Hybrid-Beamforming-Based Millimeter-Wave Cellular Network Optimization

Jia Liu[†] Elizabeth Bentley[*]
[†]Dept. of Electrical and Computer Engineering, The Ohio State University
[*]Air Force Research Laboratory, Information Directorate

***Abstract*—Massive MIMO and millimeter-wave communication (mmWave) have recently emerged as two key technologies for building 5G wireless networks and beyond. To reconcile the conflict between the large antenna arrays and the limited amount of radio-frequency (RF) chains in mmWave systems, the so-called hybrid beamforming becomes a promising solution and has received a great deal of attention in recent years. However, existing research on hybrid beamforming focused mostly on the physical layer or signal processing aspects. So far, there is a lack of theoretical understanding on how hybrid beamforming could affect mmWave *network optimization*. In this paper, we consider the impacts of hybrid beamforming on utility-optimality and queueing delay in mmWave cellular network optimization. Our contributions in this paper are three-fold: i) we develop a joint hybrid beamforming and congestion control algorithmic framework for mmWave network utility maximization; ii) we reveal a pseudoconvexity structure in the hybrid beamforming scheduling problem, which leads to simplified analog beamforming protocol design; and iii) we theoretically characterize the scalings of utility-optimality and delay with respect to channel state information (CSI) accuracy in digital beamforming.**

I. Introduction

In recent years, millimeter wave communication (mmWave) has emerged as a promising technology for building 5G wireless networks and beyond. The excitements of mmWave communications are primarily due to: i) the rich unlicensed spectrum resources in 60 GHz bands; ii) the ease of packing large antenna arrays into small form factors (a consequence of the short wavelengths); and iii) a much simplified interference management thanks to the highly directional "pencil-beam-like" mmWave signals. Moreover, recent field tests (see, e.g., [1], [2], etc.) have shown that the large directivity gains of mmWave transceivers can offset the high atmospheric attenuation in mmWave bands, dispelling the common concern that mmWave is not suitable for outdoor communications. The potential of mmWave networks has also stimulated many standardization activities (e.g., IEEE 802.15.3 wireless personal area networks, 802.11ad wireless local area networks, and fast-growing interests in mmWave cellular networks [3]).

However, the highly directional propagation of mmWave signals and the special mmWave hardware requirements also introduce several unique technical challenges for network systems. A main technical challenge is the beamforming architecture design, which lies at the heart of mmWave directional networking. Although large antenna arrays can be easily deployed in mmWave systems, the high power consumption of mixed mmWave signal components significantly limits the number of radio-frequency chains (RF chains), rendering full digital beamforming (requiring one RF chain per antenna) impractical [4]. Moreover, most of the digital beamforming schemes in traditional MIMO systems require full channel state information (CSI), which is difficult to acquire in mmWave systems due to the fast fading in mmWave spectrum and the low signal-to-noise ratio (SNR) before beamforming [5]. Because of the RF chain limitations in mmWave systems, analog beamforming approaches have been proposed (see, e.g., [6], [7]). The basic idea of analog beamforming is to control the phase shifters of antenna elements, so that the energy of the transmitted data stream is concentrated in a single direction to obtain a high directivity gain. Compared to digital beamforming, analog beamforming can be achieved by only one RF chain without requiring any CSI at the transmitter. However, analog beamforming can only transmit in a single beam direction and cannot leverage any spatial multiplexing capability of the large mmWave antenna array.

In light of the limitations of analog and digital beamformings, there is a growing consensus that the more suitable architecture for mmWave cellular networks is the *hybrid beamforming* architecture, which exploits the large mmWave antenna arrays and yet only requires a limited number of RF chains [5], [8]–[11]. Hybrid beamforming enjoys the best of both worlds: On one hand, it uses analog beamforming to offer spatial division and directivity gains to combat large mmWave channel attenuations. On the other hand, digital beamforming provides multiplexing gains for the lower dimensional *effective channels*, for which the CSI is relatively easier to acquire. It has been shown in [5], [12] that hybrid beamforming achieves a data rate performance comparable to full digital beamforming with 8 to 16 times fewer RF chains.

So far, however, the existing works on mmWave hybrid beamforming are mostly concerned with problems at the physical layer or signal processing aspects. To date, there remains a lack of theoretical understanding on how hybrid beamforming could affect the performances of mmWave network control, scheduling, and resource optimization algorithms. In this paper, our goal is to fill this gap by conducting an in-depth study on the impacts of hybrid beamforming on throughput and delay performances in mmWave cellular network optimization.

Specifically, in this paper, we focus on the algorithmic

This work has been supported by NSF grants CNS-1527078, 1446582, CCF-1618318, and an AFOSR SFFP'16 award. Distribution A. Approved for public release: Distribution unlimited: 88ABW-2016-4435 on 13 Sep 2016.

design and the throughput-delay analysis for the celebrated queue-length-based congestion control and scheduling framework (QCS) (see, e.g., [13], [14], and [15] for a survey) in hybrid-beamforming-based mmWave cellular networks. Our main results and technical contributions are as follows:

- We develop an accurate analytical model that captures the essence of hybrid beamforming in mmWave cellular networks, while being tractable enough to enable network-level understanding and analysis. Based on this analytical model, we formulate the problem of joint hybrid beamforming and congestion control for network utility maximization. We show that the joint hybrid beamforming and congestion control optimization is non-convex by nature, which creates challenges for the algorithmic designs in the MaxWeight scheduling component in the QCS framework.
- By exploiting the special problem structure of the mmWave MaxWeight scheduling component, we show that the non-convex scheduling subproblem admits a pseudoconvex approximation under a wide range of hybrid beamforming parameters of practical interests. Moreover, our analysis reveals that, to solve the scheduling subproblem, one only needs to adjust the analog beamwidth at the base station (BS), while the analog beamwidth adjustment at the mobile station (MS) side is unnecessary. This insight greatly simplifies the analog beamforming training protocol design.
- We investigate the impact of CSI inaccuracy on network performance with hybrid beamforming, where we assume that the true CSI is quantized by Q bits. We reveal a pair of interesting phase transition phenomena in utility-optimality and delay in the following sense: There exists a critical value $Q^\sharp$ such that: i) if $0 < Q < Q^\sharp$, then the deviations of steady-state queue-length grows linearly and congestion control rate is bounded by a constant; ii) If $Q \geq Q^\sharp$, the deviations of queue-lengths and congestion control rates have the same scaling laws as in the full CSI case.

Collectively, these results not only deepen our theoretical understanding of mmWave network optimization with hybrid beamforming, but also provide insights for low-complexity analog beam training and effective CSI quantization in practice. The remainder of this paper is organized as follows: In Section II, we introduce network models and the problem formulation. Section III presents the mmWave congestion control and scheduling framework, as well as the algorithmic design for analog beam training. Section IV studies the impacts of inaccurate CSI on digital beamforming. Section V provides numerical results and Section VI concludes this paper.

II. Network Model and Problem Formulation

Notation: We use boldface to denote matrices/vectors. $\mathbf{A}^\dagger$ denotes the conjugate transpose of $\mathbf{A}$. We use $\|\cdot\|$ and $\|\cdot\|_1$ to denote L^2- and L^1-norms, respectively. We let $\mathbf{I}$ denote the identity matrix, whose dimension is conformal to the context. We let $\mathbb{R}$ and $\mathbb{C}$ denote real and complex spaces, respectively.

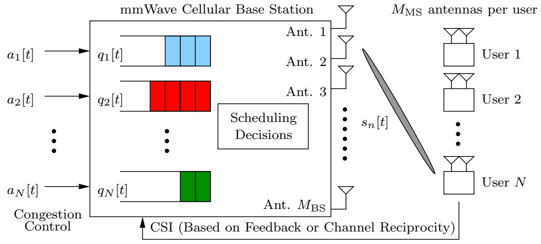

Fig. 1. A mmWave cellular downlink with a M_{BS}-antenna base station and N M_{MS}-antenna users.

