Towards Efficient Medium Access for Millimeter-Wave Networks

Jie Zhao, *Student Member, IEEE*, Dongliang Xie, *Member, IEEE*, Xin Wang, *Member, IEEE*, and Arjuna Madanayake, *Member, IEEE*

***Abstract*— The need of highly directional communications at mmWave frequencies introduces high overhead for beam training and alignment, which makes the medium access control (MAC) a grand challenge. To harvest the gain for high performance transmissions in mmWave networks, we propose an efficient and integrated MAC design with the concurrent support of three closely interactive components: 1) an accurate and low-cost beam training methodology with a) multiuser, multi-level, bi-directional coarse training for fast user association and beam alignment and b) adaptive fine beam training with compressed channel measurement and multi-resolution block-sparse channel estimation in response to the channel condition and the learning from past measurements; 2) an elastic virtual resource scheduling scheme that jointly considers beam training, beam tracking and data transmissions while enabling burst data transmissions with the concurrent allocation of transmission rate and duration; and 3) a flexible and efficient beam tracking strategy to enable stable beam alignment with beamwidth adaptation and mobility estimation. Compared with literature studies, our performance results demonstrate that our design can effectively reduce the training overhead and thus significantly improve the throughput. Compared to 802.11ad, the training overhead can be reduced more than 60%, and the throughput can be more than 75% higher. In low SNR case, the throughput gain can be more than 90%. Our scheme can also achieve about 50% higher throughput in the presence of user mobility.**

***Index Terms*— Millimeter wave, directional MAC, directional antenna, resource allocation, channel estimation.**

I. Introduction

MILLIMETER-WAVE (mmW or mmWave) communication is receiving tremendous interest from academia, industry and federal agencies as a promising technique to provide Gigabit data rate demanded by the exponential growth of wireless applications. A key challenge of mmWave communications is the low signal range as a result of the large isotropic path loss. Fortunately, the small wavelength of mmWave signals also enables a large number of antennas to be placed in small dimensions (e.g. at the base station, in the skin of a cellphone, or even within a chip), which provides a high beamforming gain to compensate for the big path loss.

The nature of highly directional transmissions in mmWave bands, however, makes the design of medium access control schemes a grand challenge [1]. New users have difficulty of associating with a small cell base station or access point (AP). If both AP and user devices are configured directionally, it could take an extremely long time to connect them and align their beams. In the measurements of basic IEEE 802.11ad [2] transmission [3], the latency for AP discovery is 5ms to 1.8s for a static client and up to 12.9s for a mobile client. On the other hand, omni-directionally transmitting/receiving training signals for beam alignment may lead to range much lower than that of data transmissions. The problem is made even harder when there are a large number of beam directions and users, and the channel reciprocity principle breaks in the presence of human blockage and environment dynamics [3].

To alleviate the training overhead, codebook-based adaptive-beam training [4]–[6] divide directions into different granularity levels. At each level, training signals are sent to all directions within a selected angular range, and a feedback message is needed to select the best beam. The feedback overhead and delay would be very high with the use of multiple rounds of feedbacks (with each round corresponding to a granularity level) and the competitions in multi-user feedbacks along each trained direction. Codebook-based scheme has been taken by 802.11ad. Alternatively, compressed sensing (CS) is exploited to estimate the sparse mmWave channels with training signals sent along random directions within the whole angular range [7]–[9]. Although the number of training directions is reduced, the channel reconstruction complexity increases exponentially with the number of measurement samples.

The big training overhead will translate into significant throughput reduction. More frequent signaling would be needed to track the directional transmissions when there exist higher channel dynamics and user mobility [10]. Despite the large amount of effort made to more efficiently find the best beam directions or allocate radio resources [11]–[13], the two are normally decoupled. Different from conventional wireless communications where only data transmissions are considered in radio resource allocations, it is necessary to concurrently schedule radio resources for channel training,

Manuscript received April 30, 2019; revised September 9, 2019; accepted October 10, 2019. The work of X. Wang was supported by NSF under Grant ECCS 1731238. The work of A. Madanayake was supported by NSF under Grant ECCS 1854798. *(Corresponding author: Dongliang Xie.)*

J. Zhao and X. Wang are with the Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794 USA (e-mail: jie.zhao@alumni.stonybrook.edu; x.wang@stonybrook.edu).

D. Xie is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: xiedl@bupt.edu.cn).

A. Madanayake is with the Department of Electrical and Computer Engineering, Florida International University, Miami, FL 33174 USA (e-mail: amadanay@fiu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2019.2947924

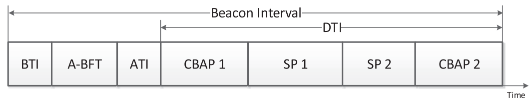


Fig. 1. BI structure in IEEE 802.11ad.

data transmissions and beam tracking, in the face of dynamics of channel conditions, user population, locations, and traffic.

In light of the challenges (training overhead, frequent signaling, resource allocation, network dynamics) above, our aim is to design an efficient and integrated MAC scheme for high performance mmWave network transmissions with the concurrent support of three closely interactive components:

a) Accurate and light-weight beam training with 1) *multi-user, multi-level, bi-directional* coarse training for fast user association and beam alignment, and 2) fine beam training with *multi-resolution block-sparse* channel estimation and compressed beam measurement, with adaptation to channel conditions and past measurement results.

b) Self-adaptive *virtual resource scheduling* to determine both user transmission opportunities and durations for facilitation of various traffic types, while trading off between beam training and data transmissions for an overall high network performance.

c) Effective beam tracking for more stable beam alignment with flexible *beamwidth adaptation* and *mobility estimation* to cope with link failures due to user motions or channel dynamics.

The rest of this paper is organized as follows. After briefly reviewing background and related work in Section II, we present our fast association and multi-level beam training approach in Section III. We further propose our multi-resolution block-sparse channel estimation technique and fine beam training design in Section IV, followed by Section V, where we develop our flexible resource scheduling and beam tracking schemes. Finally, we analyze the simulation results in Section VI, and conclude the paper in Section VII.

II. Background, Related Work, and Basic Framework

A. Background

The standards IEEE 802.11ad [2] and IEEE 802.15.3c [14] are proposed at physical layer (PHY) and medium access control layer (MAC) to enable operation in frequencies around 60 GHz mmWave band. Figure 1 shows the MAC layer superframe of IEEE 802.11ad protocol, referred as Beacon Interval (BI)). AP provides the basic timing for DEVs through beacon and announce frames, such as Beacon transmission interval (BTI) to transmit one or more beacons in different directions, Association beamforming training (A-BFT) for devices to communicate with AP and train their antenna beams and Announcement transmission interval (ATI) for AP to exchange management information with associated devices. A data transmission interval (DTI) contains service periods (SPs) to transmit data using time division multiple access (TDMA) and contention-based access periods (CBAPs) for devices to compete in transmissions using Carrier sense multiple access with collision avoidance (CSMA/CA).

Although 802.11ad provides a basic MAC framework and signaling sequences, there is no specific consideration for more efficient directional finding and transmissions. With the concurrent consideration of *beam training* and *resource allocation*, we propose a detailed design of the MAC scheme with three major components: quick and low-cost AP association and beam training, adaptive and joint scheduling of radio resources for training and transmission under channel and demand changes, and efficient beam tracking during mobility. To facilitate practical application of our work, we can fit our schemes into the 802.11ad framework, although our schemes are general and do not depend on any protocols.

B. Related Work

To compensate for the high path loss, codebook-based beamforming schemes have been proposed [4]–[6] and taken by 802.11ad. However, the signaling overhead and delay would be very high to train a large number of beams and in the presence of many users.

As an alternative, compressed sensing (CS) techniques have been proposed to estimate mmWave channels to facilitate beam alignment [7]–[9], [15]–[18], taking advantage of the sparse feature of channels at mmWave frequencies. These studies, however, did not fully consider the clustering of transmission paths [19] in channel reconstruction. Instead, taking into account the path clustering effect, we model our channel as block-sparse and propose a *multi-resolution block-sparse* method to more accurately estimate the channel. As an additional benefit, our proposed method allows for concurrent use of compressed measurements from different levels to improve the accuracy of reconstructing CS channel and reduce the total number of samples, which further reduces the computational complexity.

