On the Efficiency of Multi-Beam Medium Access for Millimeter-Wave Networks

Jie Zhao, *Member, IEEE*, and Xin Wang, *Member, IEEE*

***Abstract*— The need of highly directional communications at mmWave band introduces high overhead for beam training and alignment, which also makes the medium access control (MAC) a grand challenge. However, the need of supporting highly directional multiple beams between transmitters and receivers makes the MAC design even harder. To harvest the gain of multi-beam mmWave communications, which benefits not only from the large bandwidth of mmWave spectrum but also the diversity of concurrent multi-user multi-beam transmissions, this paper studies the medium access control (MAC) layer related issues in multi-beam mmWave networks, including (1) efficient multi-beam training schemes to enable lower overhead thus faster AP association and beam alignment, (2) block-sparse mmWave channel estimation in different beam resolutions, and (3) effective concurrent radio resource allocation to facilitate better multi-user multi-beam transmissions. Simulation results demonstrate that the proposed schemes outperform existing techniques in improving the efficiency of mmWave communications thus achieving significantly higher network performances. To our best knowledge, we are the first to comprehensively consider both efficient training for beam alignment and resource scheduling in the MAC design to enable highly directional multi-user multi-beam concurrent transmissions in a mmWave network.**

***Index Terms*— Millimeter wave, directional antenna, resource allocation, multi-beam, compressed sensing.**

I. Introduction

mmWave communication is expected to provide Gigabit data rate demanded by the exponential growth of wireless applications. However, there is a need of highly directional communications at mmWave band to compensate for the big path loss in mmWave band. The progresses of circuit design and array signal processing allow for the generation of a single beam or multiple beams simultaneously. The highly directional transmissions introduce high overhead for beam training and alignment, which makes the medium access control (MAC) a grand challenge. In a single-beam mmWave network [1], if both AP and user devices are configured directionally, it could take an extremely long time to connect them and align their beams. For example, in the measurements of basic 60GHz IEEE 802.11ad standard [2] transmission, the latency for AP discovery is 5ms to 1.8s for a static device and up to 12.9s for a mobile device [3]. This problem is made even more difficult when there are a large number of beam directions and when there exist environment dynamics such as user mobility and blockage.

In order to alleviate the training overhead in single-beam mmWave networks, in codebook-based beam training [4]–[6], directions are divided into different angular range levels. At each level, training signals are sent to all directions and a feedback message is needed to select the best beam. The feedback overhead and delay would be very high with the competition in multi-user feedbacks along each training direction and the need of multiple rounds of feedbacks for different levels. The codebook-based beam training is recommended in 802.11ad standard. As an alternative, compressed sensing (CS) is exploited to estimate the sparse mmWave channels with training signals only sent along random directions within the whole angular range [7]–[9] at the cost of high channel reconstruction complexity. More frequent signaling would be needed to track the directional transmissions when there exist higher channel dynamics and user mobility [10]. The big training overhead will translate into throughput degradation. Despite the large amount of effort on finding the best beam direction or allocating radio resources [11]–[13], the two are often decoupled. There is a critical need to concurrently schedule radio resources for beam training, data transmissions, and beam switching, in the face of dynamics of channel conditions, user population, locations, and traffic.

Despite the possibility of generating multiple beams, there are very limited efforts on the MAC design for multiple beams. The IEEE standard 802.11ay [14] enhances the 802.11ad standard to support single-user and multi-user multiple-input-multiple-output (SU-MIMO and MU-MIMO). However, the training still follows simple code-book based platform, and the transmission duration is only divided into different fields without specifying how resources will be allocated. The limited studies on multi-user multi-beam mmWave communications are either on signal processing [15] and PHY layer [16]–[18] or consider simple scenarios such as single user multi-beam transmissions [19] and omni-directional receiving at the user side [20].

Different from literature studies, we consider the coexistence of multiple users, each is able to exploit multiple beams at a time. Unlike single-user multi-beam concurrent transmission (beam diversity) or device-to-device single-beam concurrent transmission (user diversity), it is a daunting task to enable multi-user multi-beam transmissions with the simultaneous consideration of the diversity from both beam domain and user domain. The challenges include 1) how to

Manuscript received 25 March 2020; revised 30 September 2020 and 15 June 2021; accepted 24 November 2021; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor A. Khreishah. Date of publication 31 December 2021; date of current version 18 August 2022. This work was supported by the National Science Foundation (NSF) Electrical, Communications and Cyber Systems (ECCS) under Grant 1731238. *(Corresponding author: Jie Zhao.)*

The authors are with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: jie.zhao@alumni.stonybrook.edu; x.wang@stonybrook.edu).

Digital Object Identifier 10.1109/TNET.2021.3137562

enable efficient simultaneous multi-beam training with reduced training overhead, 2) how to effectively allocate beams to different users for concurrent beam transmissions with minimum interference and enhanced diversity, and 3) how to cope with ambient fluctuations such as user movement and channel dynamics.

As an effort to tackle the challenges above, our aim is to design an efficient and integrated MAC scheme for high performance multi-beam mmWave network transmissions. Our framework is accommodated by the support of two critical and closely interactive components:

(a) ***Accurate and light-weight multi-beam training*** with 1) effective simultaneous training of multi-level multi-user multi-beam for fast user association and beam alignment, and 2) fine beam training with multi-resolution block-sparse CS channel estimation and adaptive training in response to channel conditions and learning from past measurements;

(b) ***Self-adaptive virtual concurrent transmission scheduling*** that jointly considers 1) multi-beam training information, 2) concurrent multi-user multi-beam data transmissions for higher efficiency use of spectral resources, and 3) mobility handling while enabling burst user transmissions for an overall high network performance.

To our best knowledge, this is the first work that comprehensively studies the MAC design with the concurrent considerations of beam training and resource scheduling in a multi-beam multi-user mmWave network.

The rest of this paper is organized as follows. After briefly reviewing background and related work in Section II, we present our user association and multi-beam alignment design in Section III. We further propose our channel estimation technique and flexible resource scheduling schemes in Section IV and Section V, respectively. Finally, we analyze the simulation results in Section VI, and conclude the paper in Section VII.

II. Background and Related Work

In this section, we first summarize the literature studies and the features of our work, and then introduce the basic mmWave channel model. Finally we provide the basic MAC framework of our design.

A. Related Work

IEEE 802.11ad [2] is proposed at physical layer (PHY) and medium access control layer (MAC) to enable operation in frequencies around 60 GHz mmWave band. In the example IEEE 802.11ad testbed in [21], Sur *et al.* design and implement a mmWave wireless access network called UbiG, which can deliver Gigabits per second (Gbps) consistently to the commercial-off-the-shelf IEEE 802.11ad devices.

As an enhancement of the 802.11ad standard, IEEE standard 802.11ay [14], [22] introduces single-user multiple-input-multiple-output (SU-MIMO) beamforming and multi-user multiple-input-multiple-output (MU-MIMO) beamforming, which are among the most important technologies to improve the quality of service (QoS) and network throughput. Although 802.11ay provides a basic MAC framework and signaling sequences, there is no specific consideration for more efficient multi-beam alignment and concurrent transmissions.

Existing work on beam training and transmission scheduling generally focuses on single-beam operations, and the two are usually not studied together. There are a limited number of studies on multi-beam communications. Ettorre *et al.* [15] propose a multi-beam antenna design along with signal processing techniques. In [16], Sun *et al.* study from PHY perspective the multi-beam antenna equal gain combining in future millimeter-wave cellular systems in the dense urban environment, while precoding is proposed to enable interference-free transmissions of k users in [17]. Kim *et al.* [19] propose a hybrid beamforming architecture and a multi-beam transmission diversity scheme for single user MIMO operation, with one beam trained at a time. Assuming users can receive signals omni-directionally [20], the coordination between a multi-beam access point (AP) and the single antenna users is much simpler. In [18], channel sparsity, GHz-scale sampling rate, and the knowledge of mmWave RF codebook beam patterns are exploited to reduce the training overhead of multi-stream directional links. However, how to enable efficient multi-stream concurrent data transmissions from MAC layer perspectives is not discussed. Ghasempour *et al.* in [23] propose a low-complexity structure to decouple beam steering and user selection, which selects users by minimizing the overlap of their idealized beam patterns from analog training. The proposed decoupling structure is validated using trace-based emulations and high resolution 60-GHz channel models, and the results show that it can achieve less than 5% performance loss compared with maximum rates available via joint user-beam selection. In contrast to existing work, we consider both multi-beam training and multi-user multi-beam concurrent transmission scheduling. To improve the transmission capacity while considering transmission fairness, we explore both the *beam domain diversity* where the base station (BS) and each user can communicate using multiple concurrent beams and the *user domain diversity* where BS can serve multiple users simultaneously.

