to reduce system complexity. For this, we leverage the concept of resolvable beams. Specifically, two beams are said to be resolvable if the main lobe of one has negligible overlap with that of the other. In what follows, we randomly generate N_{RF} such resolvable beams (one corresponding to each RF chain) in each time slot. Recall that an array with N_t antennas can form at most N_t resolvable beams. The set of these beams, called the resolvable *beam book*, is defined in the normalized angular domain as

$$\mathcal{X} \triangleq \{\mathbf{a}(\psi), \psi \in \mathbf{\Psi}\}, \mathbf{\Psi} \triangleq \left\{\left(-1+\frac{2i}{N_t}\right)\middle| i=0,\ldots,N_t-1\right\}, \tag{43}$$

where $\mathbf{a}(\psi)$ is the array steering vector corresponding to angle ψ, and $\mathbf{\Psi}$ is referred to as the resolvable angle book. At time t, a set of N_{RF} random analog precoders is selected as

$$\mathcal{Q}_{\mathrm{RF}}(t) \triangleq \{\mathbf{f}_{\mathrm{RF},n}(t)\}_{n=1}^{N_{\mathrm{RF}}} = \mathsf{Samp}\left(\mathcal{X}, N_{\mathrm{RF}}\right), \tag{44}$$

where $\mathsf{Samp}(\mathcal{A}, N)$ denotes a operation that uniformly samples and returns N distinct elements at random from the set $\mathcal{A}$. If $|\mathcal{A}| < N$, the operation returns the full set $\mathcal{A}$.

B. Transmission Mechanism

Since the digital precoder must operate across multiple RF chains, we adopt a slightly modified protocol that requires only a few additional pilot transmissions compared to the one described in Section III-A, as detailed below.

1) Generation of N_{RF} random analog beams: The BS configures N_{RF} distinct analog beams $\{\mathbf{f}_{\mathrm{RF},n}\}_{n=1}^{N_{\mathrm{RF}}}$ as per (44).

2) Pilot-transmission for UE selection: The BS broadcasts P pilot symbols across time to all UEs. The received signal at UE-k during the pth pilot transmission is given by:

$$y_{k,p} = \sqrt{\vartheta_k}\sqrt{\frac{N_t}{L}}\sum_{\ell=1}^{L}\sum_{n=1}^{N_{\mathrm{RF}}}\alpha_{k,\ell}\mathbf{a}_{\mathrm{BS}}^H(\psi_{k,\ell})\mathbf{f}_{\mathrm{RF},n}s_{n,p} + z_{k,p}, \tag{45}$$

where $s_{n,p}$ is the pth pilot symbol transmitted through the nth RF chain. For simplicity, digital precoding weights ν_n are set to 1 during the pilot transmission. Stacking the received signals across P pilot transmissions at UE-k, we obtain:

$$\mathbf{y}_{k,\mathcal{P}} = \sqrt{\vartheta_k}\sqrt{\frac{N_t}{L}}\mathbf{S}_{\mathcal{P}}\boldsymbol{\zeta}_k + \mathbf{z}_k, \tag{46}$$

where $\mathbf{y}_{k,\mathcal{P}} \triangleq [y_{k,1},\ldots,y_{1,P}]^T$ is the received signal vector, $\mathbf{S}_{\mathcal{P}} \in \mathbb{C}^{P\times N_{\mathrm{RF}}}$ is the pilot matrix, $\mathbf{z}_k$ is the noise vector at UE-k, and $\boldsymbol{\zeta}_k$ is the effective channel vector defined as

$$[\boldsymbol{\zeta}_k]_n = \sum_{\ell=1}^{L}\alpha_{k,\ell}\mathbf{a}_{\mathrm{BS}}^H(\psi_{k,\ell})\mathbf{f}_{\mathrm{RF},n}. \tag{47}$$

The quality of $\boldsymbol{\zeta}_k$ captures how well the random beams align with the channel at UE-k. As a metric for the OBF scheme, we use the squared Euclidean norm $\|\boldsymbol{\zeta}_k\|_2^2$.

3) Effective channel response estimation: When $P \geq N_{\mathrm{RF}}$, the pilot matrix $\mathbf{S}_{\mathcal{P}}$ can be designed to have full column rank. In this case, UE-k computes the maximum-likelihood (ML) estimate of the effective low-dimensional channel vector as:

$$\hat{\boldsymbol{\zeta}}_k = \sqrt{\frac{L}{N_t}}\left(\mathbf{S}_{\mathcal{P}}^H\mathbf{S}_{\mathcal{P}}\right)^{-1}\mathbf{S}_{\mathcal{P}}^H\mathbf{y}_{k,\mathcal{P}} \in \mathbb{C}^{N_{\mathrm{RF}}}. \tag{48}$$

4) Best-UE identity feedback: Each UE computes $\xi_k \triangleq \left\|\hat{\boldsymbol{\zeta}}_k\right\|_2^2$ and starts a timer whose expiry time is inversely proportional to ξ_k. The UE k^* with the highest channel energy ξ_{k^*} has its timer expire first and initiates a feedback transmission to the BS indicating its identity k^* and $\hat{\boldsymbol{\zeta}}_{k^*}$.

5) UE scheduling and channel feedback: The BS identifies the UE whose feedback packet arrives first, denoted by

$$k^* \triangleq \arg\max_{k\in\{1,\ldots,K\}}\xi_k, \tag{49}$$

using the timer-based feedback scheme [32] described above. The selected UE-k^* also feeds back its estimated effective channel vector $\hat{\boldsymbol{\zeta}}_{k^*}$ to the BS, along with its identity. This vector is used by the BS to configure the digital precoder and enables MRT-based precoding in the current time slot.

Remark 1. *The scheme for multiple RF chains described above requires only low-dimensional CSI feedbck (that scales only with number of RF chains, which is much smaller than the number of BS antennas) estimation to enable MRT-based digital precoding at the opportunistically scheduled UE. This in turn achieves the full BF gain of fully digital BF without any per-antenna CSI estimation or hybrid precoder optimization.*

Based on the foregoing discussions, In Table I, we compare the implementation complexity of the associated signaling overheads with conventional CSI-based hybrid beamforming schemes. We see that OBF-schemes incur significantly lower complexity while still promising near-optimal performance.