Various efforts are made to only allocate radio resources in mmWave networks [11]–[13], [20], and existing work mostly focus on scheduling concurrent device-to-device communications in Wireless Personal Area Networks. Instead, we investigate uplink/downlink transmission scheduling between base station/access point and devices. We concurrently and adaptively schedule radio resources for channel training, data transmissions and beam tracking. Rather than coordinating users to transmit in each slot [21], our virtual scheduling enables the burst transmissions of packets, a major format to transmit high volume data in mmWave communications. The joint determination of transmission resources and duration makes the scheduling problem much harder, and is often bypassed by literature work.

User mobility and environmental dynamics makes it more difficult to achieve beam alignment in mmWave networks, and beam tracking is often needed to avoid transmission interruption. Based on the observation that 60 GHz channel profiles at nearby locations are highly-correlated, Zhou *et al.* [10] propose a beam-forecast scheme to reconstruct the channel profile and predict new optimal beams. Highly relying on a specific geometry model, the prediction accuracy may be compromised in practical networks. Authors in [22] design, implement and evaluate MOCA, a protocol for Mobility

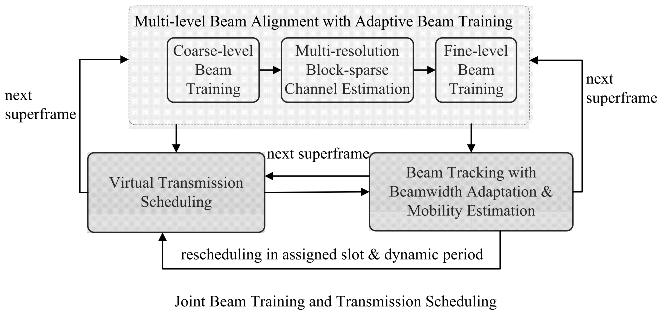


Fig. 2. Framework overview.

resilience and Overhead Constrained Adaptation for directional 60 GHz links, where mobility-induced link breakage is quickly identified and recovered with the change of beamwidth and data rate. The new beamwidth is selected from a predetermined fixed set, and the throughput will reduce when using a larger beamwidth for transmissions to alleviate the impacts of mobility. Rather than using a larger beamwidth for compromised transmission quality, to effectively handle channel dynamics and user mobility at low cost, we flexibly adapt the beamwidth for rapid reconnection in case of link failures and search for the new fine beam direction based on the estimation of user mobility levels.

C. Basic MAC Framework

To address the challenge of mmWave transmissions, we propose a MAC framework with integrated beam alignment and transmission scheduling in Fig. 2. To reduce the big overhead for beam alignment, we divide the training process into coarse level and fine level. Beams are first generated following two-level codebooks to find the possible signal directions at coarse angular ranges, with different strategies to reduce the signaling overhead. Then the finer beam training is pursued with a selected number of additional training signals randomly transmitted within the angular ranges detected with good signal quality. The mmWave channel is estimated following compressed sensing at multiple resolution levels, and the channel condition at coarser level is applied to determine the weights for the finer level to improve the channel estimation accuracy and speed. Based on the channel conditions, AP and devices are scheduled for higher transmission performance and efficient beam tracking to cope with network dynamics.

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

- First, 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) feedback aggregation to reduce signaling overhead, (2) traffic-aware adaptation of the number of contention slots, (3) compressive measurement with novel block-sparse estimation of the mmW channel at hierarchical beam resolution and (4) elastic fine beam training that jointly works with transmission scheduling in response to channel condition and learning from past training results.
- Second, to efficiently manage radio resources, we propose a virtual transmission scheduling scheme with (1) concurrent determination of transmission opportunities and duration while trading off among beam training, data transmissions and beam tracking, (2) virtual slot aggregation adaptive to heterogeneous traffic types, user demands and resource availability.
- Third, to ensure low-overhead beam alignment and alleviate link failures under user mobility and channel dynamics, we propose an efficient beam tracking scheme that achieves quick user rediscovery and disconnection remedy by (1) dynamic beamwidth adjustment and (2) flexible user movement prediction.

III. AP Association and Multi-Level Beam Alignment

To harvest the gain of mmWave communications, it calls for highly efficient training schemes to enable lower-overhead thus faster AP association and beam alignment.

The AP association and multi-resolution beam alignment component in our basic MAC framework is shown in Figure 2. To avoid high feedback overhead as in conventional codebook-based schemes, we consider two levels of *coarse training* to quickly associate users with APs. Rather than only concentrating on beamforming uplink or downlink, or assuming the existence of channel reciprocity, we consider *bi-directional* training between AP and devices. Finally, to align beams at the finest resolution desired, we will further exploit *multi-resolution* and *block-sparse* channel estimation, which will be introduced in details in Section IV.

In this section, we first present the two-level coarse training and then provide the analysis on the impacts of beam resolution on transmission range.

We use some terms and major signaling flows from 802.11ad to facilitate better understanding, and also provide the possibility of incorporating our design into the 802.11ad framework. Our scheme, however, is general and not constrained to run within 802.11 networks. The differences of our design from 802.11ad are: (a) we emphasize the coordination of training between uplink and downlink and the overhead reduction exploiting the information from the previous round of signaling, (b) we allow AP to transmit feedbacks in a batch for devices within one sector to reduce the header overhead, and (c) we determine the number of contention slots in each AP sector according to the number of associated users known from the previous signaling procedures, which alleviates the collision while avoiding the waste of radio resources.

A. Multi-Level Beam Training

We apply three levels of beamwidth following the terms of 802.11ad: quasi-omni-directional level (QOL), sector beam sweep (SBS), and fine beam steering (FBS). An example of the hierarchical beam levels is given in Figure 3. At the quasi-omni-directional level, the beamwidth will be configured to the widest possible allowed by the system to alleviate the deafness problem in receiving.

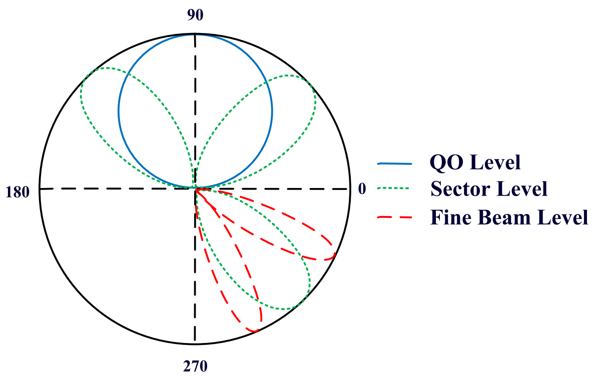


Fig. 3. Hierarchical beam levels example.

The fine beam is the desired beamwidth to use for a mmWave system to achieve high data rates.

We use antenna directions and antenna weight vectors (AWVs) interchangeably, although an AWV not only determines the main-lobe direction of the beam but also the beamwidth. We also use device and user interchangeably for ease of presentation. We consider the association and beam alignment between devices and the AP in a cell. Due to the space limitation, we won't discuss device-to-device communications. Our 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 each of its QOL directions to find the best sending SBS sector and receiving QOL direction. Then AP configures itself to listen from each SBS direction. Devices successfully receiving beacons from AP will contend for response within S_1 slots. To facilitate reverse channel training, in each SBS direction that AP listens to, a device will send along all its QOL directions the following information: its association request, the best SBS sector for AP to transmit, and its best receiving QOL direction. A device will then prepare itself at the best receive QOL direction. To reduce the feedback overhead, rather than sending a feedback to every device right away as in 802.11ad, we allow AP to send an aggregated 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; and uplink, the best QOL sending direction from a device and the best receive 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 sweeps along the set of SBS directions within the angular range of its best QOL receive direction. To determine the best transmission sector from a device, AP only listens to responses in the best receive sectors selected by devices earlier. In each AP receiving sector, multiple associated devices will contend to get a response slot among S_2 slots. Instead of using an equal number of contention slots for each AP receiving sector as in 802.11ad, we set S_2 for each sector proportional to the number of associated users that is learned from Step 1. This will reduce the collisions in the sectors with more users while avoiding wasting time slots unnecessarily in sectors with very few 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 along 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.