For more efficient beam training, in codebook-based beamforming [4]–[6] that is also taken by 802.11ad, the feedback overhead and delay would be very high to train a large number of directions as a result of the waiting duration to accommodate competitions among users for feedbacks in each direction trained and too many rounds of training at different angular range levels. Alternatively, compressed sensing (CS) techniques estimate mmWave channels to facilitate beam alignment [7]–[9], taking advantage of the sparse feature of mmWave channels. However, they only considered the sparse feature of mmWave channels, but not the clustering feature of transmission paths [24]. Similarly in [25], Hassanieh *et al.* propose a phased array mmWave system that can find the best beam alignment without scanning the entire space. However, they did not consider multi-beam alignment, which is much more challenging. Instead, we develop a multi-beam training architecture to accelerate the training process and a *multi-resolution* channel estimation method that exploits the path clustering feature to improve the training performance and benefits from samples from different levels of multi-beam measurements to reduce the total number of training samples thus the computational complexity for CS channel construction.

Radio resource allocations in mmWave networks [11]–[13], [26] mostly focus on scheduling concurrent device-to-device communications. Instead, we investigate uplink/downlink transmission scheduling between base station/access point and devices, each can exploit multiple beams at the same time. The joint determination of transmission resources and duration makes the scheduling problem difficult, and are often bypassed by literature work. Our previous work [27] has studied the joint beam training and transmission scheduling problem for single-beam mmWave networks, but the problem is much harder with the daunting task of coordinating the training and transmissions using multiple beams in the multi-user realm. We schedule radio resources for concurrent multi-user multi-beam transmissions with joint consideration of training information, data transmissions, and mobility handling. Moreover, rather than coordinating users to transmit in each slot [28], we propose virtual scheduling to enable the scheduling of burst transmission which is expected to be the major transmission format for mmWave communications.

The contributions of our work are many folds and can be summarized as follows:

First, while the beam training and transmission scheduling are already a grand challenge for single-beam mmWave networks, we investigate the two problems in a even more complicated scenario where multiple users concurrently perform transmissions using multiple beams. This makes our work essentially different from those studying single-beam cases. To our best knowledge, this is the first work that comprehensively studies the MAC design with the join consideration of beam training and resource scheduling in a multi-beam multi-user mmWave network.

Second, existing CS-based channel estimation work in mmWave realm usually consider the channel to be sparse but fails to take into account the path clustering feature in mmWave channels, which implies the path gain vector is not only sparse but also has blocks/chunks of nonzero elements. Instead, we design a block sparse CS recovery scheme that can exploit the block sparsity to reduce the channel estimation time while further increasing the estimation accuracy.

Third, to enable fast AP association and beam alignment in both uplink and downlink directions, we propose multi-user multi-resolution beam training with various innovative components over existing standards, including (1) simultaneous multi-beam training, (2) division of feedback slots into scheduled ones and contended ones, (3) compressive measurement with block-sparse estimation of the mmWave channel at hierarchical beam resolutions, and (4) elastic fine multi-beam training that jointly works with transmission scheduling in response to channel condition and learning from past training results.

Fourth, to facilitate efficient radio resource management and tackle the challenges in concurrent multi-beam transmissions, we propose two transmission scheduling schemes that are adaptive to heterogeneous traffic types, user demands and resource availability with (1) concurrent determination of multi-beam transmission opportunities and durations for multi-user transmissions, (2) virtual slot aggregation and shuffling that supports burst transmissions and reduces signaling overhead, and (3) mobility handling that allows user to promptly recover from abrupt disconnection due to movement and channel dynamics. As far as we know, there are very limited efforts studying the scheduling for multi-beam concurrent transmissions among multiple users, which is a complicated problem as it not only needs to coordinate transmissions among multiple users and multiple beams, but also among the concurrent parties that will interfere with each other.

B. System and Channel Model

We consider a general cell-based transmission environment, where each cell consists of one base station or access point, and multiple users or devices. Different from existing work on mmWave beam alignment which generally considers single beam pattern, we focus on multi-beam transmissions where both base station and users can generate multi-beam patterns. The base station (BS) and a user are assumed to be equipped with N_{tx} and N_{rx} antennas. So, the number of beams M_{tx} generated by the base station is no greater than N_{tx}, i.e., $M_{tx} \leq N_{tx}$, and the number of user beams $M_{rx} \leq N_{rx}$.

The mmWave channel is found to be sparse [24]. A channel is composed of K clusters, each consisting of L subpaths. The channel matrix $\mathbf{H}$ can be expressed as

$$\mathbf{H} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} a_{k\ell} \cdot \mathbf{D}_{rx}(\theta_{k\ell}^{rx}) \cdot \mathbf{D}_{tx}^{H}(\theta_{k\ell}^{tx}), \tag{1}$$

which we rewrite in the matrix form as

$$\mathbf{H} = \mathbf{D}_R \, \mathrm{diag}(\mathbf{a}) \mathbf{D}_T^H, \tag{2}$$

where $\mathrm{diag}(\mathbf{a})$ is a diagonal matrix with each item taken from a concatenated column vector $\mathbf{a}$ $(1 \times KL)$. a_{kl} denotes the complex path gain of the ℓ-th path in the k-th cluster. Then

$$\mathbf{a} = [\underbrace{a_{11}, a_{12}, \ldots, a_{1L}}_{\text{cluster 1}}, \underbrace{a_{21}, a_{22}, \ldots, a_{2L}}_{\text{cluster 2}}, \ldots, \underbrace{a_{K1}, a_{K2}, \ldots, a_{KL}}_{\text{cluster } K}]^T, \tag{3}$$

There are KL elements in $\mathbf{a}$. One of the major differences between our work and other CS-based mmWave channel modeling and estimation work is that we take into consideration the *block sparse* properties of the mmWave channel due to path clustering effects and exploit this feature to better estimate the mmWave channel.

The matrices $\mathbf{D}_T$ and $\mathbf{D}_R$ contain the TX and RX array response vectors as follows:

$$\mathbf{D}_T = [\mathbf{D}_{tx}(\theta_{11}^{tx}), .., \mathbf{D}_{tx}(\theta_{1L}^{tx}), .., \mathbf{D}_{tx}(\theta_{K1}^{tx}), .., \mathbf{D}_{tx}(\theta_{KL}^{tx})], \tag{4}$$

$$\mathbf{D}_R = [\mathbf{D}_{rx}(\theta_{11}^{rx}), .., \mathbf{D}_{rx}(\theta_{1L}^{rx}), .., \mathbf{D}_{rx}(\theta_{K1}^{rx}), .., \mathbf{D}_{rx}(\theta_{KL}^{rx})]. \tag{5}$$

The dimensions of $\mathbf{D}_R$, $\mathrm{diag}(\mathbf{a})$ and $\mathbf{D}_T$ are $N_{rx} \times KL$, $KL \times KL$ and $N_{tx} \times KL$. The directional response column vector of the TX antenna, $\mathbf{D}_{tx}(\theta_{k\ell}^{tx})$, for the ℓ-th sub-path in the k-th cluster at the angle of departure (AoD) $\theta_{k\ell}^{tx}$ is

$$\begin{aligned} \mathbf{D}_{tx}(\theta_{k\ell}^{tx}) &= \left[D^{(1)}(\theta_{k\ell}^{tx})), \ldots, D^{(m)}(\theta_{k\ell}^{tx}), \ldots, D^{(N_{tx})}(\theta_{k\ell}^{tx})\right] \\ &= \left[1, \ldots, e^{j\cdot(m-1)\cdot w_{k\ell}^{tx}}, \ldots, e^{j\cdot(N_{tx}-1)\cdot w_{k\ell}^{tx}}\right]^T, \end{aligned} \tag{6}$$

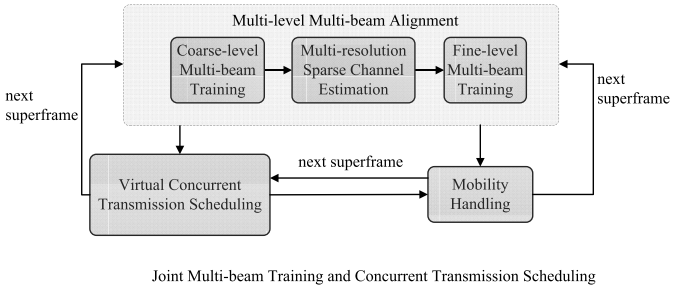


Fig. 1. Framework overview.

where $D^{(m)}(\theta_{k\ell}^{tx}) = e^{j\cdot(m-1)\cdot\frac{2\pi d_t}{\lambda}\sin\theta_{k\ell}^{tx}}$ is the response of the m-th TX antenna, d_t is the distance between two adjacent transmitting antenna elements, $\lambda = \frac{c}{f}$ is the wavelength in meters, f is the carrier frequency of the signal in Hz, c is the speed of light (3×10^8 meters/sec). We have the spatial frequency $w_{k\ell}^{tx} = \frac{2\pi d_t}{\lambda}\sin\theta_{k\ell}^{tx}$. The RX antenna's directional response column vector $\mathbf{D}_{rx}(\theta_{k\ell}^{rx})$ for the path at an angle of arrival (AoA) $\theta_{k\ell}^{rx}$ can be written similarly.