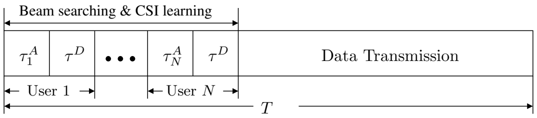

Fig. 2. Frame structure of a time-slot in mmWave cellular networks with hybrid beamforming.

1) Hybrid-Beamforming-Based mmWave Downlink: As shown in Fig. 1, we consider a mmWave cellular downlink system with N users. The BS and each user have M_{BS} and M_{MS} antennas, respectively. The mmWave downlink adopts a hybrid beamforming architecture with $M_{\mathrm{RF}}^{\mathrm{B}}$ and $M_{\mathrm{RF}}^{\mathrm{M}}$ RF chains at the BS and each user's MS, respectively (see Fig. 3). The system operates in a time-slotted mode. The time-slots are indexed by $t \in \{0, 1, 2, \ldots\}$. As shown in Fig. 2, each time-slot is of period T and contains two phases. The first phase is further divided into N mini-slots corresponding to the N users. Each mini-slot contains two parts τ_n^A and τ_n^D. In τ_n^A, both the BS and user n perform analog beam search to refresh their beam directions to mitigate link breakage caused by user n's movements [3], [16]. In τ_n^D, the BS estimates the CSI of user n for digital beamforming. In the data transmission phase, based on the analog beam and digital CSI training results, the BS picks one of the N users and steers analog beams to this user. Likewise, the scheduled user also steers analog beams toward the BS. Further, by leveraging the learned CSI to perform spatial multiplexing, the BS and a scheduled user communicate via K data streams. For mmWave systems in practice, we usually have: i) $K \leq M_{\mathrm{RF}}^{\mathrm{B}} \leq M_{\mathrm{BS}}$; ii) $K \leq M_{\mathrm{RF}}^{\mathrm{M}} \leq M_{\mathrm{MS}}$; iii) $M_{\mathrm{RF}}^{\mathrm{M}} \leq M_{\mathrm{RF}}^{\mathrm{B}}$; and iv) $M_{\mathrm{MS}} \leq M_{\mathrm{BS}}$.

a) Analog beamforming process: In time-slot t, the analog beamformers on the BS and user sides are determined by a beam training process, during which the BS and user n search over all possible direction combinations within their corresponding sectors[1], as shown in Fig. 4 (this exhaustive beam training process has been adopted in IEEE 802.11ad and IEEE 802.15.3c standards). Let T_p denote the time required for transmitting and receiving a pilot symbol. Let ψ_n^{B} and ψ_n^{M} denote the sector-level beamwidth at the BS and user n, respectively. Also, let $\theta_B[t]$ and $\theta_M[t]$ denote the beam-level beamwidth at the BS and user n's MS, respectively. Then, the beam search time τ_n^A can be computed as: $\tau_n^A = \frac{\psi_n^{\mathrm{B}}}{\theta_B[t]} \frac{\psi_n^{\mathrm{M}}}{\theta_M[t]} T_p$.

In this paper, we adopt a widely used sectored antenna pattern model (see, e.g., [17]–[19]): We assume that the gains

[1] In this paper, we assume that both the BS and user know the sectors of each other's location in each time-slot. This is a reasonable assumption because the sector information can be inferred with high accuracy from the beam direction in the previous time-slot and the mobility/trajectory information of the user.

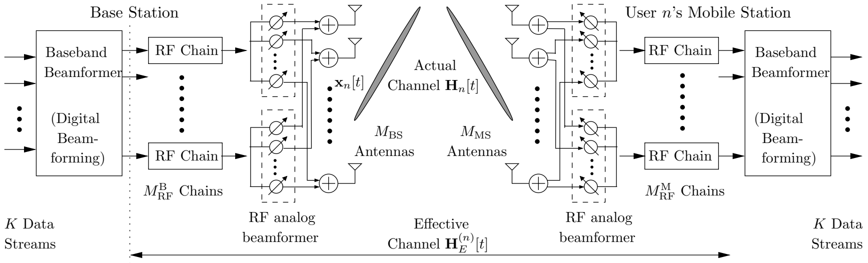

Fig. 3. Block diagram of a mmWave cellular network with hybrid beamforming.

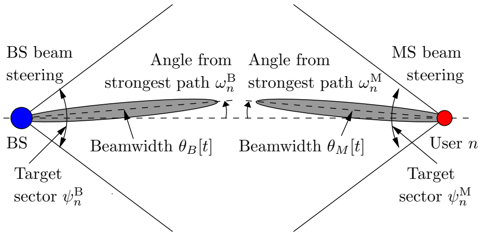
BS beam steering
Angle from strongest path $\omega_n^{\rm B}$
Angle from strongest path $\omega_n^{\rm M}$
MS beam steering
BS
Beamwidth $\theta_B[t]$
Beamwidth $\theta_M[t]$
User n
Target sector $\psi_n^{\rm B}$
Target sector $\psi_n^{\rm M}$

Fig. 4. The analog beamforming training procedure.

are a constant for all angles within the main lobe and equal to a smaller constant in the side lobes. As shown in Fig. 4, we let $\omega_n^{\rm B}$ and $\omega_n^{\rm M}$ represent the angles deviating from the strongest path between the BS and user n, respectively (the strongest path needs not be line-of-sight and Fig. 4 is only for illustrative purposes). Let $g_n^{\rm B}(\omega_n^{\rm B}, \theta_B[t])$ and $g_n^{\rm M}(\omega_n^{\rm M}, \theta_M[t])$ denote the transmission and reception gains at the BS and user n. In this paper, we adopt the following widely used antenna radiation pattern model [17]–[19]:

$$g_n^{\rm B}(\omega_n^{\rm B}, \theta_B[t]) = \begin{cases} \frac{2\pi-(2\pi-\theta_B[t])\eta}{\theta_B[t]}, & \text{if } |\omega_n^{\rm B}| \leq \frac{\theta_B[t]}{2}, \\ \eta, & \text{otherwise}, \end{cases} \tag{1}$$

$$g_n^{\rm M}(\omega_n^{\rm M}, \theta_M[t]) = \begin{cases} \frac{2\pi-(2\pi-\theta_M[t])\eta}{\theta_n[t]}, & \text{if } |\omega_n^{\rm M}| \leq \frac{\theta_M[t]}{2}, \\ \eta, & \text{otherwise}, \end{cases} \tag{2}$$

where $\eta \in [0, 1)$ is the side lobe gain. In practice, $\eta \ll 1$ for narrow beams (i.e., $\theta_B[t]$ and $\theta_M[t]$ are small). This model captures the essential features of antenna patterns (e.g., directive gains, front-to-back ratio, half-power beamwidth, etc. [19]). Once the optimal directions for transmission and reception have been determined, the communication link can be established, and data transmission phase starts. The beam training is finished when the BS and the user's beams are aligned with the strongest path, i.e., the conditions $|\omega_n^{\rm B}| \leq \frac{\theta_B[t]}{2}$ in (1) and $|\omega_n^{\rm M}| \leq \frac{\theta_M[t]}{2}$ in (2) are satisfied.

b) Digital beamforming process: Once the analog beam search is completed, the analog beamformers are known. Therefore, we can estimate the CSI of the effective channel $\mathbf{H}_E^{(n)}[t]$, which is assumed to take $\tau^D = \beta T_p$ amount of time (cf. Fig. 2), where $\beta > 0$ is some constant. With the learned CSI, the BS and user n jointly choose baseband beamformers based on some digital beamforming strategies, such as singular value decomposition (SVD), zero-forcing (ZF), etc. One particularly interesting case arises when $M_{\rm RF}^{\rm B} \gg M_{\rm RF}^{\rm M}$. In this case, the row vectors in the effective channel $\mathbf{H}_E^{(n)}[t]$ are asymptotically orthogonal to each other as $M_{\rm RF}^{\rm B}$ gets large. Thanks to this nice property, one can use the so-called conjugate beamforming, which has been shown to be asymptotically capacity-achieving in the high SNR regime [20]. We will further discuss conjugate beamforming in Section IV.

