Multi-Beam Power Allocation in Dynamic Massive MIMO Cloud Radio Access Networks

Son Dinh and Hang Liu
Dept. of Elec. Eng. & Comp. Sci.
The Catholic University of America
Washintong, DC, USA

Xianfu Chen
VTT Technical Research Centre of Finland
Oulu, Finland

Feng Ouyang
Johns Hopkins Applied Physics Laboratory
Laurel, MD, USA

ICC 2021 - IEEE International Conference on Communications | 978-1-7281-7122-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICC42927.2021.9500838

***Abstract*—In this paper, we investigate a cloud radio access network (Cloud-RAN) in which both fronthaul and radio access links use massive MIMO millimeter-wave (mmWave) transmissions. Such an all-mmWave Cloud-RAN architecture provides a flexible and cost-effective means for deployment of next-generation (5G and beyond) cellular networks to meet the demands of fast-growing mobile data traffic. Nevertheless, the design of transmit power allocation schemes for multiple massive MIMO beams on both the fronthaul and access links in a mmWave Cloud-RAN is challenging. In particular, the traffic and wireless channel states of multiple mobile terminals (MTs) change over time, while their statistics may not be known a priori. We formulate the joint fronthaul-access link massive MIMO beam power allocation problem as a Markov decision process (MDP) with an objective to optimize the long-term quality of service to all the MTs in a Cloud-RAN. A reinforcement learning algorithm is designed, which learns the optimal beam power allocation policy on the fly and adapts to the network dynamics. Further, by leveraging the structure of the underlying problem, a post-decision state is introduced and a function decomposition technique is developed to reduce the search space during the learning process. The evaluation results validate the convergence of our proposed scheme and demonstrate its superior performance over the state-of-the-art baselines.**

***Keywords*—Cloud radio access network (Cloud-RAN), massive MIMO, beam power allocation, millimeter-wave communications, reinforcement learning.**

I. Introduction

Millimeter wave (mmWave) communications and massive multiple-input multiple-output (MMIMO) are the key technologies for next-generation (5G and beyond) cellular networks to solve spectrum scarcity problem and meet the demands of fast-growing mobile data traffic. On the other hand, cloud radio access network (Cloud-RAN) [1, 2] has been considered as an innovative cellular architecture to improve network performance and significantly reduce capital and operating costs of wireless service providers. With Cloud-RAN, baseband signal processing and control plane functions are centralized in the cloud and realized with a generic computing platform. For downlink transmissions, the processed baseband signals are sent over a long fronthaul from the central unit (CU) in the cloud to remote radio heads (RRH) that perform digital-to-analog conversion, RF signal processing and amplification before sending out to mobile terminals (MTs) via antennas. A reverse process is performed for uplink signal receiving. Optical fiber links are often used for the fronthaul transmission between the CU and RRHs, but they are costly, and it may be impossible to install fiber for some RRH locations. mmWave links can be an cost-effective alternative for the fronthaul.

We consider an all-mmWave Cloud-RAN in this paper with both the fronthaul and the radio access links operating in mmWave bands. A mmWave Cloud-RAN raises a set of new technical challenges such as high pathloss and unfavorable channel characteristics of mmWave frequency. MMIMO [3, 4] is a crucial technology to enable 5G communications in mmWave bands and achieve the necessary service coverage and network capacity. In a MMIMO Cloud-RAN systems, the number of antennas at the CU and RRH is large. Advanced signal processing techniques can be used to leverage the large number of antennas and concurrently generate multiple directional signal beams, each focusing a great amount of radiated signal energy on an intended receiver. MMIMO beamforming enables the CU/RRH to simultaneously transmit/receive multiple directional signal beams on the same frequency channel with high array beamforming gains. The short wavelength of mmWave frequency allows implementing antenna arrays with a large number of elements in a reasonable size, which can form highly directional beams to combat high propagation loss and increase channel throughput.

One of the fundamental challenges in MMIMO beamforming is how to allocate the transmit power to multiple beams, namely, determining the signal amplitude and throughput of each beam. Conventionally, beamforming algorithms aim to maximize the total throughput of multiple data beams [5, 6], given an allowed value of the total transmit power. However, in a Cloud-RAN, MTs connect to different RRHs. They have heterogeneous and dynamic traffic demands and channel conditions on the fronthaul and access links. Specifically, the packet arrivals and channel states vary over time, while their statistics may not be known a priori. The traditional power allocation schemes that tend to achieve one-shot deterministic optimization based on the current channel conditions cannot predict future network dynamics and achieve overall optimal system performance on a long timescale. Casting the MMIMO beam power allocation as a dynamic stochastic optimization problem is more reasonable and compelling.

In this paper, we propose an optimization scheme that jointly allocates the transmit power of the multiple beams on both the fronthaul and access links in a MMIMO mmWave Cloud-RAN under time-varying stochastic traffic and channel conditions. The MMIMO beam power allocation problem is formulated as a Markov decision process (MDP) with the objective to optimize the quality of service (QoS) of the MTs in terms of queueing

delay and packet drop. An online reinforcement learning algorithm is designed which adapts to the network dynamics for a long-term expected optimal system performance. In addition, the algorithm does not assume to know the traffic statistics and wireless channel state transition probabilities a priori. Further, by leveraging the structure of the underlying problem, a post-decision state is introduced and a function decomposition technique is developed to reduce the search space and computation complexity in the reinforcement learning. To the best knowledge of the authors, this is the first work to solve the joint multi-beam power allocation optimization problem over two wireless MMIMO hops in a Cloud-RAN by employing a reinforcement learning-based approach.

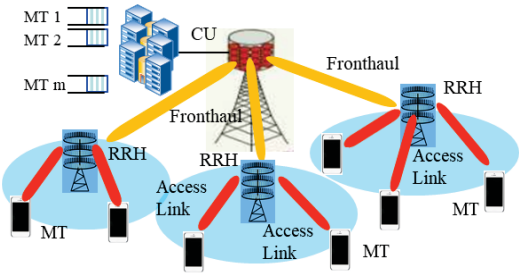


Fig. 1. An all-mmWave massive MIMO cloud-RAN.

II. System Modle and Massive MIMO Transmission

As illustrated in Fig. 1, we consider a MMIMO all-mmWave Cloud-RAN, which is composed of three main components: i) a central unit (CU) that performs baseband processing and RAN control such as MMIMO beam power allocation, data transmission scheduling, and other control plane functions; ii) the mmWave fronthaul through which the CU communicates with the RRHs, and iii) a set of RRHs serving the MTs over mmWave access links. We assume that a wireless service provider (WSP) has a slice of virtualized Cloud-RAN resources including virtual CU and RRH resources, a frequency channel for MMIMO fronthaul transmissions, and a frequency channel for MMIMO access link transmissions. The fronthaul and access links operate on different mmWave frequency channels. The focus of this work is to jointly optimize MMIMO transmissions of the multiple beams from CU to RRHs and the multiple beams from each RRH to MTs. Our analysis hereinafter concentrates on MMIMO beam power allocation for the downlink under stochastic traffic arrivals and channel states. The analysis can be extended to the uplink transmissions in a reverse way.

Considering the fronthaul downlink transmissions first, the CU is equipped with N_c antennas and simultaneously transmits K MMIMO beams to a set $\mathcal{K} = \{1, \dots k \dots, K\}$ of RRHs on the fronthaul mmWave channel, with each beam being allocated to one RRH. MMIMO is based on the fact that for large N_c, simple receiver architectures yield near-optimal performance. For ease of explanation with an analytical expression, we assume that each RRH k receives its signal beam from the CU using a single RF chain. The mmWave channel vector can be modelled as [7],

$$\mathbf{h}_{c,k} = \frac{\beta_{c,k}}{1+b_{c,k}^{e}}[1,\ e^{-j\pi\varphi_{c,k}}, \dots,\ e^{-j\pi(N_c-1)\varphi_{c,k}}], \quad (1)$$

where $b_{c,k}$ is the distance between the CU and RRH k, e is the pathloss exponent, $\varphi_{c,k}$ is the normalized direction, and $\beta_{c,k}$ is the fading attenuation coefficient. Note that we do not assume that the distribution of $\beta_{c,k}$ is known a priori. The signal-to-interference-plus-noise ratio (SINR) is as follows if conjugate beamforming is used at the CU [8],

$$SINR_{c,k} = \frac{|\mathbf{h}_{c,k}\mathbf{h}_{c,k}^{\mathrm{H}}|^2\alpha_{c,k}P_c}{\sum_{j\in\mathcal{K}\setminus k}|\mathbf{h}_{c,k}\mathbf{h}_{c,j}^{\mathrm{H}}|^2\alpha_{c,j}P_c+\sigma_{c,k}^2}, \quad (2)$$

where P_c is the total transmitted power of the CU, and $\sigma_{c,k}^2$ is the noise power. $\sum_{j\in\mathcal{K}\setminus k}|\mathbf{h}_{c,k}\mathbf{h}_{c,j}^{\mathrm{H}}|^2\alpha_{c,j}P_c$ is the interference of other RRHs' beams. $\alpha_{c,k}$, $k \in \mathcal{K}$ is the power allocation coefficients that is subject to the constraint,

$$\sum_{k=1}^{K}\alpha_{c,k} \le 1. \quad (3)$$

The achievable data rate between the CU and RRH k can then be expressed as,

$$R_{c,k} = B_f \log_2(1 + SINR_{c,k}), \quad (4)$$

where B_f is the bandwidth of the fronthaul channel. We can see from Eq. (4) that the achievable fronthaul data rate to RRH k depends on the transmit power allocated to its signal beam and the channel state between the CU and RRH.

An RRH $k \in \mathcal{K}$ simultaneously sends multiple downlink signal beams to a set $\mathcal{M}_k = \{1, \dots m \dots M_k\}$ of associated MTs, with each beam being allocated to a MT using MMIMO on the mmWave access channel. Similarly, the mmWave channel vector between RRH k and MT m can be modelled as

$$\mathbf{h}_{k,m} = \frac{\beta_{k,m}}{1+b_{k,m}^{e}}[1,\ e^{-j\pi\varphi_{k,m}}, \dots,\ e^{-j\pi(N_{k,t}-1)\varphi_{k,m}}], \quad (5)$$

where $b_{k,m}$ is the distance from RRH k to MT m, $\beta_{k,m}$ is the fading attenuation coefficient, $\varphi_{k,m}$ is the normalized direction, and $N_{k,t}$ is the number of downlink transmit antennas at RRH k. The distribution of $\beta_{k,m}$ is also unknown a priori. MT m treats the signal beams of other MTs sent from the RRHs, $j \in \mathcal{K}$ as noise and decodes its own message with the following SINR:

$$SINR_{k,m} = \frac{|\mathbf{h}_{k,m}\mathbf{h}_{k,m}^{\mathrm{H}}|^2\alpha_{k,m}P_k}{\sum_{j\in\mathcal{K},i\in M_j\setminus m}|\mathbf{h}_{j,i}\mathbf{h}_{k,m}^{\mathrm{H}}|^2\alpha_{j,i}P_j+\sigma_{k,m}^2} \quad (6)$$

where P_k is the total transmitted power of RRH k, and $\sigma_{k,m}^2$ is the noise power. $\alpha_{k,m}$ is the beam power allocation coefficient for RRH k to MT m that is subject to the constraint,

$$\sum_{m=1}^{M_k}\alpha_{k,m} \le 1. \quad (7)$$

The achievable data rate from RRH k to MT m can then be expressed as,

$$R_{k,m} = B_a \log_2(1 + SINR_{k,m}), \quad (8)$$

where B_a is the bandwidth of the access channel.

III. Problem Formulation and Reinforcement Learning Solution

A. Problem Formulation

We consider that the time horizon is discretized into decision epochs, each of which is of a fixed duration δ. Let A_m^t be the number of new packets randomly arrived at the CU in epoch t for MT m. The packet arrival process is assumed to be independent among the MTs and the distribution of A_m^t is not known a priori. There are a transmission queue maintained for each MT at the CU and another queue at the RRH for each associated MT. Let $\mathcal{M} = \{1, \dots m \dots, M\}$ denote the whole set of MTs. Further, let $x_{k,m} = 1$ if m is associated with k, i.e. $m \in$

$\mathcal{M}_k$, and $x_{k,m} = 0$ otherwise. $Q_{c,m}^t$ and $Q_{k,m}^t$ denote the lengths of the queues for MT m at the CU and RRH k, respectively, at the beginning of epoch t. Their maximal queue sizes are $Q_c^{(\max)}$ and $Q_k^{(\max)}$, at the CU and RRH, respectively. The queue evolution of MT m at the CU and RRH can be expressed, respectively, as

$$Q_{c,m}^{t+1} = \max\{\min\left\{Q_{c,m}^t + A_m^t - C_{c,k,m}^t, Q_c^{(\max)}\right\}, 0\}, \quad (9)$$

$$Q_{k,m}^{t+1} = \max\{\min\left\{Q_{k,m}^t + x_{k,m}(C_{c,k,m}^t - C_{k,m}^t), Q_k^{(\max)}\right\}, 0\}, \quad (10)$$

where $C_{c,k,m}^t$ and $C_{k,m}^t$ are the number of packets for MT m, $\forall m \in \mathcal{M}_k$ that are transmitted from the CU to RRH k and from RRH k to MT m, respectively, in epoch t. $C_{c,k,m}^t = \min\{\delta R_{c,k,m}^t/L,\ Q_{c,m}^t + A_m^t\}$ with $R_{c,k,m}^t$ to be the data rate allocated for MT m transmission in the fronthaul from the CU to RRH k in epoch t. L is the packet length. $C_{k,m}^t = \min\{\delta R_{k,m}^t/L, x_{k,m}(Q_{k,m}^t + C_{c,k,m}^t)\}$ with $R_{k,m}^t$ to be the data rate from RRH k to MT m. Considering the set of MTs associated with RRH k, $\forall m \in \mathcal{M}_k$ (i. e. $x_{k,m} = 1$), the total number of packets transmitted from CU to RRH k, $C_{c,k}^t$, should be the sum of the packets transmitted for all its associated MTs, that is,

$$C_{c,k}^t = \sum_{m\in\mathcal{M}_k} C_{c,k,m}^t \quad (11)$$

Let $\mathbf{Q}_c^t = \{Q_{c,m}^t : m \in \mathcal{M}\}$ and $\mathbf{Q}_a^t = \{Q_{k,m}^t : k \in \mathcal{K}, m \in \mathcal{M}_k\}$ denote the queue states at the CU and RRHs for the MTs. In addition, let $\mathbf{H}_c^t = \{\mathbf{h}_{c,1}^t, \dots \mathbf{h}_{c,k}^t \dots, \mathbf{h}_{c,K}^t\}$ denote the mmWave channel states of the fronthaul links between the CU and RRH $k \in \mathcal{K}$ at epoch t, and $\mathbf{H}_a^t = \{\mathbf{H}_1^t, \dots \mathbf{H}_k^t \dots, \mathbf{H}_K^t\}$ with $\mathbf{H}_k^t = \{\mathbf{h}_{k,1}^t, \dots \mathbf{h}_{k,m}^t \dots, \mathbf{h}_{k,M}^t\}^T$ be the mmWave channel states of access links between RRH $k \in \mathcal{K}$ and MTs, $m \in \mathcal{M}$. The global Cloud-RAN network state at each decision epoch t encapsulates all the queue and channel states for the fronthaul and access links, $\boldsymbol{\chi}^t = \{\mathbf{Q}_c^t, \mathbf{Q}_a^t, \mathbf{H}_c^t, \mathbf{H}_a^t\} = \{\boldsymbol{\chi}_m^t : m \in \mathcal{M}\}$ where $\boldsymbol{\chi}_m^t = \{Q_{c,m}^t, Q_{k,m}^t, \mathbf{h}_{c,k}^t, \mathbf{h}_{k,m}^t\}$ characterizes the local network state for a MT m over the two-hop wireless transmissions between the CU and m via a RRH k.

For a Cloud-RAN state $\boldsymbol{\chi}^t$ at the beginning of a decision epoch t, a MMIMO beam power allocation, $\boldsymbol{\alpha}(\boldsymbol{\chi}^t) = \{\alpha_{c,k}(\boldsymbol{\chi}^t), \alpha_{k,m}(\boldsymbol{\chi}^t) : k \in \mathcal{K}, m \in \mathcal{M}_k\}$ for the fronthaul and access link transmissions is made. Note that the data rate $R_{c,k}^t$ from the CU to RRH k and $R_{k,m}^t$ from RRH k to MT m depends on the fronthaul and access MMIMO beam power allocation as well as the channel conditions as shown in (2), (4), (6) and (8). The number of packets transmitted, i.e. transmission scheduling for different MTs over the fronthaul and access links, $C_{c,k}^t$, $C_{c,k,m}^t$, and $C_{k,m}^t$ in epoch t are thus determined by the MMIMO beam power allocation, $\boldsymbol{\alpha}(\boldsymbol{\chi}^t)$, given the channel conditions.

The packets will experience a network delay. Let $d(\boldsymbol{\chi}^t, \boldsymbol{\alpha}(\boldsymbol{\chi}^t)) = \sum_{m\in\mathcal{M}} d_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t))/M$ denote the average delay of M MTs at epoch t with $d_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t))$ being the delay of MT m, which is a function of the network state and MMIMO beam power allocation. Moreover, we can schedule the packet transmission from the CU to RRHs to ensure there is no buffer overflow at the RRHs. However, due to the limited buffer size at the CU and random packet arrivals, packets may be dropped at the CU. The number of packets dropped in decision epoch t for MT m due to buffer overflow can be expressed as $f_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t)) = \sum_{A_m^t} \Pr(A_m^t) \max\{0, (Q_{c,m}^t + A_m^t - C_{c,k,m}^t - Q_c^{(\max)}\}$ where $\Pr(A_m^t)$ is the probability that A_m^t packets of MT m arrives in epoch t, and the average packet drop number of M MTs per epoch is $f(\boldsymbol{\chi}^t, \boldsymbol{\alpha}(\boldsymbol{\chi}^t)) = \sum_{m\in\mathcal{M}} f_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t)/M$. To ensure the QoS, we consider there is a maximal tolerance threshold, $d^{(\max)}$ for the delay, i.e. $d_m \le d^{(\max)}$. Correspondingly, let $f^{(\max)}$ be the maximal tolerance threshold for the packet drop, i.e. $f_m \le f^{(\max)}$. We define the instantaneous Cloud-RAN network cost function under the state $\boldsymbol{\chi}^t$ and the MMIMO beam power allocation decision $\boldsymbol{\alpha}(\boldsymbol{\chi}^t)$ at decision epoch t as,

$$U(\boldsymbol{\chi}^t, \boldsymbol{\alpha}(\boldsymbol{\chi}^t)) = \sum_{m\in\mathcal{M}}[d_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t))/d^{(max)} + \omega f_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t))/f^{(max)}]/M$$
$$= \sum_{m\in\mathcal{M}}[\omega_d d_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t)) + \omega_f f_m(\boldsymbol{\chi}_m^t, \alpha_k(\boldsymbol{\chi}^t))], \quad (12)$$

where ω is a weight that trades off the importance of the delay and packet drop, and $\omega_d = 1/Md^{(max)}$ and $\omega_f = \omega/Mf^{(max)}$.

Stochastic and time-varying packet arrivals and wireless channel states bring challenges, and traditional one-shot optimization schemes, e.g. classical convex optimization method in an epoch [9, 10], cannot capture the network dynamics and achieve the stable and optimal performance on a longer timescale. Therefore, we develop a MMIMO beam power allocation policy $\boldsymbol{\alpha}$ that minimizes the expected long-term network cost. We define the discounted expected value of the Cloud-RAN network cost $U(\boldsymbol{\chi}^t, \boldsymbol{\alpha}(\boldsymbol{\chi}^t))$ over a sequence of network states $\boldsymbol{\chi}^t$ as follows [11],

$$V(\boldsymbol{\chi}, \boldsymbol{\alpha}) = \mathrm{E}\,[(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}U(\boldsymbol{\chi}^t, \boldsymbol{\alpha}(\boldsymbol{\chi}^t))|\,\boldsymbol{\chi}^1], \quad (13)$$

where $\gamma \in [0, 1)$ is a discount factor that discounts the network cost in the future, and $(\gamma)^{t-1}$ denotes the discount to the (t −1)-th power. $\boldsymbol{\chi}^1$ is the initial network state. $V(\boldsymbol{\chi}, \boldsymbol{\alpha})$ is also termed as the state value function of the Cloud-RAN in state $\boldsymbol{\chi}$ under MMIMO beam power allocation policy $\boldsymbol{\alpha}$. The CU strategically decides the MMIMO beam power allocation based on $\boldsymbol{\alpha}$ after observing the network state $\boldsymbol{\chi}^t$ at the beginning of a decision epoch t. The objective is to design an optimal MMIMO beam power allocation policy $\boldsymbol{\alpha}^*$ that minimizes the expected discounted long-term network cost, subject to the power constraints at the CU and RRHs, QoS constraints, and the data rate constraint (11), that is,

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, V(\boldsymbol{\chi}, \boldsymbol{\alpha}); \quad (14)$$

$$\text{s.t.}\quad \sum_{k\in\mathcal{K}} \alpha_{c,k} \le 1;\quad \sum_{m\in M_k} \alpha_{k,m} \le 1;\ d_m \le f^{(\max)}; f_m \le d^{(\max)};\ C_{c,k} = \sum_{m\in\mathcal{M}_k} C_{c,k,m} \quad (15)$$

$V^*(\boldsymbol{\chi}) = V(\boldsymbol{\chi}, \boldsymbol{\alpha}^*)$ is the optimal state value function. We consider that the Cloud-RAN network state in the subsequent epoch depends only on the state attained in the present epoch and the power allocation policy $\boldsymbol{\alpha}$. The random process of Cloud-RAN state $\boldsymbol{\chi}^t$ can be modelled as a controlled Markov process across the epochs. The stochastic MMIMO power allocation optimization problem in (14) is thus a MDP with the discounted cost criterion. For a MDP, the optimal MMIMO power allocation policy that achieves the minimum state value function can be obtained by solving a Bellman's optimality equation as follows [11, 12],

$$V^*(\boldsymbol{\chi}) = \min_{\boldsymbol{\alpha}}\{(1-\gamma)\,U(\boldsymbol{\chi}, \boldsymbol{\alpha}(\boldsymbol{\chi})) + \gamma\sum_{\boldsymbol{\chi}'} \Pr\{\boldsymbol{\chi}'|\boldsymbol{\chi}, \boldsymbol{\alpha}(\boldsymbol{\chi})\}V^*(\boldsymbol{\chi}')\}, \quad (16)$$

where $\boldsymbol{\chi}' = \{\mathbf{Q}_c', \mathbf{Q}_a', \mathbf{H}_c', \mathbf{H}_a'\}$ is the Cloud-RAN state in the subsequent epoch, and $\Pr\{\boldsymbol{\chi}'|\boldsymbol{\chi}, \boldsymbol{\alpha}(\boldsymbol{\chi})\}$ represents the state transition probability that produces the next state $\boldsymbol{\chi}'$ after making the MMIMO beam power allocation $\boldsymbol{\alpha}(\boldsymbol{\chi})$ in state $\boldsymbol{\chi}$. Traditional approaches to solve (16) are based on value iteration, policy iteration, and dynamic programming [13, 14, 15]. However, these methods require full knowledge of the network state transition probabilities and packet arrival statistics that cannot be known beforehand for our problem. In the following section, we will simplify the problem and develop a reinforcement learning solution without assumption of knowing the network statistics in advance.

B. Problem Simplification

The MMIMO beam power optimization problem in (16) is complex because the Cloud-RAN network state space with multiple RRHs and MTs as well as two-hop wireless transmissions is very large. The action space is also large thanks to power allocation for multiple fronthaul and access link beams. To solve it, we first simplify the problem by introducing a post-decision state and reducing the number of system states through decomposition. Then we propose an algorithm to obtain the optimal MMIMO beam power allocation policy with no requirement for prior knowledge of random packet arrival statistics and network state transitions by employing online reinforcement learning [16, 17].

First, we define an intermediate state called post-decision state for each decision epoch. A decision epoch is considered consisting of three phases, MMIMO beam power allocation decision at the beginning of an epoch, packet transmissions over the fronthaul and access links according to the allocated transmit power for each MMIMO beam, and then new packet arrivals at the end of an epoch. The post-decision state for each epoch is the state after the CU and RRHs finishes their data transmissions before the new packet arrivals. Note that the three phases and the post-decision state are used to derive the optimal MMIMO beam power allocation. In practice, the packets may arrive at any time, and the CU and RRHs can send the packets during the whole epoch.

At the current epoch, the post-decision state is defined as $\widetilde{\boldsymbol{\chi}} = (\widetilde{\mathbf{Q}}_c, \widetilde{\mathbf{Q}}_a, \widetilde{\mathbf{H}}_c, \widetilde{\mathbf{H}}_a)$, where the wireless channel states of the post-decision will remain the same as those at the beginning of the epoch, that is, $\widetilde{\mathbf{H}}_\mathbf{c} = \{\tilde{\mathbf{h}}_{c,k} : k \in \mathcal{K}\}$ with $\tilde{\mathbf{h}}_{\mathbf{c},\mathbf{k}} = \mathbf{h}_{\mathbf{c},\mathbf{k}}$, and $\widetilde{\mathbf{H}}_\mathbf{a} = \{\tilde{\mathbf{h}}_{\mathbf{k},\mathbf{m}} : k \in \mathcal{K}, m \in \mathcal{M}\}$ with $\tilde{\mathbf{h}}_{\mathbf{k},\mathbf{m}} = \mathbf{h}_{\mathbf{k},\mathbf{m}}$. They are independent of the beam power allocation decision. The queue state of post-decision is $\widetilde{\mathbf{Q}}_\mathbf{c} = \{\tilde{Q}_{c,m} : m \in \mathcal{M}\}$ with $\tilde{Q}_{c,m} = \max\{Q_{c,m} - C_{c,k,m}, 0\}, k \in \mathcal{K}, m \in \mathcal{M}_k$ and $\widetilde{\mathbf{Q}}_\mathbf{a} = \{\tilde{Q}_{k,m} : k \in \mathcal{K}, m \in \mathcal{M}_k\}$ with $\tilde{Q}_{k,m} = \max\{Q_{k,m} + C_{c,k,m} - C_{k,m}, 0\}$. The probability of Cloud-RAN network state transition from $\boldsymbol{\chi}$ to $\boldsymbol{\chi}'$ can then be expressed as,

$$\Pr\{\boldsymbol{\chi}'|\boldsymbol{\chi}, \boldsymbol{\alpha}(\boldsymbol{\chi})\} = \Pr\{\boldsymbol{\chi}'|\widetilde{\boldsymbol{\chi}}\}\Pr\{\widetilde{\boldsymbol{\chi}}|\boldsymbol{\chi}, \boldsymbol{\alpha}(\boldsymbol{\chi})\} = \prod_{k\in\mathcal{K}, m\in\mathcal{M}} \Pr\{\mathbf{h}_{c,k}'|\mathbf{h}_{c,k}\}\Pr\{\mathbf{h}_{k,m}'|\mathbf{h}_{k,m}\}\Pr\{A_m\}, \quad (17)$$

where $\Pr\{\widetilde{\boldsymbol{\chi}}|\boldsymbol{\chi}, \boldsymbol{\alpha}(\boldsymbol{\chi})\} = 1$ and $A_m = Q'_{c,m} - \tilde{Q}_{c,m}$. We can control packet transmission to ensure that no packet drop occurs in the transition to the post-decision state, i.e. the packet drop due to buffer overflow may only happen when the new packets arrive. We then factor the cloud-RAN cost function U in (12) into two parts, which correspond to delay d_m and packet drop f_m. The optimal state value function satisfying (16) can hence be rewritten by,

$$V^*(\boldsymbol{\chi}) = \min_{\boldsymbol{\alpha}}\{(1-\gamma)\textstyle\sum_{m\in\mathcal{M}} \omega_d d_m(\boldsymbol{\chi}_m, \boldsymbol{\alpha}(\boldsymbol{\chi})) + \tilde{V}^*(\widetilde{\boldsymbol{\chi}})\}, \quad (18)$$

where $\tilde{V}^*(\widetilde{\boldsymbol{\chi}})$ is the optimal post-decision state value function that satisfies Bellman's optimality equation,

$$\tilde{V}^*(\widetilde{\boldsymbol{\chi}}) = (1-\gamma)\textstyle\sum_{m\in\mathcal{M}} \omega_f f_m(\boldsymbol{\chi}, \boldsymbol{\alpha}^*(\boldsymbol{\chi})) + \gamma\sum_{\boldsymbol{\chi}'} \Pr\{\boldsymbol{\chi}'|\widetilde{\boldsymbol{\chi}}\}V^*(\boldsymbol{\chi}'). \quad (19)$$

From (18), we find that the optimal state value function can be obtained from the optimal post-decision state value function by performing minimization of the delay cost function over all feasible MMIMO beam power allocation decisions. The optimal beam power allocation is thus expressed as follows, which should satisfy the constraints in (15).

$$\boldsymbol{\alpha}^* = \operatorname*{argmin}_{\boldsymbol{\alpha}}\{(1-\gamma)\textstyle\sum_{m\in\mathcal{M}} \omega_d d_m(\boldsymbol{\chi}_m, \boldsymbol{\alpha}(\boldsymbol{\chi})) + \tilde{V}^*(\widetilde{\boldsymbol{\chi}})\}. \quad (20)$$

The packet arrival statistics of MTs are independent each other. We can then decompose the optimal post-decision state value function, mathematically, that is

$$\tilde{V}^*(\widetilde{\boldsymbol{\chi}}) = \textstyle\sum_{m\in\mathcal{M}} \tilde{V}_m^*(\widetilde{\boldsymbol{\chi}}_m), \quad (21)$$

where $\widetilde{\boldsymbol{\chi}}_m = \{\widetilde{Q}_{c,k}, \widetilde{Q}_{k,m}, \tilde{\mathbf{h}}_{c,k}, \tilde{\mathbf{h}}_{k,m}\}$ characterizes the local post-decision network state for MT m over the two-hop wireless transmission links between the CU and m via a RRH k. Given the optimal control policy $\boldsymbol{\alpha}^*$, according to (19) and (21), the post-decision state value function $\tilde{V}_m^*(\widetilde{\boldsymbol{\chi}}_m)$ satisfies,

$$\tilde{V}_m^*(\widetilde{\boldsymbol{\chi}}_m) = (1-\gamma)\omega_f f_m^*(\boldsymbol{\chi}_m) + \gamma\textstyle\sum_{\boldsymbol{\chi}'_m} \Pr\{\boldsymbol{\chi}'_m|\widetilde{\boldsymbol{\chi}}_m\} V_m^*(\boldsymbol{\chi}'_m), \quad (22)$$

where $\Pr\{\boldsymbol{\chi}'_m|\widetilde{\boldsymbol{\chi}}_m\} = \Pr\{\mathbf{h}_{c,k}'|\mathbf{h}_{c,k}\}\Pr\{\mathbf{h}_{k,m}'|\mathbf{h}_{k,m}\}\Pr\{A_m\}$ from (17). The optimal state value function of MT m in the subsequent decision epoch, $V_m^*(\boldsymbol{\chi}'_m)$ can be expressed as follows according to (18) and (21),

$$V_m^*(\boldsymbol{\chi}'_m) = (1-\gamma)\omega_d d_m^*(\boldsymbol{\chi}'_m) + \tilde{V}_m^*(\widetilde{\boldsymbol{\chi}}'_m) \quad (23)$$

where $\widetilde{\boldsymbol{\chi}}'_m$ is the local post-decision states of MT m in the subsequent decision epoch.

The problem is greatly simplified through the above linear decomposition of the post-decision state value function. First, a complex post-decision Bellman's optimality equation (19) is broken into M much simpler MDPs. Furthermore, in order to derive a MMIMO beam power allocation policy based on the global cloud-RAN state $\boldsymbol{\chi} = \{\boldsymbol{\chi}_m : m \in \mathcal{M}\}$ with $\boldsymbol{\chi}_m = (Q_{c,m}, Q_{k,m}, \mathbf{h}_{c,k}, \mathbf{h}_{k,m})$, at least $\prod_{k\in\mathcal{K}, m\in\mathcal{M}_k}(|Q_{c,m}| \times |Q_{k,m}||\mathbf{h}_{c,k}||\mathbf{h}_{k,m}|)$ state values should be kept. Using linear decomposition, only $M|Q_{c,m}||Q_{k,m}||\mathbf{h}_{c,k}||\mathbf{h}_{k,m}|$ values need to be stored, significantly reducing the search space in the MMIMO beam power decision making. By replacing the post-decision state value function in (20) with (21), we can obtain an optimal power allocation policy $\boldsymbol{\alpha}^*$ under a Cloud-RAN state $\boldsymbol{\chi}$.

C. Reinforcement Learning Algorithm

As discussed above, the number of new packet arrivals at the end of an epoch as well as the channel states for the next epoch are unknown beforehand. In this case, instead of directly computing the post-decision state value functions in (22), we propose an online reinforcement learning algorithm to learn $\tilde{V}_m^*(\widetilde{\boldsymbol{\chi}}_m), m \in \mathcal{M}$ on the fly. Based on the observations of the network state $\boldsymbol{\chi}_m^t = (Q_{c,m}^t, Q_{k,m}^t, \mathbf{h}_{c,k}^t, \mathbf{h}_{k,m}^t)$, the number of packet arrival A_m^t, the decision of MMIMO beam power

allocation on the fronthaul and access links, the resulted packet drop $f_m(\boldsymbol{\chi}_m^t)$ at the current epoch t, and the resulted network state $\boldsymbol{\chi}_m^{t+1} = (Q_{c,m}^{t+1}, Q_{k,m}^{t+1}, \mathbf{h}_{c,k}^{t+1}, \mathbf{h}_{k,m}^{t+1})$ at the next epoch $t+1$, the post-decision state value function for MT m can be updated by,

$$\tilde{V}_m^{t+1}(\tilde{\boldsymbol{\chi}}_m^t) = (1-\varepsilon^t)\tilde{V}_m^t(\tilde{\boldsymbol{\chi}}_m^t) + \varepsilon^t[(1-\gamma)\omega_f f_m(\boldsymbol{\chi}_m^t) + \gamma V_m^t(\boldsymbol{\chi}_m^{t+1})] \quad (24)$$

where $\varepsilon^t \in [0,1)$ is the learning rate. The MMIMO beam power allocation $\boldsymbol{\alpha}^t = [\alpha_{c,k}^t, \alpha_{k,m}^t : k \in \mathcal{K}, m \in \mathcal{M}_k\}$ at epoch t is determined as,

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\{\sum_{m\in\mathcal{M}}(1-\gamma)\omega_d d_m(\boldsymbol{\chi}_m^t) + \tilde{V}_m^t(\tilde{\boldsymbol{\chi}}_m^t)\}$$
$$\text{s.t. } \sum_{k\in\mathcal{K}} \alpha_{c,k}^t \le 1; \sum_{m\in M_k} \alpha_{k,m}^t \le 1; d_m^t \le d^{(\max)}; d_m^t \le f^{(\max)}; \; C_{c,k}^t = \sum_{m\in\mathcal{M}_k} C_{c,k,m}^t. \quad (25)$$

The state value function of MT m at epoch $t+1$ is evaluated by,

$$V_m^t(\boldsymbol{\chi}_m^{t+1}) = (1-\gamma)w_d d_m(\boldsymbol{\chi}_m^{t+1}) + \tilde{V}_m^t(\tilde{\boldsymbol{\chi}}_m^{t+1}) \quad (26)$$

The online learning algorithm for estimating the optimal post-decision state value function and determining the optimal MMIMO beam power allocation policy is summarized in Algorithm 1.

Algorithm 1. Online Learning Algorithm for Optimal Post-Decision State Value Function

1. Initialize the post-decision state value functions $\tilde{V}_m^t(\tilde{\chi}_m^t)$, $\forall \tilde{\chi}_m^t$ and $\forall m \in \mathcal{M}$ for $t = 1$.
2. At the beginning of decision epoch t, the CU observes the network state, $\boldsymbol{\chi}^t = \{\boldsymbol{\chi}_m^t : m \in \mathcal{M}\}$ with $\boldsymbol{\chi}_m^t = (Q_{c,m}^t, Q_{k,m}^t, \mathbf{h}_{c,k}^t, \mathbf{h}_{k,m}^t)$, $k \in \mathcal{K}$, $m \in \mathcal{M}_k$ and determines the MMIMO beam power allocation for the fronthaul and access links, $\boldsymbol{\alpha}^t = \{\alpha_{c,k}^t, \alpha_{k,m}^t : k \in \mathcal{K}, m \in \mathcal{M}_k\}$ according to (25).
3. After transmitting the packets according to the above decision, the CU observes the post-decision state, $\tilde{\boldsymbol{\chi}}^t = \{\tilde{\boldsymbol{\chi}}_m^t : m \in \mathcal{M}\}$, where $\tilde{\boldsymbol{\chi}}_m^t = (\tilde{Q}_{c,k,m}^t, \tilde{Q}_{k,m}^t, \tilde{\mathbf{h}}_{c,k}^t, \tilde{\mathbf{h}}_{k,m}^t)$ with $\tilde{Q}_{c,m}^t = \max\{Q_{c,m}^t - C_{c,k,m}^t, 0\}$, $\tilde{Q}_{k,m}^t = \max\{Q_{k,m}^t + C_{c,k,m}^t - C_{k,m}^t, 0\}$, $\tilde{\mathbf{h}}_{c,k}^t = \mathbf{h}_{c,k}^t$, and $\tilde{\mathbf{h}}_{k,m}^t = \mathbf{h}_{k,m}^t$.
4. With $\boldsymbol{A}^t = \{A_m^t : m \in \mathcal{M}\}$ new packets arrived at the end of decision epoch t, the network state transits to $\boldsymbol{\chi}^{t+1} = \{\boldsymbol{\chi}_m^{t+1} : m \in \mathcal{M}\}$ where $\boldsymbol{\chi}_m^{t+1} = (\tilde{Q}_{c,m}^t + A_m^t, Q_{k,m}^{t+1}, \mathbf{h}_{c,k}^{t+1}, \mathbf{h}_{c,k}^{t+1})$ at the following epoch t+1.
5. Calculate $V_m^t(\boldsymbol{\chi}_m^{t+1})$, $\forall m \in \mathcal{M}$ according to (26) and updates the post-decision state value functions $\tilde{V}_m^{t+1}(\tilde{\chi}_m^t)$, $\forall m \in \mathcal{M}$ according to (24).
6. Decision epoch index is updated by $t \leftarrow t+1$.
7. Repeat from step 2 to 6 until a predefined stopping condition is satisfied.

IV. Numerical Results

In this section, we evaluate the performance of the proposed reinforcement learning-based MMIMO beam power allocation scheme. For the purpose of performance comparisons, the following three baseline schemes are also simulated.

1) Baseline 1: Equal Beam Power Allocation - The CU allocates equal transmit power to the MMIMO beams for the fronthaul transmissions from the CU to RRHs as well as the access links from each RRH to its associated MTs.
2) Baseline 2: Data Rate Aware - The CU allocates the power to the MMIMO beams for maximizing the total data rate in epoch t for the fronthaul transmissions from the CU to the RRHs as well as for the access link transmissions from each RRH to its associated MTs without considering the queue states of the MTs.
3) Baseline 3: One-Shot Optimization - At each epoch t, the CU allocates the power to the MMIMO beams for the fronthaul and access link transmissions to minimize the total delay of the MTs at the end of epoch t.

We simulated multiple network scenarios with different numbers and locations of RRHs and MTs, packet arrivals, transmit power, and other system parameters. Due to the page limit, we present the results for a typical setting to compare the performance of the proposed reinforcement learning-based MMIMO beam power allocation scheme to that of the baselines and gain the insights. The system parameters are set as follows, unless stated otherwise. The Cloud-RAN has three RRHs, located at 150 m from the CU along three directions. Each RRH has three associated MTs, randomly allocated in a circle with a radius of 30 m around the RRH. The total transmit power is 3 W at the CU and 1 W at each RRH. The channel bandwidth is 500 MHz for the mmWave fronthaul channel and the access link channel. The number of transmit antennas at CU and RRHs is 128, and the pathloss exponent e is 2.4. The epoch duration is 10 ms, and the packet size is 1000 bytes. In the simulations, the mmWave channel model as in [18] is adopted, where three states are characterized, namely, the line-of-sight (LOS), the non-LOS (NLOS), and the blocking. The channel states, $\mathbf{h}_{c,k}^t$ and $\mathbf{h}_{k,m}^t$ of different RRHs k and MTs m are independent and evolve according to a Markov chain model. We choose the learning rate as $\varepsilon^t = \varepsilon_0/(\log t + 1)$ with $\varepsilon_0 = 0.6$ because it yields a good balance between the convergency time and learning accuracy through our simulation experiments.

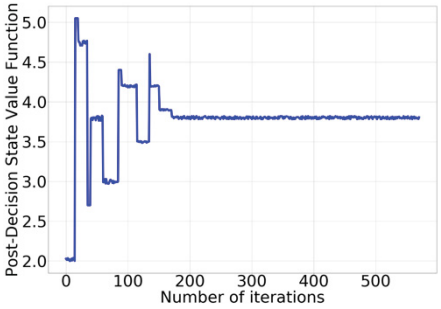
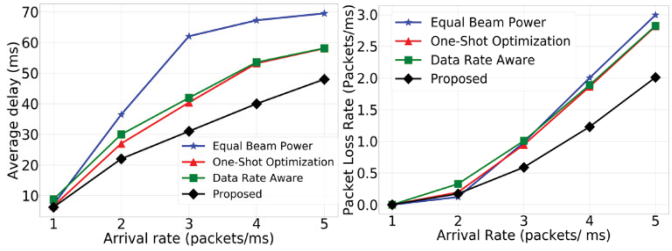


Fig. 2. Convergence of the post-decision state value function with the proposed learning process.

First, we validate the convergence property of the proposed online reinforcement learning algorithm. We consider a scenario that the number of packets arrived at the CU for each of MTs follows an independent Poisson arrival process with an average arrival rate of 2 packets/msec. As shown in Fig. 2, the algorithm first spends a short time period on learning and then converges to a stable status. The algorithm converges at a reasonable speed, around a couple hundreds of epochs.

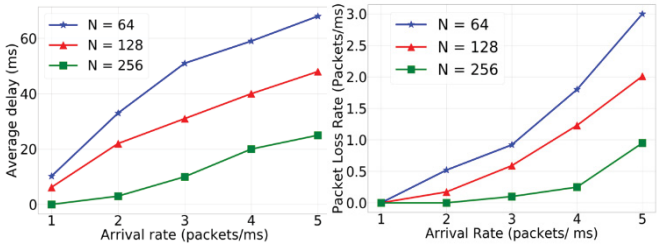
In Figs. 3 (a) and (b), we compare the average packet delay and the average packet drop rate for different schemes when the packet arrivals of the MTs follow independent Poisson arrival process and the average packet arrival rate for each of the MTs



(a) (b)

Fig. 3. (a) Average packet delay and (b) average packet loss rate vs. average packet arrival rate for different schemes.

changes. The curves indicate that our proposed reinforcement learning-based scheme outperforms all the three baseline schemes. This is because the proposed online learning scheme can capture the dynamic network state transitions. It not only considers the current packet transmission performance but also takes into account the expected long-time performance in the future when making the power allocation to the multiple MMIMO beams and determining the number of packets to send for multiple MTs. However, the baseline schemes makes shortsighted or static MMIMO beam power allocation decisions and may cause more packets to be delayed in the buffer or even dropped due to buffer overflow.



(a) (b)

Fig. 4. (a) Average packet delay and (b) average packer loss rate vs. average packet arrival rate for different numbers of MMIMO transmit antennas.

Figs. 4(a) and (b) show the effects of the number of MMIMO transmit antennas at the CU and RRHs on the average packet delay and the average packet drop rate as the proposed reinforcement learning-based MMIMO beam power allocation scheme is employed. We assume that the CU and RRHs have the same number N of MMIMO antenna elements for transmit beamforming. When the number of antennas at the CU and RRHs is larger, the antenna gains will be higher, so is the achievable data rate. Thus, the average delay and packet drop rate would be lower.

V. Conclusions

This paper proposes a scheme to jointly optimize multi-beam power allocation over both the wireless fronthaul and the access links in a massive MIMO mmWave Cloud-RAN with stochastic traffic and channel conditions. More particularly, the optimization problem is formulated as a MDP with the objective to minimize the overall queueing delay and packet drops of multiple MTs by taking into consideration of their heterogeneous and dynamic traffic and channel states. By leveraging the structure of the underlying problem, we introduce a post-decision state and exploit decomposition techniques to reduce the search space and computation complexity, and an efficient online reinforcement learning scheme is developed to achieve the optimal MMIMO beam power allocation policy. The proposed scheme does not assume a priori knowledge of random user packet arrival statistics and wireless channel state transition probabilities. The evaluation results show that our proposed scheme outperforms the baseline schemes.

References

[1] Wang, Xinbo, et al. "Virtualized cloud radio access network for 5G transport." *IEEE Comm. Magazine*, vol. 55, no. 9, pp. 202-209, 2017.

[2] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer," *IEEE Netwo*rk., vol. 29, no. 1, pp. 35–41, 2015.

[3] M. A. Albreem, M. Juntti and S. Shahabuddin, "Massive MIMO Detection Techniques: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3109-3132, 2019.

[4] J Hoydis, S ten Brink, M Debbah, "Massive MIMO in the UL/DL of cellular networks: how many antennas do we need," *IEEE J.Sel. Areas Commun*., vol. 31, no. 2, pp. 160–171, 2013.

[5] T. Kim et al., "Tens of Gbps Support with mmWave Beamforming Systems for Next Generation Communications," *IEEE Globecom*, 2013.

[6] W. Roh, et al, "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE Comm. Mag*., vol. 52, no. 2, pp. 106–113, 2014.

[7] L. Lu, G. Li, A. Swindlehurst, A. Ashikhmin and R. Zhang: An overview of massive MIMO: Benefits and challenges," *IEEE J. of Sel. Topics in Signal Processing*, vol.8, no. 5, pp. 742–758, 2014.

[8] H. Yang and T. Marzetta, "Performance of Conjugate and Zero-Forcing Beamforming in Large-Scale Antenna Systems," *IEEE J. Sel. Areas in Comm.* vol. 31, no. 2, pp. 172–179, 2013.

[9] H. Li, J. Cheng, Z. Wang and H. Wang, "Joint Antenna Selection and Power Allocation for an Energy-efficient Massive MIMO System," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 257-260, Feb. 2019.

[10] L. Jiao, Y. Wu, J. Dong and Z. Jiang, "Toward Optimal Resource Scheduling for Internet of Things Under Imperfect CSI," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1572-1581, 2020.

[11] S. M. Ross, *Introduction to stochastic dynamic programming*, Academic press, 2014.

[12] R. Howard, *Dynamic Programming and Markov Processes*, MIT Press, 1960.

[13] M. L. Puterman and M. C. Shin, "Modified policy iteration algorithms for discounted Markov decision problems," *Management Science*, vol. 24, no. 11, pp. 1127–1137, 1978.

[14] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, Belmont, MA, 1995.

[15] D. Adelman and A. J. Mersereau, "Relaxations of weakly coupled stochastic dynamic programs," *Oper. Res.*, vol. 56, no. 3, pp. 712–727, Jan. 2008.

[16] X. Chen, P. Liu, H. Liu, C. Wu, Y. Ji, "Multipath Transmission Scheduling in Millimeter Wave Cloud Radio Access Networks," *IEEE ICC'18*, Kansas City, MO, May 2018.

[17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[18] T. Bai, V. Desai and R. W. Heath, "Millimeter wave cellular channel models for system evaluation," *IEEE ICNC*, 2014.