Throughout this work, we use antenna pattern and antenna weight vector (AWV) interchangeably, where a beam pattern can be a multi-beam type that consists of multiple concurrent beams. We also use device and user interchangeably, AP and base station interchangeably, for ease of presentation.

We will introduce channel estimation later.

C. Basic MAC Framework

Our basic MAC framework (Figure 1) consists of two major components that run with close interaction.

1) Multi-Level Multi-Beam Training: Our beam training is the integration of three techniques. In order to join a mmWave network and find the proper beams for data transmission, a device first needs to perform initial association and beam training with the AP. To avoid high feedback overhead as in conventional codebook-based schemes, we consider multiple levels of simultaneous training to quickly associate users with APs in our multi-user multi-beam training design (Section III). Second, we exploit multi-beam compressed sensing to speed up the sample collection for training. Finally, to align beams at the finest resolution desired, we exploit *multi-resolution-based* compressive channel estimation (Section IV), taking advantage of block sparse feature of the channel and samples from different levels of training for low-overhead and accurate channel estimation.

2) Multi-Beam Concurrent Transmission Scheduling: To enable efficient radio resource allocation for multi-user multi-beam concurrent transmissions, based on the candidate beams selected during the training, we develop two virtual transmission scheduling schemes to properly select the set of beams across users for high transmission performance. We also propose a flexible mobility handling scheme to cope with environmental dynamics (Section V), which will also influence beam training in later superframes.

III. AP Association and Multi-Level Multi-Beam Alignment

Compared to single-beam transmission, the existence of multiple beams makes it easier for the transmitter and receiver to intercept the communications signals of each other. However, in order for the communicating pairs to effectively transmit and correctly decode the signals for higher transmission performance, there is a need to determine the optimal alignments of the transmission and receiving beams. This requires the establishment of connections between transmission beams and receiving beams, and estimation of the channel conditions. This is hard with highly directional transmissions and the goal of this step is to reduce the overall training overhead.

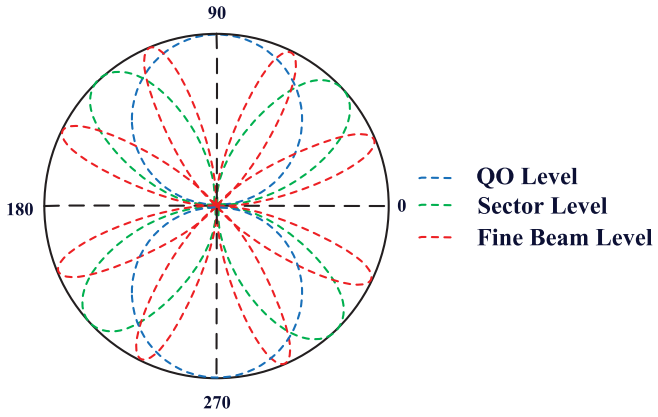


Fig. 2. Hierarchical beam levels example.

In this section, we first present the two-level coarse training, which will be used together with the compressive channel sensing for low overhead and high accuracy beam training. Different from IEEE 802.11ay, we emphasize the training coordination between uplink and downlink and the overhead reduction exploiting the information from the previous round of signaling. We allow AP to transmit feedbacks in a batch for devices within one sector to reduce the header overhead, and coordinate feedbacks from users in each AP sector according to the number of associated users known from the previous signaling round without the need of high-overhead random access.

We use some terms from 802.11ay to facilitate better understanding, and also provide the possibility of incorporating our design into the 802.11ay framework. Our scheme, however, is general and not constrained to run within 802.11 networks.

A. Multi-Level Multi-Beam Training

We apply three levels of beamwidth following the terms of 802.11ad and 802.11ay: quasi-omni-directional level (QOL), sector beam sweep (SBS), and fine-beam steering (FBS). Multi-beam patterns are available at the receiving end, depending on if the transmission is downlink or uplink. As multiple transmission signals cannot be differentiated in the training period without channel knowledge, a sender only transmits along one direction at a time.

An example of the hierarchical beam levels is given in Figure 2. At the quasi-omni-directional level, the beamwidth is configured to the widest possible allowed by the system to alleviate the deafness problem in receiving. The fine beam is the desired beamwidth to use for a mmWave system.

We consider the association and multi-beam alignment between devices and the AP in a cell. Due to the space limitation, we won't discuss device-to-device communications.

Our multi-beam alignment procedures can be completed with the following steps:

Step (1): ***Bi-directional training for quick association between AP and devices.***

An AP will send beacon messages periodically for new and existing devices to associate with and align their beams. To facilitate quick AP association while not compromising the link budget significantly, we will configure AP at SBS level and devices in QOL. Rather than performing the training for each device at a time, the training will be performed for all devices simultaneously. AP will send beacons in each SBS direction. Within a direction, a (new) device can listen from a multi-QOL-beam pattern that contains multiple QOL directions to record proper sending SBS sectors and receiving QOL directions. Then AP configures itself to listen from a multi-SBS-beam pattern that contains multiple SBS directions.

In order to send feedback information to the transmitter, multiple devices need to compete in channel access. When the number of devices is large, the overhead can be very high. Different from 802.11ad and 802.11ay, we divide the duration for feedbacks into two parts: *scheduled response slots* and *contention slots.* Scheduled slots are used by AP to schedule the transmissions from devices that it is already associated with or is aware of from the previous round of transmissions. The scheduling information can be sent in the previous announcement interval or in the current beacon. Contention slots are exploited by devices that were abruptly disconnected from AP (due to abnormal reasons such as severe channel conditions and blocking) or new comers which are unknown to AP. The use of scheduled slots will significantly reduce the total duration for feedbacks as well as speed up the feedbacks thus the overall beam-alignment process.

To facilitate reverse channel training, in the multi-SBS-beam pattern that AP listens to, a device will subsequently send *along all its QOL directions one by one* the following information: its association request, the SBS sector it selects for AP to transmit, and its best receiving QOL direction. A device will then prepare itself at the best reception multi-QOL-beam direction. To reduce the feedback overhead, rather than sending a feedback to every device right away as in 802.11ad and 802.11ay, AP will send an aggregate feedback to the group of devices in each of the selected SBS directions after receiving device messages from all its sectors.

AP and devices now obtain a preliminary association with the information: downlink, the best transmission sector of AP and the best QOL receive direction of a device; uplink, the best QOL sending direction from a device and the best reception sector at AP.

Step (2) ***Bi-directional training to find the best sector pair between AP and each device.*** To further search for the best receiving sector direction for each device, AP sends training signals again in best sectors selected from the previous step, while each associated device only receives from the multi-SBS-beam pattern containing a set of SBS directions within the angular range of its best QOL reception direction. To determine the best transmission sector from a device, AP only listens to responses in the multi-SBS-beam pattern that consists of several best reception sectors selected by devices earlier. When AP receives with the multi-SBS-beam pattern, multiple associated devices will exploit response slots to respond to AP. Similarly to Step (1), the response slots are composed of two parts: scheduled slots and contention slots. In the contention slots, devices contend to get a response slot among S_2 ones. Rather than using an equal number of contention slots for the feedbacks to each sector, we set S_2 in a multi-SBS-beam pattern where S_2 for a beam sector is set to be proportional to the number of newly associated users that is learned from Step (1). Existing users will either not need training in this superframe or can have their feedbacks transmitted in scheduled response slots.

This helps reduce the contention in sectors with more new users while avoiding wasting time slots unnecessarily in sectors with very few new users. The value of S_2 can be sent to devices along with AP feedbacks in the Step (1). If successfully obtaining a slot, a device will send a response on the link quality and the best receive sector from AP for each of the set of sector-level directions within the range of its best QOL direction. AP will immediately feedback to the device its best transmission sector to AP.

Step (3) ***Determining the best fine-level multi-beam transmission and receiving directions.*** Finally, AP and devices need training to find the best multi-beam alignment at the fine-beam level. Similar back-and-forth measures can be taken; however, due to the potentially large number of fine-beam patterns, the overhead can be unbearable. We will further reduce the overhead by exploiting the sparse estimation of the mmWave channel, which will be introduced later in Section IV.