Step 3 (Determining the Best Fine-Level Transmission and Receiving Directions): Finally, AP and devices need further training to find the best 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 compressive measurement and block-sparse estimation of the mmWave channel, which will be introduced later in Section IV.

B. Analysis of Beamwidth and Transmission Range

To analyze the directive gains of the antennas, we exploit a sectored antenna model which considers the front-to-back ratio, and the half-power beamwidth. The gains remain the same for all angles in the main lobe and are smaller in the side lobe in the ideal sector antenna pattern. Let $\theta^{\mathbf{u}}$ and $\theta^{\mathbf{v}}$ be the angles that are deviated from the boresight of the steering angles of TX and RX, $B_\theta^{\mathbf{u}}$ and $B_\theta^{\mathbf{v}}$ be beamwidths of the TX and RX antenna patterns, we have the directive gain of TX

$$G^{\mathbf{u}}\left(\theta^{\mathbf{u}}, B_\theta^{\mathbf{u}}\right) = \begin{cases} \dfrac{2\pi - (2\pi - B_\theta^{\mathbf{u}})z}{B_\theta^{\mathbf{u}}}, & \text{if } |\theta^{\mathbf{u}}| \le \dfrac{B_\theta^{\mathbf{u}}}{2} \\ z, & \text{otherwise}, \end{cases} \quad (1)$$

where $0 \le z < 1$ is the gain in the side lobe, with $z \ll 1$ for narrow beams. Likewise, the directive gain of RX can be expressed as

$$G^{\mathbf{v}}\left(\theta^{\mathbf{v}}, B_\theta^{\mathbf{v}}\right) = \begin{cases} \dfrac{2\pi - (2\pi - B_\theta^{\mathbf{v}})z}{B_\theta^{\mathbf{v}}}, & \text{if } |\theta^{\mathbf{v}}| \le \dfrac{B_\theta^{\mathbf{v}}}{2} \\ z, & \text{otherwise}. \end{cases} \quad (2)$$

The number of antennas impacts the finest beamwidth to achieve thus the maximum gain of the beam. The channel gain $G^{\mathbf{H}}(d)$ is affected by the TX-RX distance d. For a beam with the TX beamwidth $B_\theta^{\mathbf{u}}$ and RX beamwidth $B_\theta^{\mathbf{v}}$, let $G^{\mathbf{u}}(B_\theta^{\mathbf{u}})$ and $G^{\mathbf{v}}(B_\theta^{\mathbf{v}})$ be the TX and RX antenna gains, then we have the Signal to Noise Ratio (SNR) as

$$SNR(B_\theta^{\mathbf{u}}, B_\theta^{\mathbf{v}}, d) = \frac{p^T G^{\mathbf{u}}\left(B_\theta^{\mathbf{u}}\right) G^{\mathbf{H}}(d) G^{\mathbf{v}}\left(B_\theta^{\mathbf{v}}\right)}{N_0}, \quad (3)$$

where p^T indicates the transmitter power and N_0 the noise power. Obviously, beamwidth impacts the effectiveness of the beamforming and consequently the transmission range.

Compared to data transmissions at the fine beam level, the coarse-level signal transmission has a lower range. However, earlier measurement studies [19] indicate that directional beamforming gain at either one side of TX or RX may be enough to combat the additional channel fading in mmWave band. We also exploit the gain at both sender and receiver to reduce the link budget loss. Additionally, the signaling message has the rate much lower than the data, and lower-bit coding would allow the coding gain to further increase the range.

IV. Multi-Resolution Block-Sparse mmWave Channel Estimation

Upon the completion of coarse-level training in Section III, the next measure to be taken is discovering the best fine-level beam directions, which 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 a large volume of fine beam pairs as in 802.11ad or introducing more levels of training at high feedback cost, we will explore the use of compressive channel estimation to facilitate low-cost beam training.

Figure 2 shows the interactions among our multi-resolution block-sparse channel estimation module and beam training component at different levels. Different from conventional CS-based channel estimation schemes [7]–[9], [15]–[18] that only consider the channel sparsity, our contributions lie in the following aspects: (a) we further explore the block-sparse feature in mmWave channels as a result of transmission path clustering for better channel estimation in Section IV-A and (b) we iteratively exploit our block-sparse channel estimation at hierarchical beam resolution for higher accuracy and lower computational complexity in Section IV-B.

A. Block-Sparse Channel Estimation

We will now describe how we exploit the path clustering feature of mmWave channels and develop the solution to channel estimation as block-sparse channel reconstruction.

For ease of presentation, we consider only the azimuth and neglect the elevation in this paper. Implementations that facilitate both horizontal and vertical beamforming can be built on top of our design. While our proposed design can be used for any kind of antenna arrays, without loss of generality, we adopt uniform linear arrays (ULAs) in this work.

In [19], the mmWave channel is found to be not only sparse but also path clustering according to the real-world measurements in New York City (NYC), from which a statistical mmWave model is derived. We adopt this channel model, where the channel is composed of K clusters within each there are L subpaths, then with the number of transmitting and receiving antennas to be N_{tx} and N_{rx}, the channel matrix can be written as

$$\mathbf{H} = \sum_{k=1}^{K}\sum_{\ell=1}^{L} a_{k\ell} \cdot \mathbf{D}_{rx}(\theta_{k\ell}^{rx}) \cdot \mathbf{D}_{tx}^{H}(\theta_{k\ell}^{tx}), \tag{4}$$

where $a_{k\ell}$ is the complex path gain for a path ℓ ($\ell = 1, 2, \ldots, L$) in the cluster k ($k = 1, 2, \ldots, K$), with $k\ell$ jointly corresponding to the ℓ-th sub-path in the k-th cluster. For the sake of consistency, in this work, we use the terms path and sub-path interchangeably. $\theta_{k\ell}^{tx}$ and $\theta_{k\ell}^{rx}$ denote the angle of departure (AoD) and the angle of arrival (AoA) for the corresponding path.

$\mathbf{D}_{tx}(\theta_{k\ell}^{tx})$, the TX antenna's directional response column vector ($N_{tx} \times 1$ dimension) for the sub-path at the angle of departure $\theta_{k\ell}^{tx}$, is expressed as:

$$\begin{aligned}
&\mathbf{D}_{tx}(\theta_{k\ell}^{tx}) \\
&= \left[D^{(1)}(\theta_{k\ell}^{tx}), D^{(2)}(\theta_{k\ell}^{tx}), \ldots, D^{(m)}(\theta_{k\ell}^{tx}), \ldots, D^{(N_{tx})}(\theta_{k\ell}^{tx})\right] \\
&= \left[1, e^{j\cdot 1\cdot w_{k\ell}^{tx}}, e^{j\cdot 2\cdot w_{k\ell}^{tx}}, \ldots, e^{j\cdot (N_{tx}-1)\cdot w_{k\ell}^{tx}}\right]^{T},
\end{aligned} \tag{5}$$

where $D^{(m)}(\theta_{k\ell}^{tx})$ is from antenna basics, the spatial frequency $w_{k\ell}^{tx}$ can be written in terms of AoDs, as $w_{k\ell}^{tx} = \frac{2\pi d_t}{\lambda} \sin\theta_{k\ell}^{tx}$. d_t is the distances between two adjacent antenna elements in the ULAs in the TX. $\lambda = \frac{c}{f}$ is wavelength in meters. f is the carrier frequency of the signal in Hz, c is the speed of light (3×10^8 meters/sec).

