BOOST: A User Association and Scheduling Framework for Beamforming mmWave Networks

Prosanta Paul, Hongyi Wu, ChunSheng Xin

Abstract—The *millimeter wave* (mmWave) band offers vast bandwidth and plays a key role for next generation wireless networks. However, the mmWave network raises a great challenge for *user association and scheduling*, due to the limited power budget and beamformers, diverse user traffic loads, user quality of service requirement, etc. In this paper, we propose a novel framework for user association and scheduling in multi-base station mmWave networks, termed the *clustering Based dOwnlink UE assOciation, Scheduling, beamforming with power allocaTion* (BOOST). The objective is to reduce the downlink network transmission time, subject to the base station power budget, number of beamformers, user traffic loads, and the quality of service requirement at users. We compare BOOST with three state-of-the-art user scheduling schemes. On average, BOOST reduces the transmission time by 37%, 30%, and 26%, and achieves a sum rate gain of 56%, 43%, and 34%, respectively.

Index Terms—mmWave networks, user association, user scheduling, clustering, power allocation.

1 Introduction

The proliferation of intelligent wireless devices is remarkable. By the prediction of Ericsson, there will be 29 billion IoT (Internet of Things) devices by 2022, and the mobile data traffic will grow 35% annually through 2024 [1]. To address the phenomenal traffic growth, a key objective of next generation wireless networks such as 5G is to provide significantly higher bandwidth. Due to the scarcity of the sub-6 GHz spectrum, it is not possible to achieve this objective by expanding traditional cellular bands within this spectrum range. Hence, the *millimeter wave* (mmWave) band between 20 and 300 GHz becomes the front line to provide vast bandwidth and low latency needed by next generation wireless networks [2]–[4].

While the mmWave band promises much higher bandwidth, it also raises a technical challenge. The much higher frequency results in an additional path loss of $20 - 25$ dB by the physics law, compared with traditional sub-6 GHz cellular bands [5]. Thus, to make mmWave communications practical, i.e., achieve a practical range under such a severe path loss, a critical technology, *beamforming*, is needed. It uses a reconfigurable antenna array, and controls the amplitude and phase of the signal at each antenna element to concentrate the transmission power on a narrow beam toward the receiver, to result in a high signal gain. As a result, it dramatically increases the *signal to interference and noise ratio* (SINR) at the receiver, to compensate the increased path loss at the mmWave band.

While beamforming significantly increases the reach of *base stations* (BSs) in mmWave networks, it raises a great challenge for *UE association and scheduling*, which is a critically important problem for mmWave networks. In mmWave networks, the UE association generally does not rely on the distance to surrounding BSs, but depends on if a UE is covered by a beam. This makes the UE association and beamforming a joint optimization problem. The problem is even more challenging as a BS usually cannot support all beams to cover its UEs simultaneously. Instead, it has to schedule its UEs/beams across multiple time slots, because of the limitation on the power budget, as well as the limited number of beamformers due to the high cost of RF chains. Note that traditional cellular networks also schedule UEs across time slots, but that is due to the limited spectrum bandwidth at sub-6 GHz cellular bands.

The UE association and scheduling also has to consider various other factors including heterogenous UE traffic loads, the quality of service requirement for UEs, as well as the beam interference, power allocation, UE fairness, etc. More importantly, those factors are intermingled, which makes the UE association and scheduling a fundamentally challenging joint optimization problem. In this paper, we propose a novel framework for UE association, scheduling, and beamforming in mmWave networks, termed the *clustering Based dOwnlink UE assOciation, Scheduling, beamforming with power allocaTion* (BOOST). The objective is to reduce the transmission time for all UEs traffic, subject to the power budget of BSs, number of beamformers of BSs, UE traffic loads, and the required minimum SINR for each UE. Note that reducing transmission time or latency is a key objective of 5G. Moreover, in the context of this work, reducing transmission time also contributes to increasing the network capacity, another key objective of 5G.

Most of existing works on UE scheduling focused on choosing a set of UEs for a time slot to maximize the sum rate given the BS power budget. Several works chose a group of UEs with orthogonal channels to maximize the

- *P. Paul, H. Wu, and C. Xin are with the Center for Cybersecurity Education and Research, and the Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, 23529. E-mail: {ppaul001, h1wu, cxin}@odu.edu.* *The work of H. Wu is supported in part by NSF under grants CNS-1828593, OAC-1829771, CNS-1528004, DGE-1723635, and CNS-1320931. The work of C. Xin is supported in part by NSF under grants CNS-1745632, CNS-1950704, and CNS-1659795.*

sum rate (e.g., see [6]–[8]). The *semi-orthogonal user selection* (SUS) chooses a UE with the largest channel gain at first, and then in each subsequent iteration, selects a UE such that its channel has the largest orthogonal component to the subspace spanned by the previously chosen UEs [9]–[11]. In [12], UEs are partitioned into groups, with each group having similar channel covariance. In [13], a UE clustering and set covering based scheme was developed to minimize the transmit power in multi-BS networks. In [14], a scheme was presented to maximize the harmonic sum of UEs SINR in order to prioritize cell-edge UEs. The switched random beamforming has also been studied in the literature [15], [16], where a BS forms a beam in a random direction. The channel *chordal distance* (Chord-Dis) based scheme aims to maximize the sum rate, where UEs are chosen based on their channel chordal distances, which are a measure of orthogonality between UE channels (e.g., see [17]–[20]). The *UE partitioning and beamforming* (MUBFP) scheme maximizes the group-average sum rate for a single BS MU-MIMO system [21].

Most of existing works focused on BS-UE channels for selecting UEs to maximize the sum rate. However, this often results in a large transmission time to deliver all UEs traffic, because to achieve a high sum rate, UEs with good channels (high SINR) would be allocated with much more network resources such as the transmit power and bandwidth. As a result, the transmission time for UEs with poor channels can be very large. Moreover, most of those works focused on one BS, while the UE association among multiple BSs assumed a simple scheme based on the received signal power. BOOST jointly considers UE association among multiple BS, UE scheduling, and beamforming altogether, to reduce the transmission time of the entire network. Our contributions are summarized below.

- We develop a clustering algorithm that capitalizes on unique features of beamforming to group UEs into clusters to reduce interference for beamforming.
- We develop a novel UE association scheme that effectively reduces interference and balance UE traffic loads between BSs, to decrease the transmission time.
- We design a scheme for joint beamforming, power allocation, and UE scheduling to reduce the transmission time, subject to the BS power budget, UEs traffic loads, and the minimum SINR at UEs.

The rest of the paper is organized as follows. The system model is discussed in Section 2. Section 3 presents the problem formulation. The BOOST framework is described in Section 4. The performance evaluation is presented in Section 5 and Section 6 concludes the paper.

2 System Model

We assume BSs are connected to a cloud or a backend management system of the network operator, through either wired or wireless backhaul connections. The BSs information, including the location, number of antennas, etc., is known to the cloud which manages UE association and scheduling. We also assume all BSs and UEs are synchronized on both time and frequency for each time slot, in which a set of UEs are scheduled for traffic transmission. Table 1 lists major notations.

TABLE 1
Major notations

m, n, k	index for BS, UE and beam, respectively
$\mathcal{M}, \mathcal{N}, \mathcal{K}$	set of BSs, UEs, and beams, respectively
$\langle m, k\rangle$	the k-th beam of the m-th BS
$\theta_{mn}, \mathbf{a}(\theta_{mn})$	angle of arrival of UE n signal from BS m, and the corresponding steering vector
b_n, $t_{n,m,k}$	traffic load of UE n, and its transmission time when served by beam $\langle m, k\rangle$
$\mathbf{h}_{mn}$	channel gain vector between BS m and UE n
$\mathbf{w}_{mk}$	beamforming weight vector of beam $\langle m, k\rangle$
p_{mk}, p_o	transmit power of beam $\langle m, k\rangle$ and power budget in each BS
K	maximum number of beamformers at a BS
$I_{n,m,k}, \lambda_{n,m,k}$, $\gamma_{n,m,k}$	interference, rate, and SINR at UE n when it is served by beam $\langle m, k\rangle$
γ_o	minimum required SINR for UEs

2.1 mmWave Channel

Let $\mathcal{N}$, $\mathcal{M}$ and $\mathcal{K}$ be the set of UEs, BSs, and beams, respectively. Let $\mathcal{N}_m \subseteq \mathcal{N}$ be the set of UEs that can be covered by BS m under the maximum range, which is determined by a beam with the minimum beamwidth and the maximum BS transmit power. Throughout the paper, we assume each BS is equipped with an antenna array with L antenna elements, and each UE has a single antenna. Nevertheless, BOOST can be easily extended to accommodate UEs with multiple antennas. Let K be the number of beamformers (or RF chains) for a BS. Each beamformer can form one beam. Note that the required processing power, design complexity and fabrication cost of a BS grows with the number of beamformers; hence a BS can only have a limited number of beamformers. Let $\mathcal{K}_m \subseteq \mathcal{K}$ denote the set of beams of BS m. For each beam $k \in \mathcal{K}_m$, let $\mathcal{N}_{mk} \subseteq \mathcal{N}_m$ denote the set of UEs that are inside beam k of BS m, called as beam $\langle m, k\rangle$.