B. Summary of Multi-Beam Training

Compared to single beam training, where only one reception beam of a device can be trained at a time, our multi-beam training can take advantage of multiple beams of the receiver (either AP or device depending on it is uplink or downlink transmission) to simultaneously receive the training signals. This significantly speeds up the training process. Particularly, in the fine beam training part where sparse estimation of the mmWave channel is exploited, multiple reception fine beams of each user can work at the same time. Compared to a single-beam mode, the gathering of the beam training samples are greatly accelerated for faster channel estimation. Receiving with multiple beams is equivalent to increasing the sampling speed in compressed sensing, or obtaining more samples within the same sampling duration for more accurate channel estimation.

IV. Multi-Resolution and Block-Sparse mmWave Channel Estimation

To find the best fine-level transmission and receiving directions, it may need a large number of training messages. The coarse level training can constrain the messages to be sent within the best transmission and receiving sectors. However, if the number of fine beams to transmit remains large, rather than measuring all fine beam combinations (as in 802.11ay) or exploiting more levels of training at high feedback cost,

we will explore compressed channel measurements and block sparsity in channel estimation to reduce the training overhead while ensuring higher channel estimation quality.

Due to transmission path clustering, mmWave channel presents the block sparse feature. Existing CS-based channel estimation methods only consider the **sparse** characteristic of mmWave channel, while our goal is to exploit the **clustering** feature besides the sparsity for more accurate channel estimation. In CS-based channel estimation, the vector of path gain $\mathbf{a}$ of mmW channel in Eq. (3) is what we are most interested in, where only a small fraction of the elements in the gain vector are nonzero. One of the effects of the path clustering feature of mmW channels is that it results in assembled nonzero elements in the path gain vector $\mathbf{a}$ (i.e., some blocks are filled with nonzero elements, whereas others are filled with all zeroes) and this additional **block sparsity** information can be further exploited in CS reconstruction to improve the reconstruction performance. Previous CS-based studies didn't consider the path clustering feature of mmWave paths. They mostly estimate the path gains of the mmW channel as just a sparse vector, but not a block sparse vector.

Different from conventional CS-based channel estimation schemes, we further explore this block feature to improve the channel estimation performance in Section IV-A. Furthermore, we propose in Section IV-B a *multi-resolution* scheme for low complexity and high accuracy channel estimation, taking advantage of measurement samples from all beam resolution levels.

A. Block-Sparse Channel Estimation

We have introduced the channel model in Section II-B. We will now describe how we exploit the path clustering feature of mmWave channels and develop the solution to channel estimation with block-sparse channel reconstruction.

For channel estimation, assume we transmit the training signals along P directions, i.e., with P TX beamforming (BF) vectors ($\mathbf{u}_p$, $p = 1, 2, \ldots, P$), and a receiver estimates the signals from Q directions with Q RX BF vectors ($\mathbf{v}_q$, $q = 1, 2, \ldots, Q$). Taking advantage of coarse-level training, these are randomly chosen from the fine beam directions within the TX's best sectors and the RX's best sectors, respectively. Then the measurements can be expressed in the matrix format as:

$$\mathbf{R}^{Q\times P} = \mathbf{V}^H\mathbf{H}\mathbf{U} \circ \mathbf{S} + \mathbf{E}, \tag{7}$$

where $\mathbf{S}$ and $\mathbf{E}$ are respectively the training signals and noise, and

$$\mathbf{V}^{N_{rx}\times Q} = [\mathbf{v}_1, .., \mathbf{v}_q, .., \mathbf{v}_Q], \quad \mathbf{U}^{N_{tx}\times P} = [\mathbf{u}_1, .., \mathbf{u}_p, .., \mathbf{u}_P]. \tag{8}$$

With the training signals transmitted at the power A, $\mathbf{R}^{Q\times P} = \sqrt{A}\mathbf{V}^H\mathbf{H}\mathbf{U} + \mathbf{E}$, which can be vectorized as

$$\begin{aligned}
\mathbf{r} &= \mathrm{vec}(\mathbf{R}) = \sqrt{A}\mathrm{vec}(\mathbf{V}^H\mathbf{H}\mathbf{U}) + \mathrm{vec}(\mathbf{E}) \\
&\overset{\text{Theorem 1 [29]}}{=\!=} \sqrt{A}(\mathbf{U}^T \otimes \mathbf{V}^H)\mathrm{vec}(\mathbf{H}) + \mathrm{vec}(\mathbf{E}) \\
&\overset{\text{Proposition 1 [30]}}{=\!=} \sqrt{A}(\mathbf{U}^T \otimes \mathbf{V}^H)\mathbf{\Psi}\mathbf{a} + \mathrm{vec}(\mathbf{E}) \\
&= \mathbf{\Phi}\mathbf{\Psi}\mathbf{a} + \mathrm{vec}(\mathbf{E}) = \mathbf{A}\mathbf{a} + \mathrm{vec}(\mathbf{E}),
\end{aligned} \tag{9}$$

where $\mathbf{\Psi} = \mathbf{D}_R^* * \mathbf{D}_T$(Khatri-Rao product) is the basis matrix, $\mathbf{\Phi} = \sqrt{A}(\mathbf{U}^T \otimes \mathbf{V}^H)$ (Kronecker product) is the measurement matrix (determined by TX and RX beam training directions). In the derivation, we have used Theorem 1 [29] and Proposition 1 [30] as follows:

Theorem 1:

$$\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}). \tag{10}$$

Proposition 1:

$$\mathrm{vec}(\mathbf{H}) = \mathbf{\Psi}\mathbf{a}, \text{ where } \mathbf{\Psi} = \mathbf{D}_\mathbf{R}^* * \mathbf{D}_\mathbf{T} \text{ (Khatri-Rao product).} \tag{11}$$

In order to differentiate between the estimated channel and the actual channel $\mathbf{a}$, we now refer the estimated channel as $\mathbf{x}$ and replace the vector $\mathbf{a}$ in the Eq. (9) with $\mathbf{x}$ and the measurements $\mathbf{r}$ with $\mathbf{y}$, then we have the compressed sensing form $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where $\mathbf{A}$ is the sensing matrix, and $\mathbf{e}$ is the noise. Different from conventional CS-based channel estimation algorithms, to enable more accurate beam alignment, we take into account the block-sparse feature of the vector $\mathbf{x}$ when reconstructing the virtual mmWave channel. We form our problems as follows:

$$\begin{aligned}
&\min \ \sum_{i=1}^{n} \|\mathbf{X}_i\|_2 \\
&\text{subject to } \ \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \mathbf{x} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n],
\end{aligned} \tag{12}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm, i is the block index, n is the number of blocks, $\mathbf{X}_i = x_{(i-1)d+1:id}$, and d is the block size. A typical solution algorithm for (12) is presented in Sec. IV of [31] as the "Recovery of block-sparse signals" Algorithm. After recovering $\mathbf{x}$, the virtual channel $\mathbf{H}$ can be estimated as in Eq. (2).

B. Multi-Resolution Channel Estimation

We have multiple levels of beamwidth: QOL, SBS, and FBS. In our channel estimation, we not only use FBS training data to estimate the mmWave channel but also exploit those in QOL and SBS to further improve the estimation accuracy.

To facilitate the channel estimation, we can discretize angular domain with $N_{tx}^g \times N_{rx}^g$ grids, so the channel can be estimated as a vector of the dimension $N_{tx}^g N_{rx}^g \times 1$ (vec($\mathbf{H}$)). As the mmWave channel is sparse, so the channel response signals only appear in a small number of grids. Rather than uniformly discretizing the angles, we uniformly divide the spatial frequencies $w_{k\ell}^{tx}$ and $w_{k\ell}^{rx}$ into N_{tx}^g and N_{rx}^g grid points, respectively. Thus, the response column vectors of the TX and RX antennas at the angular grid n and m are respectively

$$\begin{aligned}
&\mathbf{D}_{tx}^n(\theta_{k\ell}^{tx}) \\
&\quad = \left[1, e^{j\cdot 1\cdot n\cdot \frac{2\pi}{N_{tx}^g}}, e^{j\cdot 2\cdot n\cdot \frac{2\pi}{N_{tx}^g}}, \ldots, e^{j\cdot (N_{tx}-1)\cdot n\cdot \frac{2\pi}{N_{tx}^g}}\right]^T, \\
&\mathbf{D}_{rx}^m(\theta_{k\ell}^{rx}) \\
&\quad = \left[1, e^{j\cdot 1\cdot m\cdot \frac{2\pi}{N_{rx}^g}}, e^{j\cdot 2\cdot m\cdot \frac{2\pi}{N_{rx}^g}}, \ldots, e^{j\cdot (N_{rx}-1)\cdot m\cdot \frac{2\pi}{N_{rx}^g}}\right]^T.
\end{aligned}$$

If $N_{tx}^g = N_{tx}$ and $N_{rx}^g = N_{rx}$, we have

$$\mathbf{\Psi} = IDFT_{N_{tx}}^* * IDFT_{N_{rx}}, \tag{13}$$

where $IDFT_N$ denotes an N-dimensional IDFT matrix.