$\mathbf{D}_{rx}(\theta_{k\ell}^{rx})$, the RX antenna's directional response column vector ($N_{rx} \times 1$ dimension) for the path at an angle of arrival $\theta_{k\ell}^{rx}$, can be similarly expressed.

We now use a concatenated column vector $\mathbf{a}$ ($1 \times KL$) to denote the complex path gains. Then

$$\begin{aligned}
\mathbf{a} = [&\underbrace{a_{11}, a_{12}, \ldots, a_{1L}}_{\text{cluster }1}, \underbrace{a_{21}, a_{22}, \ldots, a_{2L}}_{\text{cluster }2}, \ldots, \\
&\underbrace{a_{K1}, a_{K2}, \ldots, a_{KL}}_{\text{cluster }K}]^{T},
\end{aligned} \tag{6}$$

Note $\mathbf{a}$ is concatenated in a manner that the first L elements are for the first cluster, and the next L elements are for the second cluster and so on. As a result of path clustering, the mmWave channel in (6) is seen to have the block properties. That is, $\mathbf{a}$ is not only sparse, but also block-sparse.

The major task of mmW channel estimation in our work is to estimate $\mathbf{a}$ efficiently. To achieve this, we first rewrite (4) in matrix format as

$$\mathbf{H} = \mathbf{D}_R \operatorname{diag}(\mathbf{a}) \mathbf{D}_T^H, \tag{7}$$

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

$$\mathbf{D}_T = [\mathbf{D}_{tx}(\theta_{11}^{tx}), .., \mathbf{D}_{tx}(\theta_{1L}^{tx}), .., \mathbf{D}_{tx}(\theta_{K1}^{tx}), .., \mathbf{D}_{tx}(\theta_{KL}^{tx})], \tag{8}$$

$$\mathbf{D}_R = [\mathbf{D}_{rx}(\theta_{11}^{rx}), .., \mathbf{D}_{rx}(\theta_{1L}^{rx}), .., \mathbf{D}_{rx}(\theta_{K1}^{rx}), .., \mathbf{D}_{rx}(\theta_{KL}^{rx})]. \tag{9}$$

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:

$$\mathbf{Y}^{Q\times P} = \mathbf{V}^H \mathbf{H} \mathbf{U} \circ \mathbf{S} + \mathbf{E}, \tag{10}$$

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

$$\mathbf{V}^{N_{rx}\times Q} = [\mathbf{v}_1, .., \mathbf{v}_q, .., \mathbf{v}_Q], \quad \mathbf{U}^{N_{tx}\times P} = [\mathbf{u}_1, .., \mathbf{u}_p, .., \mathbf{u}_P]. \tag{11}$$

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

$$\begin{aligned} \mathbf{y} &= \mathrm{vec}(\mathbf{R}) = \sqrt{A}\mathrm{vec}(\mathbf{V}^H \mathbf{H} \mathbf{U}) + \mathrm{vec}(\mathbf{E}) \\ &\overset{\text{Theorem 1 [23]}}{=\!=\!=\!=} \sqrt{A}(\mathbf{U}^T \otimes \mathbf{V}^H)\mathrm{vec}(\mathbf{H}) + \mathrm{vec}(\mathbf{E}) \\ &\overset{\text{Proposition 1 [24]}}{=\!=\!=\!=} \sqrt{A}(\mathbf{U}^T \otimes \mathbf{V}^H)\boldsymbol{\Psi}\mathbf{a} + \mathrm{vec}(\mathbf{E}) \\ &= \boldsymbol{\Phi}\boldsymbol{\Psi}\mathbf{a} + \mathrm{vec}(\mathbf{E}) = \mathbf{A}\mathbf{a} + \mathrm{vec}(\mathbf{E}), \end{aligned} \tag{12}$$

where $\boldsymbol{\Psi} = \mathbf{D}_R^* * \mathbf{D}_T$(Khatri-Rao product) is the basis matrix, $\boldsymbol{\Phi} = \sqrt{A}(\mathbf{U}^T \otimes \mathbf{V}^H)$ (Kronecker product) is the measurement matrix (determined by TX and RX beam training directions). In the derivation, we have used Theorem 1 [23] and Proposition 1 [24] as follows:

Theorem 1: $\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{X})$.

Proposition 1: $\mathrm{vec}(\mathbf{H}) = \boldsymbol{\Psi}\mathbf{a}$, where $\boldsymbol{\Psi} = \mathbf{D}_R^* * \mathbf{D}_T$ (Khatri-Rao product).

In order to differentiate between the estimated channel and the actual channel $\mathbf{a}$, we now refer the estimated $\mathbf{a}$ as $\mathbf{x}$. Replacing the vector $\mathbf{a}$ in the Eq. (12) with $\mathbf{x}$, we have the compressed sensing form $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where $\mathbf{y}$ is the measurement results, $\mathbf{A}$ 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:

$$\min \sum_{i=1}^{n} \|\mathbf{X}_i\|_2, \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{y}, \ \mathbf{x} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n], \tag{13}$$

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. Figure 4 depicts the block-sparse model of (13). A typical solution algorithm for (13) is presented in Sec. IV of [25] as the "Recovery of block-sparse signals" Algorithm. After recovering $\mathbf{x}$, the virtual channel $\mathbf{H}$ can be estimated as in Eq. (7).

B. Multi-Resolution Channel Estimation

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

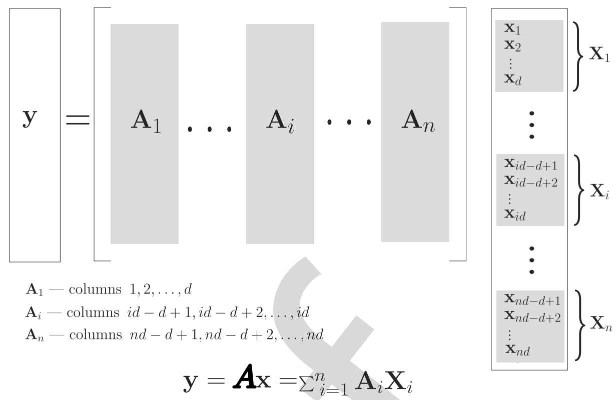


Fig. 4. Block-sparse model.

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 $N_{tx}^g N_{rx}^g \times 1$ ($\mathrm{vec}(\mathbf{H})$). 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

$$\boldsymbol{\Psi} = IDFT_{N_{tx}}^* * IDFT_{N_{rx}}, \tag{14}$$

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

$$\dim(\mathbf{H}_{QOL}) < \dim(\mathbf{H}_{SBS}) < \dim(\mathbf{H}_{FBS}). \tag{15}$$

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 levels of resolution with weighting factors* to take advantage of the multi-level training samples we have obtained in Section III. To be more specific, we transform (13) into the following weighted recovery problem under the same constraints:

$$\min \sum_{i=1}^{n} w_i \|\mathbf{X}_i\|_2, \tag{16}$$

where w_i is the weighting factor for block $\mathbf{X}_i$, and is set to be the inverse of the number of non-zero elements (supports) contained in the signal block $\mathbf{X}_i$. By assigning smaller weights to the blocks that consist of more non-zero elements and vice versa, the optimization will penalize more heavily those blocks with larger weights and fewer supports, thus leaving more residual signals to be reconstructed for those blocks that contain more information (i.e., with small weights and more supports). In this way, our block weighting approach improves the CS reconstruction performance and is exploited in the multi-resolution (i.e. different block sizes) channel estimation process to be presented later in this section. Although block weights are introduced, (16) can still be solved by the algorithm from [25], which is mentioned earlier as a solution to (13), by substituting $\mathbf{X}_i$ in (13) with $w_i\mathbf{X}_i$.

The major difference between (13) and the proposed (16) is that we set w_i to the inverse of the magnitude of the coarse direction reconstructed from the previous step, where channel is estimated in a more coarse resolution (i.e., the block in the current step corresponds to the resolution used in the previous step). By assigning smaller weights to the blocks that have higher recovered magnitude in the previous step, the optimization will penalize more heavily those blocks with less information, thus leaving more residual signals to be reconstructed for the blocks that contain more information. With the channel estimation at multiple resolution in different block sizes, our block weighting approach can improve both the CS reconstruction accuracy and speed.