Regardless the choice of digital beamforming schemes, the digital beamforming process converts $\mathbf{H}_E^{(n)}[t]$ into $K \leq \min\{M_{\rm RF}^{\rm B}, M_{\rm RF}^{\rm M}\}$ spatial channels (depending on the rank of $\mathbf{H}_E^{(n)}[t]$). We let $g_n^{(k)}[t]$ denote the effective gain of the k-th spatial channel. Based on the models of hybrid analog/digital beamforming, we have that the hybrid beamforming achievable rate of user n can be computed as:[2]

$$r_n(\theta_B[t], \theta_n[t]) = \left(1 - \frac{\tau^A + N\tau^D}{T}\right) \sum\nolimits_{k=1}^{K} \log_2 \bigg(1 + \frac{P_{\max}}{KN_0} g_n^{\rm B}(\omega_n^{\rm B}, \theta_B[t]) g_n^{\rm M}(\omega_n^{\rm M}, \theta_M[t]) g_n^{(k)}[t]\bigg), \tag{3}$$

where $\tau^A \triangleq \sum_{n=1}^{N} \tau_n^A$ and $P_{\max}$ denotes the maximum transmission power at the BS. Then, for a given channel state in time-slot t, we let $\mathcal{C}_n[t]$ denote the instantaneous achievable rate region under a chosen digital beamforming scheme:

$$\mathcal{C}_n[t] \triangleq \left\{ r_n(\theta_B[t], \theta_n[t]) \;\middle|\; \begin{array}{l} \theta_B[t] \in (0, \psi_n^{\rm B}], \\ \theta_M[t] \in (0, \psi_n^{\rm M}]. \end{array} \right\}. \tag{4}$$

It can be seen from (3) that the beamwidths $\theta_B[t]$ and $\theta_M[t]$ need to be chosen judiciously: On one hand, from (1) and (2), $g_n^{\rm B}(\omega_n^{\rm B}, \theta_B[t])$ and $g_n^{\rm M}(\omega_n^{\rm M}, \theta_M[t])$ increase as $\theta_B[t]$ and $\theta_M[t]$ decrease, leading to a higher SNR and hence a higher data rate. However, the smaller the beamwidths $\theta_B[t]$ and $\theta_M[t]$, the shorter the transmission phase, i.e., there exists a trade-off between data rate and transmission time.

2) Queueing Model: As shown in Fig. 1, the BS maintains a separate queue for each user. Let $a_n[t]$ denote the number of packets injected into queue n in time-slot t. The arrival processes $\{a_n[t]\}$, $\forall n$, are controlled by a congestion controller. We assume that there exists a finite constant $A^{\max}$ such that $a_n[t] \leq A^{\max}$, $\forall n, t$. Let $\mathbf{s}[t] \triangleq [s_1[t], \ldots, s_N[t]]^\top$ denote the scheduled service rate vector in time-slot t (the scheduling algorithm that determines $\mathbf{s}[t]$ will be presented in Section III). Then, the queue-length of user n evolves as: $q_n[t+1] = \big(q_n[t] - s_n[t] + a_n[t]\big)^+$, $\forall n$, where $(\cdot)^+ \triangleq \max(0, \cdot)$. Let $\mathbf{q}[t] = [q_1[t]], \ldots, q_N[t]]^\top$. In this paper, we adopt the following notion of queue-stability (same as in [13], [14]): We say that a network is *stable* if the steady-state total queue-length is finite, i.e., $\limsup_{t\to\infty} \mathbb{E}\{\|\mathbf{q}[t]\|_1\} < \infty$.

3) Problem Formulation: Let $\bar{a}_n \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} a_n[t]$ denote the average controlled arrival rate of user n. We associate each user n with a utility function $U_n(\bar{a}_n)$, which

[2] In this paper, equal power allocation is used for lower rate evaluation complexity in the effective MIMO channel. It has been shown that the rate loss of equal power allocation is negligible under high SNR. Also, equal power allocation is asymptotically capacity-achieving in high SNR regime [21].

is assumed to be strongly concave, increasing, and twice continuously differentiable. $U_n(\bar{a}_n)$ represents the utility gained by user n when data is injected at rate $\bar{a}_n$. Then, the joint congestion control and scheduling (JCS) optimization problem for network utility maximization can be written as:

$$\begin{aligned}\textbf{JCS:}\ \text{Maximize}\ \ & \textstyle\sum_{n=1}^{N} U_n(\bar{a}_n)\\ \text{subject to}\ \ & \text{Queue-length stability constraints,}\\ & s_n[t] \in \mathcal{C}_n[t],\ a_n[t] \in [0, A^{\max}]\ \forall n, t.\end{aligned}$$

In Section III, we will first consider the algorithmic design for solving Problem JCS under perfect CSI. Then, in Section IV, we will conduct an in-depth investigation on the impacts of CSI inaccuracy on throughput and delay.

III. Algorithmic Design under Perfect CSI

Based on the Lagrangian dual formulation, it can be shown that Problem JCS can be solved by the queue-length-based congestion control and scheduling (QCS) framework (see, e.g., [13]–[15]) in the following sense: The congestion control rate $\bar{\mathbf{a}}$ achieves an optimality gap $O(\epsilon)$ at the price of an $O(1/\epsilon)$ queue-length, where $\epsilon > 0$ controls the utility-optimality gap. Hence, the utility-optimality gap can be made arbitrarily small by decreasing ϵ. Now, consider the following QCS algorithm specialized for hybrid beamforming in mmWave networks:

A. The QCS Algorithm Specialized for Hybrid Beamforming

Algorithm 1: Queue-Length-Based Congestion Control and Scheduling in mmWave Cellular Network.

Initialization: Choose parameters $\epsilon > 0$. Set $t = 0$.

Main Loop:

1. *MaxWeight Scheduler:* In time $t \geq 1$, given queue-lengths $\mathbf{q}[t]$ and CSI $\mathbf{H}[t]$, the scheduler chooses a service rate vector $\mathbf{s}[t]$ from $\mathcal{C}_n[t]$ by hybrid beamforming such that:
$$\mathbf{s}[t] = \underset{r_n \in \mathcal{C}_n[t], \forall n}{\arg\max} \left\{ \sum_{n=1}^{N} q_n[t] r_n \right\}, \qquad (5)$$
where $\mathcal{C}_n[t]$ is defined in (4).
2. *Congestion Controller:* Given queue-lengths $\mathbf{q}[t]$, the congestion controller chooses data injection rates $a_n[t]$, $\forall n$, which are integer-valued random variables satisfying:
$$\mathbb{E}\{a_n[t] | q_n[t]\} = \min\left\{ U_n^{'-1}\left(\epsilon q_n[t]\right), A^{\max} \right\}, \qquad (6)$$
$$\mathbb{E}\{a_n^2[t] | q_n[t]\} \leq A_2^{\max} < \infty, \quad \forall q_n[t], \qquad (7)$$
where $U_n^{'-1}(\cdot)$ represents the inverse function of first-order derivative of $U_n(\cdot)$. In (6) and (7), $A^{\max}$ and $A_2^{\max}$ are some predefined sufficiently large positive constants.
3. *Queue-Length Updates:* Update the queue-lengths as $q_n[t+1] = (q_n[t] - s_n[t] + a_n[t])^+$, $\forall n$. Let $t = t+1$. Go to Step 1 and repeat the process.