Different beamwidth adopted by AP and devices affects the values of N_{tx}^g and N_{rx}^g. Denote BW_{tx} and BW_{rx} as the beamwidth of AP and a device, one option is to let both $BW_{tx} * N_{tx}^g$ and $BW_{rx} * N_{rx}^g$ cover the whole angular space, and another is to reconstruct $\mathbf{H}_{FBS}$ only within the sector space detected to have stronger signals in the coarse level training. With the first method, a larger beamwidth will correspond to a discretized channel with a smaller dimension, so we have $dim(\mathbf{H}_{QOL}) < dim(\mathbf{H}_{SBS}) < dim(\mathbf{H}_{FBS})$. As samples are not uniformly taken from all angular directions, straight-forward channel reconstruction may not be accurate. Instead, we propose to reconstruct the channel recursively *at different grid resolutions* taking advantage of the multi-level training samples we have obtained. The gain of the channel estimated based on samples of a larger beamwidth will be used as block weights for the channel constructed with samples of the finer beamwidth. Following the training process, we have

Step (a) **QOL channel reconstruction:** After QOL beam training, reconstruct vec($\mathbf{H}_{QOL}$).

Step (b) **SBS channel reconstruction:** After SBS beam training, based on QOL results in Step (1), adjust the weights at the SBS level: the SBS elements contained in nonzero QOL blocks with larger recovered magnitude are assigned with smaller weights, and then reconstruct vec($\mathbf{H}_{SBS}$).

Step (c) **FBS channel reconstruction**: After FBS beam training, according to SBS results in Step (2), adjust the weights at the FBS level: the FBS elements contained in larger nonzero SBS blocks are assigned with smaller weights, and then reconstruct vec($\mathbf{H}_{FBS}$). We can then obtain the mmWave channel matrix $\mathbf{H}_{FBS}$ for further beam alignment.

Compared with conventional CS-based channel estimation, we not only exploit the block properties in mmWave channels, but also take advantage of the multi-level beam training results to significantly reduce the number of measurements thus the complexity in reconstructing the mmWave channel, which speeds up the training.

C. Procedures for Multi-Fine-Beam Training

With the coarse multi-beam training in Section III, AP and devices have known some information about each other, including the best transmission and receiving sectors, for both downlink and uplink transmissions. We will add the following procedures for compressive multi-fine-beam training:

Step (3.1) **Downlink multi-fine-beam training**: To facilitate synchronization, each device (DEV) initially listens at the multi-SBS-beam pattern that contains its best receiving sectors to intercept system parameters. For the multi-fine-beam training, within each best transmitting sector selected in the SBS phase, AP first sends beacons along P^T randomly selected fine-beam directions, one at a time. During the transmission of each fine beam, the set of devices that select the corresponding transmission sector will each listen with its multi-fine-beam pattern from Q^R fine-beam directions in its best DEV receiving sector. If the number of beams of a device is smaller than Q^R, it will turn to different random directions to collect samples from Q^R directions. After collecting samples from $P^T Q^R$ directions, a DEV can estimate the channel and the best fine-beam directions for AP transmission and DEV receiving.

Step (3.2) **Uplink feedback training**: AP first configures itself to receive from the multi-SBS-beam pattern containing selected best receiving sectors, and each associated device will send uplink feedbacks with the best AP TX fine beam selected, SNR, suggested beam directions, etc. Each device will transmit from Q^T fine-beam directions within its best transmitting sector. As the set of devices to associate with AP is known, the beacons in Step (3.1) will contain the order of uplink transmissions from devices to avoid their uplink competition.

Sampling from the learning of past measurements: Although we cannot completely follow the channel reciprocity rule, there may be correlation in uplink and downlink channels. To further improve the channel estimation quality while reducing the number of samples, a device can select Q^T fine-beam directions close to its best downlink receiving direction. Similarly, for each uplink fine-beam transmission, the P^R directions AP listens to can be close to the best downlink transmission beam direction. This provides a guidance for the multi-beam receive pattern for AP and the users. In addition, with the downlink channel estimated, a device can suggest a few directions for uplink training based on the sequence of eigenvalues of the channel in its feedback. With all samples, AP then estimates the uplink channel to find the best theoretical fine-beams.

Step (3.3) **Downlink feedback**: AP can broadcast the selected beam pairs for uplink transmission along its set of best fine-beam directions determined.

V. Joint Beam Training and Transmission Scheduling

In this work, we explore both the *beam domain diversity* where the base station (BS) and each user can communicate using multiple concurrent beams and the *user domain diversity* where BS can serve multiple users simultaneously. We aim to improve the total network throughput by not only enabling concurrent transmissions among users but also allowing each user to effectively select (multiple) beams (Figure 3) for transmission. After training, the quality (e.g. SNR, transmission rate) of each beam pair between AP and a user is known, and this information will be exploited in our scheduling of concurrent multi-beam transmissions. To achieve our goal, we would like to select more and better beams for concurrent transmissions. However, more beams will also increase the chance of interference. So beams should be carefully selected to reduce their interference impact. In addition, in frame-based transmissions, a user is often assigned a period of time with multiple time slots to transmit. It remains a big challenge to select the user and corresponding beams and also determine the transmission duration for each selected user.

We propose in this section two scheduling strategies to accommodate concurrent transmissions among multiple users, along with our proposed slot shuffling and aggregation technique to further improve the efficiency of concurrent transmissions. We also introduce the basic procedures to handle mobility.

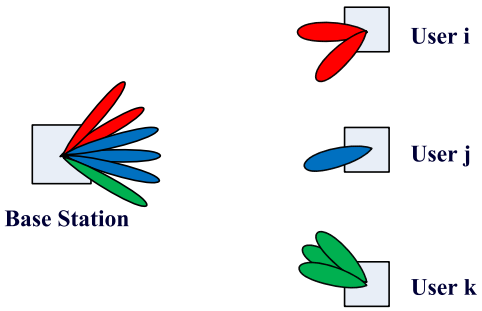


Fig. 3. Multi-user multi-beam concurrent transmission example.

A. Scheduling of Concurrent Multi-Beam Transmissions

After estimating the downlink channel in fine-beam training, a user i can select T_i high quality beams for AP transmission and R_i for receiving, and put them in its *candidate beam set*. It will inform AP its selected transmission beams. The candidate beam set for AP transmission is the union of beams selected by all users. Similarly, for uplink transmission, AP will select and inform a user its best set of beams for transmission. For ease of presentation, we will present the downlink scheduling first and then the uplink scheduling. The schedule is announced at the beginning of Data Transmission Interval.

1) Resource Allocation Option 1: Beam-Driven User Selection (BUS): As a transmitter can generate multiple beams and the candidate set contains only a subset of the beams, we could allow all beams from the candidate set to transmit. However, as a beam may be the candidate selected by multiple users for transmission or receiving, we need to determine which users should be assigned the beam in a time slot. In our proposed Beam-driven User Selection strategy, we schedule individual beam transmissions based on their qualities, and a user is selected if one or more of its beams are selected. We first present how scheduling is carried out in each time slot, and we then introduce how we can aggregate transmission slots scheduled for each user to enable low-cost burst transmission through our virtual scheduling procedure.

(1) For each beam B_i ($i = 1, 2, \ldots, I$) in the AP's candidate beam set, there may be more than one user towards whom this beam has a good quality. We schedule this beam to transmit to the user x_i, where we select x_i based on the following modified largest weighted delay first criteria:

$$x_i = \arg\max_{x_i} a(x_i)W(x_i)r(x_i, B_i)/\overline{R}(x_i), \quad i = 1, 2, \ldots, I,$$

where $x_i \in$ users and has B_i in its candidate beam set. I is the number of AP transmission beams to be scheduled, $a(x)$ is the priority parameter for a user x (determined by the service type, QoS requirements etc.) and $W(x)$ is the queuing delay. For delay-constrained traffic, we have

$$\text{Prob}[W(x) > T(x)] \leq \varepsilon(x), \tag{14}$$

where $\varepsilon(x)$ is a specified probability that the delay exceeds the threshold $T(x)$. Then the priority parameter $a(x)$ can be defined as $a(x) = -\log \varepsilon(x)/T(x)$. A smaller $\varepsilon(x)$ suggests a larger $a(x)$ that implies higher priority. $\varepsilon(x)$ can be set to 1 for delay tolerant applications. The transmission rate $r(x_i, B_i)$ between the beam B_i of AP and the user x_i can be estimated as

$$r(x_i, B_i) = B_w \log_2(1 + SINR(x_i, B_i)), \tag{15}$$

where B_w is the bandwidth, and the signal-to-interference-plus-noise ratio (SINR) is affected by not only received signal power and noise power but also interferences from other users' beams. The calculation of SINR during the scheduling of each beam will consider the interference from the set of beams already scheduled and a new beam is added into this set after each step. We will also update the rates of scheduled beams to consider the interference from the new beam.