Rather than directly estimating channels with the CS-based scheme, the use of multi-level of training largely reduces the number of samples needed thus the overhead for CS recovery for channel estimation. Further, compared to the direct finding of all beams with the traditional ℓ_1-norm optimization, the leverage of results from block-sparse CS reconstruction (16) helps to significantly reduce number of iterations needed for the channel estimation process to converge. Therefore, our algorithm can more efficiently run over the practical platforms and devices.

Following the training process, the recursive steps for our multi-resolution channel estimation approach are:

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

Step (b) *SBS channel reconstruction:* After SBS beam training, according to QOL results in Step (a), adjust the weights at the SBS level: the SBS elements contained in QOL blocks with larger magnitude (recovered in the previous step) 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 (b), adjust the weights at the FBS level: the FBS elements contained in SBS blocks with larger magnitude (recovered in the previous step) 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, our multi-resolution block-sparse mmWave channel estimation methodology not only jointly exploits the sparsity and block properties in mmWave channels, but also takes advantage of the multi-level beam training results to significantly reduce the number of measurements. This will further reduce the complexity in recovering the mmWave channel, and speed up the training.

C. Procedures for Fine Beam Training

With the coarse beam training in Section III, AP and devices have known the best transmission and receiving sectors for both downlink and uplink transmissions. We will add the following procedures for compressive fine beam training to Step (3) in Section III:

Step 3.1 (Downlink Fine Beam Training): To facilitate synchronization, each device initially listens at its best receiving sectors to intercept system parameters. For the 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. During the transmission of each fine beam, the set of devices which select the corresponding transmission sector will each listen from Q^R randomly selected fine beam directions in their respective best DEV receiving sectors. After collecting samples from P^TQ^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 selected best receiving sectors, for each associated devices will send uplink feedbacks with the best measured AP TX fine beam, 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. 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 beam pairs.

D. Analysis of Beam Training Overhead

Our beam training involves three levels of beamwidth: QOL, SBS and FBS. We use B_Q^U, B_S^U and B_F^U to represent the AP beamwidth at each level, and use B_Q^V, B_S^V and B_F^V to represent the device beamwidth. We let B_W^U and B_W^V denote the overall angular search space for the AP and the device. We first quantify the training overhead of the beam training scheme discussed in Sections III and IV-C. Let T_p denote the time to transmit a pilot training signal, $T_s = \beta_s T_p$ denote the time duration of a contention slot ($\beta_s \geq 1$). A training signal consists of a sequence of training symbols. As the number of symbols impacts the training time, it can be adapted to trade off between the training time and the gain in finding higher gain channels for higher transmission rates. The overhead in each step of beam training is analyzed as follows:

Step 1: $T_1/T_p = \left\lceil \frac{B_W^U}{B_S^U} \right\rceil \cdot \left\lceil \frac{B_W^V}{B_Q^V} \right\rceil + \beta_s S_1 \left\lceil \frac{B_W^V}{B_Q^V} \right\rceil \left\lceil \frac{B_W^U}{B_S^U} \right\rceil + \alpha_t \left\lceil \frac{B_W^U}{B_S^U} \right\rceil$; S_1, the number of device response slots, is an integer ($S_1 \geq 1$) and can be adapted according to the traffic pattern over the previous N_{past} superframes, and set according to the moving average of the associated number of devices. α_t ($\alpha_t \leq 1$) is the fraction of AP transmitting sectors that are identified by devices to be their best SBS sectors, along which AP can send them the messages. The overhead of Step (1) consists of the following items: The first one is the result of the training time taken for AP to send beacons in each sector and devices to receive in each quasi-omni-directional beam; The second item denotes the time for uplink training, where in each of AP's receiving sector, every device in quasi-omni-directional mode needs to compete for sending uplink training signals; The third item is the time taken by AP to send aggregated feedbacks in part of the sectors selected by devices. Similar illustrations can be made for other steps too.

Step 2: $T_2/T_p = \alpha_t \left\lceil \frac{B_W^U}{B_S^U} \right\rceil \cdot \left\lceil \frac{B_Q^V}{B_S^V} \right\rceil + \sum_j S_{2,j} \left\lceil \frac{B_Q^V}{B_S^V} \right\rceil$, $j = 1, 2, \ldots, \alpha_r \left\lceil \frac{B_W^U}{B_S^U} \right\rceil$; α_r ($\alpha_r \leq 1$) is the fraction of AP receiving sectors that are identified by devices to be their best SBS sectors, along which AP can receive from them the feedbacks. j is a set index indicator (not actual sector ID number) that denotes the index of AP sector in the set of best AP receiving sectors. $S_{2,j}$ denotes the number of S_2 response slots for the j-th sector in the set of AP reception sectors identified to be the best. $S_{2,j}$ can be set to be proportional to the number of devices in that sector.

Step 3: $T_3/T_p = \alpha_t \left\lceil \frac{B_W^U}{B_S^U} \right\rceil P^T Q^R + N_{dev} Q^T P^R + N_{bfb}^{AP}$, where N_{dev} is the number of devices, and N_{bfb}^{AP} is the number of the AP's best fine beams for the transmission of the feedbacks.

The total training overhead, $T_{BT} = T_1 + T_2 + T_3$, is obtained from Step (1) to (3), where $\lceil \cdot \rceil$ factors (system paramters) can usually be pre-determined by the system.

V. Joint Beam Training and Transmission Scheduling

An important MAC function is to efficiently coordinate radio resource usage among multiple users. The transmission scheduling for mmWave communications is made difficult with its need of a large amount of training to find the transmission opportunities, which we target to study in this section. Following the basic structure of 802.11ad, each superframe (Beacon Interval) consists of durations for beam training as well as frames for data transmissions. There is a tradeoff in determining the durations of the two, and we will concurrently consider both in our scheduling to achieve a high transmission performance.

The basic structure of our joint beam training and scheduling scheme is depicted in Figure 2. In this section, we introduce our design for these important components, the merits of which include (1) adaptive beam training in response to channel quality, (2) resource scheduling with joint allocation of transmission opportunities and durations that can support heterogeneous traffic conditions, user types and demands, and (3) beam tracking with beamwidth adaptation and mobility estimation. The major differences of our design from the literature are: (a) our transmission scheduling concurrently considers multiple factors to achieve overall network performance improvement, reduces the control overhead, and enables burst transmissions with virtual scheduling and aggregation of transmission slots, and (b) with various adaptations, our adaptive beam training and tracking schemes are resilient to network dynamics.

A. Adaptive Beam Training

A training signal consists of a sequence of training symbols, and training signals can be sent along many directions. The channel is dynamic and the number of training samples needed is uncertain. The training can be increased at both temporal and spatial directions to achieve more accurate channel estimation and find the best direction for higher transmission rates, while higher training time will compromise the overall transmission throughput. To reduce the training time while ensuring the desired transmission quality, we propose to adapt the training period based on the channel measurement quality.

After receiving the beacon signals from AP in step (3.1), a device will determine if it will require AP to send additional training signals based on the average SNR of the received signals. If it is lower than a pre-determine threshold, the device will request additional training in its feedback in step (3.2). The P_{add} and Q_{add} additional fine beams for AP to send and the device to receive from can be determined based on SNR as follows:

$$P_{add} = \left\lceil \eta_1 \cdot \frac{SNR_{TH} - SNR}{SNR_{TH}} \right\rceil,$$

$$Q_{add} = \left\lceil \zeta_1 \cdot \frac{SNR_{TH} - SNR}{SNR_{TH}} \right\rceil,$$

where η_1, ζ_1 control the adaptation speed, SNR_{TH} is the threshold. If multiple requesting devices share the same transmission sector, AP will set P_{add} to the highest number required, and send along randomly selected directions within the sector. Similarly, AP can also request a device to send additional uplink training signals.