Let d_{mn} and PL_{mn} denote the distance and average LOS path loss, respectively, between BS m and UE n. According to the 3GPP UMi-street canyon LOS model, the large scale path loss in dB is given as [22]

$$\text{PL}_{mn}[dB] = 10\eta \log_{10}\left(\frac{d_{mn}}{d_0}\right) + 20\log_{10}\left(\frac{4\pi d_0 \times 10^9}{c}\right) + 20\log_{10}(f) + \mathcal{X}_{\sigma_{SF}}, \quad (1)$$

where η is the path loss exponent of the LOS path, d_0 is the close-in free space reference distance which is usually 1, c is the speed of light, f is the system frequency in GHz, and $\mathcal{X}_{\sigma_{SF}}$ is a zero-mean Gaussian random variable with a standard deviation σ_{SF} in decibels. Let θ_{mn} be the *angle of arrival* (AoA) of UE n, i.e., the direction of the downlink signal to UE n with regard to the BS m antenna, and α_{mn} be the complex power gain of small-scale fading between BS m and UE n, which is modeled using a complex Gaussian distribution, i.e., $\alpha_{mn} \sim \mathcal{CN}(0, 1)$. We assume a block fading channel between UEs and BSs, as in the existing cellular and WLAN standards [12]. The channel characteristics is about the same in the order of seconds. The channel vector between BS m with L antenna elements and UE n ($n \in \mathcal{N}_{mk}$) with a single antenna is given as [15]

$$\mathbf{h}_{mn} = \sqrt{\frac{L}{\text{PL}_{mn}}}\alpha_{mn}\mathbf{a}(\theta_{mn}), \quad (2)$$

where $\mathbf{a}(\theta)$ is the response vector of the BS antenna when the UE AoA is θ. The response vector for a *uniform linear array* (ULA) at a BS is given as

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{L}}\left[1, \quad e^{-j2\pi\theta}, \ldots, e^{-j2\pi(L-1)\theta}\right]^T, \tag{3}$$

where $(\bullet)^T$ denotes the transpose of a vector. The concatenated channel matrix formed by all UEs covered by BS m is therefore written as $\mathbf{H}_m = \left[\mathbf{h}_{m,1}, \ldots, \mathbf{h}_{m,|\mathcal{N}_m|}\right]$. The received signals at all UEs covered by BS m are given as

$$\mathbf{Y}_m = \mathbf{H}_m^H \mathbf{W}_m \mathbf{x}_m + \mathbf{z}, \tag{4}$$

where $(\bullet)^H$ denotes the complex conjugate transpose, $\mathbf{x}_m \in \mathbb{C}^{|\mathcal{N}_m|\times 1}$ is the vector of transmitted signals to UEs of BS m, $\mathbf{z} \in \mathbb{C}^{|\mathcal{N}_m|\times 1}$ is the total noise, including the interference and the Additive White Gaussian Noise (AWGN) with zero mean and unit variance, $\mathbf{W}_m = \left[\mathbf{w}_{m,1}, \ldots, \mathbf{w}_{m,|\mathcal{K}_m|}\right] \in \mathbb{C}^{L\times|\mathcal{K}_m|}$ is the beamforming weight matrix of BS m, formed by concatenating individual weight vectors $\mathbf{w}$.

2.2 Downlink Transmission Time

In BOOST, each BS m constructs a set of unit-power beams, denoted as $\{\mathbf{w}_{m,1}, \ldots, \mathbf{w}_{m,|\mathcal{K}_m|}\}$, with the power of each beam $||\mathbf{w}||^2 = 1$. As illustrated in Fig. 1(d), each beam serves a group of UEs in a time slot. Let $\hat{\theta}_{mk}$ denote the direction of the main lobe of beam $\langle m, k\rangle$, formed by $\mathbf{w}_{mk}$ to serve the UEs in $\mathcal{N}_{mk}$. For a beam covering multiple UEs, the beam direction $\hat{\theta}_{mk}$ may not be aligned exactly with the AoA of a UE. Let $\theta_{n,m,k} = |\hat{\theta}_{mk} - \theta_{mn}|$ denote the difference between the beam direction $\hat{\theta}_{mk}$ and the UE AoA θ_{mn}. By the cosine antenna pattern of the ULA antenna, the effective channel gain of UE n covered by beam $\langle m, k\rangle$ can be written as [23], [24]

$$|\mathbf{h}_{mn}^H \mathbf{w}_{mk}|^2 \approx \begin{cases} \frac{L|\alpha_{mn}|^2}{\mathrm{PL}_{mn}} \cos^2\left(\frac{L\pi\theta_{n,m,k}}{2}\right) & \text{if } \theta_{n,m,k} \le \frac{1}{L}, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $(\bullet)^H$ denotes the complex conjugate transpose, α_{mn} is the complex gain of the LOS path between BS m and UE n, and PL_{mn} is the average LOS path loss between BS m and UE n. In (5), the UE channel gain is at maximum when $\theta_{n,m,k} = 0$, i.e., when the beam direction perfectly matches the UE AoA. The cosine antenna pattern provides a good approximation for the main lobe gain [23], [24]. Note that the main impact of sidelobes is to cause interference to other beams, which can be efficiently suppressed if there is a sufficient angular distance between two beams, e.g., using the *linearly constrained minimum variance* beamforming technique (to be discussed later). Therefore, in this paper, we do not consider sidelobes.

Let p_{mk} be the transmit power of beam $\langle m, k\rangle$. Let $I_{n,m,k}$ denote the total interference at UE n ($n \in \mathcal{N}_{mk}$) from all beams. Let σ^2 be the thermal noise at a UE, which is modeled as $\sigma^2 = N_o + 10\log(v) + \mathrm{NF}$, where v is the system bandwidth, N_o is the noise power spectral density and NF denotes the noise figure at the UE. The received SINR at UE n from beam $\langle m, k\rangle$ is given as

$$\gamma_{n,m,k} = \frac{p_{mk}|\mathbf{h}_{mn}^H \mathbf{w}_{mk}|^2}{I_{n,m,k} + \sigma^2}, \quad \forall n \in \mathcal{N}_{mk}. \tag{6}$$

The total interference $I_{n,m,k}$ is the summation of intra-BS interference and inter-BS interference given as follows.

$$I_{n,m,k} = \underbrace{\sum_{i\in\mathcal{K}_m\setminus k} p_{mi}|\mathbf{h}_{mn}^H \mathbf{w}_{mi}|^2}_{\text{intra-BS interference}} + \underbrace{\sum_{j\in\mathcal{M}\setminus m}\sum_{l\in\mathcal{K}_j} p_{jl}|\mathbf{h}_{jn}^H \mathbf{w}_{jl}|^2}_{\text{inter-BS interference}}, \tag{7}$$

where $\mathcal{M}$ denotes the set of BSs and $\mathcal{K}_m$ denotes the set of beams of BS m.

To achieve a certain quality of service for a UE, the SINR at the UE has to be greater than or equal to a minimum SINR γ_o, i.e.,

$$\gamma_{n,m,k} \ge \gamma_o, \quad \forall n \in \mathcal{N}_{mk}. \tag{8}$$

In order to maintain the minimum γ_o at all UEs in a beam, the required transmit power p_{mk} is obtained using (6)-(8) as

$$p_{mk} = \max_{n\in\mathcal{N}_{mk}} \frac{(I_{n,m,k} + \sigma^2)\gamma_o}{|\mathbf{h}_{mn}^H \mathbf{w}_{mk}|^2}. \tag{9}$$

The transmit power vector for all beams of BS m is given as

$$\mathbf{p}_m = \left[p_{m1}, \ldots, p_{m|\mathcal{K}_m|}\right]. \tag{10}$$

In each time slot, the transmit power from all beams of a BS must not be larger than the power budget p_o, i.e.,

$$\sum_{k\in\mathcal{K}_m} p_{mk} \le p_o, \quad \forall m \in \mathcal{M}. \tag{11}$$

We assume each UE n has a downlink traffic load b_n that needs to be delivered with a minimum SINR γ_o. Let Δ_t be the length of each time slot. The minimum number of time slots needed to carry the traffic load b_n to the UE is $\left\lceil \frac{b_n}{v_{n,m,k}\Delta_t \log_2(1+\gamma_{n,m,k})} \right\rceil$, where $v_{n,m,k}$ is the sum of spectrum from all downlink sub-carriers assigned to UE n and $\gamma_{n,m,k}$ is the received SINR.

3 Problem Statement

As stated in the preceding section, we assume all BSs in the network are synchronized on time and frequency at each time slot. BSs schedule a set of UEs for downlink transmission in each slot. Our goal is to minimize the total number of time slots required to complete delivery of all UEs traffic loads, subject to the BS power budget, the minimum SINR requirement at UEs, and the number of beamformers of BSs. If there are more than one UE in a beam, the throughput capacity of the downlink channel is shared among all UEs. The UEs can use a multiplexing scheme to share resources, e.g., the *time division multiple access* (TDMA), *orthogonal frequency division multiple access* (OFDMA), or *non-orthogonal multiple access* (NOMA). In the ensuing discussions, we assume OFDMA is used. Nevertheless, BOOST can be extended for NOMA and TDMA.

If UE n is scheduled in beam $\langle m, k\rangle$, the data rate for UE n in slot t, denoted as $\lambda_{n,m,k}^{[t]}$, is given as

$$\lambda_{n,m,k}^{[t]} = v_{n,m,k}^{[t]} \log_2(1 + \gamma_{n,m,k}^{[t]}), \tag{12}$$

where $v_{n,m,k}^{[t]}$ is the sum of spectrum from all downlink subcarriers assigned to UE n in the beam, and $\gamma_{n,m,k}^{[t]}$ is the SINR at UE n in time slot t. Let $x_{nmk}^{[t]} \in \{0,1\}$ be a binary variable to indicate if UE n is covered by beam $\langle m,k \rangle$ in time slot t. If $x_{nmk}^{[t]} = 1$, then the remaining traffic of UE n to be transported in the next slot $t+1$ is given as

$$b_n^{[t+1]} = \max(0, b_n^{[t]} - \Delta_t \lambda_{n,m,k}^{[t]}). \quad (13)$$

To reduce the number of slots required to transmit traffic from all UEs, we formulate the objective function to minimize the remaining UE traffic in each slot given in (13). For the ease of description, we drop the time slot index t in the formulation below. We want to find the values of x_{nmk}, $\mathbf{p}_m$ (transmit power) in (10), and $\mathbf{W}_m$ (beamforming weight matrix) in (4) for each BS. Let $\mathbf{X}$ denote the matrix $[x_{nmk}]$. The UE association and scheduling can be formulated as a nonlinear programming problem as below

$$\min_{\mathbf{X},\mathbf{p}_m,\mathbf{W}_m} \sum_{n\in\mathcal{N}} [b_n - x_{nmk} v_{n,m,k} \Delta_t \log_2(1+\gamma_{n,m,k})] \quad (14)$$

$$\text{s. t.} \quad \text{C1}: \forall n \in \mathcal{N}, \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}_m} x_{nmk} \le 1$$

$$\text{C2}: \forall n \in \mathcal{N}, m \in \mathcal{M}, k \in \mathcal{K}_m, \gamma_{n,m,k} \ge x_{nmk}\gamma_o$$

$$\text{C3}: \forall m \in \mathcal{M}, \sum_{k\in\mathcal{K}_m} p_{mk} \le p_o$$

$$\text{C4}: \forall m \in \mathcal{M}, |\mathcal{K}_m| \le K$$

$$\text{C5}: \forall m \in \mathcal{M}, k \in \mathcal{K}_m, \sum_{n\in\mathcal{N}_{mk}} x_{nmk} v_{n,m,k} \le v$$

$$\text{C6}: \forall n \in \mathcal{N}, m \in \mathcal{M}, k \in \mathcal{K}_m, \quad v_{n,m,k}\Delta_t \log_2(1+\gamma_{n,m,k}) \le x_{nmk} b_n.$$

Constraint C1 guarantees each UE is associated with only one BS and only one beam. C2 ensures the received SINR at each UE is greater than the threshold. C3 ensures the power allocated in all beams of a BS cannot surpass the power budget p_o. C4 ensures the number of beams in a BS is less than the number of beamformers K in a BS. C5 ensures that the sum of allocated spectrum for all scheduled UEs in a beam is not more than the available system bandwidth v. C6 guarantees that the allocated spectrum to a UE is no more than it needs to carry its traffic load in a time slot. The second term of (14) is the weighted sum rate maximization problem which has been proven NP-hard [21], [25]. Therefore, in this paper, we develop a heuristic framework, BOOST, to solve the optimization problem (14).

4 BOOST Framework

In the literature, the UE discovery can be usually conducted through two approaches: beam sweeping [26], [27] or coexisting macrocells [28]. With the first approach, each BS sweeps the whole angular space and transmits initial signals, to discover UEs. This approach may result in a large delay. With the second approach, the co-existing macrocells such as LTE towers are used to discover UEs and send the information to the cloud of the network operator, which can achieve a low delay. Given that major cellular operators all have deployed LTE systems, this is a practical approach.

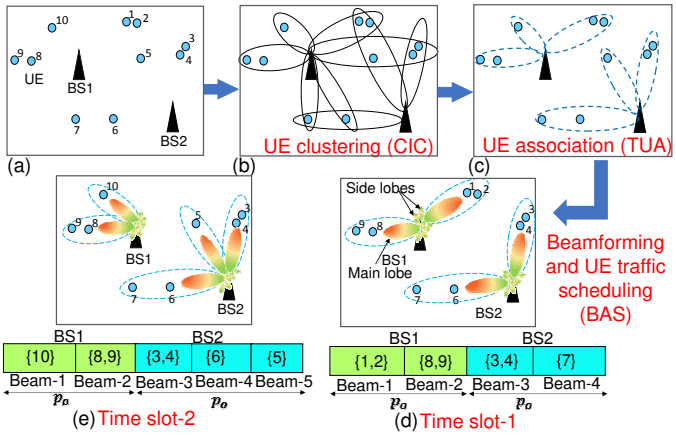

Fig. 1. BOOST framework: (a) a mmWave network with 2 BSs and 10 UEs, (b) UE clustering to reduce intra-BS interference, (c) UE association to balance traffic and reduce inter-BS interference, (d-e) scheduling UE traffic into two slots and forming beams, constrained by the power budget and UE SINR requirement.

Through UE discovery, the UEs information, including locations and AoAs, is known to the network operator. BOOST resides in the cloud of the operator to optimize UEs association and scheduling. Specifically, for given UE traffic loads, BOOST finds the optimal number of beams for each BS, assigns UEs to be covered by each beam, computes the beamforming vectors, and allocates spectrum for UEs and power for all beams in each slot, until the traffic of all UEs is delivered. The objective is to reduce the *network transmission time*, i.e., the total number of time slots required to deliver the UEs traffic, subject to the BS power budget, number of beamformers for BSs, and minimum UE SINR.

As illustrated in Fig. 1, BOOST contains three stages: 1) *UE clustering for interference control* (CIC), 2) *traffic balancing UE association* (TUA), and 3) *beamforming and UE traffic scheduling* (BAS). In our experiments, we have observed that if a group of UEs are close to each other, then forming multiple beams to individual UEs creates significant interference between those beams. Therefore, for each BS, we first group its UEs into clusters with a CIC algorithm, based on their AoAs and spatial distances. The clusters have a radial shape with the BS as the center, as illustrated in Fig. 1(b). In the CIC stage, the UEs of a BS refer to the UEs that can be reached with the maximum range of the BS through beamforming. Many UEs may be reached by two or more BSs. Hence, different BSs can have *overlapping clusters* due to such UEs. For example, in Fig. 1(b), CIC forms nine clusters for a network with ten UEs and two BSs. The three clusters of BS2 are all overlapped with clusters of BS1. There are 7 *overlapping UEs* included in these overlapping clusters.

While the BS association for non-overlapping UEs is straightforward, since each of them can be reached by only one BS, it is a challenge to properly associate overlapping UEs such that the network transmission time is minimized while the inter-BS interference is avoided or significantly reduced. As illustrated in Fig. 1(c), TUA smartly associates overlapping UEs to the clusters based on the cluster transmission time, to achieve this objective. The CIC and TUA algorithms are re-run when there is a significant change to the channel. Typically the channel characteristics between a

UE and a BS remains about the same in the order of seconds. Therefore, the CIC and TUA algorithms do not result in significant overhead.

Due to the power budget constraint, and the minimum SINR requirement, a BS often cannot form beams to cover all clusters simultaneously, with one beam to cover one cluster. Hence, BAS has to schedule the beams of a BS across a number of time slots. For example, the clusters in Fig. 1(c) have to be covered across two slots illustrated in Fig. 1(d) and (e). In each slot, BAS computes beamforming vectors for selected clusters, allocates the transmit power for each beam, and allocates spectrum for each UE in the beam to carry its traffic. At last, some beams may each have to be scheduled across multiple time slots, as the traffic transportation for the covered UE set cannot be finished in one slot. For example, in Fig. 1(d) and (e), the beams covering UE sets {3, 4}, {6, 7}, and {8, 9} are scheduled across two time slots while the beams covering sets {1, 2}, {5}, and {10} are scheduled in one slot. The objective of BAS is to minimize the network transmission time, i.e., the number of slots to finish traffic delivery for all beams/UEs. Next, we describe each component of BOOST in details.

4.1 UE Clustering for Interference Control (CIC)

Clustering is generally NP-hard [29]. Hence, in practice, heuristic algorithms are widely used. The challenge for clustering UEs is that unlike classic clustering that typically uses the spatial distance, we have to take beamforming into account. Furthermore, we also have to design the CIC algorithm to help achieving the objective of BOOST, reducing latency. In this paper, we design a heuristic clustering algorithm to reduce the interference between beams, and increase the average SINR at UEs in each cluster, both of which help to increase capacity and reduce latency. Moreover, UEs are grouped based on both AoAs (with regard to the BS) and the spatial distance.

Algorithm 1 describes the CIC algorithm. At the beginning each UE is treated as a cluster with the corresponding AoA for beamforming direction. We use the equal power allocation for all clusters of a BS to compute the UE SINR. In each successive iteration, the CIC algorithm identifies two best candidate clusters such that merging them could increase the UE SINR (line 7–9). We select two candidate clusters with the minimum angular distance, which also needs to be not larger than $1/L$ (line 4), because the beamforming gain approaches zero by (5) if it is larger than $1/L$. This process continues until we cannot find such candidate clusters or all UEs are processed. The CIC algorithm output is the set of UE clusters in $\mathcal{K}$. Note that we do not limit the number of beams, nor consider the inter-BS interference between beams in the CIC stage, which will be taken care by TUA and BAS.

Both CIC and TUA use the cosine pattern channel gain given in (5) to compute the UE interference and SINR by (7) and (6), respectively, while BAS computes the actual channel gain and SINR using the beamforming vector.

4.2 Traffic Balancing UE Association (TUA)

In BOOST, the CIC algorithm groups UEs into clusters for each BS independently, assuming the maximum range for the BS. Nevertheless, we should not directly form beams to cover those clusters, because clusters from different BSs can overlap with each other, resulting in significant inter-BS interference. TUA selects a unique BS (or the corresponding cluster) for each overlapping UE (belonging to more than one cluster), so that each overlapping UE goes to one cluster only and there is no overlapping any more between clusters. The objective of TUA is to reduce the transmission time, which is closely relevant to the traffic load. This is achieved through two techniques. First, to reduce the network transmission time, BOOST smartly balances UE traffic load between BSs. Second, to reduce the transmission time for a cluster, BOOST balances transmission times of UEs in the cluster, i.e., avoids the situation where some UEs in the cluster have completed traffic transmission, and other UEs have significant residual traffic to be transmitted. In other words, we would like the transmission time for every UE in the cluster to be the same. Hence we allocate spectrum for each UE in proportion to its traffic load and received SINR. Next, we find the transmission time for a cluster.

Let $\mathcal{S}_{mk} \in \mathcal{K}_m$ denote the k-th UE cluster of BS m obtained by the CIC algorithm. The UEs in cluster $\mathcal{S}_{mk}$ are covered by the corresponding beam $\langle m, k\rangle$ from BS m. Let $\gamma_{n,m,k}$ and b_n denote the received SINR and traffic load of UE $n \in \mathcal{S}_{mk}$, respectively.

Theorem 1. *Let v denote the total system bandwidth. If the system bandwidth is allocated to UEs of $\mathcal{S}_{mk}$ in proportion to their traffic loads and SINR, i.e., $\frac{b_n}{\log_2(1+\gamma_{n,m,k})}$, the transmission time for cluster $\mathcal{S}_{mk}$, denoted as $t(\mathcal{S}_{mk})$, is equal to the transmission time of every UE, and is given as*

$$t(\mathcal{S}_{mk}) = \sum_{i \in \mathcal{S}_{mk}} \frac{b_i}{v \log_2(1+\gamma_{i,m,k})}. \tag{15}$$

Algorithm 1: UE Clustering for Interference Control

Input: UE set $\mathcal{N}_m$ of BS m, AoAs $\hat{\boldsymbol{\theta}}$, transmit power p_o

1 Let each UE be a cluster, with the set of clusters $\mathcal{K}_m = \{\{1\}, \ldots \{|\mathcal{N}_m|\}\}$, and their AoAs $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_{|\mathcal{N}_m|}\}$

2 $\mathcal{X} \leftarrow \emptyset$

3 **while** $\mathcal{N}_m \setminus \mathcal{X}$ *is not empty* **do**

4 Select two clusters $(i, j) = \underset{(s,k)\in\mathcal{K}_m, s\neq k}{\operatorname{argmin}} |\hat{\theta}_s - \hat{\theta}_k|$ such that $|\hat{\theta}_s - \hat{\theta}_k| \leq 1/L$

5 **if** (i, j) *is empty* **then**

6 break

7 Find the UE SINR in the combined cluster $h = (i, j)$ and other clusters in $\mathcal{K}_m$ except i and j, each with power $\frac{p_o}{|\mathcal{K}_m|-1}$ using (6)

8 **if** *min. SINR for UEs in cluster h is greater than the min. SINR in both clusters i and j* **then**

9 Use cluster h to replace clusters i, j in $\mathcal{K}_m$, and accordingly update the AoAs $\hat{\boldsymbol{\theta}}$.

10 **else**

11 Add all UEs in clusters i and j to $\mathcal{X}$

12 **Output:** $\mathcal{K}_m$ as the set of clusters of BS m

Proof. Let $v_{n,m,k}$ denote the spectrum share for UE $n \in \mathcal{S}_{mk}$, the required transmission time for UE n, denoted as $t_{n,m,k}$, is given by

$$t_{n,m,k} = \frac{b_n}{v_{n,m,k} \log_2(1+\gamma_{n,m,k})}. \quad (16)$$

If we allocate spectrum for each UE in proportion to its traffic load and received SINR, then the spectrum share for UE n can be written as

$$v_{n,m,k} = \frac{\beta_n v}{\sum_{i \in \mathcal{S}_{mk}} \beta_i}, \quad (17)$$

where β_n is the coefficient used for spectrum share of UE n, and is computed using its traffic load and SINR as below

$$\beta_n = \frac{b_n}{\log_2(1+\gamma_{n,m,k})}. \quad (18)$$

After substituting β_n into (17), the spectrum share for UE n is derived as follows

$$v_{n,m,k} = \frac{v b_n \prod_{i \in \mathcal{S}_{mk}, i \neq n} \log_2(1+\gamma_{i,m,k})}{\sum_{i \in \mathcal{S}_{mk}} b_i \left[\prod_{j \in \mathcal{S}_{mk}, j \neq i} \log_2(1+\gamma_{j,m,k})\right]}. \quad (19)$$

After substituting $v_{n,m,k}$ into (16), the transmission time for UE n, $t_{n,m,k}$, becomes

$$t_{n,m,k} = \frac{b_n}{v_{n,m,k} \log_2(1+\gamma_{n,m,k})} = \sum_{i \in \mathcal{S}_{mk}} \frac{b_i}{v \log_2(1+\gamma_{i,m,k})}. \quad (20)$$

It is clear that by (20), the transmission time to complete traffic delivery to every UE in $\mathcal{S}_{mk}$ is the same, i.e., $t_{1,m,k} = \cdots = t_{|\mathcal{S}_{mk}|,m,k}$. Hence, (20) also denote the transmission time $t(\mathcal{S}_{mk})$ for cluster $\mathcal{S}_{mk}$. ■

Next, we use the cluster transmission time in (15) as the cost metric for UE association, to balance traffic amount and UEs between BSs, with the objective to reduce the network transmission time. For a BS, some of its clusters may be overlapped with clusters of other BSs, while others are not, as illustrated in Fig. 2. Note that the clusters of the same BS do not overlap with each other. Let $\mathcal{L}$ denote the set of UE clusters of all BSs which is obtained by the CIC algorithm. Let $\mathcal{R}$ denote the set of *non-overlapping* clusters in $\mathcal{L}$, i.e.,

$$\mathcal{R} = \{S_i \mid S_i \in \mathcal{L} \text{ and } \forall S_j \in \mathcal{L} \backslash S_i, j \neq i, S_i \cap S_j = \emptyset\}. \quad (21)$$

Let $\mathcal{R}_m \subseteq \mathcal{R}$ denote the set of non-overlapping clusters of BS m. We define the cost of BS m, denoted as c_m, as the transmission time of the BS, which is the maximum transmission time of its clusters. This is used in the TUA algorithm to determine which BS an overlapping UE should be associated to. However, the challenge is that we do not know the transmission times of overlapping clusters until the association of overlapping UEs is completed. Hence, during the execution of the TUA algorithm, we set the BS cost as the maximum transmission time of its clusters that have been processed so far. Initially, the BS cost is set as the maximum transmission time of the non-overlapping clusters, i.e.,

Algorithm 2: Traffic Balancing UE Association (TUA)

Input: Clusters $\mathcal{L}$ from CIC, UE set $\mathcal{N}$
Output: Clusters $\bar{\mathcal{L}}$ with UEs uniquely associated

1 $\bar{\mathcal{L}} \leftarrow \mathcal{L} \setminus \mathcal{R}$. $\mathcal{X} = \bigcup_{S \in \mathcal{R}} S$
2 Compute the initial cost vector $\mathbf{c}$ of $|\mathcal{M}|$ BSs by (22)
3 **while** $\mathcal{X} \neq \mathcal{N}$ **do**
4 Select a cluster $S^* = \underset{S \in \bar{\mathcal{L}}}{\operatorname{argmax}} \sum_{n \in S} b_n$ and $S^* \setminus \mathcal{X} \neq \emptyset$
5 Select a UE $n^* = \underset{n \in S^* \setminus \mathcal{X}}{\operatorname{argmax}} \|\mathbf{h}_n\|^2$
6 Let $\{\bar{S}_1, \ldots, \bar{S}_j, \ldots, \bar{S}_J \mid n^* \in \bar{S}_j \in \bar{\mathcal{L}}\}$ denote the clusters in $\bar{\mathcal{L}}$ that include n^*
7 For $1 \leq j \leq J$, $S_j = \{n^*\} \bigcup \{n | n \in \bar{S}_j, \|\mathbf{h}_n\|^2 > \|\mathbf{h}_{n^*}\|^2\}$
8 **if** *n^* is an overlapping UE, i.e. $J > 1$* **then**
9 For $1 \leq j \leq J$, compute the temporary cost of the BS of S_j, $c'_{m(S_j)} = \max(c_{m(S_j)}, t(S_j))$
10 $j^* = \underset{1 \leq j \leq J}{\operatorname{argmin}} c'_{m(S_j)}$
`// Associate UE` n^* `to BS` $m(S_{j^*})$
11 For any $S \in \bar{\mathcal{L}}, S \neq S_{j^*}, S \leftarrow S \setminus S_{j^*}$
12 Update BS $m(S_{j^*})$ cost $c_{m(S_{j^*})} = c'_{m(S_{j^*})}$
13 $\mathcal{X} = \mathcal{X} \bigcup S_{j^*}$
14 **else**
15 If $t(S_1) > c_{m(S_1)}$, let $c_{m(S_1)} = t(S_1)$
16 $\mathcal{X} = \mathcal{X} \bigcup S_1$

$$c_m = \max_{s \in \mathcal{R}_m} (t(s)), \text{ or } 0 \text{ if } \mathcal{R}_m = \emptyset. \quad (22)$$

It is updated whenever an overlapping cluster of the BS is processed in the TUA algorithm. That is, c_m dynamically changes whenever a UE is associated to BS m.

TUA only needs to be applied on UEs in overlapping clusters. TUA associates an overlapping UE (belonging to more than one cluster) to the right cluster, such that the interference between BSs is reduced. *In the ensuing discussions, for the ease of description, we interchangeably refer to UE association to a cluster as to the BS of the cluster*. Algorithm 2 illustrates the TUA algorithm. The main idea is that for each overlapping UE n, we associate it with the cluster that results in the lowest BS cost. This in turn results in a smaller network transmission time. In the algorithm, $\mathcal{X}$ denotes the UEs that have been associated so far. In each iteration, TUA picks a cluster $\mathcal{S}^*$ that has the highest traffic load. Then it chooses an unprocessed UE n^* in cluster S^* with the highest channel gain. Next, TUA finds all other clusters that overlap due to UE n^*, denoted as $\{\bar{S}_1, \ldots, \bar{S}_j, \ldots, \bar{S}_J \mid n^* \in \bar{S}_j \in \bar{\mathcal{L}}\}$. UE n^* is associated to the cluster that results in the lowest BS cost (lines 9–13). While associating n^*, we bundle it with the UEs with a higher channel gain as they are likely to be associated together. Hence, in Algorithm 2, $\{S_1, \ldots, S_j, \ldots, S_J\}$ are used instead of $\{\bar{S}_1, \ldots, \bar{S}_j, \ldots, \bar{S}_J\}$.

Fig. 2 illustrates the main idea of TUA, where S_1 and S_2 are overlapping clusters, with UE b being an overlapping UE. Fig. 2(b) illustrates the BS costs before processing UE

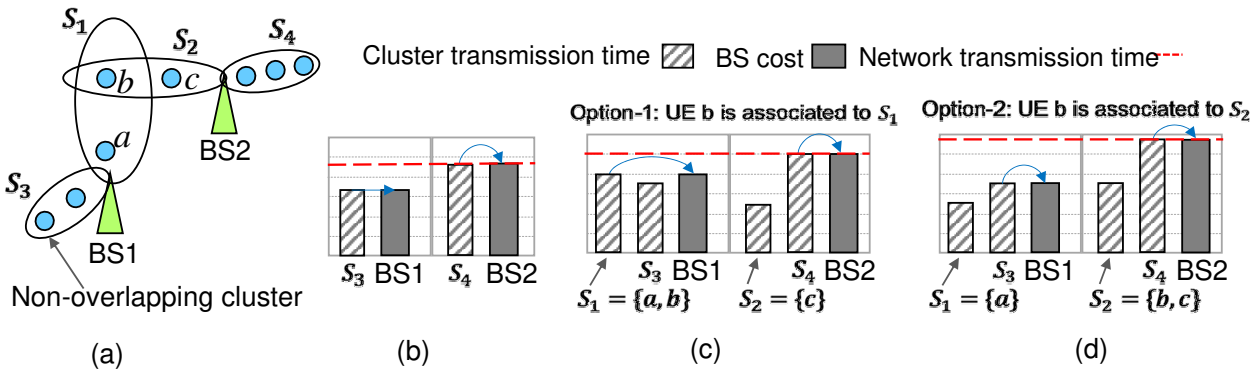

Fig. 2. UE association by TUA, (a) a 2-BS network, (b) before processing UE b, (c) associating b to S1, (d) associating b to S2

b, which are the transmission times of clusters S_3 and S_4, respectively. Fig. 2(c) and (d) illustrate the BS cost change if we associate b to cluster S_1 or S_2. In option 1, the BS1 cost changes to the transmission time of cluster S_1, as it is larger than the transmission time of cluster S_3. However, this is still lower than the BS2 cost. Hence the network transmission time is still the BS2 cost, which slightly increases due to the power splitting to cover UE c. In option 2, UE b is associated to cluster S_2. To cover both b and c, BS2 needs to split more power, and hence the transmission time of S_4 increases further, which results in a higher network transmission time compared with option 1. Hence TUA selects option 1.

4.3 Beamforming and UE Traffic Scheduling (BAS)

Given the UE clusters obtained by TUA, the BAS algorithm schedules which clusters into a time slot. Then it computes the BS beamforming weight vectors, and allocates power for each beam. Through scheduling as much traffic as possible into every slot, BAS aims to reduce the number of time slots required to deliver traffic of all UEs subject to the BS power budget, UE SINR requirement, and UE traffic loads.

4.3.1 Beamforming Weight Vector

We use the *linearly constrained minimum variance* (LCMV) scheme to compute beamforming weight vectors [30, p.513]. LCMV is able to suppress the power response toward the directions of other beams if there is a sufficient angular space from those beams. To achieve a complex gain g^* in the UE direction θ, the weight vector $\mathbf{w}$ needs to satisfy $\mathbf{a}(\theta)^H\mathbf{w} = g^*$, where $\mathbf{a}(\theta)$ is given in (3). Let $\mathbf{A_\theta} = [\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_k)]$ be the constraint matrix for total k UEs of a BS, with AoAs $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$. Let $\mathbf{f}$ be the k-dimension single column response vector of the k UEs. The covariance matrix of the k UE signals $\mathbf{R} = \mathbb{E}\left[\mathbf{h}\mathbf{h}^H\right] = \frac{L|\boldsymbol{\alpha}|^2}{\mathrm{PL}}\mathbf{A}_\theta\mathbf{A}_\theta^H$, where $\mathbf{h}$ is a matrix consisted of the k UE channel vectors given in (2), L is the number of antenna elements of a BS, PL is the path loss vector of the k UEs, and $\boldsymbol{\alpha}$ is the complex gain vector of the k UE signals. The LCMV beamforming to minimize the transmit power is formulated as

$$\min\{\mathbf{w}^H\mathbf{R}\mathbf{w}\} \text{ subject to } \mathbf{A}_\theta^H\mathbf{w} = \mathbf{f}. \tag{23}$$

The closed-form solution of (23) is obtained as follows using the Lagrange multiplier method

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{A}_\theta(\mathbf{A}_\theta^H\mathbf{R}^{-1}\mathbf{A}_\theta)^{-1}\mathbf{f}. \tag{24}$$

By choosing an appropriate weight vector for each UE cluster, it is possible to approximately eliminate or significantly reduce the interference between beams, as long as they are sufficiently separated in the angular space. The time complexity of the LCMV beamforming in (24) is $\Theta(\max(kL^2, L^{2.373}))$ assuming $L \geq k$, where $L^{2.373}$ is the matrix inverse time for an $L \times L$ matrix.

4.3.2 Interference Controlled Power Allocation

While BAS schedules a cluster into a time slot, it may select to cover only a subset of UEs in the cluster. Let $\mathcal{F}$ denote the list of UE sets to be scheduled in a slot. Let $\mathcal{F}_i$ denote the ith set in $\mathcal{F}$. Let m_i denote the corresponding BS for set $\mathcal{F}_i$. For the UE sets $\mathcal{F}_i$ ($1 \leq i \leq |\mathcal{F}|$), we construct the constraint matrix $\mathbf{A}_\theta$, the corresponding covariance matrix $\mathbf{R}$, and the response vector $\mathbf{f}$. Then we use (24) to compute $|\mathcal{F}|$ number of normalized weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{|\mathcal{F}|}$, where $\mathbf{w}_i$ is for set $\mathcal{F}_i$. By (9), the transmit power allocated to the beam to cover set $\mathcal{F}_i$ is computed as

$$p_i = \max_{n \in \mathcal{F}_i}\left(\frac{(I_n + \sigma^2)\gamma_o}{|\mathbf{h}_{m_i,n}^H\mathbf{w}_i|^2}\right), \quad \forall\, 1 \leq i \leq |\mathcal{F}| \tag{25}$$

where $\mathbf{h}_{m_i,n}$ is the channel vector between BS m_i and UE n, γ_o is the required minimum SINR threshold for all UEs, and σ^2 is the noise power in the AWGN channel at the UE.

The power computed by (25) is the minimum power to ensure the SINR threshold for UEs. If a BS has more power than needed for its beams, the following theorem states that the residual power can be allocated to the beams in proportion to their minimum power, while achieving higher SINRs for all UEs.

Theorem 2. *Let $\mathbf{p}_m$ denote the power vector of the beams of BS m computed by (25) and p_{mi} denote the power of beam $< m, i >$. Let $\gamma_{n,m,i}(\mathbf{p}_m)$ denote the SINR of UE n in beam $< m, i >$ under power allocation $\mathbf{p}_m$. If $\sum_i p_{mi} < p_o$, the power allocation $\mathbf{p}'_m = \frac{p_o}{\sum_i p_{mi}}\mathbf{p}_m$ meets the BS power constraint p_o, while the UE n SINR $\gamma_{n,m,i}(\mathbf{p}'_m)$ is larger than the original SINR $\gamma_{n,m,i}(\mathbf{p}_m)$.*

Proof. The total transmit power of the BS under $\mathbf{p}'_m$ is $\sum_i \frac{p_o}{\sum_i p_{mi}} p_{mi} = \frac{p_o}{\sum_i p_{mi}} \sum_i p_{mi} = p_o$. Hence it meets the

power budget requirement. With BS power vector $\mathbf{p}_m$, the SINR at UE n is given by (6) as $\gamma_{n,m,i}(\mathbf{p}_m) = \frac{p_{mi}|\mathbf{h}_{mn}^H \mathbf{w}_{mi}|^2}{I_{n,m,i}+\sigma^2}$, where σ^2 is the noise power and $I_{n,m,i}$ is the sum of intra-BS and inter-BS interference at UE n given by (7). As the CIC algorithm ensures that AoAs of UEs in different clusters of the same BS are spatially separated by at least $1/L$, where L is the number of antennas at BS, the constraint in (23) guarantees negligible interference between UEs in different clusters of the same BS. That means the normalized intra-BS interference at UE n from other beams $\langle m, j\rangle$, $j \neq i$, is approximately 0, i.e.,

$$\sum_{j\in\mathcal{K}_m\setminus i} |\mathbf{h}_{mn}^H \mathbf{w}_{mj}|^2 = \sum_{j\in\mathcal{K}_m\setminus i} \frac{L|\alpha_{mn}|^2}{\mathrm{PL}_{mn}} |\mathbf{a}^H(\theta_{mn})\mathbf{w}_{mj}|^2 \approx 0,$$

where α_{mn} is the signal path complex gain, PL_{mn} is the path loss, and $\mathbf{a}^H(\theta_{mn})$ is the BS m antenna response vector toward the direction of UE n. For clusters from different BSs, the inter-BS interference becomes very high if the angular spatial separation between clusters is small. Line 12 of Algorithm 3 (to be discussed) effectively prevents such significantly interfering beams of different BSs from scheduled in the same slot. This technically reduces the inter-BS interference to be close to zero. Thus, the normalized inter-BS interference at UE n from beams $\langle k, l\rangle$, $k \neq m$, is

$$\sum_{k\in\mathcal{M}\setminus m} \sum_{l\in\mathcal{K}_k} \frac{L|\alpha_{kn}|^2}{\mathrm{PL}_{kn}} \underbrace{|\mathbf{a}^H(\theta_{kn})\mathbf{w}_{kl}|^2}_{\approx 0} \approx 0.$$

As a result, increasing power in one beam does not cause significant interference to another beam since both interference $I'_{n,m,i}$ with power $\mathbf{p}'_m$ and interference $I_{n,m,i}$ with power $\mathbf{p}_m$ are very small. If $\sum_i p_{mi} < p_o$, we have $p'_{mi} > p_{mi}$ by the allocation of residual power. Therefore, we have

$$\gamma_{n,m,i}(\mathbf{p}'_m) = \frac{p'_{mi}|\mathbf{h}_{mn}^H \mathbf{w}_{mi}|^2}{I'_{n,m,i}+\sigma^2} \approx \frac{p'_{mi}|\mathbf{h}_{mn}^H \mathbf{w}_{mi}|^2}{\sigma^2} > \frac{p_{mi}|\mathbf{h}_{mn}^H \mathbf{w}_{mi}|^2}{\sigma^2} \approx \frac{p_{mi}|\mathbf{h}_{mn}^H \mathbf{w}_{mi}|^2}{I_{n,m,i}+\sigma^2} = \gamma_{n,m,i}(\mathbf{p}_m).$$

■

4.3.3 UE Scheduling

The UE scheduling has two components. First, for each BS, it selects which UE clusters for transmission in the current slot, subject to the BS power budget and number of beamformers. We use a greedy scheme to select clusters with larger traffic loads. Second, for a selected cluster for the current slot, depending on its total traffic load, the UE scheduling may select all UEs in the cluster, or a subset of UEs, or even a single UE for transmission. We introduce two propositions for selecting UEs in a cluster.

Proposition 1. *Let Δ_b denote the traffic load that can be delivered with SINR γ_o in a slot duration. If the total traffic load of a cluster $\mathcal{S}$ is less than Δ_b, i.e., $\sum_{n\in\mathcal{S}} b_n \leq \Delta_b$, the transmission time for cluster $\mathcal{S}$ can be reduced by $(|\mathcal{S}|-1)$ slots if a single beam is formed to cover all UEs in $\mathcal{S}$ rather than one beam per UE (multi-user beamforming).*

Proof. If *one beam per UE* is formed for all UEs simultaneously, the interference between beams is typically very high as those UEs are very close to each other. Hence, the minimum SINR cannot be met. To avoid interference, each UE has to be scheduled into a different slot. Note that even though the transmission of the traffic of one UE needs only a fraction of slot duration, it is usually not possible to form a beam for another UE in the middle of a slot [31]. This results in total $|\mathcal{S}|$ slots. In contrast, if one beam is formed to cover all UEs in the cluster, the total traffic can be transmitted in one slot, reducing the transmission time by $(|\mathcal{S}|-1)$ slots. ■

Proposition 2. *For a cluster $\mathcal{S}$, if $\max_{n\in\mathcal{S}} b_n < \Delta_b$ and $\sum_{n\in\mathcal{S}} b_n > \Delta_b$, i.e., the traffic load of any single UE is smaller than Δ_b but the total load is larger than Δ_b, then a subset of UEs $\mathcal{S}' = \underset{s\subseteq\mathcal{S}}{\operatorname{argmin}} \left([\Delta_b - \sum_{n\in s} b_n] \geq 0\right)$ can be scheduled into the current slot. This helps to reduce the cluster $\mathcal{S}$ transmission time.*

Proof. The total UE traffic load of $\mathcal{S}'$ is not larger than Δ_b. Hence by Proposition 1, the traffic transmission to all UEs in $\mathcal{S}'$ can be completed in one slot, by forming one beam to cover all UEs. Moreover, scheduling $\mathcal{S}'$ into the current slot minimizes the remaining traffic of cluster $\mathcal{S}$ to be transmitted in future slots. If such scheduling is repeatedly

Algorithm 3: Beamforming and UE Traffic Scheduling

Input: UE clusters $\bar{\mathcal{L}}$, UE traffic $\mathbf{b}^{[t]}$ at slot t
Output: Remaining UE traffic $\mathbf{b}^{[t+1]}$ for slot $t+1$

1 $\mathcal{F} \leftarrow \{\emptyset\}$. $\quad \mathcal{L} = \bar{\mathcal{L}} \cup R$
2 **while** $\mathcal{L} \neq \emptyset$ **do**
3 $\quad s^* = \underset{s}{\operatorname{argmax}} \left(\sum_{n\in s, s\in\mathcal{L}} b_n^{[t]}\right)$
4 $\quad$ **if** # *of clusters in $\mathcal{F}$ for the BS of $s^* < K$* **then**
5 $\quad\quad$ Let $s' = s^*$ and select a UE in s' with max. traffic, i.e., $n^* = \underset{n\in s'}{\operatorname{argmax}} \left(b_n^{[t]}\right)$
6 $\quad\quad$ **if** $b_{n^*}^{[t]} \geq \Delta_b$ **then**
7 $\quad\quad\quad s' = \{n^*\}$
8 $\quad\quad$ **else if** $\sum_{n\in s^*} b_n^{[t]} > \Delta_b$ *and* $b_{n^*}^{[t]} < \Delta_b$ **then**
9 $\quad\quad\quad s' = \underset{s\subseteq s^*}{\operatorname{argmin}} \left([\Delta_b - \sum_{n\in s} b_n^{[t]}] \geq 0\right)$
10 $\quad\quad \mathcal{F} \leftarrow \mathcal{F} \cup \{s'\}$ `// Add s' to F`
11 $\quad\quad$ Compute $\mathbf{w}_1, \ldots, \mathbf{w}_{|\mathcal{F}|}$ and $\mathbf{p} = [p_1, \ldots, p_{|\mathcal{F}|}]$ by (24) and (25)
12 $\quad\quad$ **if** *sum of power of all beams of any BS is greater than p_o* **then**
13 $\quad\quad\quad \mathcal{F} \leftarrow \mathcal{F} \setminus s'$
14 $\quad \mathcal{L} \leftarrow \mathcal{L} \setminus s^*$
15 For all $m \in \mathcal{M}$, update $\mathbf{p}_m$ as $\mathbf{p}_m = \frac{p_o}{\sum_i p_{mi}} \mathbf{p}_m$ by Theorem 2
16 Compute the SINR $\gamma_n^{[t]}$, bandwidth $w_n^{[t]}$, and remaining traffic $b_n^{[t+1]}$ for each UE in the clusters in $\mathcal{F}$ by (6), (19), and (13)
17 Remove a UE from $\mathcal{N}$ and all clusters in $\bar{\mathcal{L}} \cup R$ if its traffic delivery can be finished in this slot

applied on cluster $\mathcal{S}$, the total cluster transmission time is effectively reduced. ■

Algorithm 3 illustrates the BAS algorithm. In each iteration, BAS chooses a cluster s^* with the maximum traffic load (line 3). It then schedules UEs of s^* to the current slot as follows. *First*, if the maximum-load UE n^* has a traffic load larger than Δ_b, then schedule only UE n^* in the current slot (line 7). *Otherwise*, if the total traffic load of all UEs in s^* is larger than Δ_b, then by Preposition 2, schedule a subset $s' \subset s^*$ to this slot (line 9). *At last*, if the total traffic load of all UEs in s^* is smaller than Δ_b, then by Preposition 1, schedule the entire s^* to this slot. The beamforming weight vectors and transmit power vectors for all clusters in $\mathcal{F}$ are then computed (line 11). Next, every BS is checked and enforced for the power budget constraint (line 12). This is a critical step to prevent significant interference between beams from different BSs to transmit in the same slot, as such interference likely results in violation of the BS power constraint. The power allocation in line 15 follows Theorem 2. The BAS algorithm continues for the next time slot, until the traffic of all UEs is scheduled.

TABLE 2
System parameters

Parameter	Value
System operating frequency	73 GHz
Number of BSs	4
Number of UEs	100
Network dimension	$[300 \times 300]$ m^2
Number of antennas in ULA	64
Adjacent antenna spacing Δ_d in ULA	$\lambda/2$
Number of beamformers K per BS	10
System bandwidth (FDD duplex mode)	400 MHz
Time slot duration Δ_t	0.125 ms
LOS path loss exponent η	2.1
Minimum required SINR γ_o at UE	15 dB
Noise power spectral density N_o	-174 dBm/Hz
Noise figure NF at UE receiver	6 dB

TABLE 3
Transmission time improvement of BOOST

UE Pattern	vs. MUBFP	vs. Chord-Dis	vs. SUS
Spread	29%	29%	24%
Grouped	41%	32%	28%
Dense	40%	30%	28%

5 Performance Evaluation

We evaluate the performance of BOOST comparatively with three state-of-the-art UE scheduling schemes in the literature: MUBFP [21], Chord-Dis [17], and SUS [10]. Those three schemes use *multiuser beamforming*, where one beam is formed for each UE. Table 2 lists system parameters used in simulations. We assume BSs have a ULA antenna with 64 elements. UEs are equipped with a single antenna. The sample networks are in a $[300 \times 300]$ m^2 outdoor area, with 4 BSs and 100 active UEs. The BSs positions are evenly distributed in the network. The UEs are distributed in three patterns; *spread*, *grouped*, and *dense*. With the spread UE distribution, UE locations are randomly generated in the network area. With the grouped UE distribution, 20 UE groups are created first and then each of the 100 UEs randomly selects a group. Each BS is assigned with 5 groups. Note that this assignment of groups to BSs is for the sole purpose of generating UE locations, irrelevant to the actual UE association. For each group, we randomly generate a 20 or 25-degree sector with regard to its BS. For each UE, we generate a random angle within the sector of its group, and a random distance between 10 m and 106 m from the BS of its group. The dense UE distribution is similar to the grouped UE distribution, but with 3 UE groups for each BS. In the experiments, the LCMV beamforming scheme is used. The *reference signal received power* based UE association is used in MUBFP, Chord-Dis, and SUS. The OFDMA scheme is assumed for spectrum sharing among UEs in a beam. The time slot duration is assumed 0.125 ms, which is one of the slot durations defined in 3GPP [32]. The required minimum SINR γ_o is set to 15 dB which is needed to use high rate modulation schemes [31]. The traffic load of each UE is randomly generated between 12.5 KB and 250 KB, following a uniform distribution. Note that this is not the UE data rate, but the traffic amount of the UE. As the UE location from the UE discovery may be inaccurate, we use a Gaussian distribution with a zero mean and 0.5 m standard deviation to model the error on the UE X and Y coordinates from the UE discovery [33]. The location error of UEs, i.e., the distance between the actual UE location and the reported location by the UE discovery, ranges from 0.1 m to 1.6 m. The experiment results using larger standard deviation values have similar trends and are omitted due to the space limitation.

Fig. 3 illustrates the average network transmission time as a function of the BS power budget p_o, for 50 experiments using different seeds for UE location generation. Each data point indicates the network transmission time to complete delivery of the traffic loads of all UEs. Overall, BOOST significantly outperforms other schemes, thanks to its effective UE clustering, association, and scheduling, which altogether make it possible to form beams with significantly reduced interference. Moreover, BOOST not only achieves lower transmission time but also a better confidence interval. The 95% confidence interval of the transmission time from the 50 experiments is 0.15, 0.54, 0.33, and 0.53 ms for BOOST, MUBFP, Chord-Dis, and SUS, respectively. From the 'spread' to the 'dense' UE distribution, the transmission time increases. This is because the interference between beams increases. Thus a smaller number of beams can be scheduled in each slot to maintain the minimum SINR requirement at UEs. Table 3 illustrates the percentage decrease of the average transmission time of BOOST, with respect to MUBFP, Chord-Dis, and SUS. Averaging over the three UE distributions, BOOST reduces the transmission time by 37%, 30%, and 26%, respectively.

Next we evaluate the sum rate of UEs with regard to the BS power budget, as plotted in Fig. 4. BOOST clearly outperforms other schemes on the sum rate as well. Table 4 illustrates the percentage increase of the average sum rate of BOOST compared with other schemes. On average, BOOST achieves 56%, 43%, and 34% gain, respectively. From the 'spread' to the 'dense' UE distribution, the sum rate decreases in all schemes, due to the higher interference as

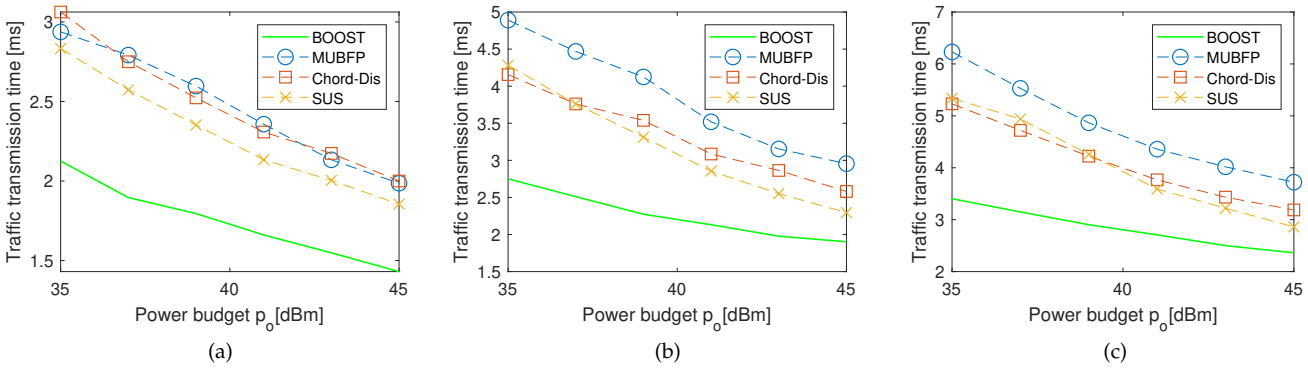

Fig. 3. Transmission time vs. BS power budget under 3 UE distributions: (a) spread, (b) grouped, and (c) dense

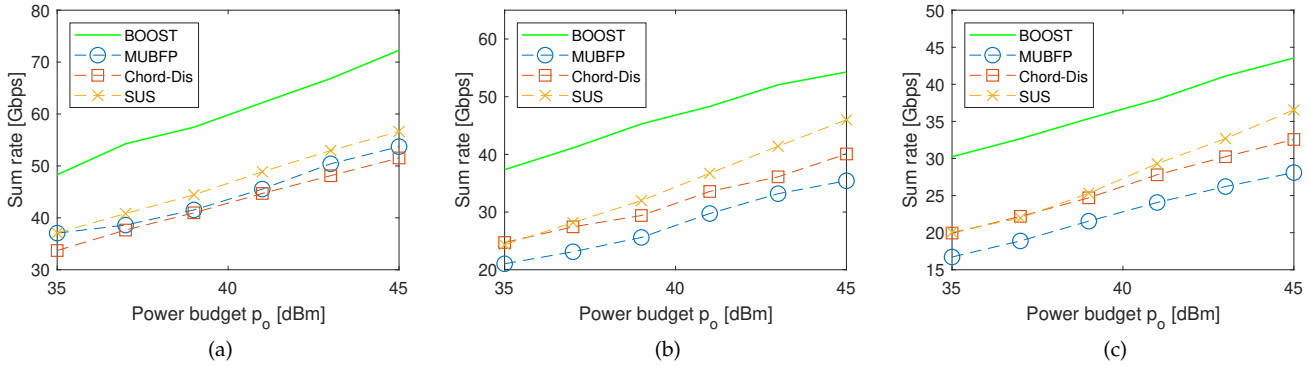

Fig. 4. Sum rate vs. BS power budget under 3 UE distributions: (a) spread, (b) grouped, and (c) dense

TABLE 4
Sum rate increase of BOOST

UE Pattern	vs. MUBFP	vs. Chord-Dis	vs. SUS
Spread	36%	41%	29%
Grouped	67%	46%	36%
Dense	65%	41%	36%

discussed above.

The above discussions are for the average performance among 50 experiments under each UE distribution. Next we examine how dispersed the results of individual experiments are. Figs. 5 and 6 plot the *probability density distribution* (PDF) of the network transmission time and UE SINR, respectively, obtained from 50 experiments under each UE distribution, given the BS power budget $p_o = 35$ dBm. The PDF results with different values of simulation parameters such as the power budget, minimum SINR, and slot duration exhibit similar trends and are omitted. From Fig. 5, BOOST has a smaller spread on both the transmission time and SINR. This indicates that BOOST can smartly adapt to different networks to achieve good performance in all scenarios. In contrast, other schemes cannot adapt to different networks well, i.e., their performance is good for some networks, but poor for others. In particular, the spread of UE SINR for BOOST is much smaller than the ones of other schemes. This demonstrates the advantage of BOOST–effectively guarantees a required SINR for each UE while wisely provides just enough SINR to all UEs, through smartly clustering and associating UEs to different clusters/beams. In contrast, other schemes result in highly dispersed SINR for different UEs.

At last, we evaluate the UE fairness on the transmission time using the Jain index, which ranges from 0 (worst fairness) to 1 (best fairness). It is at the maximum when the transmission time for all UEs is the same regardless of their traffic loads and locations in the network. Fig. 7 plots the mean and 95% confidence interval of the Jain index from all experiments. BOOST clearly outperforms other schemes, on both the Jain index value and the confidence interval. Furthermore, the fairness of BOOST is similar among all three UE distributions. Together with the much better confidence interval in all three cases, again, this indicates that BOOST can smartly adapt to different networks and achieves similar fairness. On the other hand, other schemes have significantly worse confidence intervals as well as dispersed fairness across three UE distributions, which means the fairness has a large variance across different networks.

6 Conclusion and Future Directions

We have developed a novel *clustering Based dOwnlink UE assOciation, Scheduling, beamforming with power allocaTion* (BOOST) framework for mmWave networks, with the objective to reduce the network transmission time for given UE traffic loads, subject to the BS power budget, minimum UE SINR requirement, and number of beamformers. We have compared BOOST with three state-of-the-art schemes on the transmission time, sum rate, SINR, and UE fairness. Overall BOOST significantly outperforms them. On average, BOOST reduces the transmission time by 37%, 30%, and 26%, and

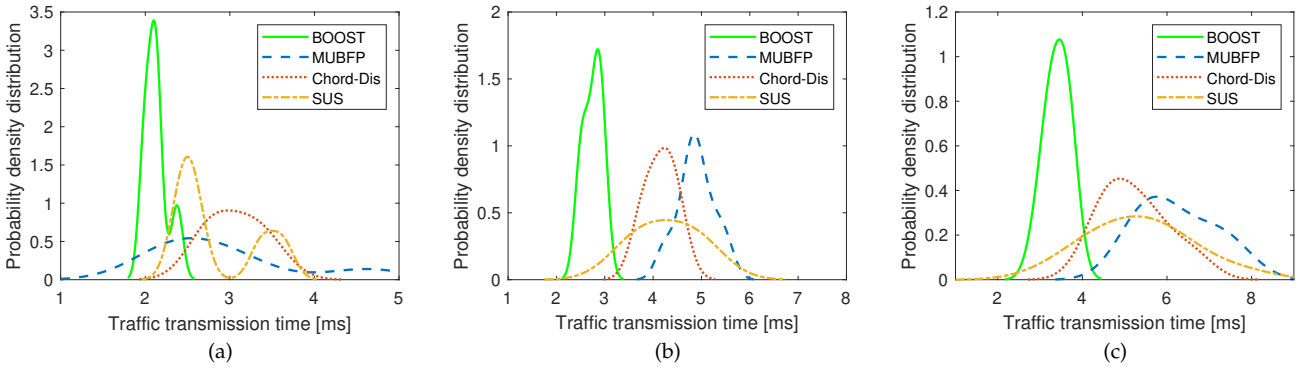

Fig. 5. PDF of transmission time under (a) spread, (b) grouped, and (c) dense UE distribution, with $p_o = 35$ dBm

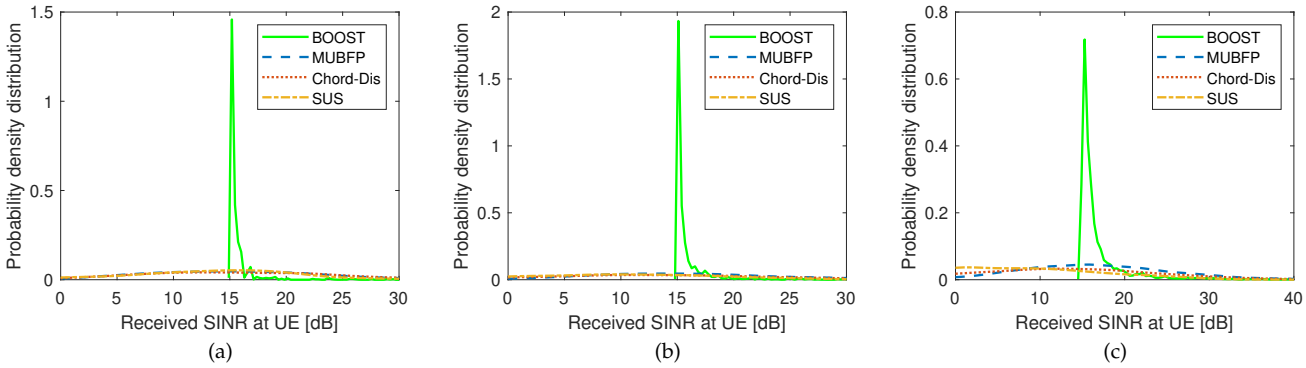

Fig. 6. PDF of UE SINR under (a) spread, (b) grouped, and (c) dense UE distribution, with power budget $p_o = 35$ dBm

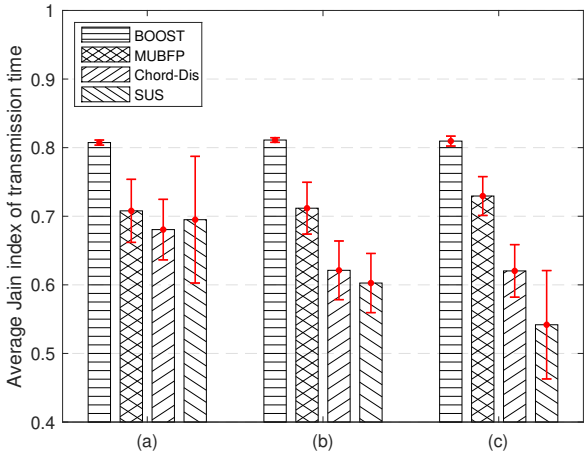

Fig. 7. Average Jain index of transmission times with 95% confidence interval under (a) spread, (b) grouped, and (c) dense UE distribution, with power budget $p_o = 35$ dBm

achieves a sum rate gain of 56%, 43%, and 34%, respectively, compared with other three schemes.

In this paper, we consider only the LOS paths between UEs and BSs. However, by the 3GPP model, the probability of a LOS path decreases with regard to the distance. Hence, for future directions, we plan to consider NLOS paths between UEs and BSs.

REFERENCES

[1] Ericsson, "Internet of things forecast," 2018. [Online]. Available: https://www.ericsson.com/en/mobility-report/internet-of-things-forecast

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[3] Z. He, S. Mao, S. Kompella, and A. Swami, "On link scheduling in dual-hop 60 GHz mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11 180–11 192, Dec. 2017.

[4] Y. Wang, S. Mao, and T. Rappaport, "On directional neighbor discovery in mmwave networks," in *Proc. IEEE ICDCS*, 2017.

[5] C. R. Anderson and T. S. Rappaport, "In-building wideband partition loss measurements at 2.5 and 60 GHz," *IEEE Trans. Wireless Commun.*, vol. 3, no. 3, pp. 922–928, May 2004.

[6] S. Huang, H. Yin, J. Wu, and V. C. Leung, "User selection for multiuser MIMO downlink with zero-forcing beamforming," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3084–3097, Sep. 2013.

[7] W.-L. Shen, K. C.-J. Lin, M.-S. Chen, and K. Tan, "SIEVE: Scalable user grouping for large MU-MIMO systems," in *Proc. IEEE Infocom*, 2015.

[8] X. Xia, S. Fang, G. Wu, and S. Li, "Joint user pairing and precoding in MU-MIMO broadcast channel with limited feedback," *IEEE Commun. Lett.*, vol. 14, no. 11, pp. 1032–1034, Nov. 2010.

[9] M. Min, Y.-S. Jeon, and G.-H. Im, "On achievable rate of user selection for MIMO broadcast channels with limited feedback," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 122–135, Jan. 2017.

[10] G. Lee and Y. Sung, "A new approach to user scheduling in massive multi-user mimo broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1481–1495, Apr. 2018.

[11] J. Mao, J. Gao, Y. Liu, and G. Xie, "Simplified semi-orthogonal user selection for MU-MIMO systems with ZFBF," *IEEE Wireless Commun. Lett.*, vol. 1, no. 1, pp. 42–45, Feb. 2012.

[12] A. Adhikary, E. Al Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, "Joint spatial division and multiplexing for mm-wave channels," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1239–1255, Jun. 2014.

[13] P. Paul, H. Wu, C. Xin, and M. Song, "Beamforming oriented topology control for mmwave networks," *IEEE Trans. Mobile Comput.*, in press.
[14] D. Lee, G. Y. Li, X.-L. Zhu, and Y. Fu, "Multistream multiuser coordinated beamforming for cellular networks with multiple receive antennas," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3072–3085, Mar. 2016.
[15] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, Feb. 2017.
[16] G. Lee, Y. Sung, and J. Seo, "Randomly-directional beamforming in millimeter-wave multiuser MISO downlink," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1086–1100, Feb. 2016.
[17] J. Choi, G. Lee, and B. L. Evans, "User scheduling for millimeter wave hybrid beamforming systems with low-resolution adcs," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2401–2414, Apr. 2019.
[18] K. Ko and J. Lee, "Multiuser MIMO user selection based on chordal distance," *IEEE Trans. Commun.*, vol. 60, no. 3, pp. 649–654, Mar. 2012.
[19] M. Taniguchi, H. Murata, S. Yoshida, K. Yamamoto, D. Umehara, S. Denno, and M. Morikura, "Indoor experiment of multi-user MIMO user selection algorithm based on chordal distance," in *Proc. IEEE Globecom*, 2013.
[20] B. Zhou, B. Bai, Y. Li, D. Gu, and Y. Luo, "Chordal distance-based user selection algorithm for the multiuser mimo downlink with perfect or partial CSIT," in *Proc. IEEE AINA*, Mar. 2011.
[21] B. Hu, C. Hua, C. Chen, X. Ma, and X. Guan, "MUBFP: Multiuser beamforming and partitioning for sum capacity maximization in MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 233–245, Jan. 2017.
[22] 3GPP TR 38.901 version 14.3.0 Release 14, "Study on channel model for frequencies from 0.5 to 100 GHz," 2018. [Online]. Available: https://www.3gpp.org/release-14
[23] X. Yu, J. Zhang, M. Haenggi, and K. B. Letaief, "Coverage analysis for millimeter wave networks: The impact of directional antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1498–1512, Apr. 2017.
[24] N. Deng and M. Haenggi, "A novel approximate antenna pattern for directional antenna arrays," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 832–835, Oct. 2018.
[25] M. F. Hanif, L.-N. Tran, A. Tölli, M. Juntti, and S. Glisic, "Efficient solutions for weighted sum rate maximization in multicellular networks with channel uncertainties," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5659–5674, Nov. 2013.
[26] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, S. S. Panwar, and T. S. Rappaport, "Directional cell discovery in millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6664–6678, Dec. 2015.
[27] Y. Li, J. G. Andrews, F. Baccelli, T. D. Novlan, and C. J. Zhang, "Design and analysis of initial access in millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6409–6425, Oct. 2017.
[28] 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.
[29] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is NP-hard," *Theoretical Computer Science*, vol. 442, pp. 13 – 21, Jul. 2012.
[30] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.
[31] 3GPP TS 36.101 version 14.5.0 Release 14, "Evolved universal terrestrial radio access (E-UTRA); user equipment (ue) radio transmission and reception," 2018. [Online]. Available: https://www.3gpp.org/release-14
[32] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 2018.
[33] M. Koivisto, A. Hakkarainen, M. Costa, P. Kela, K. Leppanen, and M. Valkama, "High-efficiency device positioning and location-aware communications in dense 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 188–195, Aug. 2017.

Prosanta Paul received his B.S. in from Khulna University of Engineering and Technology, Bangladesh, in 2009, and M.S. in Electrical Engineering from South Dakota School of Mines and Technology, South Dakota, in 2014. He is currently pursuing his Ph.D. degree in the Department of Electrical and Computer Engineering, Old Dominion University. His research interests include Millimeter wave communications, wireless communications and networking, cognitive radio networks, and Internet of Things.

Hongyi Wu is the Batten Chair of Cybersecurity and the Director of the Center for Cybersecurity Education and Research at Old Dominion University (ODU). He is also a Professor in Department of Electrical and Computer Engineering and holds joint appointment in Department of Computer Science. Before joining ODU, he was an Alfred and Helen Lamson Endowed Professor at the Center for Advanced Computer Studies (CACS), University of Louisiana at Lafayette (UL Lafayette). He received the B.S. degree in scientific instruments from Zhejiang University, Hangzhou, China, in 1996, and the M.S. degree in electrical engineering and Ph.D. degree in computer science from the State University of New York (SUNY) at Buffalo in 2000 and 2002, respectively. His research focuses on networked and intelligent cyber-physical systems for security, safety, and emergency management applications. He chaired several conferences such as IEEE Infocom 2020, IEEE WoWMoM 2016, and IEEE Globecom Wireless Communication Symposium 2015. He also served on the editorial board of several journals including IEEE Transactions on Mobile Computing, IEEE Transactions on Parallel and Distributed Systems and IEEE Internet of Things Journal. He received NSF CAREER Award in 2004, UL Lafayette Distinguished Professor Award in 2011, and IEEE Percom Mark Weiser Best Paper Award in 2018. He is a Fellow of IEEE.

ChunSheng Xin is a Professor in the Center for Cybersecurity Education and Research, and the Department of Electrical and Computer Engineering, Old Dominion University. He received his Ph.D. in Computer Science and Engineering from the State University of New York at Buffalo in 2002. His interests include cybersecurity, wireless communications and networking, cyber-physical systems, and Internet of Things. His research has been supported by almost 20 NSF and other federal grants, and results in more than 100 papers in leading journals and conferences, including three Best Paper Awards, as well as books, book chapters, and patent. He has served as Co-Editor-in-Chief/Associate Editors of multiple international journals, and symposium/track chairs of multiple international conferences including IEEE Globecom and ICCCN. He is a senior member of IEEE.