The parameters $a(x)$, $W(x)$, and $\overline{R}(x)$ will be updated after assigning each slot to ensure that other users have the chance of transmissions. *As the slots are only assigned and users have not transmitted yet, so our parameter update is virtual.*

(2) The physical header for each transmission is usually very large. If a user is scheduled to transmit in several time slots and user transmissions for each slot are carried out independently, a header needs to be added for each transmission. Additional signaling may be needed for synchronization of each transmission. To reduce the big overhead for headers and signaling, after Step (1) is completed for all transmission slots, we perform shuffling and aggregation of slots assigned over a beam for each user, so a beam can be used to transmit to a user continuously. For example, beam B_i may be assigned to user x_i in the 3rd, 7th and 10th slot, and then the shuffling can allow user x_i to exploit beam B_i continuously in the 3rd, 4th and 5th slot. As all candidate beams of AP have been assigned in every transmission slot (although maybe to different users), the interference relationship among beams remain unchanged throughout all transmission slots. So this virtual shuffling and aggregation won't affect the interference relationship while improving the efficiency of concurrent multi-user transmissions.

2) Resource Allocation Option 2: Beam Selection for Maximal Sum Rate (BSMSR): The increase of the total number of beam transmissions could potentially increase the total transmission rate, but also the overall interference. Rather than scheduling transmissions for each beam in the candidate beam set of AP, in each time slot, we propose another option to only select part of beams that contribute to the largest rate improvement while considering the interference impact of the new beam on the already scheduled beams. The same beam may have different transmission rates for different users. Different from the Option 1 which determines the user based on its association with the beam, our *Beam Selection for Maximal Sum Rate* strategy looks for the best beam while considering its effect on the user transmission rate. This greedy scheduling idea will roughly achieve a sub-optimal network throughput.

To determine the beams to transmit in each slot, the following procedures are taken:

(1) In order to avoid starving users with poor communication conditions, we randomly select the *first* user x among all associated with AP. Based on the estimated transmission rate of each beam in the candidate beam set of x, B_{k_x} ($k_x = 1, 2, \ldots, |\mathbf{B}_x|$), we select the best beam to add into

the scheduled beam set:

$$B_{k_x} = \arg\max_{B_{k_x}} r(x, B_{k_x})) = \arg\max_{B_{k_x}} B_w \log_2(1 + SNR(x, B_{k_x})),$$

where B_w is the transmission bandwidth and $SNR(x_j, B_{k_{x_j}})$ is the estimated signal to noise ratio (SNR) for the signal received from beam B_{k_x} of user x.

(2) Estimating the transmission rate of each user x_j ($j = 1, 2, \ldots, J$) on each *remaining* beam in its candidate beam set $B^*_{k_{x_j}}$ that excludes scheduled beams in previous steps. Then the next user and beam selected is:

$$(x_j, B^*_{k_{x_j}}) = \arg\max_{(x_j, B^*_{k_{x_j}})} B_w \log_2(1 + SINR(x_j, B^*_{k_{x_j}})). \quad (16)$$

Different from Step (1), the rate is estimated based on signal-to-interference-plus-noise ratio (SINR) to take into account interferences from beams already scheduled:

$$SINR = \frac{\text{received signal power of the beam of interest}}{\text{noise + interferences from other users' beams}}.$$

(3) Once a beam is scheduled for transmission to a user, it is removed temporarily from the candidate beam set of all users until the next slot to schedule. Repeat Step (2) until either all beams are scheduled or the sum rate of all scheduled beams decreases with the addition of new beam.

After each step, the estimated transmission rates and SINRs need to be updated by taking into account all concurrent beams scheduled. For a given user, the estimation of SINR will change in the course of scheduling, because the user may be assigned more beams or more beams from interfering users are scheduled.

When the scheduling is stopped, we can perform virtual shuffling and aggregation of slots similar to that in Step (2) of Option 1, for the same purpose of better facilitating burst transmissions.

3) Uplink Scheduling: We have previously introduced the downlink scheduling. Our two downlink scheduling options can be easily extended in uplink scheduling. However, interference has much more significant impacts in uplink transmissions than in downlink transmissions, since all signals are received at AP.

B. Mobility Handling

Although multi-beam transmissions have made mmWave communications much more resilient to user mobility and blockage compared with single-beam transmissions, in practical applications, mobility and blockage can still significantly degrade the network performance if not properly handled. If a user is abnormally disconnected from AP due to movement and channel condition change, it will waste resources allocated during the scheduling. To facilitate a user to quickly recover from abrupt disconnection due to movement and channel dynamics, we propose the following beam switching scheme for downlink transmissions (uplink is similar):

Step (1) If a user is abruptly disconnected, the user will promptly check its candidate beam set and neighboring beams to find alternative beams which have not been scheduled for transmission in its transmission duration allocated (which possibly consists of a few time slots scheduled).

Step (2) Based on the information gathered in multi-beam training and transmission scheduling, we will check if adding an alternative beam to transmit to the disconnected user can improve the sum rate of the network for the remainder of the use's original scheduled transmission slots, considering the interference brought by the new beam to other users.

Step (3) If there are alternative beams, the user chooses the one that can obtain the largest sum rate improvement, and inform AP to initiate a fast reconnection using this beam. If the fast connection is successful, the user can complete transmission by exploiting the new beam for the remainder of its transmission duration allocated. If the fast reconnection is not successful, no alternative beam can be found to improve the sum rate, or there are no other beams allocated to the user, the user will not be accommodated in the current superframe and it needs to re-associate with AP in the next superframe. Although the user is disconnected temporarily, without transmission, its interference to other users is eliminated. The disconnected user will also have higher priority in the transmission scheduling process later.

VI. Performance Evaluation

We evaluate the performance with the comparison of the following schemes: (1) Proposed-BUS (proposed multi-beam training and scheduling of Beam-driven User Selection), (2) Proposed-BSMSR (proposed multi-beam training and scheduling of Beam Selection for Maximal Sum Rate), (3) Baseline-CS (Proposed multi-beam training but with another CS channel estimation solution in [32]), (4) HOL (Proposed multi-beam training but with Head-of-Line delay based slot-by-slot scheduling adapted from [28] for mmWave networks), (5) Baseline-802.11ay-like (Baseline that adopts basic features in beamforming and resource scheduling from IEEE 802.11ay standard [14]), (6) Proposed-BUS-MH (Proposed-BUS with mobility handling), and (7) Proposed-BSMSR-MH (Proposed-BSMSR with mobility handling). Note that Baseline-802.11ay-like is not a complete implementation of 802.11ay, although it includes the basic features from 802.11ay such as 60GHz frequency, Beacon Interval structure, multi-resolution beamforming training, and radio resource scheduling.

A. Settings

In our performance studies, we consider the scenario with one AP and multiple devices. The mmWave channel is generated based on the model derived from NYC measurements in [24]. More default parameters are presented in Table I. We studied the following performance metrics: (1) Training overhead (averaged temporal cost in a superframe to complete the beam training), (2) End-to-end delay (averaged latency for packets of users from source to destination), and (3) Network throughput (total throughput of users). The results are averaged among a long period (200 seconds). We use the traditional random waypoint (RWP) model to simulate the mobility of

TABLE I
DEFAULT PARAMETERS

Parameter	Description
length of a BI or superframe	default 200ms
# AP/DEV antenna	128/64
Bandwidth/Carrier frequency	1 GHz/60 GHz
# QO level/# sectors per QO level	4/4
# fine beams per setor, AP/DEV	8/4
AP-DEV distance	20 to 100m

users. To be more specific, we let the original position and direction of a user be randomly and independently chosen, and the mobility level is controlled by simulating each user with a random speed and maintaining a pre-determined average of speedss among the users. The device will be disconnected from the AP if the mobility causes abrupt beam misalignment or blocking.