If a device or AP has collected training signals from two rounds, it can compare the difference between the channel estimation based on the total training signals obtained in both rounds to determine if more training is needed. In this case, P_{add} and Q_{add} for the next round are determined by

$$P_{add} = \left\lceil \eta_2 \frac{\|\mathbf{H} - \mathbf{H}_{prev}\|_1}{\Delta_{\mathbf{H}}} \right\rceil, \quad Q_{add} = \left\lceil \zeta_2 \frac{\|\mathbf{H} - \mathbf{H}_{prev}\|_1}{\Delta_{\mathbf{H}}} \right\rceil,$$

where η_2 and ζ_2 are the adaptation factors, $\mathbf{H}$ and $\mathbf{H}_{prev}$ are the estimated channels in the current round and the previous round, $\Delta_{\mathbf{H}}$ is the threshold for channel estimation difference, and the triggering condition is $\|\mathbf{H} - \mathbf{H}_{prev}\|_1 > \Delta_{\mathbf{H}}$.

B. Virtual Transmission Scheduling

The data transmission interval is composed of three components: $T_{DTI} = T_{ran} + T_{sp} + T_{dp}$. T_{sp} denotes the durations for scheduled periods, where the scheduling of data transmissions for different users in the network significantly impacts the network throughput. The contention-based *random access period* T_{ran} can be used to send unscheduled uplink data and some short messages. The traffic can be bursty and the mmWave channel is subject to low coherent time and channel blocking. We also introduce a *dynamic period* T_{dp} to accommodate the immediate needs of user data transmissions or beam tracking, which is developed in Section V-C.

As random access will introduce high overhead, IEEE 802.11ad allows the use of TDMA kind of service period, but without giving a detailed scheme how the radio resources can be scheduled for use. Coordinating transmissions among users with heterogeneous quality requirements in the presence of different types of traffic and blocking-prone wireless channels is a grand challenge. The simple slot allocation for continuous voice transmissions used in conventional cellular networks cannot be applied to the dynamic packet transmissions.

There are two major issues to address for the data transmission scheduling in T_{sp}: 1) How to select the users to transmit, and 2) How to determine the transmission durations to allocate to the selected users. To accommodate user requests while also meeting the resource constraint, there are a large number of options. It is difficult to select the users and also determine the transmission duration for each user at the same time in practical scheduling. We propose a *self-adaptive virtual resource scheduling* scheme based on user requests, application types, and practical resource availability.

To accommodate different types of applications, we divide the scheduled transmission period T_{sp} into two logic parts: *reserved period* and *allocated period*, in other words, $T_{sp} = T_{sp}^{res} + T_{sp}^{allo}$. A reserved period T_{sp}^{res} is used to support users which require long-term and periodic transmissions in every superframe, such as real-time multi-media streaming and updates of monitoring data. Admission control is needed and can be performed based on any rule of the service providers. In this paper, we consider a scheme with the limit of N_{ds}^{res} streams to admit in the reserved period, with each data stream occupying at most N_{ts}^{res} transmission slots. For a required transmission rate, the number of time slots needed to support an application will adapt as the channel condition changes with two options: 1) adapting the number of time slots allocated to the admitted users in each superframe based on the estimated channel condition, and 2) keeping the number of time slots unchanged, but letting the guaranteed applications to compete in getting the remaining resources needed. We can ensure enough time slots to support the minimum rate required by each application through the option 1 and allocate additional resources based on the option 2.

Users with elastic traffic will compete for resources in the allocated period T_{sp}^{allo}. The sector set selected for transmissions after training is denoted as $\mathcal{I}$ and the user set in the i-th sector is $\mathcal{J}_i$, then $x_{i,j(i)}$ denotes the $j(i)$-th user to transmit in the i-th sector. We use $x_{i,j(i)}^T$ and $x_{i,j(i)}^R$ to differentiate between the uplink transmission to AP and downlink transmission from the AP. The rate $r(x)$ of a data stream x can be estimated from the channel measurement. If the minimum data rates needed for uplink and downlink transmissions cannot be accommodated due to poor channel condition, we consider the user experiences a outage and set the effective user data rate to zero. Transmission Slot (TS) is the basic unit in our temporal resource scheduling, and a data link can take multiples of TS. To maximize the network performance, we need to schedule the data streams ($x^T(i, j(i))$ and $x^R(i, j(i))$) and their allocated TSs.

It is difficult to simultaneously determine which users to transmit and the transmission duration as there is a coupling between the transmission priority and the resources already allocated. We propose a novel virtual scheduling scheme with two steps: (a) efficient resource allocation to determine which user to transmit in each time slot, and (b) slot shuffling to allocate each user with continuous time slots by aggregating all its slots assigned *virtually* in the scheduled period. This allows each user to transmit data as a burst to reduce the control overhead without incurring synchronization and adding a transmission header in each slot.

In each slot, if we straight-forwardly select the user with the highest channel rate and priority to transmit, all resources may be allocated to one user, at the cost of resource starvation for others. The greedy focus on one metric neglects the trade-offs among different performance factors for different users and the network. Instead, we aim to maximize the overall network performance by considering the fairness jointly determined by multiple factors: priority, delay, and data rate. We assign each slot virtually to the user with the largest weighted data rate according to the following schedule:

$$x = \arg\max_x a(x)W(x)r(x)/\overline{R}(x), \tag{17}$$

where

$$x \in \{x_{i,j(i)}^T, x_{i,j(i)}^R\}, \quad i \in \mathcal{I},\ j(i) \in \mathcal{J}_i. \tag{18}$$

(17) can be solved with a heuristic algorithm that searches through the candidate streams to look for $\{x_{i,j(i)}^T, x_{i,j(i)}^R\}$ to maximize the objective function. Since the range of candidate beams has been narrowed down with our multi-level beam training, the search space of the candidate beams is small. As the beams are chosen from a discrete space, the complexity of our algorithm is low. $a(x)$ is the priority parameter for a data stream x (determined by the service type, QoS requirements etc.) and $W(x)$ is the queuing delay. For delay-constrained traffic, we have

$$\text{Prob}[W(x) > T(x)] \le \varepsilon(x), \tag{19}$$

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 parameters $a(x)$, $W(x)$, and

$\overline{R}(x)$ (the average transmission rate of user 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*. The transmission rate r can be estimated as

$$r = B_w \log_2(1 + SNR(B_\theta^{\mathbf{u}}, B_\theta^{\mathbf{v}}, d)), \tag{20}$$

where B_w is the bandwidth. As presented in Section III-B, the signal-to-noise ratio SNR is affected by the antenna numbers, the channel conditions, the TX-RX distance, and the transmission and reception beamwidths. The slots assigned to the same user can be used together to perform burst transmission. Rather than determining the user to transmit and the transmission duration together, our scheduling scheme significantly reduces the complexity with the virtual scheduling of transmission in each slot and the aggregation of slots into a duration. Our scheduling scheme can support users with different number of antennas.

C. Beam Tracking With Beamwidth Adaptation and Mobility Estimation

Featured by highly directional transmissions, two major challenges faced by mmWave communications are channel dynamics and user mobility, which can cause frequent disconnections thus degraded network performance. To cope with these problems, we introduce two important components to facilitate beam tracking, *beamwidth adaptation* and *mobility estimation.* Upon disconnection, additional low-cost training of new beam directions may help the user to recover from disconnection, unfulfilled transmissions may be rescheduled to transmit in the remaining time of the duration assigned to the device and the dynamic resource block (T_{dp} in Section V-B).

Beam quality can be tracked with testing signals piggybacked at the end of data packets. Upon detecting a significant reduction of the beam quality or disconnection, our proposed *Beamwidth Adaptation* will be triggered:

(1) A sender will quickly switch to train two beams adjacent to the original beam direction using the time slot scheduled for the corresponding device if its remaining time is enough or using the time in the dynamic period.

(2) If a user moves too fast and gets out of the coverage of its backup beams, especially when the beamwidths of TX and RX beams are very narrow, we propose to train one further left beam and one further right beam with the beamwidth doubled to speed up the searching.

(3) If a user is found in one of the two double-width beams, we continue to train and find the best fine beam.

This searching process can continue, and the number of additional beams to search depends on the system configuration. If a user constantly moves, when reaching its scheduled time slot, its direction may largely deviate from the optimal direction found through beam training at the beginning of the superframe. The frequent and large-range beam search after the disconnection will incur a high training overhead. To better handle user mobility, we propose another *Mobility Estimation* scheme to predict the user direction based on the beam search range over the past N_p superframes:

$$\theta_{dev} = T_{lat} \sum_{i=1}^{N_p} |\theta_{dev}^i| / \sum_{i=1}^{N_p} T_{lat}^i, \tag{21}$$

where the angular deviation θ_{dev}^i of a mobile user in the i-th past superframe can be known from the beam tracking process, T_{lat}^i is the time taken to search for the new beam direction in the i-th past superframe and T_{lat} is the time duration from the end of training in the current superframe to the slot time assigned to the user. $\sum_{i=1}^{N_p} |\theta_{dev}^i| / \sum_{i=1}^{N_p} T_{lat}^i$ is an estimation of the averaged angular moving speed of the user. The sign $(+/-)$ of θ indicates whether the angular deviation is left or right, and we let the sign of θ_{dev} be the same as that in the previous superframe. With this estimation, in the time slot scheduled for the user, BS will first deviates its steering direction by θ_{dev} so that the signal can have a better chance to reach the mobile user. In case there is an estimation inaccuracy and thus the link breakage, the range of the beam searching will be much smaller.

VI. Simulations and Results

In this section, we evaluate the performance of our proposed schemes. As comparison, we will demonstrate the performances of the following schemes: (1) Proposed-adaptive (proposed scheme with adaptive training), (2) Proposed-nonadaptive (proposed scheme without adaptive training), (3) CS-nonadaptive (nonadaptive beamforming with baseline CS [18]), (4) HOL (since we can't find related uplink/downlink scheduling work to compare in mmWave realm, we adapt Head-of-Line delay based slot-by-slot scheduling in [21] for mmWave networks), (5) 802.11ad (codebook-based training, IEEE standard in [2]), (6) Proposed-nonCS (proposed multi-level beam training without CS-based channel estimation assistance), (7) Proposed-BT-BA (proposed-adaptive with Beam Tracking and Beamwidth Adaptation), (8) Proposed-BT-BA-ME (Proposed-BT-BA with Mobility Estimation) and (9) Proposed-w/o-BT-BA (proposed-adaptive with no BT or BA).

A. Settings

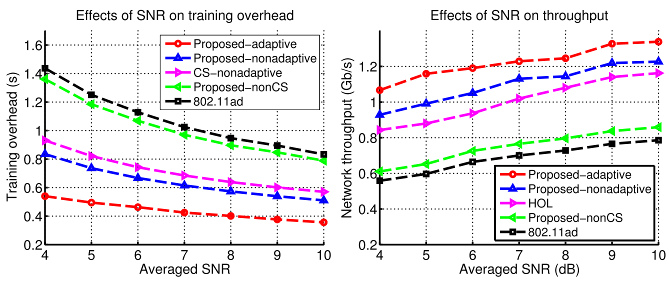
In our performance studies, we consider the scenario with one AP and multiple devices. The mmWave channel is simulated from the model derived from NYC measurements in [19]. The user traffic (both downlink and uplink) is generated as follows: user arrivals conform to Poisson distribution; traffic load paramters for different users are uniformly distributed between 400 and 500 packets per second; packet size ranges from 5 to 10 KB. 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) and (2) Network throughput (total throughput among all users). The results are averaged among a long period (200 seconds).

B. Effect 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

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	4
# sector beams per QO level	4
# fine beams per setor, AP/DEV	8/4
# fine beams per setor, AP/DEV	8/4
# of users	20



(a) Training overhead (b) Network throughput

Fig. 5. Effects of SNR.

ensure a given quality of channel estimation. Figure 5(a) shows that when the channel condition is better, the beam training overhead is reduced. The gain of our scheme improves when the training overhead is larger at lower SNR. When SNR is 4dB, from 802.11ad to Proposed-nonCS, we observe an improvement of 7.29% in terms of training overhead. This shows the benefits of our design efforts on top of 802.11ad in Section III-A, including adapting the number of response slots to reduce collision, allowing AP to feedback in groups instead of device by device etc. Proposed-adaptive performs 62.41% better than 802.11ad and Proposed-nonadaptive 41.89%. The results demonstrate the benefits of our proposed beam training and adaptive schemes in reducing the training overhead and improving the training quality. Compared with 802.11ad, Propose-nonCS works differently in the coarse-level beam training but works in the same way for finding the optimal fine beams. It can be seen that our CS-based schemes perform much better than Proposed-nonCS, which confirms that the training overhead is majorly affected by fine beam training and our CS-based schemes significantly reduce the fine beam training overhead by exploiting our CS-based multi-resolution channel estimation scheme. Compared with CS-nonadaptive, our proposed schemes perform better with much lower training overhead. Proposed-adaptive and Proposed-nonadaptive outperform CS-nonadaptive by 41.99% and 10.31%, respectively. Different from conventional CS-nonadaptive schemes, besides adaptive beamforming, we also exploit the block features of mmWave channels and take advantage of coarse training to largely reduce unnecessary measurements, and exploit multi-resolution channel estimation which take advantage of samples from different levels of measurements and block sparsity of mmWave channel for higher quality channel reconstruction.

As expected, in Figure 5(b), the throughput increases with the SNR, thanks to higher achievable data rates and reduced training overhead. At SNR of 4dB, compared with 802.11ad, we observe a throughput improvement of 90.96% for Proposed-adaptive, 66.18% for Proposed-nonadaptive and 9.31% for Propose-nonCS. We again see that the Proposed-nonCS outperforms 802.11ad by reducing the training overhead involved in coarse-level beams and our CS-based schemes significantly outperform Propose-nonCS and 802.11ad by further reducing the fine beam training overhead thus improving the throughput. The comparison also confirms that the advantages of our proposed schemes and the adaptive beam training in enabling more efficient radio resource allocation. Proposed-adaptive outperforms Proposed-nonadaptive in throughput because (1) it can reduce training overhead and (2) nonadaptive scheme may not train sufficient number of beams or find the best quality beam to accurately estimate channel, especially under low SNR. Compared with HOL, Proposed-adaptive and Proposed-nonadaptive improve the throughput by 31.67% and 14.58%, respectively. Our joint training and transmission scheduling scheme performs better by concurrently scheduling radio resources for beam training, data transmissions and beam tracking. Also, our virtual scheduling allows for burst transmissions in multiple slots, reducing the overhead for synchronization and attached packet header in each slot.

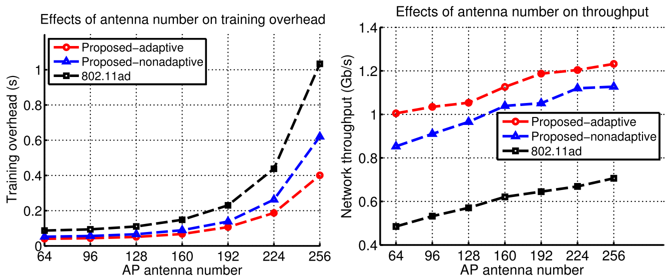
From Figures 5a and 5b, we can clearly see the tradeoffs between beam training duration and network throughput. As beam training overhead increases, there is likely a shorter period for data transmissions, which affects the throughput. Since our scheme jointly schedules beam training and data transmissions based on network conditions, we are able to better trade off between training and transmissions to achieve higher performance than the other schemes compared.

C. Effect of Antenna Number

The number of antennas greatly affects the number of possible beams to be measured thus the training overhead. The network throughput is significantly impacted by the beam training overhead. The larger the training overhead, the less the time available for data transmission thus reducing the throughput. On the other hand, more antennas will also introduce higher beamforming gain in transmissions.

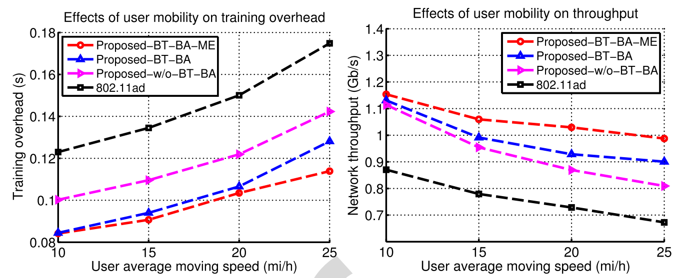
In Figure 6a, the training overhead grows exponentially with the number of antennas. With a larger antenna number, there will be many more possible beams to be trained. When the number of AP antennas is 256, compared with 802.11ad, we observe an overhead reduction of 61.25% when Proposed-adaptive is used and 40% overhead reduction when using Proposed-nonadaptive. This demonstrates the effectiveness of our proposed schemes in reducing the training overhead and the adaptive beam training further reduces the overhead.

In Figure 6b, the throughput increases when the number of antennas increases, but the gain doesn't seem to fully reflect the gain from antenna number. Obviously, the higher training overhead compromises the beamforming gain, which further confirms that it is important to control the training overhead. We also see that when AP has 256 antennas, Proposed-adaptive performs 74.41% better than 802.11ad and



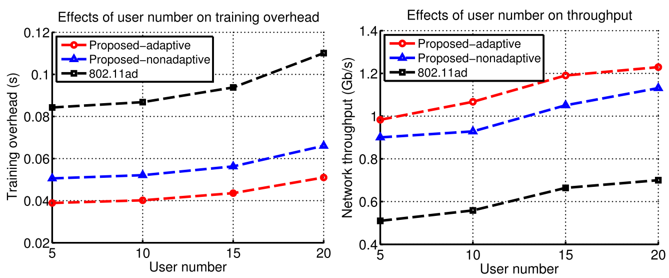
(a) Training Overhead (b) Network throughput

Fig. 6. Effects of antenna number.



(a) Training overhead. (b) Network throughput

Fig. 8. Effects of user mobility.



(a) Training overhead. (b) Network throughput

Fig. 7. Effects of user number.

Proposed-nonadaptive 59.56%, which indicates the benefits of our proposed beam training and adaptive schemes in reducing the training overhead for higher throughput.

D. Effect 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 periods. While keeping each user's traffic load the same, we vary the number of users.

Figure 7a shows that the overall training overhead increases with the number of users, as longer time is needed to complete the channel training for more users. The performance of both of our proposed schemes outperform 802.11ad, and the improvement increases at higher number of users. At a user number of 10, compared to 802.11ad, our Proposed-adaptive and Proposed-nonadaptive have an overhead reduction of 53.53% and 40.03%, respectively. In Figure 7b, the network throughput increases with the number of users, which is not difficult to understand since more users are joining the network for data transmission. Both of our proposed schemes significantly increase the network throughput under different number of users. At the user number of 10, Proposed-adaptive and Proposed-nonadaptive outperform 802.11ad by 94.17% and 66.11%, respectively. This demonstrates the effectiveness of our MAC schemes in accommodating more network traffic.

E. Effect of User Mobility

The highly directional transmissions of mmWave networks make the network performances sensitive to the movement of users. We vary the mobility levels of users and study the benefits and tradeoffs of our proposed beam tracking, beamwidth adaptation and mobility estimation schemes.

In Figure 8a, our proposed schemes significantly outperform 802.11ad, and the improvement increases as the users move faster. Our Proposed-BT-BA-ME further improves from Proposed-BT-BA and Proposed-w/o-BT-BA with the use of mobility estimation. At the average moving speed of 25 mi/h, the Proposed-w/o-BT-BA, Proposed-BT-BA, Proposed-BT-BA-ME outperform 802.11ad by 18.59%, 26.71% and 34.86%, respectively. While user mobility causess link disconnections, our flexible beam training and beamwidth adaptation with mobility prediction reduce the training overhead and delay to realign the beam. The beamwidth-adaption is very effective in tracking the beams under mobility while the mobility estimation helps to further improve the performance.

In Figure 8(b), the network throughput degrades as the users' mobility level increases, which shows the sensitivity of mmWave networks to user movement. At the user average moving speed of 25 mi/h, the Proposed-w/o-BT-BA, Proposed-BT-BA and Proposed-BT-BA-ME outperform 802.11ad by 20.31%, 33.90% and 46.85%, respectively. The results validate the benefits of our proposed beam tracking components and their effectiveness in reducing the training overhead to maintain connectivity for mobile users. The reduction of tracking and training overhead further allows more resources for data transmissions to improve the throughput.

VII. CONCLUSION

With its potential of supporting multi-Gbps data transmissions, millimeter-wave technique is a promising candidate for next-generation wireless communications. However, the need of highly directional transmission brings great challenges in the design of medium access control in mmWave networks. This paper addresses the need of a low-cost multi-user beam training scheme with the concurrent use of multi-level coarse training and multi-resolution block-sparse channel estimation for fine beam alignment. We also jointly allocate radio resources for beam training and data transmissions, design an efficient virtual scheduling scheme based on user application types and demands, and incorporate flexible beam tracking scheme for low-overhead beam re-alignment in the presence of user mobility and channel dynamics. Simulation results show the significant benefits of our proposed design compared with 802.11ad and also the tradeoffs in 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–Local and Metropolitan Area Networks–Specific Requirements-Part 11: WLAN MAC and PHY Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band*, IEEE Standard 802.11ad-2012, 2012.

[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," in *Proc. IEEE Global Telecommun. Conf.*, Nov./Dec. 2009, pp. 1–6.

[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. Alkhateeby, G. Leusz, 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. Personal, 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—Local and Metropolitan Area Networks—Specific Requirements—Part 15.3: Amendment 2: Millimeter-Wave-Based Alternative Physical Layer Extension*, IEEE Standard 802.15.3c-2009, amendment to IEEE std 802.15.3-2003, Oct. 2009.

[15] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 164–174, Nov. 2010.

[16] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath, Jr., "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in *Proc. 48th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 957–961.

[17] R. Méndez-Rial, C. Rusu, A. Alkhateeb, N. González-Prelcic, and R. W. Heath, Jr., "Channel estimation and hybrid combining for mmWave: Phase shifters or switches?" in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2015, pp. 90–97.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] M. K. Haider and E. W. Knightly, "Mobility resilience and overhead constrained adaptation in directional 60 ghz wlans: Protocol design and system implementation," in *Proc. 17th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, New York, NY, USA, Jul. 2016, pp. 61–70. doi: 10.1145/2942358.2942380.

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

[24] 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.

[25] 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.

Jie Zhao 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, New York, USA. His research interests include millimeter-wave communications, cognitive radio networks, and networked sensing and detection.

Dongliang Xie received the Ph.D. degree from the Beijing Institute of Technology, China, in 2002. He is currently a Professor with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT), China. His research interests include resource-constrained wireless communication and information-centric network, including architecture of ubiquitous and heterogeneous network, complex network analysis, content retrieval, and service management.

Xin Wang received the B.S. and M.S. degrees in telecommunications engineering and wireless communications 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 was a Member of Technical Staff in the area of mobile and wireless networking at 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. 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. 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 has served in executive committee and technical committee for numerous conferences and funding review panels. She achieved the NSF Career Award in 2005 and the ONR Challenge Award in 2010. She serves as an Associate Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING.

Arjuna Madanayake received the B.Sc. degree in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 2002, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Calgary, Alberta, Canada. From 2010 to 2018, he was a tenured Faculty Member with the University of Akron, Akron, OH, USA. In 2018, he joined the Faculty of FIU in August 2018. He is currently an Associate Professor with Florida International University (FIU), Miami, FL, USA. His research interests are in array signal processing, circuits, systems, electronics, fast algorithms, and computer architecture.