C. Asymptotic Analysis of the Average Throughput

We begin by noting that in (42), all L paths are not necessarily resolvable. That is, for a given UE-k, the set of (normalized) AoDs $\{\psi_{k,\ell}\}_{\ell=1}^{L}$ may contain angles that are spaced closer than $2/N_t$, making them indistinguishable. To address this, we select a subset of resolvable paths, i.e., angles separated by at least $2/N_t$, and denote the number of such paths as $L_{\mathrm{res}} \leq L$. By grouping all paths within a beamwidth of $2/N_t$ into a single effective path, we simplify (42) as[4]

$$y_{k,p} = \sqrt{\vartheta_k}\sum_{n=1}^{N_{\mathrm{RF}}}\sqrt{\frac{N_t}{L_{\mathrm{res}}}}\sum_{\ell=1}^{L_{\mathrm{res}}}\alpha_{k,\ell}\mathbf{a}_{\mathrm{BS}}^H(\psi_{k,\ell})\mathbf{f}_{\mathrm{RF},n}s_{n,p} + z_{k,p}. \tag{50}$$

As a result, the norm of the effective channel at UE-k is

$$\left\|\hat{\boldsymbol{\zeta}}_k\right\|_2^2 = \sum_{n=1}^{N_{\mathrm{RF}}}\left|\sqrt{\frac{N_t}{L_{\mathrm{res}}}}\sum_{\ell=1}^{L_{\mathrm{res}}}\alpha_{k,\ell}\mathbf{a}_{\mathrm{BS}}^H(\psi_{k,\ell})\mathbf{f}_{\mathrm{RF},n}\right|^2. \tag{51}$$

Then, the average throughput of the CSA-RP OBF scheme described in Sec. IV-B can be obtained as

$$S^{(K)} \triangleq \mathbb{E}\left[\log_2\left(1+\max_{k\in\{1,\ldots,K\}}\rho\left\|\hat{\boldsymbol{\zeta}}_k\right\|_2^2\right)\right]. \tag{52}$$

As in Sec. III-C, an exact characterization of (52) is analytically intractable and offers limited insight. Therefore, similar to Sec. III-C, we again analyze the asymptotic regime as $K \to \infty$, which, as shown in Sec. V, yields a close

[4]We re-normalize the channel gain by replacing $\sqrt{L}$ with $\sqrt{L_{\mathrm{res}}}$ to preserve the energy of the overall channel.

TABLE I: Comparison of signaling overhead and complexity.

Signaling Type	Conventional Hybrid BF	OBF (Single RF)	OBF (Multiple RF)
CSI Estimation	High ($\sim \mathcal{O}(N_t)$)	Minimal ($\sim \mathcal{O}(1)$)	Low ($\sim \mathcal{O}(N_{\mathrm{RF}})$)
CSI Feedback	High-dimensional full CSI	UE identity only	UE identity + low-dim vector
Joint-Precoder Optimization	Required	Not required	Not required
Overall Complexity	$\sim \mathcal{O}(N_t K)$	$\sim \mathcal{O}(1)$	$\sim \mathcal{O}(N_{\mathrm{RF}} K) (\ll \mathcal{O}(N_t K))$

approximation even for moderate K. Before proceeding, we note from (44) that the BS-generated beams in $\mathcal{Q}_{\mathrm{RF}}(t)$ and the channel paths at UE-k in $\{\mathbf{a}_{\mathrm{BS}}(\psi_{k,\ell}(t))\}_{\ell=1}^{L_{\mathrm{res}}}$ are both directional, and their interaction results in either alignment or misalignment. Since the entries of $\boldsymbol{\zeta}_k$ inherently capture this behavior, we have the following result that characterizes the beam alignment event.

Lemma 1. *Consider a system with an N_t-antenna BS with N_{RF} chains and L_{res} resolvable paths at UE-k. Define the random variable $M_k(t)$ as the number of BS beams generated at time t that align with the L_{res} channel paths of UE-k, i.e.,*

$$M_k(t) \triangleq \sum_{\ell=1}^{\bar{l}_N} \mathbb{1}\left\{\mathbf{a}_{\mathrm{BS}}(\psi_{k,\ell}(t)) \in \mathcal{Q}_{\mathrm{RF}}(t)\right\}. \tag{53}$$

Then $M_k(t) \sim \mathrm{HypGeo}\big(N_t, \bar{l}_N, \bar{L}_N\big)$, where $\bar{l}_N \triangleq \min\{N_{\mathrm{RF}}, L_{res}\}$, $\bar{L}_N \triangleq \max\{N_{\mathrm{RF}}, L_{res}\}$, i.e., $M_k(t)$ follows Hypergeometric distribution with parameters N_t, $\bar{l}_N$, $\bar{L}_N$.

Proof. For an antenna array, the beam pattern can be modeled as *flat-top*[5], which is defined as [37, Eq. 5]

$$g(\phi) \triangleq \begin{cases} 1, & |\phi| \le \frac{\Delta\phi_{ft}}{2}, \\ 0, & \text{otherwise}, \end{cases} \tag{54}$$

where ϕ is the angle of interest and $\Delta\phi_{ft}$ is the beam width of the array's main lobe. That is, the response of an array in any direction has a constant gain over its main lobe, and the gain is 0 outside the main lobe. Thus, for any two $\psi_1, \psi_2 \in \boldsymbol{\Psi}$,

$$\mathbf{a}_{\mathrm{BS}}^H(\psi_1)\mathbf{a}_{\mathrm{BS}}(\psi_2) = \begin{cases} 1, & \text{if } \psi_1 = \psi_2, \\ 0, & \text{if } \psi_1 \neq \psi_2. \end{cases} \tag{55}$$

Now recall from (51) that

$$\begin{aligned} \left\|\hat{\boldsymbol{\zeta}}_k\right\|_2^2 &= \sum_{n=1}^{N_{\mathrm{RF}}} \left| \sqrt{\frac{N_t}{L_{\mathrm{res}}}} \sum_{\ell=1}^{L_{\mathrm{res}}} \alpha_{k,\ell} \mathbf{a}_{\mathrm{BS}}^H(\psi_{k,\ell}) \mathbf{f}_{\mathrm{RF},n} \right|^2 \\ &\overset{(a)}{=} \sum_{n=1}^{N_{\mathrm{RF}}} \left| \frac{N_t}{\sqrt{L_{\mathrm{res}}}} \sum_{\ell=1}^{L_{\mathrm{res}}} \alpha_{k,\ell} \mathbf{a}_{\mathrm{BS}}^H(\psi_{k,\ell}) \mathbf{a}_{\mathrm{BS}}(\phi_n) \right|^2, \end{aligned} \tag{56}$$

where (a) holds because $\mathcal{Q}_{\mathrm{RF}}$ consists of array steering vectors pointing in N_{RF} distinct angles, $\{\phi_n\}_{n=1}^{N_{\mathrm{RF}}}$. Now, in (56), we observe that N_{RF} beams generated by the BS and the angles of L_{res} paths in the channel at UE-k are independently sampled from $\boldsymbol{\Psi}$. In this view, we now characterize the distribution of $M_k(t)$, that counts the number of matching beams between the beams formed by the BS and angles of the channel at UE-k, as described in the theorem's statement.

[5]This characterization accurately models the interaction between any two resolvable beams and more generally captures the directional response of practical antenna arrays, particularly when the BS employs a large number of antennas, as is common in mmWave systems.

To derive the probability mass function (PMF) of $M_k(t)$, let $\bar{l}_N = L_{\mathrm{res}}$. Then $\{M_k(t) = m\}$ corresponds to the event that m (out of L_{res}) paths match with m of the N_{RF} beams, and the remaining $N_{\mathrm{RF}} - m$ beams match with other $N_t - L$ angles that do not appear in the channel response. Thus, the PMF is computed as

$$p_{M_k(t)}(m) \triangleq \mathsf{Pr}\left(M_k(t) = m\right) = \frac{\binom{L_{\mathrm{res}}}{m}\binom{N_t - L_{\mathrm{res}}}{N_{\mathrm{RF}} - m}}{\binom{N_t}{N_{\mathrm{RF}}}}, \tag{57}$$

where the support of the PMF is given by

$$m_{\min} \triangleq \max\{0, N_{\mathrm{RF}} + L_{\mathrm{res}} - N_t\} \le m \le m_{\max} \triangleq \bar{l}_N. \tag{58}$$

Here, the lower bound ($m_{\min}$) holds because the Binomial coefficient $\binom{x}{y}$ is defined only when $x \ge y \ge 0$, and the upper bound ($m_{\max}$) holds because at most $\bar{l}_N$ beams can match between BS and UE-k.

Similarly, if $\bar{l}_N = N_{\mathrm{RF}}$, we can obtain the PMF as

$$p_{M_k(t)}(m) = \frac{\binom{N_{\mathrm{RF}}}{m}\binom{N_t - N_{\mathrm{RF}}}{L_{\mathrm{res}} - m}}{\binom{N_t}{L_{\mathrm{res}}}}, \; m = m_{\min}, \ldots, m_{\max}. \tag{59}$$

However, using basic combinatorics, we can show that the probabilities in (57) and (59) are identical, and hence we can encapsulate both of them using a common expression:

$$p_{M_k(t)}(m) = \frac{\binom{\bar{l}_N}{m}\binom{N_t - \bar{l}_N}{\bar{L}_N - m}}{\binom{N_t}{\bar{L}_N}}, \; m = m_{\min}, \ldots, m_{\max}. \tag{60}$$

where $\bar{L}_N \triangleq \max\{N_{\mathrm{RF}}, L_{\mathrm{res}}\}$. Finally, recognizing that the above PMF corresponds to a Hypergeometric distribution with parameters $(N_t, \bar{l}_N, \bar{L}_N)$ completes the proof. □

From Lemma 1, we see that the *beam alignment probability* follows the PMF of a Hypergeometric random variable. This arises from a sequence of binary events indicating whether each BS-formed beam aligns with a channel path. Furthermore, the lemma shows how a tractable characterization of beam alignment can be obtained using discrete probability theory by leveraging the resolvable characteristics of mmWave channels. We have the following theorem.

Theorem 5. *Consider an mmWave system where an N_t-antenna BS equipped with N_{RF} RF chains serves K UEs using the CSA-RP OBF scheme described in Sec. IV-B. Then, as $K \to \infty$ the achievable average throughput, $S^{(K)}$, in (52) scales as in (61) on the top of a page, where the symbols are as defined earlier.*

$$S^{(K)} \approx \log_2\left(1+\rho\frac{N_t}{L_{\text{res}}}\left(\log K + \Big(\min\{N_{\text{RF}},L_{\text{res}}\}-1\Big)\log\log K - \log\Gamma\Big(\min\{N_{\text{RF}},L_{\text{res}}\}\Big)\right)\right). \tag{61}$$

Proof. From (52), the key step is to characterize $\max_{k\in[K]}\left\{\left\|\hat{\boldsymbol{\zeta}}_1\right\|_2^2, \left\|\hat{\boldsymbol{\zeta}}_2\right\|_2^2, \dots, \left\|\hat{\boldsymbol{\zeta}}_K\right\|_2^2\right\}$. To this end, we note that

$$\left\|\hat{\boldsymbol{\zeta}}_k\right\|_2^2 = \sum_{n=1}^{N_{\text{RF}}}\left|\sqrt{\frac{N_t}{L_{\text{res}}}}\sum_{\ell=1}^{L_{\text{res}}}\alpha_{k,\ell}\mathbf{a}_{\text{BS}}^H(\psi_{k,\ell})\mathbf{f}_{\text{RF},n}\right|^2 \tag{62}$$

$$\overset{(a)}{=} \frac{N_t}{L_{\text{res}}}\sum_{n=1}^{N_{\text{RF}}}\left|\sum_{\ell=1}^{L_{\text{res}}}\alpha_{k,\ell}\cdot\mathbb{1}\left\{\phi_n(t)=\psi_{k,\ell}\right\}\right|^2 \tag{63}$$

$$\overset{(b)}{=} \frac{N_t}{L_{\text{res}}}\sum_{n=1}^{N_{\text{RF}}}\left|\alpha_{k,\ell_n^*}\mathbb{1}\left\{\phi_n(t)=\psi_{k,\ell}\right\}\right|^2 \tag{64}$$

$$\overset{(c)}{=} \frac{N_t}{L_{\text{res}}}\sum_{n=1}^{M_k(t)}\left|\alpha_{k,\ell_n^*}\right|^2 \overset{(d)}{=} \frac{N_t}{L_{\text{res}}}\sum_{n=1}^{M_k(t)}\left|\alpha_{k,n}\right|^2, \tag{65}$$

where in (a), we used the resolvability condition of the channel beams and the randomly generated beams by the BS, in (b), we used the fact that at most one of the channel paths can be aligned with the nth beam generated by the BS; ℓ_n^* denotes the path index that matches with the nth beam formed at the BS, in (c), we used the definition of the number of matching beams as given in Lemma 1, and in (d) we used the fact that $\alpha_{k,n}$ are i.i.d. and ℓ_n^* are necessarily distinct indices, so the left- and right-hand-sides are equal in distribution.

It remains to characterize

$$E_K \triangleq \max_{k\in[K]}\left\{\sum_{n=1}^{M_1(t)}\left|\alpha_{1,n}\right|^2, \sum_{n=1}^{M_2(t)}\left|\alpha_{2,n}\right|^2, \dots, \sum_{n=1}^{M_K(t)}\left|\alpha_{K,n}\right|^2\right\}. \tag{66}$$

Towards that end, we first observe that $\tilde{G}_k \triangleq \sum_{n=1}^{M_k(t)}\left|\alpha_{k,n}\right|^2$ is the sum of a random (hypergeometric-distributed) number of exponential random variables. As a result, obtaining an exact characterization of its distribution is analytically intractable. However, as argued earlier, our objective is not to characterize the distribution of $\tilde{G}_k$ itself, but rather to understand the behavior of its maximum order statistic across users.

In particular, when $K\to\infty$, the maximum of K i.i.d. random variables is governed by the asymptotic behavior of their right tail. Consequently, it suffices to study the complementary cumulative distribution function (CCDF) of $\tilde{G}_k : \bar{F}_{\tilde{G}_k}(g) \triangleq \Pr\left(\tilde{G}_k > g\right)$ as $g\to\infty$ [38]. We have

$$\bar{F}_{\tilde{G}_k}(g) \overset{(e)}{=} \sum_{m=1}^{\bar{l}_N} p_{M_k(t)}(m)\bar{F}_{\tilde{G}_k}(g\big|M_k(t)=m) \tag{67}$$

$$\overset{(f)}{=} g^{\bar{l}_N-1}\sum_{m=1}^{\bar{l}_N}\sum_{i=0}^{m-1}p_{M_k(t)}(m)\left(\frac{g^i}{g^{\bar{l}_N-1}}\right)\frac{e^{-g}}{i!} \tag{68}$$

$$\xrightarrow{(g)} p_{M_k(t)}\left(\bar{l}_N\right)\frac{g^{\bar{l}_N-1}e^{-g}}{(\bar{l}_N-1)!} \overset{(h)}{\sim} p_{M_k(t)}(\bar{l}_N)\bar{F}_{\tilde{G}_k^{\text{tail}}}(g), \tag{69}$$

where in (e), we used the law of total probability, in (f), we recognized that conditioned on $M_k(t)=m$, $\tilde{G}_k$ is the sum of m exponential random variables and therefore follows a Gamma distribution with shape parameter m and rate 1, and then used the expression for its CCDF; in (g), we noted that as $g\to\infty$, $g^i/g^{\bar{l}_N-1}\to 0$ for all $i<\bar{l}_N-1$, and hence, only the dominant term corresponding to $m=\bar{l}_N$ survives, and finally in (h), we defined the random variable $\tilde{G}_k^{\text{tail}} \triangleq \tilde{G}_k\big|\{M_k(t)=\bar{l}_N\}$ for which the CCDF under $g\to\infty$ simplifies to

$$\bar{F}_{\tilde{G}_k^{\text{tail}}}(g) = g^{\bar{l}_N-1}\sum_{i=0}^{\bar{l}_N-1}\left(\frac{g^i}{g^{\bar{l}_N-1}}\right)\cdot\frac{e^{-g}}{i!} \sim \frac{g^{\bar{l}_N-1}e^{-g}}{(\bar{l}_N-1)!}, \tag{70}$$

where we again used $g^i/g^{\bar{l}_N-1}\to 0$ as $g\to\infty$ for $i<\bar{l}_N-1$. The key observation is that, in the limit $K\to\infty$, the maximum of $\{\tilde{G}_k\}_{k=1}^K$ is effectively achieved by the subset of users satisfying $M_k(t)=\bar{l}_N$. Let $K'\le K$ denote the number of such users, with $K'=\varphi K$ for some $\varphi\in[0,1]$. It suffices to analyze the maximum of K' i.i.d. Gamma random variables each with shape $\bar{l}_N$ and rate 1. From standard results in extreme value theory [36], we have $\lim_{K'\to\infty}(E_{K'}-d_{K'}) \xrightarrow{d} \Lambda$, where Λ follows the Gumbel distribution and $d_{K'}$ is a centering sequence. This implies that $E_{K'}$ asymptotically concentrates about $d_{K'}$. Using known results for Gamma distributions [39, Table 3.4.4], we can show

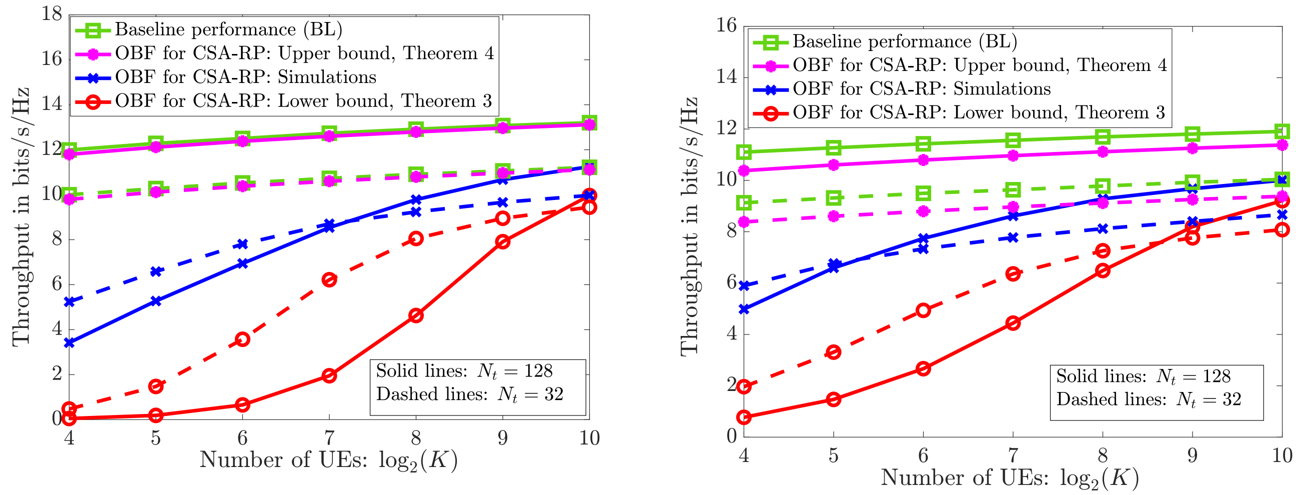
$$E_K \xrightarrow{d} \ln(K) + (\bar{l}_N-1)\ln(\ln(K)) - \ln(\bar{l}_N) + \mathcal{O}(1), \tag{71}$$

where the $\mathcal{O}(1)$ term accounts for the fraction φ. Substituting (71) into (65) and then into (52) completes the proof. $\square$

From Theorem 5, we observe that compared to the single-RF-chain case, employing multiple RF chains yields an additional SNR gain that scales linearly with $\bar{l}_N = \min\{N_{\text{RF}},L_{\text{res}}\}$, i.e., with the minimum of the number of RF chains and the number of resolvable paths. This improvement stems from the increased degrees of freedom due to multiple beams, which enable *multipath diversity* on top of the multi-user diversity gain. In nutshell, under the developed OBF, the achievable SNR scales a) *linearly* in the number of antennas, N_t, indicating the full BF gain, b) *logarithmically* in the number of UEs, K, indicating the *multi-user* diversity, and c) *linearly* in $\min\{N_{\text{RF}},L_{\text{res}}\}$, indicating the *multi-path* diversity gain. This demonstrates that OBF schemes readily scales with system parameters at low-complexity. Further, when $N_{\text{RF}}=1$, the scaling in Theorem 5 reduces to the single-RF-chain case, recovering the $\mathcal{O}(\log\ln K)$ behavior reported in Sec. III-C.

V. Simulation Results

In this section, we validate the theoretical analysis using Monte Carlo simulations and provide insights into the key trade-offs associated with OBF-based schemes in the mmWave communication bands. The mmWave channels are modeled according to (2), where the complex path gains are generated independently as $\mathcal{CN}(0,1)$ and the normalized AoDs are drawn uniformly at random from $[-1,1)$. We first investigate the performance under a single-RF-chain configuration and subsequently extend the study to multiple RF chains.



(a) $L = 1$.

(b) $L = 4$.

Fig. 3: Average throughput vs. K for $N_t = 32, 128$, $\rho = 10$ dB.

A. Single RF Chain

1) Experiment 1: Throughput vs. number of UEs, K: Figure 3a illustrates the average throughput as a function of the number of UEs K (in log-domain) for a single-RF-chain with $N_t = 32$ and 128 BS antennas, a SNR of 10 dB, and a single-path channel ($L = 1$). The performance of the proposed CSA-RP scheme is compared against the baseline (BL) scheme, which assumes perfect CSI at the BS and computes the optimal analog precoder for the UE having the best channel gain.

The BL scheme provides the maximum throughput since it performs optimal beamforming for every UE and subsequently schedules the best UE. Importantly, the CSA-RP scheme approaches this near-optimal performance as the number of UEs increases. This is in line with Theorem 2, which showed that the probability that the randomly chosen CSA-RP precoder aligns with at least one of the UEs' channel converges to 1 as K grows. This is because the CSA-RP precoder is structurally matched to the array response of the geometric mmWave channel model. Hence, the analytical upper bound $R^{(K)} \approx \log_2(1 + N_t \log K)$ is close to the BL throughput.

To further understand this, Fig. 3a also plots the analytical lower and upper bounds derived in Theorems 3 and 4, respectively. However, this upper bound does not explicitly account for two practical effects: (i) only a fraction ϵK of UEs are nearly aligned with the randomly chosen beam, and (ii) even these UEs experience a beamforming gain slightly below N_t. These effects are captured in the analysis through the parameters ϵ and δ. Specifically, the simulated CSA-RP throughput can be approximated as $\log_2(1 + (N_t - \delta)\log(\epsilon K))$, and hence the gap between the analytical upper bound and the simulated throughput can be expressed as $\log_2(1 + N_t \log K) - \log_2(1 + (N_t - \delta)\log(\epsilon K)) \approx \left(\frac{\delta}{N_t} - \frac{\log \epsilon}{\log K}\right)\log 2$. Here, $\delta \in [0, N_t]$ quantifies how far the effective BF gain deviates from the ideal value N_t, while $\epsilon \in (0, 1]$ represents the fraction of UEs whose channels are nearly aligned with the random precoder. As $\delta \to 0$, the beamforming gain approaches N_t, and as $\epsilon \to 1$, a larger fraction of UEs experience near-alignment. Consequently, as K grows large, this gap vanishes, explaining why the CSA-RP throughput converges to both the analytical upper bound and the BL performance.

The analytical lower bound in Theorem 3 explicitly relates δ to ϵ_0, where ϵ_0 quantifies the angular mismatch between the randomly chosen precoder direction and the physical AoDs. Smaller values of ϵ_0 imply tighter angular alignment, leading to smaller δ, but at the cost of reducing the fraction of UEs that satisfy this alignment condition. The derived lower bound lies below and exhibits the same scaling trend as the simulated throughput, validating the accuracy of the bound.

We also note yet another insight from Fig. 3a: for smaller values of K, the system with $N_t = 32$ achieves higher throughput than that with $N_t = 128$. This occurs because increasing N_t narrows the beamwidth, which in turn increases the number of UEs required to ensure that a randomly generated beam points toward at least one UE. In particular, for $N_t = 128$, roughly at least 128 users are required in the system before the performance surpasses that of the $N_t = 32$ case. Beyond this regime, the larger array gain offered by $N_t = 128$ dominates, resulting in superior throughput.

2) Experiment 2: Impact of multipath channels: Figure 3b examines the effect of multipath propagation on the performance of the CSA-RP scheme by considering $L = 4$ paths at the UEs. All other system parameters remain the same as in Fig. 3a. This experiment highlights how the presence of multiple channel paths influences both the achievable throughput and the tightness of the analytical bounds derived earlier. Most of the observations carry over from $L = 1$ case; some differences due to the multiple paths are explained below.

Similar to the single-path case, the CSA-RP scheme achieves near-optimal performance as the number of UEs increases. However, due to the multipath structure, this convergence occurs faster (with a smaller number of UEs), in line with Theorem 2. Further, Fig. 3b reveals that the gap between the analytical upper bound on the throughput of the CSA-RP scheme and the BL throughput is higher compared to Fig. 3a. Although each individual path adheres to a geometric

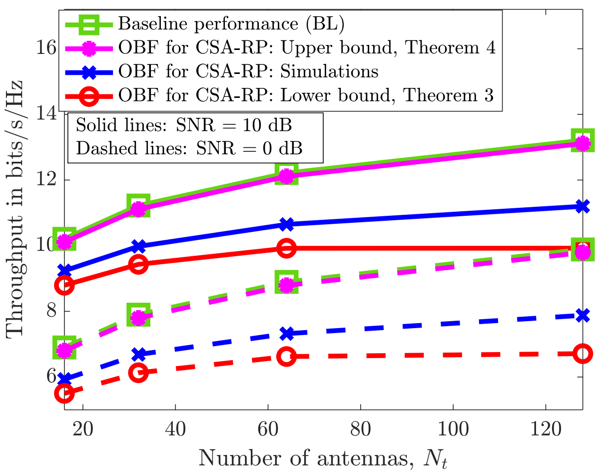


Fig. 4: Throughput vs. N_t for $\rho = 0, 10$ dB, $K = 1000$, $L = 1$.

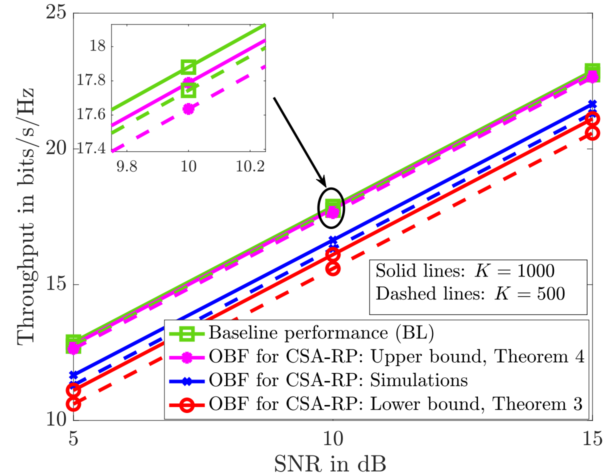


Fig. 5: Throughput vs. SNR for $K = 500, 1000$, $N_t = 32$.

channel model, the superposition of multiple paths results in an overall channel that does not exhibit the phase progression characteristic of a single array response vector. Consequently, the composite channel no longer has the same structure as a geometric channel. As a result, the randomly generated analog precoder at the BS does not match the structure of the full multipath channel. Instead, the random beam aligns with only one of the channel paths for a subset of UEs in most realizations. Since the path gains are normalized such that the total average channel energy is unity, the average SNR gain achievable through random analog BF with opportunistic user selection scales only as N_t/L. In contrast, in the BL system, the BS computes the optimal analog BF vector for each UE individually. In this case, the phase angles of the BF vector are not constrained to follow an arithmetic progression, as in the single-path scenario, but are instead optimized to precisely match the multipath channel. This enables the BL scheme to achieve the full array gain of N_t, which is higher than that of the CSA-RP scheme.

Yet another insight can be obtained by examining the crossover behavior between different number of BS antennas. In Fig. 3b, for $L = 4$, the simulated throughput for $N_t = 128$ begins to outperform that of $N_t = 32$ after approximately 32 users. However, for the analytical lower bound of the CSA-RP scheme, this crossover occurs only after around 256 UEs. The underlying reason is the same as in Experiment 1: increasing N_t results in narrower beams, which require a larger number of UEs to ensure beam alignment. Nevertheless, compared to the $L = 1$ case, this crossover is observed with a significantly smaller number of UEs. This is because the presence of $L = 4$ paths effectively provides approximately 4 times the number of beam-alignment opportunities in the system. In other words, from the perspective of opportunistic BF, the system behaves as if the number of UEs has increased by a factor of 4. This observation is consistent with the expression in (21), which increases with the number of paths. So, even with a small number of UEs, a system with $N_t = 128$ can achieve better performance than one with $N_t = 32$, in the $L = 4$ case.

3) Experiment 3: Throughput vs. number of BS antennas: Figure 4 plots the average throughput as a function of the number of BS antennas N_t for a single-path channel ($L = 1$) with a fixed number of users $K = 1000$. Results are shown for two SNR values, 0 dB and 10 dB. For every doubling of N_t, the throughput of the BL scheme and the analytical upper bound of the CSA-RP scheme increase by approximately 1 bit/s/Hz. This behavior is expected, since both the BL scheme and the analytical upper bound incur no loss in array gain and hence scale logarithmically with N_t.

In contrast, the analytical lower bound on the throughput of the CSA-RP scheme increases slowly with each doubling of N_t when the number of UEs is finite. This behavior follows from the expression in (28), which depends on the effective array gain term $(N_t - \delta)$. From (31), the parameter δ scales as $\mathcal{O}(N_t^2)$, causing the effective array gain in the lower bound to increase slowly for large antenna dimensions. Further, the simulated throughput of the CSA-RP scheme also increases at a rate of slightly smaller than 1 bit/s/Hz for each doubling of N_t. This loss relative to the BL scheme arises because the analog precoder is configured using randomly chosen phase angles, and for a perfect BF condition, we need a large number of UEs. Nevertheless, the CSA-RP scheme still achieves substantial throughput while completely avoiding excessive training overhead, large feedback overhead, and computationally expensive optimization of the analog precoding phase angles.

4) Experiment 4: Throughput versus SNR: In Experiment 4, Fig. 5 plots the average throughput as a function of the SNR for a single-RF-chain system with $N_t = 32$ BS antennas and a single-path channel ($L = 1$). Results are shown for two values of total UEs: $K = 500$ and $K = 1000$. Clearly, the simulated throughput of the BL scheme, the analytical throughput of the CSA-RP scheme (including both the upper and lower bounds), and the simulated throughput of the CSA-RP scheme increase by approximately 1 bit/s/Hz for every 3 dB increase in SNR, as expected. Furthermore, increasing the number of users from $K = 500$ to $K = 1000$ results in a marginal improvement in throughput because, in the single-RF-chain setting, the multi-user diversity gain scales as $\log_2(\log K)$, leading to diminishing returns as K becomes large.

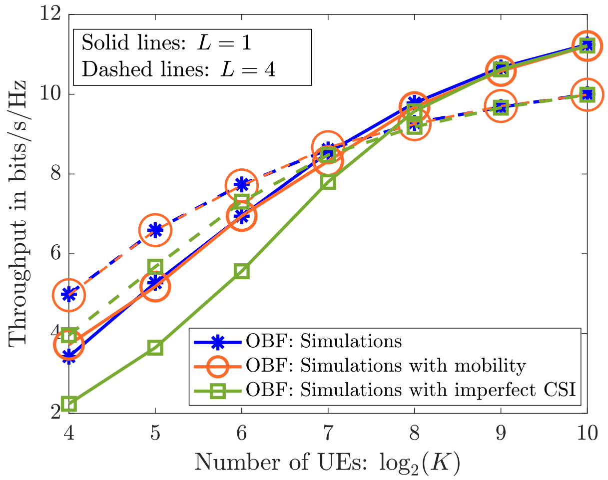


Fig. 6: Performance under realistic system scenarios.

5) Experiment 5: Performance with Realistic scenarios: We now evaluate the robustness of the OBF scheme under practical system impairments and dynamic scenarios. In particular, we consider (i) imperfect CSI-based scheduling and (ii) the impact of UE mobility. For imperfect CSI, recall from Sec. III-A that each UE estimates its effective received SNR, $\{\gamma_k\}_{k=1}^K$, which may be imperfect in practice. To model this, we first define $g_k(t) = \frac{1}{\sqrt{N_t}} h_k(t)\mathbf{F}_{\mathrm{RF}}(t)$ as the normalized true effective channel at UE-k. Then, let an estimator-agnostic expression for the estimated received SNR at UE-k be given by $\hat{\gamma}_k(t) = \rho_k N_t |\hat{g}_k(t)|^2$, where $\hat{g}_k(t)$ is an estimate of the normalized channel and satisfies

$$g_k(t) = \varsigma_0 n_k(t) + \sqrt{1-\varsigma_0^2}\,\hat{g}_k(t). \tag{72}$$

Here, $n_k(t) \sim \mathcal{CN}(0,1)$ represents the estimation noise, and $\varsigma_0 \in [0,1]$ controls the estimation accuracy. In particular, $\varsigma_0 = 0$ corresponds to perfect CSI, while $\varsigma_0 = 1$ represents a fully noisy estimate. Based on these imperfect estimates, the BS schedules the UE given by

$$\hat{k}(t) = \arg\max_{k\in 1,\ldots,K} \hat{\gamma}_k(t). \tag{73}$$

Next, to model mobility, we consider time-varying channel angles at the UEs. Let the angle vector for UE-k across L paths at time t be $\boldsymbol{\theta}(t) = [\phi_{k,1}(t), \ldots, \phi_{k,L}(t)]^T$. Following [40], its evolution is modeled as

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(t-1) + \mathbf{u}(t), \tag{74}$$

where $\mathbf{u}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_u)$ with covariance $\mathbf{Q}_u = \left(\frac{0.5\pi}{180}\right)^2 \mathbf{I}_L$. Accordingly, the channels at UEs evolve over time as per (2).

In Fig. 6, we evaluate the performance of OBF under imperfect CSI (with $\varsigma_0 = 0.1$) and UE mobility for $N_t = 128$ with $L = 1$ and $L = 4$. The key observations are as follows:

1) The performance degradation due to imperfect CSI is minimal compared to the perfect CSI case, and the gap diminishes as the number of UEs increases. Importantly, the result demonstrates that the OBF scheme remains robust and continues to effectively exploit multi-user diversity even under realistic CSI imperfections.
2) Under UE mobility, the performance of OBF remains comparable to the static case. This is because the proposed scheme operates on a slot-by-slot basis and does not rely on temporal channel correlation. Since the analytical results hold for arbitrary channel realizations, time variations do not affect the fundamental conclusions.

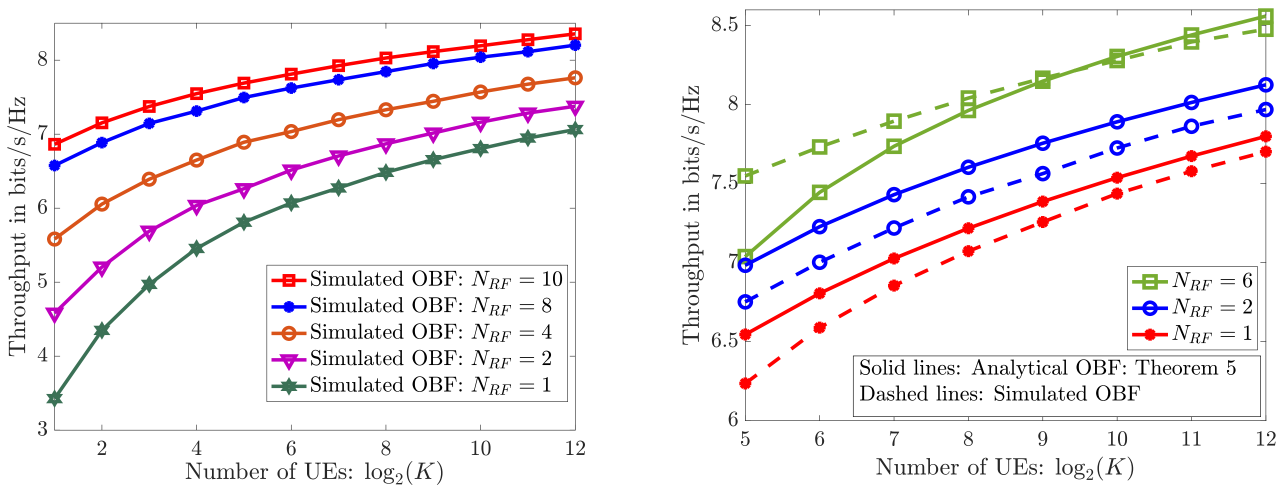
These results confirm that the developed OBF framework remains robust under practical non-idealities. Finally, the performance trends across different values of L can be interpreted as follows. When the number of UEs is small, the presence of multiple paths increases the likelihood that a random beam aligns with at least one path of at least one UE, thereby improving the probability of achieving array gain and resulting in higher throughput compared to the single-path case. However, when the number of UEs is large, it becomes highly likely that at least one UE achieves near-perfect alignment even in the single-path scenario. In this regime, under the CSA-RP design, with multiple-paths, the effective received SNR is primarily contributed to by a single path, which constitutes only a fraction of the total channel energy. Consequently, the single-path case outperforms the multi-path scenario. This behavior is consistent with the observations in Fig. 3.

B. Multiple RF Chains

1) Experiment 6: Throughput scaling with the number of RF chains: Figure 7a illustrates how the throughput scales with the number of RF chains, N_{RF}, for different total number of UEs. The throughput increases monotonically with N_{RF} across all UE regimes; however, the actual gain in the throughput itself depends on the number of UEs in the system.

In sparse-UE regimes ($K = 2, 4, 8, 16$), increasing N_{RF} yields substantial throughput gains. For instance, when $K = 2$, increasing N_{RF} from 1 to 10 improves the throughput from 3.5 to 6.9 bits/s/Hz. Similarly, for $K = 4$ and $K = 8$, the throughput increases from 4.4 to 7.2 bits/s/Hz and from 5.0 to 7.4 bits/s/Hz, respectively. These pronounced improvements can be attributed to the fact that, in sparse-user regimes, multiuser diversity is limited and each additional RF chain enables the BS to form additional beams and provide *multi-path diversity* in the system. As the number of UEs increases, the throughput continues to improve with N_{RF}, but the incremental gains gradually diminish. For example, at $K = 32$, increasing N_{RF} from 1 to 10 improves the throughput from 5.8 to 7.7 bits/s/Hz, which represents a lower relative gain than the above. When the number of UEs is even higher, say at $K = 256$, the relative gain further reduces, with the throughput increasing from 6.5 to only 8.03 bits/s/Hz. This diminishing return occurs because, as the number of UEs grows, opportunistic user selection becomes increasingly effective, allowing the system to exploit multiuser diversity even with fewer RF chains, indicating a reduced dependence on hardware resources.

Thus, deploying multiple RF chains is most beneficial in sparse-user regimes, where large relative throughput gains can be achieved by exploiting multi-path diversity. In contrast, in high-user regimes, the system can achieve near-optimal performance through multiuser diversity, making single-RF-



(a) Performance with $N_{\rm RF} = 1, 2, 4, 8, 10$ and $L = 10$.

(b) Performance with $N_{\rm RF} = 1, 2, 6$ and $L = 6$.

Fig. 7: Average throughput versus number of UEs, K with multiple RF chains and $N_t = 16$, $\rho = 10$ dB.

chain architectures a more hardware-efficient choice with reduced implementation complexity.

2) Experiment 7: Validation of the analytical scaling behavior: Figure 7b compares the simulated throughput with the analytical results for $N_{\rm RF} = 1, 2$, and 6, with $L = 6$ resolvable paths and $N_t = 16$ BS antennas. This plot validates the accuracy of the analytical rate expression given in Theorem 5 and illustrates how its constituent terms govern the throughput behavior across different user regimes with multiple RF chains.

As shown in Fig. 7b, both the simulated and analytical throughput increase with the number of RF chains, confirming that the analytical expression correctly captures the scaling behavior. For larger values of $N_{\rm RF}$, such as $N_{\rm RF} = 6$, the constant term $\log \Gamma(N_{\rm RF})$ in (61) becomes more pronounced, while the $\log K$ and $(N_{\rm RF} - 1) \log \log K$ terms grow slowly for small and moderate values of K. This causes visible mismatches between the analytical and simulated throughput when the number of UEs is not sufficiently large. However, as K increases, the growth of the $\log K$ and $(N_{\rm RF} - 1) \log \log K$ terms eventually dominates the constant offset $\log \Gamma(N_{\rm RF})$. Consequently, in the large-user regime, the analytical and simulated results align closely, demonstrating that the asymptotic scaling laws accurately characterize the throughput behavior for large K. Next, for $N_{\rm RF} = 2$, the additional $(N_{\rm RF} - 1) \log \log K$ term rate expression starts to dominate over the magnitude of $\log \Gamma(N_{\rm RF})$ term. This balance results in a close agreement between the analytical and simulated throughput even in the small-user regime. Finally, for $N_{\rm RF} = 1$, the rate expression (61) reduces to a pure $\log K$ scaling without any additional terms, and the analytical and simulated results match closely as K increases. This recovers the throughput scaling of the CSA-RP opportunistic BF with single RF chain. These results confirm that the analytical characterization accurately captures throughput scaling laws across different values of $N_{\rm RF}$, and validate the theory presented for the multiple-RF-chain case.

VI. Conclusion

In this paper, we investigated OBF based on randomized analog precoding for hybrid A-D multi-user mmWave systems, with the goal of mitigating extensive training overhead, large feedback requirements, and the need for computationally intensive analog precoder optimization. By analyzing two OBF schemes, namely, the Full-RP and the CSA-RP, we showed that the CSA-RP scheme achieves throughput close to that of an ideal BL system with coherent BF and best-UE selection using a significantly smaller number of UEs compared to the Full-RP scheme. Motivated by these, we focused on the CSA-RP scheme and derived the asymptotic average throughput for both single- and multiple-RF-chain system configurations using tools from extreme value theory. Using this, we demonstrated how OBF can effectively harness full array gain along with multi-user and multi-path diversity gains in the system. Simulation results validated the analytical findings and revealed that multiple-RF-chain architectures provide superior performance in sparse-user regimes with substantially lower time and computational complexity. Overall, the results establish OBF as an attractive, low-complexity, and low-feedback solution for the practical implementation of mmWave IoT networks.

References

[1] K. Verma and C. R. Murthy, "Opportunistic beamforming based hybrid analog-digital multi-user mmwave communications: Design and analysis," in *Proc. IEEE Nat. Conf. Commun. (NCC)*, 2023, pp. 1–6.

[2] R. W. Heath *et al.*, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[3] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[4] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2017.

[5] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[6] 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.

[7] 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.

[8] 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.

[9] J. Song, J. Choi, and D. J. Love, "Codebook design for hybrid beamforming in millimeter wave systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 1298–1303.

[10] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, "Codebook-based hybrid precoding for millimeter wave multiuser systems," *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5289–5304, Oct. 2017.

[11] J. Wang *et al.*, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, 2009.

[12] V. Venkateswaran *et al.*, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4131–4143, 2010.

[13] L. Zhao, K. Ying, K. Kang, and H. Qian, "On the digital predistortion of wideband mmwave communication systems with beam squint," *IEEE Internet Things J.*, vol. 12, no. 13, pp. 24 083–24 093, Jul. 2025.

[14] L. Ling, Z. Lian, Z. Ma, L. Zhang, and Y. Su, "Channel modeling and performance analysis for RIS-assisted mmwave communications," *IEEE Internet Things J.*, vol. 12, no. 3, pp. 3188–3201, Feb. 2025.

[15] T. Du, J. Yang, C.-K. Wen, S. Xia, and S. Jin, "General simultaneous localization and mapping scheme for mmwave communication systems," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 22 521–22 536, Jun. 2024.

[16] Q. Xiao, Y. He, G. Yu, J. Yuan, and R. Yin, "Localization-assisted fast and robust beam optimization for mmwave communications," *IEEE Internet Things J.*, vol. 12, no. 9, pp. 12 383–12 396, May 2025.

[17] J. Zhao, L. Yang, M. Xia, and M. Motani, "Unified analysis of coordinated multipoint transmissions in mmwave cellular networks," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12 166–12 180, Jul. 2022.

[18] X. Ye, L. Fu, and J. M. Cioffi, "Joint codebook selection and MCS adaptation for mmwave eMBB services based on deep reinforcement learning," *IEEE Internet Things J.*, vol. 11, no. 19, pp. 31 545–31 560, Oct. 2024.

[19] R. Song, J. Shen, H. Yuan, X. Gao, X. Ding, Y. Liu, and D. B. da Costa, "Line-of-sight MIMO systems: DoF analysis and hybrid beamforming design," *IEEE Internet Things J.*, vol. 11, no. 19, pp. 31 791–31 804, Oct. 2024.

[20] Y. Liu, M. Li, W. Deng, and M.-J. Zhao, "Adaptive position-aware near-field beam training for millimeter-wave XL-MIMO," *IEEE Internet Things J.*, vol. 12, no. 18, pp. 38 185–38 197, Sep. 2025.

[21] 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, 2019.

[22] P. Wang, J. Fang, X. Yuan, Z. Chen, and H. Li, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14 960–14 973, 2020.

[23] X. Li and A. Alkhateeb, "Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, Nov. 2019, pp. 800–805.

[24] H. Ju, S. Jeong, S. Kim, and B. Shim, "Transformer-aided parametric CSI feedback for mmwave massive MIMO systems," in *IEEE Int. Conf. Commun (ICC)*. IEEE, 2023, pp. 3596–3601.

[25] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019.

[26] A. Morsali, A. Haghighat, and B. Champagne, "Deep learning-based hybrid analog-digital signal processing in mmwave massive-MIMO systems," *IEEE Access*, vol. 10, pp. 72 348–72 362, 2022.

[27] H. Hojatian, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, "Unsupervised deep learning for massive MIMO hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7086–7099, Nov. 2021.

[28] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE*, vol. 1, June 1995, pp. 331–335.

[29] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[30] G. Lee, Y. Sung, and J. Seo, "Randomly-directional beamforming in millimeter-wave multiuser MISO downlink," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1086–1100, Feb. 2016.

[31] G. Lee, Y. Sung, and M. Kountouris, "On the performance of random beamforming in sparse millimeter wave channels," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 560–575, Apr. 2016.

[32] V. Shah, N. B. Mehta, and R. Yim, "Optimal timer based selection schemes," *IEEE Trans. Commun.*, vol. 58, no. 6, pp. 1814–1823, 2010.

[33] R. Talak and N. B. Mehta, "Optimal timer-based best node selection for wireless systems with unknown number of nodes," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4475–4485, 2013.

[34] C. Wang, Z. Li, T.-X. Zheng, H. Chen, and X.-G. Xia, "Robust hybrid precoding design for securing millimeter-wave IoT networks under secrecy outage constraint," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 13 024–13 038, Aug. 2021.

[35] L. Bai, T. Li, Z. Xiao, and J. Choi, "Performance analysis for SDMA mmWave systems: Using an approximate closed-form solution of downlink sum-rate," *IEEE Access*, vol. 5, pp. 15 641–15 649, 2017.

[36] H. A. David and H. N. Nagaraja, *Order statistics*. John Wiley & Sons, 2004.

[37] S. Singh *et al.*, "Interference analysis for highly directional 60-GHz mesh networks: The case for rethinking medium access control," *IEEE/ACM Trans. Netw.*, vol. 19, no. 5, pp. 1513–1527, Oct. 2011.

[38] N. Mehrnia and S. Coleri, "Wireless channel modeling based on extreme value theory for ultra-reliable communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1064–1076, Feb. 2022.

[39] P. Embrechts *et al.*, *Modelling extremal events*, ser. Stochastic Modelling and Applied Probability. Springer Berlin, Heidelberg, 1997.

[40] C. Zhang, D. Guo, and P. Fan, "Tracking angles of departure and arrival in a mobile millimeter wave channel," in *2016 IEEE Int. Conf. Commun (ICC)*, May 2016, pp. 1–6.