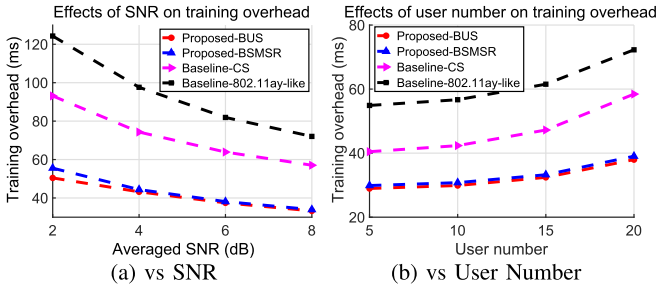
Due to the daunting complexity of evaluating our MAC design itself, we assume idealized pencil-beam shape for antenna patterns in our performance studies, i.e., we don't specifically consider perfectly practical beam shapes like side lobes or grating lobes. In real-world scenarios, the effects of this will depend on the number of antennas and their placement. The number and impacts of side lobes will increase with the number of antenna elements, which reduces the gain of each beam. With larger space among elements, the amplitude of grating lobes also increases. As a result, using too many antennas and excessive space among elements can lead to higher interferences among users, which suggests that the antennas need to be properly configured according to the environment and use cases when it comes to real-life mmWave applications.

B. Effects of SNR

Noise conditions in wireless mmWave networks greatly impact the data transmission quality thus network performances. At lower SNR, more training samples are needed to ensure the quality of channel estimation. Beam training overhead and beamforming quality also impact the end-to-end delay and network throughput.

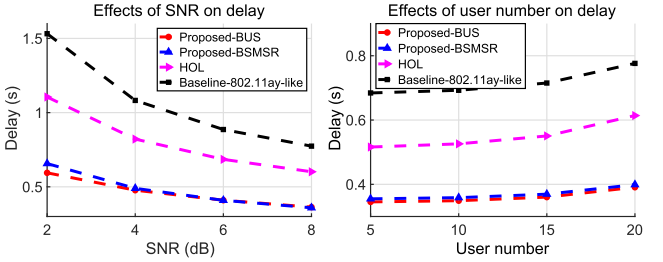
In Figures 4a, 5a, and 6a, the channel condition is better, the beam training overhead and end-to-end delay are both reduced, and the network throughput is improved. The gain of our schemes over others is larger at lower SNR. When SNR is 2dB, we observe that (1) Proposed-BUS and Proposed-BSMSR have 59.42% and 55.32% lower training overhead compared to Baseline-802.11ay-like, and 45.84% and 40.36% lower overhead compared with Baseline-CS; (2) Proposed-BUS and Proposed-BSMSR have 61.21% and 57.13% smaller end-to-end delay compared to Baseline-802.11ay-like, and 46.23% and 40.59% lower delay compared with HOL; (3) Compared with Baseline-802.11ay-like, Proposed-BUS and Proposed-BSMSR are able to enhance the network throughput by 64.61% and 90.49%, respectively. The improvement is 41.92% and 64.24% when compared with HOL.

The improvement of our proposed two schemes over others (Baseline-CS, Baseline-802.11ay-like) shows the benefits of our design efforts. Our multi-beam training and multi-resolution channel estimation techniques, as designed, reduce



(a) vs SNR (b) vs User Number

Fig. 4. Training overhead.



(a) vs SNR (b) vs User Number

Fig. 5. End-to-end delay.

the training overhead and improve the training quality. The proposed transmission scheduling schemes efficiently enable concurrent transmissions, reduce the end-to-end delay, and enhance the network throughput.

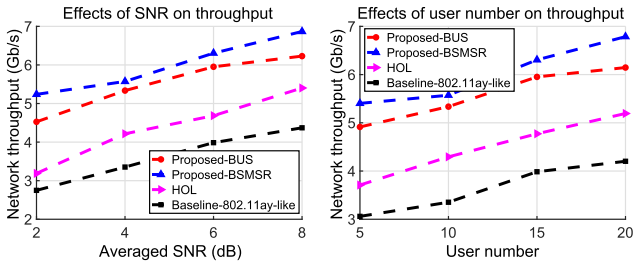
Between Proposed-BUS and Proposed-BSMSR, Proposed-BUS performs slightly better in terms of end-to-end delay, as the delay is one of the scheduling constraint factor in Proposed-BUS. In contrast, Proposed-BSMSR outperforms Proposed-BUS on network throughput because it is aimed to maximize the sum rate of users. This shows one of the many trade-offs in practical mmWave network design.

C. Effects of User Number

The number of users in the network has a significant impact on the network performances as it affects the efficiency of beam training and AP association thus achievable data transmission rate and the allocation of different data transmission slots. While keeping each user's traffic load the same, we vary the number of users.

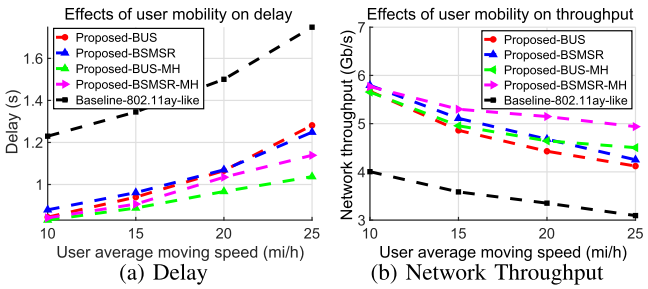
From Figures 4b, 5b, and 6b, we observe an increase in training overhead, end-to-end delay and network throughput as the number of users becomes larger. At a user number of 20: (1) In terms of training overhead, Proposed-BUS performs 47.24% better than Baseline-802.11ay-like and 28.40% better than Baseline-CS, respectively. The improvement for Proposed-BSMSR is 45.46% and 26.13%; (2) For end-to-end delay, Proposed-BUS outperforms Baseline-802.11ay-like by 49.52% and HOL by 33.07%, respectively. Likewise, the delay reduction for Proposed-BSMSR is 48.12% and 31.22%; (3) Proposed-BUS and Proposed-BSMSR gain a similar network throughput enhancement of 60.60% and 45.64% over Baseline-802.11ay-like and HOL, respectively.

Again, the observations above confirm the advantages of the proposed design: less training overhead, smaller delay, and better throughput.



(a) vs SNR (b) vs User Number

Fig. 6. Network throughput.



(a) Delay (b) Network Throughput

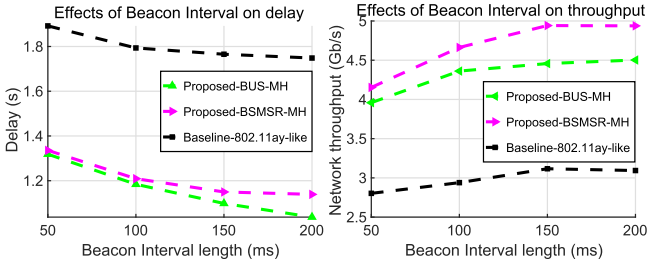
Fig. 7. Effect of user mobility.

D. Effects of User Mobility

The highly directional transmissions of mmWave networks make the network performance sensitive to the movement of users. We evaluate the performance under different mobility levels of users from 10 to 25 miles/hour, which is a typical speed range from fast walking to running. Figures 7a and 7b show that the network performance for all schemes deteriorate with the increase of mobility level. However, our proposed beam switching scheme can effectively alleviate the impact of mobility and achieve higher performance. The performance improvement becomes larger as the mobility grows. In addition, as the number of users increase, the performances of the two schemes with Mobility Handling degrade more slowly than the others without mobility handling, which indicates that our mobility handling scheme makes the network more resilient when there are a high number of users. The alleviation of the mobility effects allows users to more quickly recover from a sudden disconnection from the AP, thus reducing the end-to-end delay and improving the throughput. These in turn help to better exploit resources available for higher transmissions performance. When the average moving speed of users is 25 miles/hour: (1) In terms of end-to-end delay, Proposed-BUS-MH outperforms Baseline-802.11ay-like and Proposed-BUS by 40.67% and 19.05%, respectively. The improvement for Proposed-BSMSR-MH is 34.86% and 8.8%, respectively; (2) Proposed-BUS-MH has network throughput higher than Baseline-802.11ay-like and Proposed-BUS by 45.54% and 9.27%, respectively, and Proposed-BSMSR-MH 59.61% and 16.17%, respectively.

E. Effects of Beacon Interval

We also studied the impacts of Beacon Interval by varying its length from 50ms to 200ms in the multi-user multi-beam scenario with user mobility. Figures 8a and 8b depict how the end-to-end delay and network throughput changes with the



(a) Delay (b) Network Throughput

Fig. 8. Effect of beacon interval.

