Pareto-Optimal Hybrid Beamforming for Finite-Blocklength Millimeter Wave Systems

Jitendra Singh, *Graduate Student Member, IEEE,* Banda Naveen, Suraj Srivastava, *Member, IEEE* Aditya K. Jagannatham, *Senior Member, IEEE* and Lajos Hanzo, *Life Fellow, IEEE*

***Abstract*—Short-packet communication (SPC) is essentially synonymous with ultra-reliable low-latency communication (uRLLC), which must meet stringent latency and reliability requirements. However, achieving efficient hybrid beamforming (HBF) in SPC-based millimeter wave (mmWave) systems is challenging due to the constraints of finite block lengths, limited number of radio frequency chains (RFCs), and owing to the complex optimization of transmit precoders (TPCs). In this work, we investigate the achievable rate region of an SPC-based mmWave downlink system. We harness the HBF for finite block lengths low-latency communication, relying on a low number of RFCs. We formulate a Pareto optimization problem for characterizing the achievable rate region, while considering the transmit power, mmWave hardware, and block length constraints. To solve this highly non-convex problem, we propose a bisection search-based block coordinate descent (Bi-BCD) algorithm, in which we optimize the RF TPC, the baseband (BB) TPC, and the block length. Specifically, we jointly optimize the RF and BB TPCs for a fixed block length, which involves both Remanian conjugate gradient (RCG) and second-order cone programming (SOCP) techniques, and then we optimize the block length by the mixed integer programming method. Subsequently, we update the achievable rate via the bisection search method. Finally, we present simulation results and quantify the efficiency of the proposed scheme.**

***Index Terms*—Short packet communication, millimeter wave, hybrid beamforming, *Pareto boundary*.**

I. INTRODUCTION

THE Ultra reliable and low latency communication (uRLLC) is a key component of next-generation wireless (NGW) networks [1], [2]. As defined by 3GPP [3], the uRLLC mode ensures 99.999% reliability for a 32-byte packet together with a latency under 1 ms, which is crucial for applications like smart grids, industrial automation, autonomous vehicles, and mission-critical communication. In conventional wireless communication, data transmission often involves high block length (HBL) packets, which typically denotes a packet size of 1500 bytes at an infinitesimally low codeword decoding error probability. As a result, HBL transmission leads to increased latency, which is not suitable for uRLLC systems. In order to support uRLLC, short packet communication (SPC), which involves transmissions of finite block length (FBL) packets. However, these aspects of SPC transmission render the classical Shannon's capacity analysis inapplicable. To this end, Polyanskiy *et al.* [4] derived an approximate rate expression for SPC, which is a function of the signal-to-noise power ratio (SNR), block length, and decoding error probability. However, due to the complicated rate expression, optimizing the beamforming is highly intractable for SPC systems. To handle this issue, the authors of [5]–[7] optimized beamforming for MU-MISO systems considering SPC. Specifically, the authors of these treatises explored the four most common formulations for beamformer design: transmit power minimization, max-min SINR, sum-rate maximization, and energy efficiency maximization, while considering the uRLLC requirements as constraints for both perfect and imperfect channel state information (CSI). To investigate the reliability degradation caused by FBL transmission, the authors of [8] derived a closed-form expression for the average decoding error probability in the SPC-aided system. While the error probability in SPC-aided systems decreases with an increase in block length, this improvement may lead to higher end-to-end latency. To address this trade-off, Yu *et al.* [9] examined the average age of information (AoI) in SPC-based massive machine-type communication (mMTC) systems. Specifically, the average AoI serves as a time-averaged measure of information freshness, reflecting how outdated the received information is.

Pareto boundary plays a crucial role in characterizing achievable rate regions of the MU systems, as it encompasses all rate tuples where increasing one user's rate necessitates a corresponding reduction in another's rate. A common method for characterizing the *Pareto boundary* involves solving a series of weighted sum-rate maximization (WSR-Max) problems [10], which are typically non-convex. Alternatively, Zhang *et al.* [11] proposed the concept of the rate profile vector for obtaining a point on the Pareto boundary. Specifically, the authors of [11] characterized the rate region of the MU-MISO system with the aid of the rate profile vector, which transforms the equivalent problem into a weighted minimum-rate maximization formulation. This approach is generally convex and offers improved computational efficiency compared to WSR-Max problems.

Due to the congestion of the frequency bands in the sub-6 GHz regime, NGW networks might consider exploiting the currently underused millimeter wave (mmWave) frequency

L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council (EPSRC) projects under grant EP/Y037243/1, EP/W016605/1, EP/X01228X/1, EP/Y026721/1, EP/W032635/1, EP/Y037243/1 and EP/X04047X/1 as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028). The work of Aditya K. Jagannatham was supported in part by the Qualcomm Innovation Fellowship; in part by the Qualcomm 6G UR Gift; and in part by the Arun Kumar Chair Professorship. The work of S. Srivastava was supported in part by IIT Jodhpur's Research Grant No. I/RIG/SUS/20240043; and in part by Telecom Technology Development Fund (TTDF) under Grant TTDF/6G/368.

J. Singh B. Naveen, and A. K. Jagannatham are with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, UP 208016, India (e-mail: jitend@iitk.ac.in; naveenb22@iitk.ac.in; adityaj@iitk.ac.in).

S. Srivastava is with the Department of Electrical Engineering, Indian Institute of Technology Jodhpur, Jodhpur, Rajasthan 342030, India (email: surajsri@iitj.ac.in).

L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

band [12]. However, a large number of antennas are required for mitigating the excessive path loss of mmWave-aided communications. In this scenario, the conventional fully-digital beamformer (FDB) of sub-6 GHz MIMO systems is highly inefficient due to the large number of RF chains (RFCs), which is equal to that of the number of antennas. This results in increased power consumption and hardware costs. To circumvent this issue, the authors of [13], [14] proposed a hybrid beamforming (HBF) architecture that requires remarkably fewer RFCs. More precisely, the HBF architecture divides the transmit precoder (TPC) into a baseband (BB) TPC and an analog/RF TPC, where the RF TPC is implemented using digitally controllable phase shifters. However, these phase shifters constrain each element of the RF TPC to be of unit modulus, which poses a stringent constraint in terms of obtaining the HBF. To handle this issue, the Riemann conjugate gradient (RCG) based HBF was employed in [13], [14], wherein the authors optimized the HBF under the unit modulus, quality of service (QoS), and transmit power constraints. Lin *et al.* [15] proposed secrecy and energy-efficient HBF schemes for satellite-terrestrial integrated networks, focusing on multibeam satellites sharing the mmWave spectrum with cellular systems. As a further advance, Lin *et al.* [16] investigated HBF designs in RIS-aided satellite-terrestrial relay networks, aiming for minimizing the total transmit power, while satisfying the user rate constraints.

The literature of mmWave systems discussed in [12]–[16] relies on transmission models associated with the HBLs and the conventional Shannon capacity formula. While these models are useful in generic contexts, they do not address the strict reliability and latency requirements of uRLLC applications. The Shannon capacity formula is inapplicable in the SPC regime due to the FBLs and non-zero error probabilities. Therefore, it is essential to investigate SPC in the context of mmWave systems. However, owing to the complex rate expression of SPC transmission [4], coupled with the hybrid design of TPC [12]–[16], the associated beamforming optimization is challenging in the SPC-aided mmWave systems. Motivated by this, we explore the achievable rate regions and propose a Pareto optimal HBF design for SPC-based mmWave MIMO systems. To achieve this objective, we formulate a sum rate maximization problem considering the minimum rate requirements of the user equipment (UEs) based on the rate profile [11] and block lengths, as well as the unit modulus and transmit power constraints. To solve this highly non-convex problem, we propose a bisection-block coordinate descent (Bi-BCD) method for optimizing the BB TPC, the RF TPC, the block lengths, and the achievable rate. The numerical results demonstrate that with the proposed algorithm, the achievable rate region of the SPC-based mmWave system approaches the performance of an ideal HBL-aided mmWave system, despite using substantially fewer RFCs and the FBL.

II. System Model

As seen in Fig.1, we consider the SPC-aided mmWave downlink, in which a BS having $N_{\rm t}$ transmit antennas

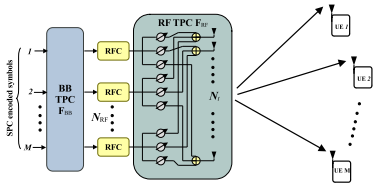


Figure 1: Illustration of an SPC-based mmWave system.

(TAs), provides uRLLC services for M single-antenna UEs. In the hybrid MIMO architecture of [12]–[14], the signal processing blocks at the BS include a BB TPC $\mathbf{F}_{\rm BB} = [\mathbf{f}_{{\rm BB},1}, \ldots, \mathbf{f}_{{\rm BB},M}] \in \mathbb{C}^{N_{\rm RF} \times M}$ followed by the RF TPC $\mathbf{F}_{\rm RF} \in \mathbb{C}^{N_{\rm t} \times N_{\rm RF}}$. Without loss of generality, the necessary condition at the BS to admit M active UEs is : $M \leq N_{\rm RF} << N_{\rm t}$. Furthermore, the BS transmits the information symbols $\{s_1, \ldots, s_M\}$ to the UEs in the SPC regime, where s_m represents the symbol of the mth uRLLC UE. Upon assuming that the transmit symbols $s_m, \forall m$, are independent and identically distributed (i.i.d.) random variables that follow the distribution $s_m \sim \mathcal{CN}(0,1)$, the received signal y_m at the mth UE is expressed as

$$y_m = \mathbf{h}_m^{\rm H} \mathbf{F}_{\rm RF} \mathbf{f}_{{\rm BB},m} s_m + \sum_{n=1, n\neq m}^{M} \mathbf{h}_m^H \mathbf{F}_{\rm RF} \mathbf{f}_{{\rm BB},n} s_n + n_m, \quad (1)$$

where n_m is the i.i.d. complex additive white Gaussian noise obeying the distribution $n_m \sim \mathcal{CN}(0, N_o)$. Moreover, $\mathbf{h}_m^H \in \mathbb{C}^{1\times N_{\rm t}}$ is the mmWave channel between the BS and the mth UE, which is modeled as [12]

$$\mathbf{h}_m = \sqrt{\frac{N_t}{N_{\rm clu} N_{\rm ray}}} \sum_{i=1}^{N_{\rm clu}} \sum_{j=1}^{N_{\rm ray}} \Omega_{ij} \mathbf{a}_{\rm t}(\phi_{ij}), \quad (2)$$

where $N_{\rm clu}$ and $N_{\rm ray}$ denote the number of scattering clusters and the number of scattering rays per cluster. The quantity Ω_{ij} represents the complex path gain of the jth ray in the ith cluster. Moreover, $\mathbf{a}_{\rm t}(\phi_{ij}) \in \mathbb{C}^{N_{\rm t}\times 1}$ is the array response vector, which is given by

$$\mathbf{a}_{\rm t}(\phi_{ij}) = \frac{1}{\sqrt{N_t}} \left[1, e^{j\frac{2\pi \bar{d}}{\lambda}\sin(\phi_{ij})}, \ldots, e^{j(N_t-1)\frac{2\pi \bar{d}}{\lambda}\sin(\phi_{ij})}\right]^T, \quad (3)$$

where ϕ_{ij} is the angle of departure (AoD), and $\bar{d}$ is the antenna spacing, which is assumed to be half of the wavelength λ. Assuming that the CSI is available at both the BS and UEs, the signal-to-interference-plus-noise ratio (SINR) of the mth UE is expressed as

$$\gamma_m = \frac{{\rm tr}\left(\mathbf{F}_{\rm RF}^H \mathbf{H}_m \mathbf{F}_{\rm RF} \mathbf{B}_m\right)}{\sum\limits_{\substack{n=1\\ n\neq m}}^{M} {\rm tr}\left(\mathbf{F}_{\rm RF}^H \mathbf{H}_m \mathbf{F}_{\rm RF} \mathbf{B}_n\right) + N_o}, \quad (4a)$$

where ${\rm tr}(\cdot)$ denotes the trace operator. The quantities $\mathbf{H}_m \in \mathbb{C}^{N_{\rm t}\times N_{\rm t}}$ and $\mathbf{B}_m \in \mathbb{C}^{N_{\rm RF}\times N_{\rm RF}}$ are given by $\mathbf{H}_m = \mathbf{h}_m \mathbf{h}_m^H$ and $\mathbf{B}_m = \mathbf{f}_{{\rm BB},m} \mathbf{f}_{{\rm BB},m}^H$, respectively. Therefore, the achievable rate R_m of the mth UE having a block length of β_m and

a non-zero decoding error probability ϵ_m, is given by [4],

$$R_m = \ln\left(1+\gamma_m\right) - \sqrt{\frac{V_m}{\beta_m}} Q^{-1}\left(\epsilon_m\right), \forall m, \tag{5}$$

where $Q(\cdot)$ is the Gaussian Q-function. The quantity V_m is the channel dispersion for the mth UE, which is given as $V_m = 1 - \frac{1}{(1+\gamma_m)^2}$. Therefore, the sum rate, $R_{\rm sum}$ of the system considered is given by

$$R_{\rm sum} = \sum_{m=1}^{M} R_m. \tag{6}$$

III. Problem formulation

This paper seeks to characterize the achievable rate region for the SPC-enabled mmWave system by optimizing the RF and BB TPCs. The achievable rate region is defined as the collection of all the rate tuples $(R_1, \ldots, R_M)$, that can be simultaneously achieved, and thus it is expressed as

$$\mathcal{R} = \bigcup_{\substack{\|\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2 \le P_{\rm t}, |\mathbf{F}_{\rm RF}(i,j)|=1, \\ \forall i,j, \sum_m \beta_m \le N, \forall m,}} \left\{(R_1, R_2, \ldots, R_M)\right\}, \tag{7}$$

where $\|\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2 \le P_{\rm t}$ is the transmit power constraint, $|\mathbf{F}_{\rm RF}(i,j)| = 1, \forall i,j$ is the unit-modulus constraint on each element of the RF TPC. The constraint $\sum_m \beta_m \le N, \forall m$, ensures that the symbols of all UEs are transmitted within a maximum block length of N symbols. Moreover, the upper right boundary of this region $\mathcal{R}$ is the *Pareto boundary*, since it consists of rate-tuples at which it is impossible to improve a particular UE's rate, without simultaneously decreasing the rate of at least one of the other UEs. Following [11], any rate-tuple located on the *Pareto boundary* of the rate region can be obtained by solving the following optimization problem along with a particular rate profile vector $\boldsymbol{\eta} = [\eta_1, \eta_2, \ldots, \eta_M]$

$$\max_{\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB}, \{\beta_m\}_{m=1}^M} \quad R_{\rm sum}, \tag{8a}$$

$$\text{s. t.} \quad R_m \ge \eta_m R_{\rm sum}, \forall m, \tag{8b}$$

$$|\mathbf{F}_{\rm RF}(i,j)| = 1, \forall i,j \tag{8c}$$

$$\|\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2 \le P_{\rm t}, \tag{8d}$$

$$\sum\nolimits_{m=1}^{M} \beta_m \le N, \tag{8e}$$

$$\beta_m \in \mathbb{Z}^+, \forall m, \tag{8f}$$

where $\eta_m \in [0,1]$ in (8b) denotes the target rate ratio between the achievable rate of the mth UE and the sum rate $R_{\rm sum}$, with $\sum_{m=1}^M \eta_m = 1$. The constraint (8f) guarantees that the block length of each UE must be a non-negative integer. The complete *Pareto boundary* of the achievable rate region $\mathcal{R}$ can be characterized by solving the optimization problem (8) in conjunction with various rate-profile vectors $\boldsymbol{\eta}$. However, the optimization problem (8) is highly non-convex due to the non-convex constraints (8b) and (8c), and owing to the tightly coupled TPC variables $\mathbf{F}_{\rm RF}$ and $\mathbf{F}_{\rm BB}$ in both the objective function and constraints. Moreover, the embedding of the finite block length β_m and the transmission error probability ϵ_m in the rate expression as defined in (5) makes the optimization problem (8) even more sophisticated and challenging. To solve the optimization problem (8), we propose an iterative optimization strategy based on the bisection search and block coordinate descent (Bi-BCD) technique. In this method, we optimize the RF TPC $\mathbf{F}_{\rm RF}$, BB TPC $\mathbf{F}_{\rm BB}$ and block length β_m for the fixed $R_{\rm sum}$ by employing the BCD method, and then subsequently updating $R_{\rm sum}$ by employing the bisection search. The complete procedure of the proposed Bi-BCD method is described next.

IV. Proposed approach

A. Joint optimization of $\mathbf{F}_{\rm RF}$ and $\mathbf{F}_{\rm BB}$ for fixed β_m

For a given $\beta_m, \forall m$, the resultant optimization problem is expressed as

$$\max_{\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB}} R_{\rm sum} \quad \text{s. t.} \quad \text{(8b), (8c), and (8d).} \tag{9}$$

To solve the above problem, we obtain the target rate $R_{\rm sum}$ by employing the well-known bisection search method. Specifically, for any given target rate $R_{\rm sum} \ge 0$, the equivalent feasible problem for (9) can be stated as

$$\min_{\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB}} \mathcal{P}\left(\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB}\right) = \|\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2 \quad \text{s. t. (8b) and (8c).} \tag{10}$$

Let us denote the optimal solution of (9) as $\{\mathbf{F}_{\rm RF}^*, \mathbf{F}_{\rm BB}^*\}$. Then it becomes evident that the problem (9) is feasible for a given sum rate $R_{\rm sum}$, when the transmit power constraint satisfies $||\mathbf{F}_{\rm RF}^*\mathbf{F}_{\rm BB}^*\|_F^2 \le P_t$. Furthermore, to handle the non-convex rate constraint (8b) in the above problem (10), we first transform it into a tractable form by using the Proposition presented below.

Proposition. *Given $\epsilon_m \ge 0$ and $\beta_m \ge 0, \forall m$, the constraint $R_m \ge \eta_m R_{\rm sum}$ is equivalent to $\gamma_m \ge \Gamma_m, \forall m$, where Γ_m is given by*

$$\Gamma_m = e^{\eta_m R_{\rm sum} + \frac{\kappa_m}{2}} - 1, \tag{11}$$

and κ_m is the generalized Lambert $\mathcal{W}$ function, which is given by

$$\kappa_m = \mathcal{W}\left(\tfrac{2Q^{-1}(\epsilon_m)}{\sqrt{\beta_m}}, \tfrac{-2Q^{-1}(\epsilon_m)}{\sqrt{\beta_m}}; -4\delta_m^2\left(\frac{Q^{-1}\left(\epsilon_m\right)}{\sqrt{\beta_m}}\right)^2\right), \tag{12}$$

where $\delta_m = e^{-\eta_m R_{\rm sum}}$.

Proof. Please refer to Appendix (A) for the detailed proof. □

By employing the above proposition, (10) can be recast as

$$\min_{\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB}} \mathcal{P}\left(\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB}\right) \quad \text{s. t.} \quad \gamma_m \ge \Gamma_m, \forall m, \text{and (8c).} \tag{13}$$

Since $\mathbf{F}_{\rm RF}$ and $\mathbf{F}_{\rm BB}$ are coupled in both the objective function and the constraint in the above optimization problem (13), we adopt the BCD approach to design them alternatively.

1) Optimize $\mathbf{F}_{\mathrm{RF}}$*:* Let $\alpha_m \triangleq \mathrm{tr}\left(\mathbf{F}_{\mathrm{RF}}^H \mathbf{H}_m \mathbf{F}_{\mathrm{RF}} \mathbf{B}_m\right) - \Gamma_m \sum_{n\neq m}^{M} \mathrm{tr}(\mathbf{F}_{\mathrm{RF}}^H \mathbf{H}_m \mathbf{F}_{\mathrm{RF}} \mathbf{B}_n)$, $\mathbf{g} = \mathrm{vec}(\mathbf{F}_{RF}) \in \mathbb{C}^{N_{\mathrm{t}} N_{\mathrm{RF}} \times 1}$, $\mathbf{\Pi} = \mathbf{F}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_{\mathrm{t}}}$, and $\mathbf{\Upsilon}_{n,m} = \mathbf{B}_n^T \otimes \mathbf{H}_m \in \mathbb{C}^{N_{\mathrm{t}} N_{\mathrm{RF}} \times N_{\mathrm{t}} N_{\mathrm{RF}}}$, where $|\mathbf{g}(l)| = 1, \forall l = \{1, \ldots, N_{\mathrm{t}} N_{\mathrm{RF}}\}$. Thus, we can rewrite $\mathcal{P}\left(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}\right) = \|\mathbf{\Pi}\mathbf{g}\|_2^2$ and $\alpha_m = \mathbf{g}^H \mathbf{\Upsilon}_{m,m} \mathbf{g} - \Gamma_m \sum_{n\neq m}^{M} \mathbf{g}^H \mathbf{\Upsilon}_{n,m} \mathbf{g}$. Therefore, the equivalent optimization problem constructed for the RF TPC is given by

$$\min_{\mathbf{g}} \ ||\mathbf{\Pi}\mathbf{g}||_2^2 \quad \text{s. t.} \quad \phi_m(\mathbf{g}) \leq 0, \forall m, \text{and} \quad |\mathbf{g}(l)| = 1, \forall l, \tag{14}$$

where $\phi_m(\mathbf{g}) = N_o \Gamma_m - \alpha_m$. The above problem (14) is a quadratically constrained quadratic program (QCQP) with an extra unit constraint. To solve this problem, we adopt the penalty-based method, where the SINR constraint $\phi_m(\mathbf{g}) \leq 0$ is added to the objective function as a penalty term. Thus, the equivalent penalized optimization problem is expressed as

$$\begin{aligned} \min_{\mathbf{g}} \ & f(\mathbf{g}) = ||\mathbf{\Pi}\mathbf{g}||_2^2 + \mu \sum_{m=1}^{M} \left(\max\left(0, \phi_m\left(\mathbf{g}\right)\right) \right)^2 \\ \text{s. t.} \quad & |\mathbf{g}(l)| = 1, \forall l, \end{aligned} \tag{15}$$

where μ is a penalty factor, which studies a tradeoff between the objective function and penalty constraint. Since the elements of $\mathbf{g}$ form a complex circle manifold as $\mathcal{G} = \{\mathbf{g} \in \mathbb{C}^{N_{\mathrm{t}} N_{\mathrm{RF}} \times 1} : |\mathbf{g}(l)| = 1, \forall 1 \leq l \leq N_{\mathrm{t}} N_{\mathrm{RF}}\}$, the above problem (15) can be efficiently solved by employing the RCG algorithm. To this end, the Euclidean gradient of the function $f(\mathbf{g})$ is formulated as

$$\nabla f(\mathbf{g}) = 2\mathbf{\Pi}^H \mathbf{\Pi}\mathbf{g} + \mu \sum_{m=1}^{K} G_m, \tag{16}$$

where the quantity G_m is defined as $G_m = 4\phi_m\left(\mathbf{g}\right) \left(-\mathbf{\Upsilon}_{m,m}\mathbf{g} + \Gamma_m \sum_{\substack{n=1 \\ n\neq m}}^{M} \mathbf{\Upsilon}_{n,m}\mathbf{g} \right)$. Consequently, one can obtain the Riemannian gradient from the corresponding Euclidean gradient $\nabla f(\mathbf{g})$. Then (15) can be solved iteratively on the Riemannian space utilizing the conjugate gradient algorithm. For further details on the RCG algorithm, motivated readers might like to consult [13], [14].

2) Optimize $\mathbf{F}_{\mathrm{BB}}$*:* When $\mathbf{F}_{\mathrm{RF}}$ is fixed, the resultant optimization problem of the BB TPC $\mathbf{F}_{\mathrm{BB}}$, is given by

$$\min_{\mathbf{F}_{\mathrm{BB}}} \ \mathcal{P}\left(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}\right) \quad \text{s. t.} \quad \gamma_m \geq \Gamma_m, \forall m. \tag{17}$$

To solve the above problem (17), we reformulate the non-convex SINR constraint as a second-order cone constraint by introducing a common phase shift to $\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},m}$ Thus, the equivalent SOCP problem constructed for (17) is given by

$$\min_{\mathbf{F}_{\mathrm{BB}}} \ \mathcal{P}\left(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}\right) \ \text{s. t.} \ \left\| \begin{array}{c} \mathbf{A}^H \mathbf{e} \\ \sqrt{N_o} \end{array} \right\|_2 \leq \sqrt{1 + \frac{1}{\Gamma_m}} t_{m,n}, \tag{18}$$

where $t_{m,n} = \mathbf{h}_m^H \mathbf{F}_{\mathrm{RF}} \mathbf{f}_{\mathrm{BB},n}$, $\mathbf{A}(m,n) = t_{m,n}$, and $\mathbf{e} \in \mathbb{C}^{M\times 1}$ is the elementary vector having a one on its mth position and zero otherwise. The above SOCP problem (18) is a convex optimization problem and can be efficiently solved using a standard convex optimization tool.

B. Optimization of $\{\beta_m\}_{m=1}^M$ *for given* $\mathbf{F}_{\mathrm{RF}}$ *and* $\mathbf{F}_{\mathrm{BB}}$

For the given $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$, the sub-problem of addressing the block length β_m is cast as

$$\max_{\{\beta_m\}_{m=1}^M} R_{\mathrm{sum}} \quad \text{s. t.} \quad \text{(8b), (8e), and (8f).} \tag{19}$$

To achieve a point on the *Pareto boundary* for the rate profile provided, we set the block length constraint to be met with equality, i.e., $\sum_{m=1}^{M} \beta_m = N$. Therefore, for a fixed target rate $R_{\mathrm{sum}} \geq 0$, the constraint (8b) can be modified as follows

$$\beta_m \geq \left(\frac{\sqrt{V_m} Q^{-1}\left(\varepsilon_m\right)}{\ln\left(1+\gamma_m\right) - \eta_m R_{\mathrm{sum}}} \right)^2. \tag{20}$$

Consequently, the optimization problem considering the modified block length is given by

$$\text{Find: } [\beta_1, \ldots, \beta_M] \quad \text{s. t.} \quad \text{(8f), (20), and } \sum_{m=1}^{M} \beta_m = N. \tag{21}$$

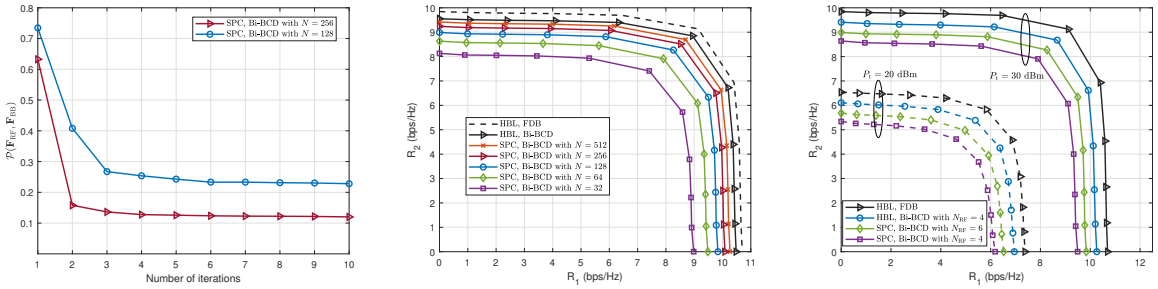
Note that for fixed values of $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$, the above problem (21) is a non-convex mixed integer programming problem, which can be solved efficiently via the approach discussed in [2]. Finally, we update the achievable sum rate R_{sum} for the fixed $\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}$ and $\{\beta_m\}_{m=1}^M$ via the bisection search method.

We summarize the complete procedure of the proposed Bi-BCD procedure solving (8) in Algorithm 1. To conceive the *Pareto boundary* of the achievable rate region for the SPC-based mmWave system, Algorithm 1 optimizes the RF TPC and BB TPC, as well as the block lengths alternately for the different rate profile vectors $\boldsymbol{\eta}$. Observe that the complexity of Algorithm 1 depends on the inner layer of the BCD method and on the outer layer of the bisection search. Moreover, the main computational complexity of the BCD method in the inner layer arises due to (16), which is approximately $\mathcal{I}_P \left[\mathcal{O}\left(\left(N_{\mathrm{t}} N_{\mathrm{RF}}\right)^2 \right) \right]$, where $\mathcal{I}_P$ represents the number of iterations. Therefore, the overall complexity of Algorithm 1 is given by $\mathcal{I}_a \left[\mathcal{I}_P \left(\mathcal{O}\left(\left(N_{\mathrm{t}} N_{\mathrm{RF}}\right)^2 \right) \right) \right]$, where $\mathcal{I}_a$ is the number of iterations required for convergence in the outer layer.

V. SIMULATION RESULTS

In this section, we present the simulation results of our proposed Bi-BCD algorithm for characterizing the rate region of the SPC-based mmWave system. We set the key parameters used in the simulation as listed in Table I. In addition, the stopping parameters are set as $\tau_1 = 1 \times 10^{-8}$, $\tau_2 = 1 \times 10^{-2}$ and the bisection tolerence as $\tau_3 = 1 \times 10^{-6}$. We compare our proposed scheme (SPC, Bi-BCD) to the following ideal benchmarks:

- *Benchmark 1 (HBL, FDB):* This corresponds to the HBL transmission via the FDB, which serves as an ideal Pareto-optimal boundary.
- *Benchmark 2 (HBL, Bi-BCD):* In this scheme, we adopt the HBL transmission, followed by the HBF scheme by employing the Bi-BCD method.



(a) (b) (c)

Figure 2: (a) Convergence behavior with $R_{\mathrm{sum}} = 10$ bps/Hz; (b) Achievable rate region of two UEs with varying N; (c) Achievable rate region of two UEs with varying P_t and N_{RF}.

Algorithm 1 Bi-BCD algorithm for solving (8)

Input: $R_{\mathrm{sum,L}} = 0$, $R_{\mathrm{sum,U}}$, $\boldsymbol{\eta}$, P_t, N, thresholds τ_1,τ_2,τ_3, $\mu > 1$, $\psi_\mu > 1$

initialize: $\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}$, $\{\beta_m\}_{m=1}^{M}$
repeat
  $R_{\mathrm{sum}} = (R_{\mathrm{sum,L}} + R_{\mathrm{sum,U}})/2$
  evaluate $\Gamma_m, \forall m$ using (11)
  repeat
    set $\kappa = 0$, $\mathcal{P}^{(\kappa)}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) = \infty$
    initialize $\mathbf{g}^i = \mathrm{vec}(\mathbf{F}_{\mathrm{RF}}^{(\kappa)})$ and set i = 0.
    while $\left(\|\nabla_{\mathcal{G}} f(\mathbf{g}^{\mathbf{i}})\|^2 \geq \tau_1\right)$ **do**
      set $\phi_m \mathbf{g}^{(\mathbf{i})} = N_o \Gamma_m - \alpha_m^{(i)}$
      update $\mathbf{g}^{(i)}$ by solving (14)
    end while
    if $\max(0, \phi_m(\mathbf{g}^{\mathbf{i}})) \leq \tau_1, \forall m$ then $\mathbf{g} = \mathbf{g}^{\mathbf{i}}$
    else update $\mu = \psi_\mu \mu$, and go to step 7.
    update $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$ as $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)} = \mathrm{reshape}(\mathbf{g})$ to $N_{\mathrm{t}} \times N_{\mathrm{RF}}$ matrix
    given $\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)}$ and $\{\beta_m^{\kappa}\}_{m=1}^{M}$, obtain $\mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}$ by solving (18)
    update $\mathcal{P}^{(\kappa+1)}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) = \|\mathbf{F}_{\mathrm{RF}}^{(\kappa+1)} \mathbf{F}_{\mathrm{BB}}^{(\kappa+1)}\|_F^2$
    set $\kappa \leftarrow \kappa + 1$
  until $\left|(\mathcal{P}^{(\kappa)}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) - \mathcal{P}^{(\kappa-1)}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})\right| \leq \tau_2$
  evaluate $\{\beta_m\}_{m=1}^{M}$ by solving (21)
  if obtained set $\{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}, \{\beta_m\}_{m=1}^{M}\}$ is feasible, set $R_{\mathrm{sum,L}} = R_{\mathrm{sum}}$.
  else set $R_{\mathrm{sum,U}} = R_{\mathrm{sum}}$.
until $R_{\mathrm{sum,U}} - R_{\mathrm{sum,L}} \leq \tau_3$
output: $\mathbf{F}_{\mathrm{RF}}$, $\mathbf{F}_{\mathrm{BB}}$, $\{\beta_m\}_{m=1}^{M}$, rate tuple $\mathcal{R} = (R_1, R_2, \ldots, R_M)$

Fig. 2a shows the convergence behavior of the function $\mathcal{P}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ in the Bi-BCD algorithm for the fixed rate of $R_{\mathrm{sum}} = 10$ bps/Hz. It can be observed that $\mathcal{P}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ exhibits monotonically decreasing behavior and reaches a local optimum within a few iterations, which verifies the convergence of the Bi-BCD algorithm. Furthermore, the function $\mathcal{P}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ monotonically decreases upon increasing block length N. This shows that more power is required for a high block length N to meet the higher QoS demands.

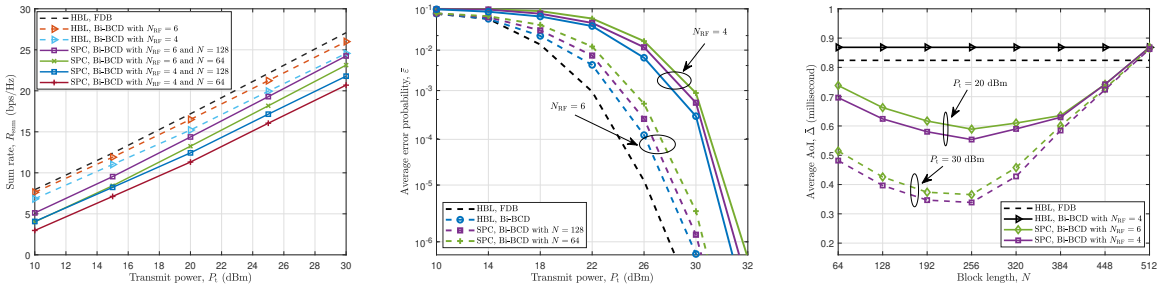
In Fig. 2b, we illustrate the *Pareto boundary* of the achievable rate region for two UEs using SPC for different block length N, while considering $P_t = 30$ dBm and $N_{\mathrm{RF}} = 6$. Since the conventional HBL transmission neglects the transmission error probability, HBL with FDB represents the upper bound for the SPC transmissions. Notably, the *Pareto boundary* of the HBL transmission via HBF, by employing the Bi-BCD algorithm, approaches the HBL with the FDB scheme. This shows the efficiency of our proposed Bi-BCD algorithm, which requires only a few RFCs. Moreover, the *Pareto boundary* of the SPC transmission via HBF monotonically increases with the block length N and gradually approaches its ideal HBL transmission at $N = 512$. This is because of the influence of FBL and the transmission error probability constraints in the SPC transmission. To demonstrate the effect of the numbers of RFCs and of the transmit power on the SPC transmission via HBF, Fig. 2c plots the achievable rate region of two UEs via varying N_{RF} and P_{t} for the fixed block length of $N = 128$. It can be observed from the figure that there is a small loss in the achievable rate boundaries of the SPC transmission via HBF with respect to the benchmarks. Thus, the proposed scheme reduces both hardware costs and power consumption by reducing the number of RFCs and still performs close to the ideal benchmarks. Moreover, as expected, the performance of the system improves upon increasing the transmit power P_{t} due to an increase in the resultant SINR at each UE.

Table I: Simulation parameters and corresponding values

Parameter	Definition	Value
N_{t}	Number of transmit antennas	64
N_{RF}	Number of RFCs	$\{4, 6\}$
M	Number of UEs	$\{2, 4\}$
N	Maximum allowable block length	512
ϵ	Decoding error probability	1×10^{-6}
N_{clu}	Number of clusters	4
N_{ray}	Number of rays in each cluster	8
P_{t}	Maximum transmit power	$\{20, 30\}$dBm
N_{o}	Noise power	-91 dBm

In Fig. 3a, we plot the sum rate versus transmit power for four UEs, $M = 4$. The figure reveals that the scheme proposed for SPC via HBF approaches the ideal benchmarks upon increasing either N_{RF} or N. Hence, to achieve a specific target sum rate for a fixed transmit power, one can further reduce the cost and power consumption by decreasing the number of RFCs and increasing the block length.

To investigate the reliability of the proposed scheme, we portray the average error probability in Fig. 3b with varying N and N_{RF}. For the fixed achievable rate R_m, we calculate the average error probability of the system as $\bar{\varepsilon} \approx$



(a) (b) (c)

Figure 3: (a) Sum rate versus transmit power with $M=4$; (b) Average error probability versus transmit power with $M=4$ and $R_m=6$ bps/Hz, $\forall m$; (c) Average AoI versus block length with $M=4$.

$\mathbb{E}_{\gamma_m}\left[Q\left(\sqrt{\frac{R_m}{V(\gamma_m)}}\left(\log_2(1+\gamma)-\frac{N}{R_m}\right)\right)\right]$ [8], where $\mathbb{E}_{\gamma_m}$ represents the expectation operator with respect to γ_m. As seen from the figure, increasing the transmit power and block length enhances the error probability performance. This improvement is anticipated, as higher transmit power and longer block length lead to a higher achievable rate. Moreover, the proposed FBL scheme achieves error probability comparable to that of Shannon capacity having HBL for both $N_{\rm RF}=4$ and 6, demonstrating the efficacy of the proposed design.

To further investigate the latency performance, Fig. 3c plots the average AoI of the system. Specifically, by following [9], the average AoI of the systems is given by $\bar{\Delta}=\frac{N(3-\bar{\varepsilon})}{2B(1-\bar{\varepsilon})}$, where $\bar{\varepsilon}$ is the average error probability and B represents the bandwidth, set to 28 GHz. As observed in the figure, the AoI initially decreases upon increasing N, reaches a minimum value, and then begins to increase as N continues to grow. This behavior occurs because, at first, increasing the block length enhances the achievable rate and decreases the error rate, leading to a reduction in AoI. However, beyond a certain point, the block length grows faster than the achievable rate, causing the AoI to increase. Thus, there exists an optimal block length that minimizes the average AoI. Additionally, the AoI decreases with an increase in transmit power and the number of RFCs, which may be attributed to the corresponding improvement in the achievable rate.

VI. CONCLUSION

The achievable rate region of an SPC-enabled mmWave system was investigated, which requires only few RFCs and finite block lengths. To design the Pareto optimal HBF for the SPC transmission, we proposed a Bi-BCD algorithm, which employs the BCD technique to design the BB, RF TPCs, as well as block length and the bisection search to obtain the rate of the users. Finally, simulation results are shown and compared to the benchmarks under different settings, which show the efficacy of our proposed Pareto-optimal HBF design in terms of the achievable rate. Future research should explore extending this work to practical scenarios associated with imperfect CSI, designing robust beamforming to address channel uncertainties as well as joint channel estimation and data detection.

APPENDIX A
PROOF OF PROPOSITION

Taking the constraint (8b) with equality into account, we have

$$\ln\left(1+\gamma_m\right)-\sqrt{V_m}\tau_m=\eta_m R_{\rm sum}, \tag{22}$$

where $\tau_m=\frac{Q^{-1}(\epsilon_m)}{\sqrt{\beta_m}}$. By employing $\delta_m=e^{-\eta_m R_{\rm sum}}$, (22) can be rewritten as

$$\ln\left[\delta_m\left(1+\gamma_m\right)\right]=\sqrt{V_m}\tau_m. \tag{23}$$

Let us define ϱ_m as $\varrho_m=\ln\left[\delta_m\left(1+\gamma_m\right)\right]$, then one can rewrite V_m as $V_m=1-\delta_m^2 e^{-2\varrho_m}$. Therefore, by substituting V_m into (23), ϱ_m is rewritten as

$$\varrho_m=\sqrt{1-\delta_m^2 e^{-2\varrho_m}}\tau_m. \tag{24}$$

By considering $\kappa_m=2\varrho_m$, and applying basic mathematical operations, (24) can be transformed as follows

$$e^{\kappa_m}\left(\kappa_m-2\tau_m\right)\left(\kappa_m+2\tau_m\right)=-4\delta_m^2\tau_m^2. \tag{25}$$

Consequently, the minimum value of γ_m can be achieved by $\Gamma_m=e^{\eta_m R_{\rm sum}+\frac{\kappa_m}{2}}-1$.

REFERENCES

[1] K. W. H. Ren and C. Pan, "Intelligent reflecting surface-aided URLLC in a factory automation scenario," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 707–723, 2022.

[2] B. Liu *et al.*, "Energy-efficient optimization in distributed massive MIMO systems for slicing eMBB and URLLC services," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10 473–10 487, 2023.

[3] G. J. Sutton *et al.*, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 3, 2019.

[4] Y. Polyanskiy *et al.*, "Channel coding rate in the finite blocklength regime," *EEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[5] S. He, Z. An, J. Zhu, J. Zhang, Y. Huang, and Y. Zhang, "Beamforming design for multiuser urllc with finite blocklength transmission," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8096–8109, 2021.

[6] S. Pala *et al.*, "Joint optimization of URLLC parameters and beamforming design for multi-RIS-aided MU-MISO URLLC system," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 148–152, 2023.

[7] T. Li *et al.*, "Robust beamforming design with finite blocklength for uRLLC," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, 2022.

[8] R. Hashemi *et al.*, "Average rate and error probability analysis in short packet communications over RIS-aided URLLC systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10 320–10 334, 2021.

[9] B. Yu, Y. Cai, and D. Wu, "Joint access control and resource allocation for short-packet-based mMTC in status update systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 851–865, 2021.

[10] X. Shang *et al.*, "Multiuser MISO interference channels with single-user detection: Optimality of beamforming and the achievable rate region," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4255–4273, 2011.
[11] R. Zhang and S. Cui, "Cooperative interference management with MISO beamforming," *IEEE Trans. Signal Process.*, vol. 58, no. 10, 2010.
[12] J. Singh *et al.*, "Energy efficiency optimization in reconfigurable intelligent surface aided hybrid multiuser mmwave mimo systems," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 581–589, 2023.
[13] M. Hui *et al.*, "Hybrid beamforming for utility maximization in multiuser broadband millimeter wave systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 16 042–16 057, 2023.
[14] J. Singh *et al.*, "Joint hybrid transceiver and reflection matrix design for RIS-aided mmWave MIMO cognitive radio systems," *IEEE Trans. Cogn. Commun. Netw.*, pp. 1–1, 2024.
[15] Z. Lin *et al.*, "Secrecy-energy efficient hybrid beamforming for satellite-terrestrial integrated networks," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6345–6360, 2021.
[16] ——, "Refracting RIS-aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 3717–3724, 2022.