Although being optimal, the QCS framework has a major limitation in that the MaxWeight scheduling problem is difficult to solve in general and could be NP-Hard for many wireless networks [15]. Surprisingly, in what follows, we will show that the physical layer properties of mmWave hybrid beamforming imply several special mathematical structures that lead to efficient solution for the MaxWeight subproblem.

B. The MaxWeight Scheduling Subproblem

To solve the MaxWeight scheduling subproblem, we start by examining the properties of the set of instantaneous hybrid beamforming achievable rates $\{r_n \in \mathcal{C}_n[t], \forall n\}$. First, we note that the BS forms only one beam to one of the N users in each time-slot, say user n. This implies that $r_{n'} = 0, \forall n' \neq n$. Hence, the MaxWeight problem in (5) can be simplified as:

$$\max_{r_n \in \mathcal{C}_n[t], \forall n} \left\{ \sum_{n=1}^{N} q_n[t] r_n \right\} = \max_{n \in \{1,\dots,N\}} \left\{ q_n[t] r_n \middle| r_n \in \mathcal{C}_n[t] \right\}$$
$$\overset{(a)}{=} \max_{n \in \{1,\dots,N\}} \left\{ q_n[t] \left[\max_{\theta_B[t], \theta_M[t]} \left\{ r_n(\theta_B[t], \theta_M[t]) \right\} \right] \right\}, \qquad (8)$$

where (a) follows from the fact that $r_n(\theta_B[t], \theta_M[t])$ does not depend on $q_n[t]$. As a result, solving the MaxWeight scheduling problem in (5) boils down to first solving the inner rate maximization problem in (8) for each user, and then choosing the user who has the largest rate-queue-length product. Toward this end, we explicitly write down the inner maximization problem in (8) for each user n as follows:

$$\begin{aligned}\text{Maximize} \left(1 - \frac{\tau^A + \tau^D}{T}\right) \sum_{k=1}^{K} \log_2 \Bigg(1 + \frac{P_{\max}}{K N_0} g_n^{\mathrm{B}}(\omega_n^{\mathrm{B}}, \theta_B[t])\\ g_n^{\mathrm{M}}(\omega_n^{\mathrm{M}}, \theta_M[t]) g_n^{(k)}[t]\Bigg) \qquad (9)\end{aligned}$$
$$\text{subject to} \quad \theta_B[t] \in (0, \psi_n^{\mathrm{B}}],\ \theta_M[t] \in (0, \psi_n^{\mathrm{M}}].$$

Unfortunately, due to the multiplication between the time fraction and the rate in the objective function, Problem (9) falls into the class of polynomial programming problems, which is *non-convex* [22]. However, it turns out that if the side lobe gain η is small, Problem (9) can be approximated by a univariate pseudoconvex problem. We state this result as follows:

Theorem 1 (Univariate pseudoconvex approximation)**.** *If the side lobe gain satisfies $\eta \ll \frac{1}{3}$, then Problem (9) can be approximated by the following univariate optimization problem:*

$$\text{Maximize} \quad \left(b_0 - \frac{b_1}{\tilde{\theta}[t]}\right) \sum_{k=1}^{K} \log_2 \left(1 + \frac{4\pi^2 c_n^{(k)}}{\tilde{\theta}[t]}\right) \qquad (10)$$
$$\text{subject to} \quad \tilde{\theta}[t] \in \left[b_1/b_0,\ \psi_n^{\mathrm{B}} \psi_n^{\mathrm{M}}\right],$$

where $b_0 \triangleq 1 - (N\beta T_p/T)$, $b_1 \triangleq \frac{T_p}{T} \sum_{n=1}^{N} \psi_n^{\mathrm{B}} \psi_n^{\mathrm{M}}$, and $c_n^{(k)} \triangleq (P_{\max}/K N_0) g_n^{(k)}[t]$ are constants. Moreover, Problem (10) is a pseudoconvex optimization problem.

Theorem 1 can be proved by substituting the antenna radiation pattern model in (1) and (2) into the objective function of Problem (9) and then exploiting the condition $\eta \ll \frac{1}{3}$ to simplify. Since Problem (10) is a maximization problem with one simple box constraint, showing its pseudoconvexity is equivalent to showing the pseudoconcavity of the objective function. We refer readers to the appendix for proof details.

Remark 1. Three remarks of Theorem 1 are in order: i) In practice, the conditions $T_p \ll T$ and $\eta \ll \frac{1}{3}$ can usually be satisfied because a pilot symbol is much shorter compared to a time-slot and the mmWave beams are sharp; ii) The pseudoconvex (which further implies strictly quasiconvex) and univariate properties suggest that Problem (10) can be solved by simple one-dimensional line search methods [22, Theorem 8.1.1] (e.g., the bisection or the golden section methods); iii) It can be seen from the proof of Theorem 1 that we have defined $\tilde{\theta}[t] \triangleq \theta_B[t]\theta_M[t]$. Note that the optimal objective value of Problem (10) is only a function of $\tilde{\theta}^*[t]$ and does not depend on the specific values of $\theta_B[t]$ and $\theta_{\mathrm{M}}[t]$, as long as their product is equal to $\tilde{\theta}^*[t]$. This implies that we can simply set $\theta_M[t]$ to some appropriate fixed value and only adjust $\theta_B[t]$ at the BS side. In other words, there is no need to jointly adjust $\theta_B[t]$ and $\theta_M[t]$. This insight greatly simplifies the protocol designs in the analog beamforming phase. □

Collectively, the results in this section provide an algorithmic solution to Problem JCS assuming that the CSI learned in τ^D (hence the digital beamforming gains $g_n^{(k)}[t]$) is accurate. However, it remains unclear how the network utility and delay performance of Algorithm 1 will be affected if the CSI is inaccurate. This problem will be addressed in the next section.

IV. The Impacts of Inaccurate CSI on the QCS Algorithm with Hybrid Beamforming

Due to the short coherence time of mmWave channels (around an order of magnitude lower than that of microwave bands since Doppler shifts scale linearly with frequencies [3]), traditional CSI feedback approach is not suitable for mmWave-based cellular networks. Also, due to the limited transmit power at the MSs and the lack of beamforming gains for the uplink pilot symbols, the accuracy of TDD-based CSI estimation based on channel reciprocity is limited. Given these CSI estimation challenges, it is likely that the CSI learned during the τ^D period (cf. Fig. 2) is inaccurate. In this paper, we assume that the number of RF chains at the BS is much greater than that at the MSs, (e.g., tens of times larger). This setting is relevant for cases where the physical size, hardware costs, and power constraint of the BS are not limiting factors of the system. The cases where the BS and MSs have comparable numbers of RF chains will be left for our future studies. In what follows, we start with the digital beamforming for effective mmWave channels with a large number of RF chains at the BS and its operations under a limited CSI model.

1) Digital Beamforming for Effective Channels with $M_{\mathrm{RF}}^{\mathrm{B}} \gg M_{\mathrm{RF}}^{\mathrm{M}}$: As mentioned in Section II, due to the near orthogonality between the rows in the effective channel in this case, the simple conjugate digital beamforming technique can be used. Recall that the received signal of user n can be written as: $\mathbf{y}[t] = \mathbf{H}_E^{(n)}[t]\mathbf{F}_D^{(n)}[t]\mathbf{u}_n[t] + \tilde{\mathbf{n}}[t]$, where $\mathbf{H}_E^{(n)}[t] \in \mathbb{C}^{M_{\mathrm{RF}}^{\mathrm{M}} \times M_{\mathrm{RF}}^{\mathrm{B}}}$ is the effective channel by taking into account the effects of analog beamforming; and $\mathbf{F}_D^{(n)}[t]$ is the transmit digital beamformer. Under conjugate beamforming, we let $\mathbf{F}_D^{(n)}[t] = \mathbf{H}_E^{(n)}[t]^\dagger$. Also, we assume that the effective channel $\mathbf{H}_E^{(n)}[t]$ is of full row rank so that $K = M_{\mathrm{RF}}^{\mathrm{M}}$ (i.e., all receiver RF chains are utilized). Then, thanks to the near orthogonality between the rows in $\mathbf{H}_E^{(n)}[t]^\dagger$, the achievable rate under conjugate beamforming can be computed as:

$$r_n[t] \approx \left(1 - \frac{\tau^A + \tau^D}{T}\right) \sum_{k=1}^{K} \log_2 \left(1 + \frac{P_{\max}}{K N_0} \|\mathbf{h}_{E,k}^{(n)}[t]\|^2\right), \quad (11)$$

where $\mathbf{h}_{E,k}^{(n)}[t]$ denotes the k-th row of $\mathbf{H}_E^{(n)}[t]$. Note that (11) and (3) are equivalent since $\mathbf{h}_{E,k}^{(n)}[t]$ has absorbed the analog beamforming gains $g_n^{\mathrm{B}}(\omega_n^{\mathrm{B}}, \theta_B[t])$ and $g_n^{\mathrm{M}}(\omega_n^{\mathrm{M}}, \theta_n[t])$.

2) CSI Inaccuracy Modeling: Given the inevitable CSI errors and to alleviate the CSI estimation burden for digital beamforming, we adopt the limited CSI model in the literature (see, e.g., [21] and references therein). Such limited CSI can be obtained by Q bits of feedback from each user. Alternatively, based on the channel reciprocity, the BS could use Q bits to rapidly quantize the uplink CSI (see Fig. 1). In either case, the value of Q depends on the CSI learning time τ^D and efficiency of the specific CSI learning algorithm. The Q-bit limited CSI for each RF chain k can be determined by a vector quantization codebook $\mathcal{B}_k \triangleq \{\mathbf{c}_k^1, \ldots, \mathbf{c}_k^{2^Q}\}$, where $\mathbf{c}_k^i \in \mathbb{C}^{M_{\mathrm{RF}}^{\mathrm{B}}}$, $i = 1, \ldots, 2^Q$, represents a codeword. Given an effective channel $\mathbf{H}_E^{(n)}[t]$, the codeword for its k-th row vector $\mathbf{h}_{E,k}^{(n)}[t]$ is chosen by picking the one that is closest to $\mathbf{h}_{E,k}^{(n)}[t]$ in the following sense [21]: $i_k^*[t] = \arg\min_{j \in \{1,\ldots,2^Q\}} \sin^2(\angle(\mathbf{h}_{E,k}^{(n)}[t], \mathbf{c}_n^j))$, where $i_k^*[t]$ denotes the index of the chosen codeword in time-slot t. Let $\widehat{\mathbf{H}}_E^{(n)}[t] \in \mathbb{C}^{M_{\mathrm{RF}}^{\mathrm{M}} \times M_{\mathrm{RF}}^{\mathrm{B}}}$ denote the estimated channel matrix by collecting all codewords $i_k^*[t]$, $\forall k$. Then, based on $\widehat{\mathbf{H}}_E^{(n)}[t]$, the BS performs conjugate beamforming to construct K spatial channels. However, due to the errors in $\widehat{\mathbf{H}}_E^{(n)}[t]$, inter-channel interference is not negligible under conjugate beamforming, and the amount of interference depends on the codebook size 2^Q and the choice of the quantization scheme.

Let $\hat{r}_n^Q[t]$ denote the actual conjugate beamforming achievable rate under the true CSI $\mathbf{H}_E^{(n)}[t]$ while the system is treating the Q-bit limited CSI $\widehat{\mathbf{H}}_E^{(n)}[t]$ as if it is accurate. Also, let $\widehat{\mathbf{H}}_{E,1}^{(n)}[t]$ and $\widehat{\mathbf{H}}_{E,2}^{(n)}[t]$ represent two estimated CSI values obtained by using Q_1 and Q_2 bits, respectively. Then, we can show that the following monotonicity result of the conjugate beamforming achievable rate holds under limited CSI, which will be used in our subsequent analysis (the proof is relegated to our online technical report [23] due to space limitation):

Lemma 1 (Monotonicity of beamforming achievable rate)**.** *If $Q_1 \leq Q_2$, then there exists a CSI quantization scheme under which $\hat{r}_n^{Q_1}[t] \leq \hat{r}_n^{Q_2}[t]$. Further, $\hat{r}_n^Q[t] \uparrow r_n[t]$ as $Q \to \infty$.*

3) Algorithmic Changes to the QCS Framework: Due to the use of Q-bit limited CSI in mmWave hybrid beamforming, the QCS algorithmic framework in Algorithm 1 also needs to be modified accordingly as follows:

Algorithm 2: Queue-Length-Based Congestion Control and Scheduling in mmWave Cellular Network with Q-Bit CSI.

```
Initialization: Choose parameters $\epsilon > 0$. Set $t = 0$.
```

Main Loop:

1. *MaxWeight Scheduler:* In time-slot $t \geq 1$, given the queue-length vector $\mathbf{q}[t]$ and the Q-bit estimated CSI $\widehat{\mathbf{H}}_E^{(n)}[t]$, $\forall n$, we let $\tilde{r}_n[t]$ be the believed conjugate beamforming achievable rate under $\widehat{\mathbf{H}}_E^{(n)}[t]$, $\forall n$. Then, the scheduler chooses a user n such that $n = \arg\max_{n' \in \{1,\ldots,N\}} \{q_{n'}[t]\tilde{r}_{n'}[t]\}$. As a result, the actual achievable service rates are $s_{Q,n}[t] = \widehat{r}_n^Q[t]$ and $s_{Q,n'}[t] = 0$, $\forall n' \neq n$.
2. *Congestion Controller:* Same as in Algorithm 1.
3. *Queue-Length Updates:* Same as in Algorithm 1.

4) Performance Analysis: To describe our main theoretical results, we first need the following deterministic problem, where we assume that the channel state process is not random and fixed at its mean. We let $\bar{\mathcal{C}}^Q \triangleq \{r_n^Q, \forall n : r_n^Q = \mathbb{E}\{\widehat{r}_n^Q[t]\}\}$ denote the mean achievable rate region. Also, the congestion control and scheduling variables are time-invariant and denoted as a_n and $s_{Q,n}$, $\forall n$, respectively. Then, the deterministic congestion control and scheduling problem can be written as:

$$\text{Maximize} \left\{ \frac{1}{\epsilon} \sum_{n=1}^{N} U_n(a_n) \;\middle|\; \begin{array}{l} a_n - s_{Q,n} \leq 0, \forall n, \\ s_{Q,n} \in \bar{\mathcal{C}}^Q, \forall n, \\ a_n \in [0, a^{\max}],\ \forall n. \end{array} \right\}. \quad (12)$$

Based on the convex approximation argument in Theorem 1, it is clear that Problem (12) is approximately convex. Thus, there exists a unique optimal solution. Associating dual variables $q_{Q,n} \geq 0$, $\forall n$ with the constraints $a_n - s_{Q,n} \leq 0$, $\forall n$, we obtain the Lagrangian as follows: $\Theta_\epsilon(\mathbf{q}_Q) \triangleq \max_{\mathbf{a},\mathbf{s}_Q \in \bar{\mathcal{C}}^Q} \{\frac{1}{\epsilon}\sum_{n=1}^N U_n(a_n) + \sum_{n=1}^N q_{Q,n}(s_{Q,n} - a_n)\}$, where the notation $\Theta_\epsilon(\cdot)$ signifies the Lagrangian's dependence on ϵ and the vector $\mathbf{q}_Q \triangleq [q_{Q,1}, \ldots, q_{Q,N}]^\top \in \mathbb{R}_+^N$ contains all dual variables. Then, the Lagrangian dual problem of Problem (12) can be written as:

$$\text{Minimize } \Theta_\epsilon(\mathbf{q}_Q), \text{ subject to } \mathbf{q}_Q \in \mathbb{R}_+^N. \quad (13)$$

Let $(\mathbf{a}_Q^*, \mathbf{s}_Q^*)$ and $\mathbf{q}_{Q,(\epsilon)}^*$ be the optimal primal and dual solutions to Problems (12) and (13), respectively. Then, it can be shown that $\mathbf{q}_{Q,(\epsilon)}^*$ satisfies the following properties (the proof is based on the Karush-Kuhn-Tucker (KKT) conditions [22] and is relegated to [23] due to space limitation):

Lemma 2 (Primal and dual solutions)**.** *$\mathbf{q}_{Q,(\epsilon)}^* = \frac{1}{\epsilon}\mathbf{q}_{Q,(1)}^*$, i.e.,, $\mathbf{q}_{Q,(\epsilon)}^*$ grows linearly and the slope depends on $\mathbf{q}_{Q,(1)}^*$. Further, if $Q_1 \leq Q_2$, then the slopes satisfy $\mathbf{q}_{Q_1,(1)}^* \geq \mathbf{q}_{Q_2,(1)}^*$. The congestion control solution $\mathbf{a}_Q^*$ is independent of ϵ and equal to the service rate $\mathbf{s}_Q^*$.*

With Lemma 2, we are now ready to present the main results. Our first result says that the steady-state queue-length vector $\mathbf{q}^\infty$ lies in a bounded neighborhood of the dual solution $\mathbf{q}_{Q,(\epsilon)}^*$ of Problem (13). Further, the size of the neighborhood manifests a phase-transition phenomenon.

Theorem 2 (Queueing delay phase transition)**.** *Under Algorithm 2 with any given Q-bit CSI quantization scheme, there exists a critical value $Q^\sharp$ independent of the performance control parameter K of Algorithm 2, such that:*

- *If $0 < Q < Q^\sharp$, then $\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{Q,(\epsilon)}^*\|\} = O(C_{(Q)}\frac{1}{\epsilon})$, where $C_{(Q)} \geq 0$ is a constant depending on the quantization codebook, and $C_{(Q)}$ decreases as Q increases;*
- *If $Q \geq Q^\sharp$, then $\mathbb{E}\{\|\mathbf{q}^\infty - \mathbf{q}_{Q,(\epsilon)}^*\|\} = O(1/\sqrt{\epsilon})$.*

Remark 2. Note that the $O(1/\sqrt{\epsilon})$-scaling of queue-length deviation when $Q \geq Q^\sharp$ is the same as that under the full CSI case [14]. This shows an unexpected insight that full CSI is not necessary to induce in order sense the original QCS queue-length scaling behavior. □

Now, let $a_{Q,n}^\infty \triangleq \mathbb{E}\{\min\{U_n^{'-1}(\epsilon q_n^\infty), a^{\max}\}\}$, $\forall n$, be the steady-state congestion control rates under a given Q-bit CSI quantization scheme and let $\mathbf{a}_Q^\infty \triangleq [a_{Q,1}^\infty, \ldots, a_{Q,N}^\infty]^\top$. The next main result characterizes the phase transition of the scaling of $\mathbf{a}_Q^\infty$'s deviation from the solution $\mathbf{a}_Q^*$ of Problem (12):

Theorem 3 (Congestion control phase transition)**.** *Under Algorithm 2 with any Q-bit CSI quantization scheme, there exists a critical value $Q^\sharp$, same as in Theorem 2, such that:*

- *If $0 < Q < Q^\sharp$, then $\|\mathbf{a}_Q^\infty - \mathbf{a}_Q^*\| = O(C_{(Q)})$, where $C_{(Q)} \geq 0$ is the same constant as defined in Theorem 2;*
- *If $Q \geq Q^\sharp$, then $\|\mathbf{a}_Q^\infty - \mathbf{a}_Q^*\| = O(\sqrt{\epsilon})$.*

Both Theorems 2 and 3 can be proved by Lyapunov stability analysis, and the details are relegated to the appendix.

V. Numerical Results

In this section, we conduct simulations to demonstrate the theoretical results in Sections III and IV. We first verify the approximation accuracy and the pseudoconvexity of Problem (10). We set SNR to 30 dB and set the T_p/T ratios to 0.01 and 0.001. We vary the side lobe gain η from 0.1 to 0.5 and the results are shown in Figs. 5 and 6. We can see that, under both T_p/T ratios, the approximation gaps shrinks as η decreases. In these examples, the gaps under $\eta = 0.1$ are almost negligible. Moreover, we note that the approximation function is indeed pseudoconcave, as predicted by Theorem 1.

Next, we examine the impacts of Q on the queue-lengths and the results are shown in Fig. 7. In our simulations, the BS and each MS have 64 and 2 RF chains, respectively. The total SNR is 30 dB. We use $\log(\cdot)$ as the utility function for each user (i.e., the proportional fairness metric [15]) and adopt random vector quantization (RVQ) as our Q-bit CSI quantization codebook [21], with $Q = 1, 2, 4, 8, 16, 32, 48$, and 64. We also draw an accompanying dash line to show the scaling trend of each curve in Fig. 7. For small Q values, we can see that the mean queue-length deviation increases faster than the square root law, roughly displaying a linear scaling with respect to ϵ as indicated in Theorem 2. For this example, the critical value of Q is 8. Once $Q \geq 8$, the queue-length deviations scale as $O(1/\sqrt{\epsilon})$, also confirming Theorem 2.

Lastly, we study the impacts of Q-bit CSI on the congestion control rates and the results are illustrated in Fig. 8. For small Q values, we can see that $\mathbf{a}_Q^\infty$ is only affected by Q and is a constant independent of ϵ. Also, $\mathbf{a}_Q^\infty$'s gap to the full CSI case shrinks as Q increases, which confirms Lemma 2 and Theorem 3. Again, we can observe that the critical value of

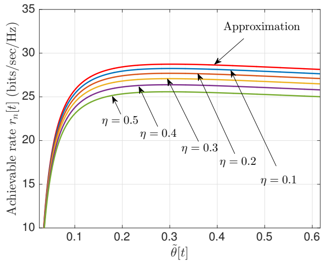

Fig. 5. The approximation gaps of Problem (10) under different analog slide lobe gains η ($T_p/T = 0.01$).

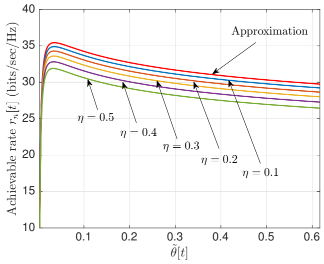

Fig. 6. The approximation gaps of Problem (10) under different analog slide lobe gains η ($T_p/T = 0.001$).

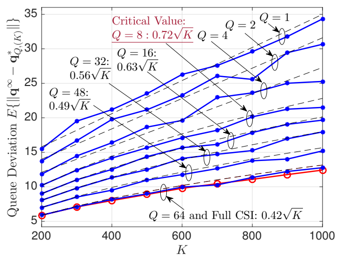

Fig. 7. Average queue-length deviation with respect to K for $Q = 1, 2, 4, 8, 16, 32, 48, 64$ bits.

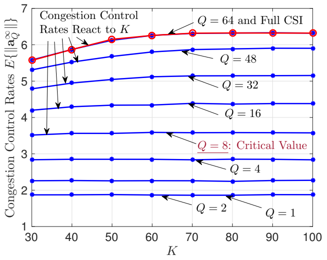

Fig. 8. The congestion control rates with respect to K for $Q = 1, 2, 4, 8, 16, 32, 48, 64$ bits.

Q is 8: When $Q \geq 8$, $\mathbf{a}^\infty_Q$ displays an $O(\sqrt{\epsilon})$ diminishing gap to $\mathbf{a}^*_Q$, which agrees with Theorem 3.

VI. Conclusion

In this paper, we studied the problem of hybrid-beamforming-based mmWave cellular network optimization. We first showed that the hybrid beamforming scheduling subproblem in this framework enjoys a hidden pseudoconvexity structure, which leads to simplified analog beam training design. We then characterized two phase transition phenomena in throughput and delay with respect to CSI accuracy in digital beamforming. Collectively, these results deepen our understanding of mmWave network optimization. Hybrid beamforming in mmWave networking is an exciting and under-explored research area. Our future directions include multi-cell mmWave networks with hybrid beamforming and the impacts of CSI inaccuracy on limited RF chains at the BS side.

Appendix

1) Proof of Theorem 1: We let $r_n[t]$ denote the objective function of Problem (9). Substituting (1) and (2) into $r_n[t]$ with the defined constants, we can rewrite the objective function as:

$$r_n[t] = \left(b_0 - \frac{b_1}{\theta_B[t]\theta_M[t]}\right) \sum_{k=1}^{K} \log_2\left(1 + \Delta_0 c_n^{(k)}\right), \quad (14)$$

where the term Δ_0 can be further written as:

$$\Delta_0 = \left(\frac{4\pi^2(1-\eta)^2}{\tilde{\theta}[t]} + \frac{2\pi(1-\eta)(\theta_B[t]+\theta_{\mathrm{M}}[t])}{\tilde{\theta}[t]} + \eta^2\right),$$

Here, we define $\tilde{\theta}[t] \triangleq \theta_B[t]\theta_M[t]$. Now, we claim that

$$\frac{4\pi^2(1-\eta)^2}{\tilde{\theta}[t]} \gg \frac{2\pi(1-\eta)(\theta_B[t]+\theta_M[t])}{\tilde{\theta}[t]} \quad (15)$$

is true if $\eta \ll \frac{1}{3}$. To see this, we first note that $\eta \ll \frac{1}{3}$ implies $4\pi \ll \frac{2\pi(1-\eta)}{\eta}$. Also, since $\theta_B[t], \theta_M[t] \in (0, 2\pi]$, we have $\theta_B[t] + \theta_M[t] \leq 4\pi \ll \frac{2\pi(1-\eta)}{\eta}$, which implies (15). Hence, it follows that $(\Delta) \approx (\frac{4\pi^2}{\tilde{\theta}[t]} + \eta^2)$, which further implies:

$$r_n[t] = (14) \approx \left(b_0 - \frac{b_1}{\tilde{\theta}[t]}\right) \sum_{k=1}^{K} \log_2\left(1 + \frac{4\pi^2 c_n^{(k)}}{\tilde{\theta}[t]}\right),$$

i.e., the objective function in (10). This completes the proof.

Next, we prove pseudoconvexity of Problem (10). We let $f(\tilde{\theta}[t])$ denote the *negative* objective function and our goal is to show that $f(\tilde{\theta}[t])$ is pseudoconvex, i.e., for any $\tilde{\theta}_1[t]$ and $\tilde{\theta}_2[t]$ in the feasible interval, if$f'(\tilde{\theta}_1[t])(\tilde{\theta}_2[t] - \tilde{\theta}_1[t]) \geq 0$, we must also have $f'(\tilde{\theta}_2[t])(\tilde{\theta}_2[t] - \tilde{\theta}_1[t]) \geq 0$.

First, let us consider the case where $\tilde{\theta}_2[t] \geq \tilde{\theta}_1[t]$. Then, showing $f'(\tilde{\theta}_2[t])(\tilde{\theta}_2[t] - \tilde{\theta}_1[t]) \geq 0$ is equivalent to showing $f'(\tilde{\theta}_2[t]) \geq 0$. Note that, in this case, the condition $f'(\tilde{\theta}_1[t])(\tilde{\theta}_2[t] - \tilde{\theta}_1[t]) \geq 0$ simply means $f'(\tilde{\theta}_1[t]) \geq 0$, i.e.,

$$f'(\tilde{\theta}_1[t]) = \sum_{k=1}^{K} \frac{1}{\tilde{\theta}_1^2} \Bigg[\underbrace{\frac{4\pi^2 c_n^{(k)}}{\ln(2)} \cdot \frac{b_0\tilde{\theta}_1[t] - b_1}{\tilde{\theta}_1[t] + 4\pi^2 c_n^{(k)}}}_{(P1)} \underbrace{- b_1 \log_2\left(1 + \frac{4\pi^2 c_n^{(k)}}{\tilde{\theta}_1[t]}\right)}_{(P2)} \Bigg] \geq 0. \quad (16)$$

It is obvious that the term $(P2)$ is an increasing function of $\tilde{\theta}[t]$. Now, consider the fractional term $\frac{b_0\tilde{\theta}_1[t]-b_1}{\tilde{\theta}_1[t]+4\pi^2 c_n^{(k)}}$ in $(P1)$, which is negative-valued according to the definitions of b_0, b_1, and the feasible interval. Also, from the definition of b_0, we have $b_0 < 1$, implying that the absolute value of the nominator is increasing at a slower rate than that of the denominator. This means that $(P1)$ is also an increasing function of $\tilde{\theta}[t]$. Hence, $f'(\tilde{\theta}[t])$ is increasing since both $(P1)$ and $(P2)$ are increasing. As a result, $f'(\tilde{\theta}_1[t]) \geq 0$ and $\tilde{\theta}_2[t] \geq \tilde{\theta}_1[t]$ imply $f'(\tilde{\theta}_2[t]) \geq 0$ and thus the case of $\tilde{\theta}_2[t] \geq \tilde{\theta}_1[t]$ is proved. The other case where $\tilde{\theta}_2[t] \leq \tilde{\theta}_1[t]$ can also be proved similarly and we omit the details in here for brevity. This completes the proof.

2) Proofs Sketch of Theorems 2 and 3: Due to space limitation, we provide a proof sketch in this paper and refer readers to [23] for further details. To prove Theorem 2, we use an α-parameterized quadratic Lyapunov function: $V_\alpha(\mathbf{q}[t]) = \frac{\epsilon^\alpha}{2}\|\mathbf{q}[t] - \mathbf{q}^*_{Q,(K)}\|^2$, where the parameter $\alpha \in \{0, 1\}$ and its value will be specified later. We first bound the conditional mean Lyapunov drift as follows:

$$\mathbb{E}\{V_\alpha(\mathbf{q}[t+1]) - V_\alpha(\mathbf{q}[t]) | \mathbf{q}[t]\} \overset{(c)}{\leq} \epsilon^\alpha \Big[-\frac{\epsilon}{\Phi} \big\|\mathbf{q}[t] - \mathbf{q}^*_{Q,(\epsilon)}\big\|^2 + D_0 \Big] + \epsilon^\alpha \mathbb{E}\Big\{(\mathbf{q}[t])^\top (\mathbf{s}^* - \mathbf{s}_Q[t]) \big| \mathbf{q}[t]\Big\}, \quad (17)$$

where $D_0 \triangleq \frac{N}{2}(A_2^{\max} + (s^{\max})^2)$ and $\mathbf{s}^* \triangleq \lim_{Q\to\infty} \mathbf{s}_Q^*$. Then, calculating the T-step Lyapunov drive, arranging terms, dividing both sides by T, and letting $T \to \infty$ yields:

$$0 \leq J + \sum_{\mathbf{q}\in\mathbb{Z}_+^N} \pi_{\mathbf{q}}^{\infty}(\mathbf{q})^{\top}(\mathbf{s}^* - \mathbf{s}_Q^{\infty}) = J + \mathbb{E}\big\{(\mathbf{q}^{\infty})^{\top}(\mathbf{s}^* - \mathbf{s}_Q^{\infty}\big\}, \quad (18)$$

where $J \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\mathbf{q}\in\mathcal{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0])\{\epsilon^{\alpha}$ $[\frac{-\epsilon}{\Phi}\|\mathbf{q}[t] - \mathbf{q}_{(\epsilon)}^*\|^2 + D_0]\}$; $\mathbf{s}_Q^{\infty} \triangleq \arg\max_{\mathbf{x}\in\mathcal{C}_{\mathbf{H}[\infty]|\widehat{\mathbf{H}}[\infty]}} (\mathbf{q}^{\infty})^{\top}\mathbf{x}$ represents the steady-state service rates with Q-bit CSI. Next, we consider two cases based on the positivity of $\mathbb{E}\big\{(\mathbf{q}^{\infty})^{\top}(\mathbf{s}^* - \mathbf{s}_Q^{\infty})\big\}$ as follows:

Case I): $Q \geq Q^{\sharp}$ such that $\mathbb{E}\big\{(\mathbf{q}^{\infty})^{\top}(\mathbf{s}^* - \mathbf{s}_Q^{\infty})\big\} \leq 0$: In this case, it follows from (18) that

$$0 \leq \lim_{T\to\infty} \frac{1}{T} \sum\nolimits_{t=0}^{T-1} \sum\nolimits_{\mathbf{q}\in\mathcal{Z}_+^N} \Pr(\mathbf{q}[t] = \mathbf{q}|\mathbf{q}[0]) \times \\ \Big\{\epsilon^{\alpha}\Big[-\frac{\epsilon}{\Phi}\big\|\mathbf{q}[t] - \mathbf{q}_{Q,(\epsilon)}^*\big\|^2 + D_0\Big]\Big\}. \quad (19)$$

We now consider the term in the second line in (19) by setting $\alpha = 0$. Then, by using similar techniques to prove the Pake's lemma [24], we can show that

$$\mathbb{E}\{\|\mathbf{q}^{\infty} - \mathbf{q}_{Q,(\epsilon)}^*\|\} \leq \Big(\beta + \frac{\eta}{\delta}\Big)\frac{1}{\sqrt{\epsilon}} = O(1/\sqrt{\epsilon}). \quad (20)$$

Case II): $Q \leq Q^{\sharp}$ such that $\mathbb{E}\big\{(\mathbf{q}^{\infty})^{\top}(\mathbf{s}^* - \mathbf{s}_Q^{\infty})\big\} > 0$: In this case, we set $\alpha = 1$. It thus follows from (17) that:

$$\mathbb{E}\big\{\Delta V_1(\mathbf{q}[t])|\mathbf{q}[t]\big\} \leq -\frac{\epsilon^2}{\Phi}\Big\|\mathbf{q}[t] - \mathbf{q}_{Q,(\epsilon)}^*\Big\|^2 + \\ \epsilon\|\mathbf{q}[t] - \mathbf{q}_{Q,(\epsilon)}^*)^{\top}\|C_{(Q)} + \epsilon D_0, \quad (21)$$

where $C_{(Q)} \triangleq \max_{\mathbf{q}:\|\mathbf{q}\|=1} \mathbb{E}\{\|\mathbf{s}_Q^* - \mathbf{s}_Q\|\mathbf{q}\}$. Again, by using the techniques to prove the Pake's lemma, we can show that

$$\mathbb{E}\{\|\mathbf{q}^{\infty} - \mathbf{q}_{Q,(\epsilon)}^*\|\} \leq \Big(\Big[(C_{(Q)}\Phi) + \\ \sqrt{(C_{(Q)}\Phi)^2 + 4D_0\Phi}\Big] + \frac{\eta}{\delta}\Big)\frac{1}{\epsilon} = O(C_{(Q)}\frac{1}{\epsilon}). \quad (22)$$

This completes the proof of Theorem 2.

3) Proof Sketch of Theorem 3: To show the results in Theorem 3, we first note that $\mathbb{E}\{a_n[t]|q_n[t]\} = \min\{U_n^{\prime -1}(\epsilon q_n[t], A^{\max})\}$ and $a_n^* = U_n^{\prime -1}(\epsilon q_n^*)$, $\forall n$. After some algebraic derivations and upper bounding, we have

$$\|\mathbf{a}_Q^{\infty} - \mathbf{a}_Q^*\| \leq \|\mathbf{a}_Q^{\infty} - \mathbf{a}_Q^*\|_1 \leq \frac{\epsilon\sqrt{N}}{\phi}\mathbb{E}\big\{\big\|\mathbf{q}^{\infty} - \mathbf{q}_{Q,(\epsilon)}^*\big\|\big\}. \quad (23)$$

Note that, in the proof of Theorem 2, we have shown the phase transition of $\mathbb{E}\big\{\big\|\mathbf{q}^{\infty} - \mathbf{q}_{Q,(\epsilon)}^*\big\|$ in (20) and (22), respectively. Thus, multiplying (20) and (22) by ϵ arrives at the result stated in Theorem 3. This completes the proof.

References

[1] Y. Zhu, Z. Zhang, Z. Marzi, C. Nelson, U. Madhow, B. Y. Zhao, and H. Zheng, "Demystifying 60 GHz outdoor picocells," in *Proc. ACM MobiCom*, Maui, HI, September 2014, pp. 5 – 16.

[2] S. Sur, V. Venkateswaran, X. Zhang, and P. Ramanathan, "60 GHz indoor networking through flexible beams: A link-level profiling," in *Proc. ACM SIGMETRICS*, Portland, OR, June 2015, pp. 71–84.

[3] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, mar 2014.

[4] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, June 2011.

[5] A. Alkhateeb, O. E. Ayach, GeertLeus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, October 2014.

[6] J. Wang et al., "Beam codebook based beamforming protocol formulti-Gbps millimeter-waveWPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, August 2009.

[7] S. Hur, T.Kim, D.Love, J. Krogmeier, T. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, October 2013.

[8] S. Han, I. Chih-Lin, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[9] X. Zhang, A. Molisch, and S. Kung, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4091–4103, Nov. 2005.

[10] V. Venkateswaran and A. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Trans. Signal Process*, vol. 58, no. 8, pp. 4131–4143, Aug. 2010.

[11] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun*, vol. 13, no. 3, pp. 1499–1513, Mar. 2013.

[12] T. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog-digital beamforming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3311–3326, May 2015.

[13] M. J. Neely, E. Modiano, and C.-P. Li, "Fairness and optimal stochastic control for heterogeneous networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 2, pp. 396–409, Apr. 2008.

[14] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1333–1344, Dec. 2007.

[15] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[16] H. Zhao, R. Mayzus, S. Sun, M. Samimi, J. K. Schulz, Y. Azar, K. Wang, G. N. Wong, F. Gutierrez, and T. S. Rappaport, "28 GHz millimeter wave cellular communication measurements for reflection and penetration loss in and around buildings in New York City," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 154–160, July 2013.

[17] A. M. Hunter, J. G. Andrews, and S. Weber, "Transmission capacity of ad hoc networks with spatial diversity," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5058–5071, Dec. 2008.

[18] J. Wildman, P. H. Nardelli, M. Latva-aho, and S. Weber, "On the joint impact of beamwidth and orientation error on throughput in wireless directional poisson networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 7072–7085, Jun. 2014.

[19] H. Shokri-Ghadikolaei, L. Gkatzikis, and C. Fischione, "Beam-searching and transmission scheduling in millimeter wave communications," in *Proc. IEEE ICC*, London, UK, Jun. 2015, pp. 1292 – 1297.

[20] E. G. Larsson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[21] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.

[22] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. New York, NY: John Wiley & Sons Inc., 2006.

[23] J. Liu and E. S. Bentley, "Understanding the impacts of hybrid beamforming on millimeter wave cellular network performances," The Ohio State University, Tech. Rep., July 2016. [Online]. Available: https://www2.ece.ohio-state.edu/~liu/mmWave_HybridBF_SCH_TR.pdf?dl=0

[24] A. G. Pakes, "Some conditions on the ergodicity and recurrence of Markov chains," *Operations Research*, vol. 17, no. 6, pp. 1058–1061, Nov. 1969.