Beacon Interval (BI) length, respectively. As the BI length increases, the network performance is generally improved for all three schemes in comparison, which can be explained by more time available for transmissions with reduced beam alignment overhead and fewer resource contentions among users.

In addition, our proposed schemes achieve higher lift than Baseline-802.11ay-like mainly because (1) they cope with multi-user multi-beam scenarios better thus benefit more from a longer BI, and (2) they have better mechanisms to deal with increased user movement in a longer time period. In terms of end-to-end delay, Proposed-BUS-MH, Proposed-BSMSR-MH, and Baseline-802.11ay-like improved by 27.10%, 21.32%, and 7.59%. With regard to network throughput, the lift is 12.08%, 15.88%, and 9.40%, respectively.

It is also observed that our proposed schemes consistently outperform in different choices of Beacon Interval length. When Beacon Interval length is 50ms, Proposed-BUS-MH and Proposed-BSMSR-MH perform better in delay than Baseline-802.11ay-like by 30.32% and 29.41%, respectively. In terms of network throughput, the performance lift is 41.24% and 48.18%, respectively.

VII. Conclusion

With its potential of supporting multi-Gbps data transmissions, millimeter-wave technique is a promising candidate for next generation wireless communications. Despite the huge potential of multi-beam communications, there are very limited efforts on the MAC design for multi-beam mmWave networks. This paper addresses the need of a low-cost multi-user multi-beam training scheme with the simultaneous training and multi-resolution adaptive block-sparse channel estimation for fine-beam alignment. We also jointly allocate radio resources for beam training and data transmissions by designing two virtual scheduling schemes to efficiently schedule concurrent multi-user multi-beam transmissions based on user application types and demands, and incorporate a flexible beam switching scheme for fast disconnection recovery in the presence of mobility and channel dynamics. Simulation results show the significant benefits of our proposed design compared with 802.11ay-based and other schemes and also the tradeoffs among various design considerations in the proposed framework.

References

[1] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3437–3458, Oct. 2015.

[2] *IEEE Standard for Information Technology–Telecommunications and Information Exchange Between Systems–Local and Metropolitan Area Networks–Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band*, IEEE Standard 802.11ad-2012 (Amendment to IEEE Std 802.11-2012, as amended by IEEE Std 802.11ae-2012 IEEE Std 802.11aa-2012), 2012, pp. 1–628.

[3] S. Sur, V. Venkateswaran, X. Zhang, and P. Ramanathan, "60 GHz indoor networking through flexible beams: A link-level profiling," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Modeling Comput. Syst.*, New York, NY, USA, Jun. 2015, pp. 71–84.

[4] T. Baykas *et al.*, "IEEE 802.15.3c: The first IEEE wireless standard for data rates over 1 Gb/s," *IEEE Commun. Mag.*, vol. 49, no. 7, pp. 114–121, Jul. 2011.

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

[6] Z. Yang, P. H. Pathak, Y. Zeng, and P. Mohapatra, "Sensor-assisted codebook-based beamforming for mobility management in 60 GHz WLANs," in *Proc. IEEE 12th Int. Conf. Mobile Ad Hoc Sensor Syst.*, Oct. 2015, pp. 333–341.

[7] P. Schniter and A. Sayeed, "Channel estimation and precoder design for millimeter-wave communications: The sparse way," in *Proc. 48th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2014, pp. 273–277.

[8] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2909–2913.

[9] D. Ramasamy, S. Venkateswaran, and U. Madhow, "Compressive adaptation of large steerable arrays," in *Proc. IEEE ITA*, Feb. 2012, pp. 234–239.

[10] A. Zhou, X. Zhang, and H. Ma, "Beam-forecast: Facilitating mobile 60 GHz networks via model-driven beam steering," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[11] X. An, S. Zhang, and R. Hekmat, "Enhanced MAC layer protocol for millimeter wave based WPAN," in *Proc. IEEE 19th Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2008, pp. 1–5.

[12] M. X. Gong, R. Stacey, D. Akhmetov, and S. Mao, "A directional CSMA/CA protocol for mmWave wireless pans," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2010, pp. 1–6.

[13] H. Shokri-Ghadikolaei, L. Gkatzikis, and C. Fischione, "Beam-searching and transmission scheduling in millimeter wave communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 1292–1297.

[14] *IEEE Standard for Information Technology–Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks–Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Enhanced Throughput for Operation in License-Exempt Bands Above 45 GHz*, IEEE Standard 802.11ay-2021 (Amendment to IEEE Std 802.11-2020 as amendment by IEEE Std 802.11ax-2021), 2021, pp. 1–768.

[15] M. Ettorre, R. Sauleau, and L. Le Coq, "Multi-beam multi-layer leaky-wave SIW pillbox antenna for millimeter-wave applications," *IEEE Trans. Antennas Propag.*, vol. 59, no. 4, pp. 1093–1100, Apr. 2011.

[16] S. Sun, G. R. MacCartney, Jr., M. K. Samimi, S. Nie, and T. S. Rappaport, "Millimeter wave multi-beam antenna combining for 5G cellular link improvement in New York City," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 5468–5473.

[17] R. Pal, K. V. Srinivas, and A. K. Chaitanya, "A beam selection algorithm for millimeter-wave multi-user MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 852–855, Apr. 2018.

[18] Y. Ghasempour, M. K. Haider, C. Cordeiro, D. Koutsonikolas, and E. Knightly, "Multi-stream beam-training for mmWave MIMO networks," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, Oct. 2018, pp. 225–239.

[19] C. Kim, T. Kim, and J.-Y. Seol, "Multi-beam transmission diversity with hybrid beamforming for MIMO-OFDM systems," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 61–65.

[20] Y. Zhao, X. Xu, Y. Su, L. Huang, X. Du, and N. Guizani, "Multi-user MAC protocol for WLANs in mmWave massive MIMO systems with mobile edge computing," *IEEE Access*, vol. 7, pp. 181242–181256, 2019.

[21] S. Sur, I. Pefkianakis, X. Zhang, and K.-H. Kim, "Towards scalable and ubiquitous millimeter-wave wireless networks," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, Oct. 2018, pp. 257–271.

[22] Y. Ghasempour, C. R. C. M. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11ay: Next-generation 60 GHz communication for 100 Gb/s Wi-Fi," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 186–192, Dec. 2017.

[23] Y. Ghasempour, M. K. Haider, and E. W. Knightly, "Decoupling beam steering and user selection for mu-MIMO 60-GHz WLANs," *IEEE/ACM Trans. Netw.*, vol. 26, no. 5, pp. 2390–2403, Oct. 2018.

[24] M. R. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[25] H. Hassanieh, O. Abari, M. Rodriguez, M. Abdelghany, D. Katabi, and P. Indyk, "Fast millimeter wave beam alignment," in *Proc. Conf. ACM Special Interest Group Data Commun.*, New York, NY, USA, Aug. 2018, pp. 432–445.

[26] L. X. Cai, L. Cai, X. Shen, and J. W. Mark, "REX: A randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 113–121, Jan. 2010.

[27] J. Zhao, D. Xie, X. Wang, and A. Madanayake, "Towards efficient medium access for millimeter-wave networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2786–2798, Dec. 2019.

[28] Y. Chen, X. Wang, and L. Cai, "HOL delay based scheduling in wireless networks with flow-level dynamics," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4898–4903.

[29] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[30] C. G. Khatri and C. R. Rao, "Solutions to some functional equations and their applications to characterization of probability distributions," *Sankhyā, Indian J. Statist. A*, vol. 30, no. 2, pp. 167–180, Jun. 1968.

[31] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

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

Jie Zhao (Member, IEEE) received the B.S. degree in telecommunications engineering from the Huazhong University of Science and Technology, Wuhan, China, and the Ph.D. degree in electrical engineering from the State University of New York at Stony Brook, NY, USA. His research interests include sparse data analytics, machine learning, millimeter-wave communications, cognitive radio networks, and networked sensing and detection.

Xin Wang (Member, IEEE) received the B.S. degree in wireless communications engineering and the M.S. degree in telecommunications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, and the Ph.D. degree in electrical and computer engineering from Columbia University, New York, NY, USA. She is currently an Associate Professor with the Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY, USA. Before joining Stony Brook, she was a member of technical staff in the area of mobile and wireless networking with Bell Labs Research, Lucent Technologies, NJ, USA, and an Assistant Professor with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, USA. Her research interests include algorithm and protocol design in wireless networks and communications, mobile and distributed computing, networked sensing and detection, and machine learning. She received the NSF Career Award in 2005 and the ONR Challenge Award in 2010. She has served in executive committee and technical committee of numerous conferences and funding review panels, and serves as an Associate Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING.