mmFlexible: Flexible Directional Frequency Multiplexing for Multi-user mmWave Networks

Ish Kumar Jain, Rohith Reddy Vennam, Raghav Subbaraman, Dinesh Bharadia
University of California San Diego, La Jolla, CA
{ikjain, rvennam, rsubbaraman, dineshb}@ucsd.edu

IEEE INFOCOM 2023 - IEEE Conference on Computer Communications | 979-8-3503-3414-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/INFOCOM53939.2023.10229065

***Abstract*—Modern mmWave systems have limited scalability due to inflexibility in performing frequency multiplexing. All the frequency components in the signal are beamformed to one direction via pencil beams and cannot be streamed to other user directions. We present a new flexible mmWave system called mmFlexible that enables flexible directional frequency multiplexing, where different frequency components of the mmWave signal are beamformed in multiple arbitrary directions with the same pencil beam. Our system makes two key contributions: (1) We propose a novel mmWave front-end architecture called a delay-phased array that uses a variable delay and variable phase element to create the desired frequency-direction response. (2) We propose a novel algorithm called FSDA (Frequency-space to delay-antenna) to estimate delay and phase values for the real-time operation of the delay-phased array. Through evaluations with mmWave channel traces, we show that mmFlexible provides a 60-150% reduction in worst-case latency compared to baselines[1].**

***Index Terms*—mmWave, beamforming, delay-phased array, frequency multiplexing, OFDMA, scheduling**

I. Introduction

Millimeter-wave (mmWave) networks have the potential to provide wireless connectivity to a growing number of users with their vast bandwidth resources. However, current mmWave systems have a significant limitation that they are unable to simultaneously serve multiple users by distributing small chunks of frequency resources to different users who are in different directions. Unlike sub-6 systems, that use Omni-antennas to radiate signal in all directions, mmWave systems use pencil beams that illuminate a small region in space, meaning that all the frequency components are directed towards a fixed direction and cannot be distributed to other directions. This inflexibility leads to two main issues, as illustrated in Figure 1(a). Firstly, it leads to high latency, as the base station (gNB) must serve different user directions in a time-division manner, causing some users to experience long wait times, which is detrimental to latency-sensitive applications. Secondly, it leads to low effective spectrum usage. When a gNB serves one device at a time, each device gets a lot of instantaneous capacity which it may fail to utilize due to limited demand. But because the gNB cannot direct the remaining frequency resources to other directions, those resources are wasted, leading to low effective spectrum usage. Furthermore, other users in other directions could have used these wasted bands to improve overall spectrum usage.

[1]Artifacts link: https://wcsng.ucsd.edu/dpa

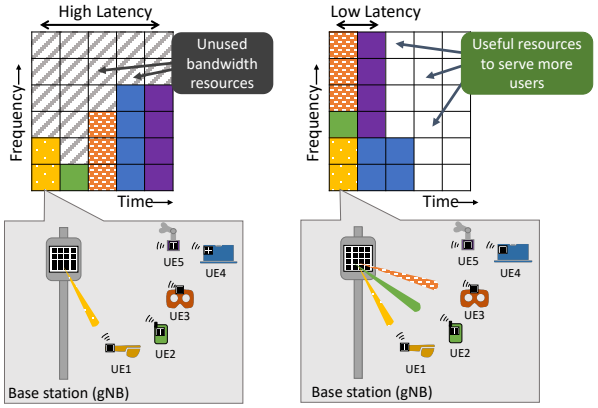


Fig. 1: mmFlexible enables efficient use of the available mmWave spectrum resources through flexible directional-frequency multiplexing, allowing multiple users to be served simultaneously with low latency and high spectrum utilization.

In this paper, we ask the question "*whether a mmWave base station can transmit or receive to any set of arbitrary directions using any set of contiguous frequency bands, creating a flexible frequency-direction beamforming response*". This is possible using massive antennas and digital beamforming, but traditional mmWave systems rely on analog phased arrays with a single RF-chain for cost and power efficiency. However, analog phased arrays cannot create such frequency-direction response, as they take a single input from the RF-chain and radiate all frequencies in the signal in one fixed direction. One naive solution is to split the phased array into multiple sub-arrays and program them to radiate in different directions, but this reduces the directivity in each direction, reducing signal strength, range, reliability, and even data rate.

Recently, new mmWave front-end architectures such as true-time delay (TTD) array [1]–[3] and leaky-wave antenna [4] have been proposed for frequency-dependent beamforming that spread different frequencies in different directions. However, these beam patterns have limitations that each user only receives a tiny fraction of the bandwidth in one timeslot, which may not meet their demand in a reasonable time frame. Additionally, these architectures do not provide control over number of beams, beam directions and ***beam-bandwidth***[2],

[2]**Beam-bandwidth** is defined as a fraction of system bandwidth that has high beamforming gain in the desired beam direction and low elsewhere.

resulting in large chunks of frequency resources being wasted in directions where there is no active user, leading to low spectrum utilization.

We propose a system called mmFlexible that performs *flexible directional-frequency multiplexing* by allocating a sub-set of contiguous frequency resources to each user, regardless of their direction, preventing spectrum wastage. This is achieved by creating multiple concurrent pencil beams (multi-beams) in different user directions, where each beam carries a separate frequency band according to the user's demand in that beam direction. A key feature of mmFlexible's multi-beam response is that it preserves beamforming gain across all beams while re-distributing power to only desired frequency-direction pairs with minimal leakage in other directions and frequency bands. This set of frequency-direction pairs can be chosen arbitrarily, providing flexibility in performing directional-frequency multiplexing. This enables efficient use of the entire frequency band (up to 800 MHz for 5G NR and 2.3 GHz for IEEE 802.11ax bands), reducing spectrum wastage and providing low latency network access, as shown in Fig. 1(b).

To implement mmFlexible, we introduce a novel mmWave analog array architecture called delay phased array (DPA) that can generate multi-beams with a flexible number of beams, with arbitrary beam directions and beam bandwidths (Refer to Figure 2). Unlike fixed delay architectures, DPA uses variable delay and phase elements at each antenna to create any desired frequency-direction response. Our insight is to use delays and phases in a complementary manner, with variable delay providing *frequency selectivity* and variable phase providing *direction steerability*, allowing DPA to create multi-beams towards multiple arbitrarily chosen frequency-direction pairs.

We architect the design of DPA and provide insights on the hardware requirements for creating the antenna array, and perform an analysis on the range and resolution of variable delay values required to generate our desired multi-beams. One of the challenges in implementing large delays on circuits is the size and complexity of the transmission lines, and integrating them onto an IC becomes even more difficult at mmWave frequencies due to bandwidth and matching constraints [5]. Our design addresses this by significantly reducing the range of delay values required, compared to traditional TTD array designs, making it practical and easy to manufacture. For example, TTD array requires monotonically increasing delays at each antenna, requiring a delay of 20 ns even for a 16-element array, which is 13x more than state-of-the-art mmWave delay designs [5]. In contrast, our DPA design consists of increasing and decreasing delay values for consecutive antennas, resulting in lower delay range requirements than traditional TTD arrays. Furthermore, the delay range is independent of the number of antenna elements, making it scalable for large arrays.

The next challenge is the software programming of DPA to meet the beamforming requirements of mmFlexible. The software should determine the appropriate values for the variable delays and phases at each antenna of DPA. Solving for these discrete values is computationally difficult as it is a non-convex and NP-hard problem. A naive solution is to pre-compute and store the delay and phase values for every possible frequency-direction pairs, but this is infeasible due to a large number of such combinations (10^{28}). To overcome this, we develop a novel FSDA (frequency-space to delay-antenna) algorithm which provides a single-shot solution for estimating the delays and phases in real time. It does this by mapping the desired frequency-space response to the delay-antenna space using a 2D transform and then extracting the corresponding delays and phases for each antenna. FSDA algorithm can be implemented in real-time using fast and efficient 2D FFT techniques. Our algorithm only requires the angles for each user and corresponding frequency resources allocated to users in the current time slot. The angles can be obtained from any standard compliant initial access protocol and so mmFlexible can be easily integrated into the standard 5G mmWave protocols.

In summary, ***we make the following contributions***:

- We propose mmFlexible, the first system that enables flexible directional-frequency multiplexing in mmWave networks, achieving higher spectrum usage, low latency, and scalability to support a large number of users.
- We design a novel mmWave front-end architecture called delay phased array that can generate multi-beams with a flexible number of beams, beam directions, and beam-bandwidths while maintaining high beamforming gains.
- We provide a new algorithm called FSDA (Frequency-space to delay-antenna) which estimates delays and phases in real-time using 2D FFT techniques and can be easily integrated with standard 5G mmWave protocols.
- We evaluate the performance of mmFlexible using real mmWave traces and show an improvement in latency by 60-150% compared to baselines. Furthermore, the multi-user sum-throughput is improved by 3.9x compared to true-time-delay array baseline [1].

II. Background and Motivation

In this section, we would provide a primer on different analog antenna array architectures, which have a single baseband radio frequency (RF) chain and mechanism to best achieve flexibility in resource allocation to multiple users. Next, we would show a realistic example (Figure 2) that none of the architecture can meet the requirements for flexible directional-frequency multiplexing, and how our flexible DPA architecture can meet these requirements.

A. Primer on phased arrays and true-time delay arrays

■ **Phased array:** The phased array takes a single input from the digital chain, split into N copies, apply appropriate phase-shift and radiate from N antennas (Fig. 3(a)). The input signal with all of its constituent frequency bands is radiated in a direction specified by the phase setting. The set of phases at each antenna constitute a weight vector w_{phase} as:

$$w_{\text{phase}}(n) = e^{j\Phi_n} \tag{1}$$

where Φ_n is the programmable phases for antenna index $n (n \in [0, N-1])$. Now, if we program the phases to create a

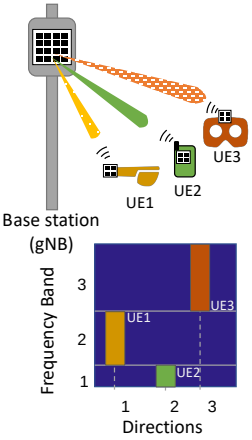


Fig. 2: Desired frequency-space beam response for 3 users in 3 different angular directions.

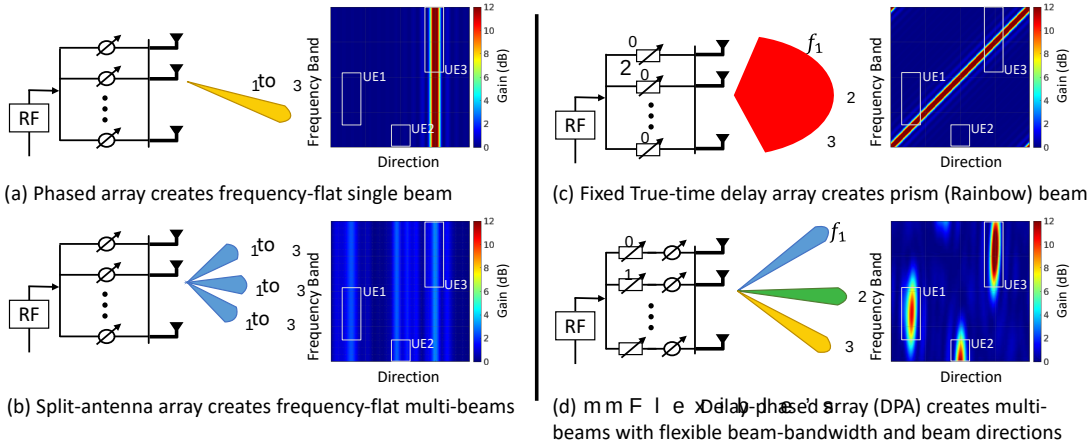


Fig. 3: Comparison of different mmWave front-end architectures and corresponding frequency-space patterns they created. The traditional phased array and split-antenna array create a frequency-flat response, while TTD array creates a frequency-selective rainbow-like pattern radiating in all directions. In contrast, DPA provides flexible frequency-direction response with programmable number of beams, beam-bandwidth and beam directions.

directional beam at an angle say 30°, then this phase shift is applied to all the frequency components in the signal radiating them along that same angle 30°. Different frequency components in the signal cannot be radiated to other directions because of the frequency-independent nature of weights in a phased array. Therefore, a phased array can serve users in only one direction at a time and cannot perform flexible directional-frequency multiplexing with many users. Split-antenna phased array (Fig. 3(b)) uses the phased array architecture split into multiple sub-arrays to create multiple beams towards different users in one TTI, but suffers from lower antenna gain, range, and throughput.

■ **True-time delay array:** The true-time delay (TTD) array uses a delay element to replace the phase shift element as shown in Fig. 3(c). The antenna weights with delay element is given by:

$$w_{\text{delay}}(t,n) = \delta(t - \tau_n) \tag{2}$$

Past work on TTD array [1]–[3], [6] have shown that this architecture creates a prism-type response by radiating all the frequencies in all the directions in a linear fashion (Fig. 3(c)) by using a fixed set of delays at each antenna given by $\tau_n = \frac{n}{B}$, for bandwidth B. This response has two major limitations: 1) most frequency bands would be wasted in space if there are no user in that direction and 2) each user gets a small frequency band and cannot scale it up arbitrarily. Due to these limitations, TTD array is not suitable for flexible directional-frequency multiplexing.

B. Current mmWave architectures are incapable for flexible directional-frequency multiplexing

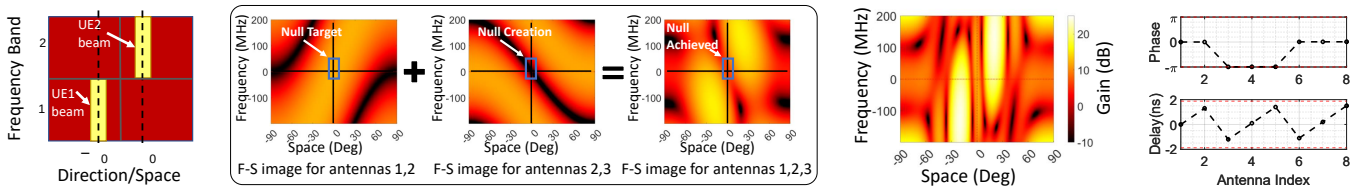
Let us take a simple network scenario to understand the performance tradeoffs for the above systems. Typically for 4K VR applications, end-end latency should be < 100 ms, Where transmission goes from the public cloud, network provider, base station (edge), and device. Over-the-air (base station to device) transmission latency comes down to a strict latency requirement of < 1 ms [7]. Consider 10 users in the network with similar traffic demand: each user requires 60 Mbps (4K VR) throughput and < 1 ms over-the-air latency due to the interactive nature of VR [7]–[10]. The aggregate throughput provided by the users is 600 Mbps, only 30% of the max physical layer throughput of a 400 MHz FR2 gNB in downlink (2.2 Gbps). We assume all users are distinctly located (in different angular directions) and RF conditions are good (e.g. LOS and no blockage). This allows us to control common external factors and exclusively evaluate architectural capabilities. We assume our DPA and split array can create up to 8 concurrent beams and TTD array creates a prism beam pattern. The comparison is shown in Table I. From our end-end evaluations (Fig. 7(a)), we can see that in edge scenarios all the baselines failed to meet the 1 ms over-the-air latency requirement. It is evident that mmFlexible is the best to support latency-critical applications while satisfying higher throughput demands.

Architecture	Latency* (ms)	Packet Loss	Throughput
phased array (TDMA)	> 1.25 ms	24.0%	54.9 Mbps
split-antenna array	> 2 ms	33.3%	47.3 Mbps
TTD array	> 3.5 ms	76.4%	18.3 Mbps
DPA	< 0.5 ms	0.0%	71.3 Mbps

TABLE I: Delay-phased array & other baselines for 10 4k VR users (*We mention over-the-air worst case latency).

III. MMFLEXIBLE'S DPA HARDWARE DESIGN

mmFlexible introduces a new mmWave front-end architecture, delay-phased array (DPA), to enable flexible directional-frequency multiplexing for mmWave networks. The DPA design addresses limitations of existing mmWave systems with phased arrays and TTD arrays by creating a multi-beam response with flexible beam directions and beam bandwidths.



(a) Desired F-S image (b) Specific Frequency-Space (F-S) signature images (c) 8 antenna F-S image (d) Delays and phases

Fig. 4: Explanation on how DPA creates a desired frequency-space image with two beams at -20° and 20°. The first beam occupies frequency band in [-200,0] MHz and the second beam occupies frequency band of (0,200] MHz. DPA can create the desired image in (c) using delay and phase values in (d).

In this section, we discuss practical implementation of DPA and show how we can build it with shorter delays.

A. Architecture of delay-phased array (DPA)

DPA architecture consists of programmable delay and programmable phase element per antenna with a single-RF chain as shown in Fig. 3(d). These elements can be programmed together to create flexible beam responses that are not possible by either of the two elements alone. Our insight is to control two knobs: delays τ_n and phase Φ_n to get the desired response. We define beam weights for DPA as:

$$w_{\text{dpa}}(t,n) = w_{\text{phase}}(n) w_{\text{delay}}(t,n) = e^{j\Phi_n}\delta(t-\tau_n) \tag{3}$$

Notice the dependence of DPA weights with time, which leads to a dependence on frequency. Upon taking FFT, the exponential term in the weights become $\Phi_n - 2\pi f\tau_n$, which is a function of frequency f at antenna index n. Therefore, the beamforming response of this weight vector will be a function of both frequency and direction. The beamforming gain of an antenna array represents the power radiated by the antenna array in different directions. The expression for beamforming gain $G(f,\theta)$ for DPA at a frequency f and direction θ is:

$$G(f,\theta) = \sum_{n=0}^{N-1} \mathcal{F}(w_{\text{dpa}}(t,n)) e^{-jn\pi\sin(\theta)} \tag{4}$$

where $\mathcal{F}$ is Fourier transform and $e^{-jn\pi\sin(\theta)}$ is the standard steering vector transformation from antenna to space $(\sin(\theta))$-domain[3]. Essentially, the response is the sum of the individual contribution from all antennas. By equating exponential terms of DPA weights to that of array response, we get: $\Phi_n - 2\pi f\tau_n = n\pi\sin(\theta)$. The variable-delay causes a slope in frequency (f) - space $(\sin\theta)$ plot, while the variable-phase causes a constant shift along the space axis. Together, they can create an arbitrary line with configurable slope and intercept in frequency-space domain. With this insight in place, we will discuss how to program the phase and delay values at each antenna to get the desired beam response in Section IV.

[3]Note that the steering vector for a linear antenna array is given by $e^{-jn2\pi\frac{d}{\lambda}\sin(\theta)}$, which depends on the array geometry (antenna spacing d) and signal wavelength λ [11]. We assume $d = \frac{\lambda}{2}$ and approximate the steering vector as $e^{-jn\pi\sin(\theta)}$.

B. Range of delay element in DPA

Before discussing DPA software programming, we emphasize that it is important to analyze the set of possible delay values that the hardware can support practically. Here we describe the requirements for the delay range and how it helps create a practical circuit board. Delay elements are implemented with variable-length transmission lines on a circuit board. Building large delay lines in IC at mmWave frequencies is prohibitive because of large size, bandwidth, and matching constraints [5]. Therefore, our design ensures that the delay range is not too large. While traditional true-time delay (TTD) array requires a delay range proportional to the number of antenna N, which is large for large antenna array (18.75 ns for 16 antenna array) [1]. In contrast, the delay range for mmFlexible is independent of the number of antennas. For the two-beam case, the delay range for mmFlexible is $\frac{3}{2B}$ (shown later in (5)), which is $3.7ns$ for 400 MHz bandwidth for 5G NR; significantly less than that required by TTD arrays. The delay range increases with the number of concurrent beams, but is independent of number of antennas, making it scalable to large arrays.

Delay control with sub-ns accuracy has been demonstrated in full-duplex circuits for interference cancellation [12]. Recently, authors of [13] have shown accurate delay control with 0.1 ns resolution and with 6-bit control (64 values until 6.4 ns), which satisfies the requirement of mmFlexible.

IV. mmFlexible's DPA Software Design

A. Requirements for mmFlexible

Our goal is to construct arbitrary frequency-direction response $G(f,\theta)$ via DPA architecture, which would be energy efficient and enable efficient resource utilization with low latency. If we carefully notice the example in section 2, we observe that the split antenna achieves the frequency-direction mapping but with a loss of 6 dB SNR, which is a corollary of the entire 400 MHz radiated in each of the four directions. It leads to our first requirement: **Req.1:** *The system must be able to transmit/receive signals in the specific frequency-direction pairs associated with each user, with minimal energy leakage in other directions and frequencies.* Furthermore, we should be able to control the amount of bandwidth assigned to each user, which leads to: **Req.2:** *Flexibility in allocating bandwidth to each user, allowing for narrow beams in space*

for higher antenna gain and wide beams in frequency to support high-demand users.

B. Meeting mmFlexible's requirements with DPA

With this intuition in place, we revisit and explain how can mmFlexible achieve these requirements through a simple two-user example in Fig. 4. Let us consider the two users are located at $-\theta_0$ and θ_0 respectively, and the base station wishes to serve these two users with equal beam-bandwidth of $B/2$ each, where B is the total system bandwidth. To support such flexible directional-frequency multiplexing, the base station must create a frequency-direction beam response shown in Fig. 4(a). We call such 2D beam patterns as *frequency-space (F-S) images* for simplicity. So how does DPA create these images and meet the above requirements?

We provide a closed-form expression for the set of delays τ_n and phases Φ_n for each antenna that would generate the above beamforming response as follow:

$$\tau_n = \left(\frac{3}{2B} n \sin(\theta_0) + \frac{3}{4B}\right) \mod \frac{3}{2B} \tag{5}$$

$$\Phi_n = \text{round}(n \sin(\theta_0))\pi \mod 2\pi \tag{6}$$

The derivation for these expression is omitted for brevity and can be found at https://wcsng.ucsd.edu/dpa.

Now, we achieve the requirements for mmFlexible by assigning a complementary F-S images to a subset of antennas. For instance, we create a positive slope in F-S image using antennas 1 and 2, and then create a complementary negative slope with antenna 2 and 3 as shown in Fig. 4(b). When the two responses are combined together, we observe a frequency-space image where they combine constructively at desired user locations while creating a null (low gain) at other locations (Meeting **Req-1**).

For the second requirement, we create such beams by choosing the number of antennas for creating positive or negative slope. The intuition is that higher number of antennas makes the corresponding signature image narrow in space. For instance, it is clear from Fig. 4(d) that three consecutive antennas (e.g. antenna 3,4,5) has increasing delays, while only two antenna (e.g. 2,3) has decreasing delays. This helps in making positive slope in F-S image narrow (3 antenna contribution), while the negative slope remains wide (2 antenna contribution). This effect causes beams that are narrow in space, but arbitrarily wide in frequency as shown in Fig. 4(c) (Meeting **Req-2**).

This intuition helps in understanding how a simple 2-beam frequency-space image is created. We use this insight to develop a novel FSDA algorithm that estimates the delay and phase values for any frequency-space image with arbitrary number of beams, beam directions and beam-bandwidths.

C. FSDA algorithm for estimating delays and phases in DPA

To create a desired frequency-direction beam response, the base station needs to estimate the corresponding delays and phases per-antenna in DPA. One naive solution is to try different discrete values in a brute-force way using a look-up table to get the desired beams. But, this solution is computationally hard and memory intensive since there is a large set of possibilities for the delays and phases at each antenna. For instance, with 64 delays and 64 phases per-antenna (assume both are 6-bit, so $2^6 = 64$), a brute-force look-up table search would require $64^N \times 64^N \approx 10^{28}$ probes to try each combination and then storing it all in memory, which is impossible to solve with even high memory and high computing machines. Our insight is that we can pose this problem as an optimization framework and solve them in computationally efficient way.

We now formulate the optimization problem with insights we have obtained from the previous subsection and from the fundamentals of digital signal processing. The goal is to relate weights (delays and phases) to the antenna gain pattern in (7) in a way that simplifies our estimation problem.

Our insight is that similar to how frequency and time are related by a Fourier transform, there is a similar transform that relates space and antenna using steering matrices. So, there are two transforms that bridges the world of antenna weights to desired gain pattern: time to frequency transform and antenna to space transform. Mathematically, we re-write gain pattern of DPA to emphasize on this 2D transform:

$$G(f, \theta) = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \mathbf{U}(f, k) w_{\text{dpa}}(k, n) V(n, \theta) \tag{7}$$

where $\mathbf{U}(f, k)$ is a discrete domain Fourier transform (DFT) and the steering matrix V is defined per-element as $V(n, \theta) = e^{-jn\pi \sin(\theta)}$. Now since the signal is actually sampled only discretely with a sampling time of T_s, our original delay weight element $w_{\text{delay}}(t, n)$ would reduce to $w_{\text{delay}}(k, n) = \delta(kT_s - \tau_n)$ as described in (7).

We then represent gain pattern by a discrete frequency-space matrix $\mathbf{G}$ and the weights as discrete time-antenna matrix $\mathbf{W}$ and relate them with the following 2D transform:

$$\mathbf{G} = \mathbf{UWV} \tag{8}$$

where $\mathbf{U}$ is time to frequency transform matrix and $\mathbf{V}$ is antenna to space transform matrix. Here we formulate $\mathbf{W}$ as $K \times N$ matrix, where K is the number of discrete time values and N number of antennas.

We follow a three step process to estimate the weight matrix $\mathbf{W}$ that create our desired frequency-space image intuitively explained in Fig. 5. There are two inputs to our algorithm: Angles and desired frequency bands for each user. These two inputs are enough to represent the given frequency-space image. As a first, we create a binary frequency-space matrix which consists of 1s at desired frequency-space locations and 0s otherwise, we denote it by $\mathbf{G}_{\text{desired}}$. We then formulate the following optimization problem:

$$\begin{aligned} \hat{\Phi_n}, \hat{\tau_n} &= \min ||\mathbf{G}_{\text{desired}} - \mathbf{UWV}||^2 \\ \text{s.t.} \quad \mathbf{W}(k, n) &= e^{j\Phi_n} \delta(kT - \tau_n) \end{aligned} \tag{9}$$

This optimization is a non-convex due to the non-linear terms such as exponential in phase and delta in delay. Moreover,

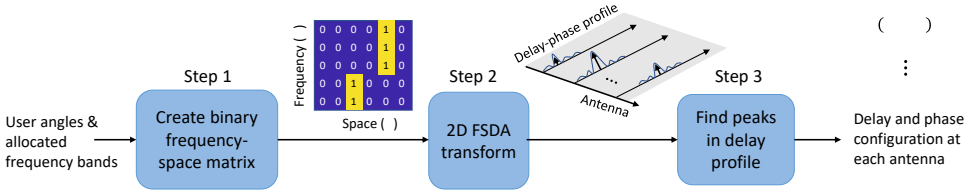


Fig. 5: Three steps of our FSDA algorithm to estimate per-antenna delays and phases that generate a desired frequency-space image. The core step is our novel 2D FSDA transform with some pre-processing and post-processing of inputs and outputs respectively.

the constraint on having a discrete set of values for delays and phases makes it NP hard. We make an approximation by relaxing the delta constraint and let the weights at each antenna take any variation over time. It means that we allow weights to take the form of a continuous profile over time at each antenna rather than a delta function which is non-zero at only one value and zero otherwise. We call it delay-phase profile at each antenna. How do we estimate this delay-phase profile?

Our insight is that we can write an inverse transform of $\mathbf{U}$ and $\mathbf{V}$ to go from frequency-space domain to time-antenna domain. The logic behind such formulation is that using appropriate discrete grid along time and space axis, we can formulate $\mathbf{U}$ and $\mathbf{V}$ as linear transforms, i.e., $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^\dagger = \mathbf{I}$ for identity matrix $\mathbf{I}$ (Note $(.)^\dagger$ is pseudo-inverse of a matrix). Therefore, it is easy to write their inverse by simply taking the psuedo-inverse. We estimate $\hat{\mathbf{W}}$ as:

$$\hat{\mathbf{W}} = \mathbf{U}^\dagger \mathbf{G}_{\text{desired}} \mathbf{V}^\dagger \tag{10}$$

The final step of our algorithm is to extract delays and phases from $\hat{\mathbf{W}}$. Note that each column in $\hat{\mathbf{W}}$ contains the delay-phase profile. We find the maximum peak in this profile and the index corresponding to this peak gives the delay and the max value at this peak gives the phase term. Note that since we did not put any restriction on the number of non-zero delay taps, we could get more than one delay-taps per-antenna. We empirically found that the estimated delay profile has only one significant peak with high magnitude than other local peaks (See Section V). Also, the intuition comes from our insights from previous section that usually one delay per-antenna suffice in creating the desired response.

■ **Weights Quantization:** The delay and phase values obtained from FSDA algorithm is still continuous in nature and must be discretized to be fed into the DPA hardware. We quantize both the phase and delay values with a 6-bit quantizer in software before feeding to the array. The quantized phase takes one of the 64 values in $[0°, 360°)$ and quantized delay varies in the range [0, 6.4ns) with an increment of 0.1 ns.

■ **Computation complexity of FSDA:** The run-time complexity of FSDA is dominated by the 2D FFT transform on a given frequency-space image. Given the frequency axis is divided into M subcarriers and the space axis into D directions, the run-time complexity is $\mathcal{O}(MD(\log(M) + \log(D))$.

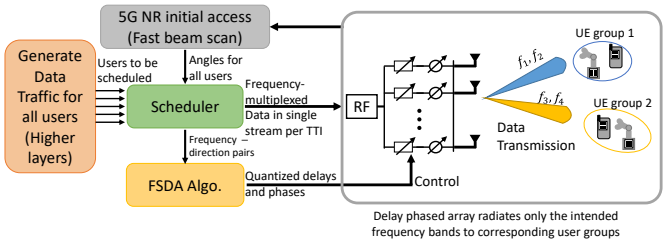


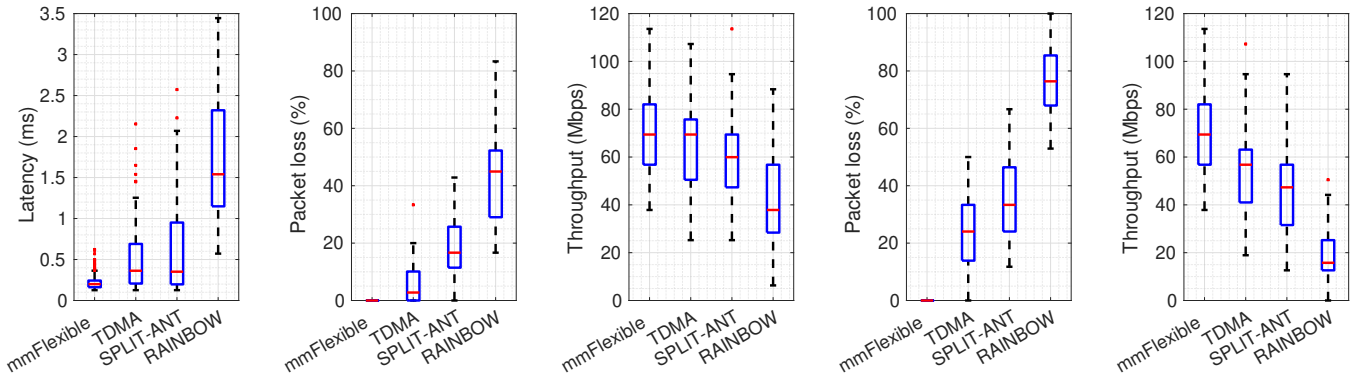
Fig. 6: Implementation overview for mmFlexible with four main components: Data traffic generator, Fast beam scan angle estimation, a scheduler and our FSDA algorithm.

V. Evaluation

A. Implementation and emulation with 28 GHz dataset

We implement an end-end system of mmFlexible with four major components, as illustrated in Fig. 6. The components of the systems are as follows:

1) ***Data traffic generator:*** We generate MAC layer packets with the throughput and latency constraints mentioned in the example in Sec. II based on [7], [9], [10]. We test with two different latency constraints: 1 ms and 0.5 ms. We use the same traffic generator for our system and all baselines.
2) ***Users angle estimation (initial access):*** We use the *channel collected from mobile 28 GHz testbed [14]*, using switched beamforming techniques [15], [16] for user's angle estimation. We leverage the existing 5G NR SSB Beam scan [17] using an exhaustive search to estimate angles. The gNB scans 64 beams in the codebook, and each UE reports the best beam index that maximizes the received signal strength from which the gNB determines the user's angle.
3) ***Data Scheduling:*** We implement a Proportional Fair (PF) scheduler [18] to allocate spectrum resources to users on a per-TTI basis. The available 120° field of view is mapped into 10 groups with each 12° half-power beam width. The scheduler uses user grouping, demand generation information, and SNRs (mapped to CQIs) to determine which user group to support and how many subcarriers to allocate for each user. Throughput is then calculated as a function of the allocated resources and channel to each user.
4) ***FSDA Algorithm:*** Our system's front-end uses DPA, requiring delays and phases as input. The FSDA algorithm provides quantized delays and phases, which are applied to the DPA to generate beams in desired directions and frequency bands. Array gain from the FSDA algorithm or



(a) Latency (b) Loss for <1ms latency (c) Thput for <1ms latency (d) Loss for <0.5ms latency (e) Thput for <0.5ms latency

Fig. 7: End to End performance gain with DPA compared to baselines under different scenarios: (a) mmFlexible is able to meet strict latency constraints ($<$ 1 ms), but the baselines are not; (b) to (d) Percentage of packet loss and Throughput performance for latency constraint $<$ 1 ms and $<$ 0.5 ms respectively.

respective baselines is then fed back to the scheduler for SNR computation.

We evaluate the performance of the baselines and mmFlexible with a PF scheduler by emulating 2000 TTIs of $125\mu s$ each, for a total duration of 0.25 seconds, across all users. The system undergoes the above four-step procedure for each TTI.

■ **Baselines:** Our paper compares the performance of mmFlexible with three baselines: TDMA, Split antennas [19], and Rainbow-link [1]. The TDMA approach has the scheduler assign one direction per TTI and beam in that direction over all frequencies. The Split antennas baseline has the scheduler assign one or multiple directions based on SNR & demand requirements support the given directions in all subcarriers. The Rainbow-link [1] transmits in all directions regardless of the number of user directions, using only subcarriers in user directions and wasting the rest.

B. End-to-end Results

1) **Latency**: Latency is the time for a packet to travel from source (gNB) to destination (UE) over the air. We evaluate the latency distribution across the baselines and present the results in Fig 7(a). We see that mmFlexible has a median latency of 0.2 ms, while TDMA and Split-antenna baselines have a higher median latency of 0.32 ms and 0.26 ms respectively. Our implementation features equal offered throughput for every user direction, which is the best case scenario for Rainbow-link operation; despite this, the median latency for the Rainbow-link is 1.5 ms because of its inability to assign bandwidth to a user that is proportional to its demand. The worst-case latency of mmFlexible is well below 1 ms. Notably, all baselines have weighted right tail distributions of latency, making their worst case much worse than 1 ms. The inability of the baseline methods to honor the latency constraint leads to dropped packets and ultimately lower link throughput and reliability.

2) **Packet loss**: If a packet's latency constraint is not met, then the packet is considered undelivered and lost. In Fig 7(b) and Fig 7(d), we see that mmFlexible is able to function without any packet loss. However, due to their inability to honor the latency constraints, TDMA, Split-antenna, and Rainbow-link baselines result in a median packet loss of 24.0%, 33.3%, and 76.4% respectively. mmFlexible's ability to serve multiple users in different directions enables it to optimally allocate resources without any power degradation and meet both throughput and latency constraints. TDMA is forced to serve one use direction at a time, resulting in a violation of latency constraints. Split-antenna baseline attempts to serve users in multiple directions simultaneously, but suffers from reduced throughput due to SNR degradation, resulting in high latency and packet loss compared to mmFlexible and TDMA. The Rainbow-link baseline allocates too few resources to each user and is the slowest and most unreliable in delivering packets.

3) **Per-user Throughput**: As shown in Fig.7(c) and Fig.7(e), mmFlexible outperforms all three baselines TDMA, Split antennas, and Rainbow-link. TDMA can only support one user group direction out of all the presented user demand directions, reducing its efficiency. However, in scenarios with heavy throughput demand, it performs similarly to mmFlexible. The mean throughput of mmFlexible is 1.3$\times$ that of TDMA throughput in the case of 0.5 ms latency requirement. The Split-antenna approach serves users in different directions by splitting its antennas, resulting in reduced overall throughput. mmFlexible provides 1.5$\times$ more throughput than the split scenario. Rainbow-link performs better only in scenarios with low throughput demand and users in all directions, but performance degrades in all other cases. mmFlexible provides 3.9$\times$ more throughput than the Rainbow-link baseline.

C. Benchmarks

We benchmark various theoretical and systems aspects of mmFlexible and compare the results with a split-antenna baseline. We chose the split-antenna baseline as it is closest to mmFlexible in creating multiple simultaneous beams in different directions. We present our results by calculating SNR for various scenarios. We use a channel dataset (28 GHz testbed [14]) and compute SNR per user for all subcarriers

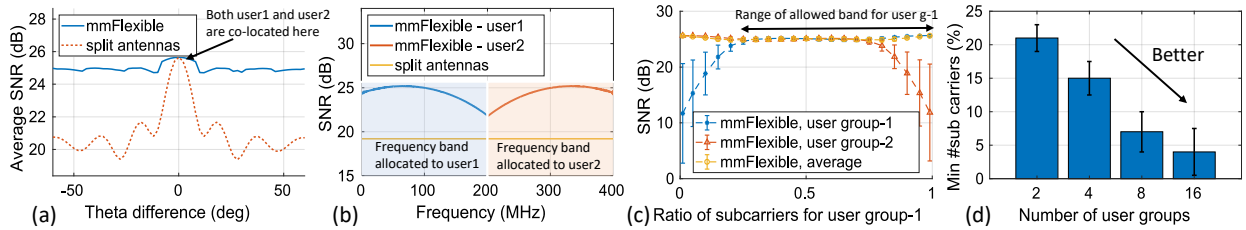


Fig. 8: The benchmark performance gain of mmFlexible compared to baselines under different scenarios: (a) Angle separation between two users (b) (c) Subcarrier allocation between users (d) How much minimum frequency resources can be allocated without SNR degradation.

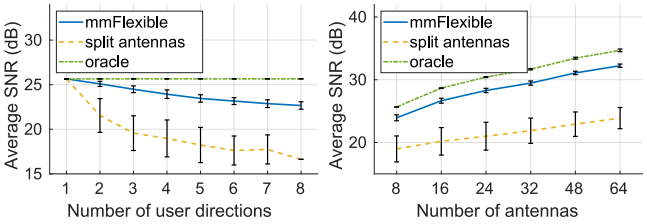


Fig. 9: Scaling the mmFlexible's performance with (a) Increasing number of user directions and (2) Increasing number of antennas.

by evaluating over LOS channel model [20] with no blockage and equal distance consideration for all users.

1) **Effect of angular separation between users on SNR:** We evaluate the performance of mmFlexible and a split-antenna approach in serving two users in different locations (θ_1, θ_2) with a single RF chain and an equal number of subcarriers. Figure 8(a) shows the mean SNR of the two users for different angular separations $(\theta_1 - \theta_2)$. We observe that mmFlexible performs optimally in all cases compared to the baseline. When the angular separation is close to zero, both users are in the same direction and can be served by a single beam. In these scenarios, both approaches converge to a single beam and perform similarly. The baseline becomes inefficient as the angular separation increases, while mmFlexible provides a stable response after 12^o separation with a 3-5 dB higher gain than the baseline approach.

2) **Does mmFlexible have interference due to multiple user transmissions from different directions?** No, mmFlexible is designed to support frequency multiplexing of users in different directions without interference. We evaluated this by comparing mmFlexible to a split-antenna baseline for two users separated by 30^o angle and equal bandwidth, as shown in Fig. 8(b). Both approaches receive data from user-1 in the 0-200 MHz bandwidth and from user-2 in the remaining 200-400 MHz bandwidth, without sharing any subcarriers between users. This results in no interference from simultaneous multiple-user transmissions for both approaches. Additionally, mmFlexible provides a higher gain to users by radiating only in allocated subcarriers as illustrated in Fig. 3(d), resulting in higher SNR than the baseline. The average SNR over all subcarriers shows that mmFlexible has a gain of 5dB compared to the split-antenna approach, even a 3dB higher gain at edge subcarriers.

3) **Impact of subcarrier allocation on SNR**: mmFlexible can reliably support both low-bandwidth *Ultra-Reliable Low-Latency Communication (URLLC)* and high-bandwidth *AR/VR* devices in different directions simultaneously. Here we evaluate the effect of subcarrier allocation on SNR for two user groups. Fig.8(c) shows the SNR achieved with subcarrier allocation to user group-1 from 1% to 99% (remaining subcarriers are allocated to user group-2). The average SNR over all users remains at 25 dB and does not depend on the proportion of subcarriers allocated to individual user groups. It is observed that the SNR of user group-1 converges to the average SNR when it is allocated more than 20% of the total subcarriers. Therefore, to achieve the benefits of mmFlexible in a URLLC link, at least 20% of the total subcarriers should be allocated to a user group. Additionally, as presented in Fig.8(d), as the number of user groups or user directions increases, the convergence point occurs at a lower percentage of subcarriers allocated. For instance, when there are 4 users in the system, then each user can be allocated with a minimum 15% of subcarriers without degrading the SNR performance. This convergence point drops down to 4% for 16 users which is favorable for URLLC applications which requires low-bandwidth and low latency.

Theoretical Baseline (Oracle): Oracle is a theoretical entity that can perfectly transmit only in desired subcarriers without any degradation in power at edge subcarriers (as shown in Fig. 2 desired frequency-space response).

4) **Impact of the number of user directions**: The performance of the mmFlexible improves as the number of supported user directions increases. We tested the system using a base station antenna array with 8 transmit/receive antennas, varying the number of user directions from 1 to 8. Fig. 9(a) illustrates that the relative gain (difference between the average SNR of mmFlexible and split antennas) increases with an increase in user directions. As the number of user directions increases, the split antenna approach divides the antennas per beam, resulting in reduced gain. Conversely, mmFlexible transmits power only in desired frequency bands and angular directions resulting in a higher gain. The Oracle creates a digital frequency filter at each antenna, resulting in a perfect frequency-space slicing which ensures that the average SNR remains constant regardless of whether it serves in one direction or multiple directions simultaneously. In contrast, mmFlexible has one delay per antenna (for hardware feasibility), which makes it difficult to create an ideal frequency-space slicing, leading to

power degradation at the edge subcarriers. Despite this, even after eight splits, the degradation is less than 2.5 dB with the Oracle, and the gain is more than 6 dB higher compared to the split antenna baseline. Error bar in Fig. 9(a) indicates average SNR variations with users in different angle separations.

5) **Impact of the number of antennas**: We show that mmFlexible performs better with the increase in the number of antennas. We evaluated this hypothesis by serving four users and varying antennas from 8 to 64. Fig. 9(b) shows gain variations from 8 antennas to 64 antennas for 4 users; it is clearly evident that mmFlexible outperforms with the increase in antennas over the split antennas baseline. The error bar in the figure shows the variations in the average SNR when serving four users at different angle separations. Similar to the Oracle, mmFlexible's performance remains constant even as the number of antennas increases because the number of frequency splits is determined solely by the number of user directions, which are the same in all cases.

VI. Related Work

mmFlexible builds upon previous work in mmWave and THz communications, but sets itself apart by introducing a system that can perform flexible directional-frequency multiplexing, while maintaining energy efficiency and high performance in terms of range, throughput, and link reliability. To the best of our knowledge, no existing literature has achieved this level of frequency slicing without compromising on performance. mmFlexible's unique approach enables efficient use of the entire frequency band, reducing spectrum wastage and providing low latency network access.

■ **Split antenna phased array:** Traditional phased array beamforming does not support flexible directional-frequency multiplexing because of a single narrow pencil beam. In the past, split-antenna arrays have been used to create concurrent multi-beams across multiple directions [21]–[25]. However, this approach often results in lower beamforming gain and throughput for each user. The array gain reduces proportionally to the number of beams, as the total available power is distributed along multiple directions and across the entire bandwidth. In contrast, mmFlexible uses a unique split-beam mechanism with frequency selectivity that radiates only in the desired frequency band, preserving high directivity, signal strength, and throughput.

■ **True-time delay array architecture:** Previous work on True-time delay arrays (TTD) has primarily focused on beam steering for ultra-wideband signals [26], [27] and more recently, single-shot beam training [2], [3], [28], compressive channel estimation [29], wideband tracking [30] and THz communication [31]. However, none of these works address the problem of flexible low-latency multi-user communication. Rainbow-link [1] uses TTD arrays for multi-user communication, but it is limited to fixed low-throughput IoT applications (limited to 7.8 MHz per user [1]) and cannot flexibly allocate a large number of subcarriers to a single direction for broadband users. mmFlexible addresses this limitation by using variable delay elements to create arbitrary frequency slicing and the ability to radiate those frequencies in any desired direction. mmFlexible's beamforming is orthogonal to previous work in this area, but it can leverage the fast beam training capability of TTD arrays.

■ **Other front-end mmWave architectures:** In [32], a new mmWave receiver architecture for frequency multiplexing was proposed, which utilizes a network of mixers at each antenna to receive different frequency components from different directions. However, this approach has a fixed hardware structure that is not scalable to support a large number of users, requiring different hardware for different numbers of users. In contrast, mmFlexible's delay-phased array (DPA) is flexible and programmable, providing a scalable solution for supporting a large number of users.

VII. Discussion and Future Work

We discuss potential future work ideas:

■ **Circuit for DPA:** Implementing circuit delays in mmWave frequencies is challenging due to non-linearity, bandwidth and matching constraints [33], [34]. In contrast, delay elements can be more accurately implemented at intermediate frequencies (IF) (sub-6 GHz) using techniques such as voltage-time converters [13] and switched-capacitor arrays [5]. Recent work [5] has shown that efficient mixers, phase-shifters, and IF true-time-delays can be used to make a DPA that meets the requirements of mmFlexible.

■ **Single RF vs Multi-RF systems:** mmFlexible works with a single-RF chain and radiating each frequency component in different directions, while past work on the single-RF system stream all frequencies in one direction. Multi-RF systems (Hybrid arrays) [35], [36] offer freedom to create multi-stream to multi-directions, but each stream carries the entire bandwidth. Thus they also cannot perform flexible directional-frequency multiplexing. Our work can be extended to the multi-RF system where each RF is connected to a DPA that creates large sectors where each DPA can serve in a sector with our flexible directional-frequency multiplexing technology.

■ **Applications beyond flexible frequency multiplexing:** Delay-phased arrays have the potential to enable a plethora of applications in communication and sensing beyond flexible directional-frequency multiplexing. For instance, the ability to create arbitrary and controllable frequency-space beams can help faster localization and tracking of multiple targets. Delay-phased arrays also enable simultaneous communication and sensing paradigms where some frequency bands are used for communication while other bands can be used for sensing. All these applications can be enabled with a simple software or firmware updates on the same underlying hardware. We leave these applications for our future work.

VIII. Acknowledgements

We are grateful to the anonymous reviewers their valuable feedback, as well as to the WCSNG group at UC San Diego for their input. The research was supported by NSF #2211805.

REFERENCES

[1] R. Li, H. Yan, and D. Cabric, "Rainbow-link: Beam-alignment-free and grant-free mmw multiple access using true-time-delay array," *IEEE Journal on Selected Areas in Communications*, 2022.

[2] H. Yan, V. Boljanovic, and D. Cabric, "Wideband millimeter-wave beam training with true-time-delay array architecture," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1447–1452.

[3] V. Boljanovic, H. Yan, C.-C. Lin, S. Mohapatra, D. Heo, S. Gupta, and D. Cabric, "Fast beam training with true-time-delay arrays in wideband millimeter-wave systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1727–1739, 2021.

[4] Y. Ghasempour, R. Shrestha, A. Charous, E. Knightly, and D. M. Mittleman, "Single-shot link discovery for terahertz wireless networks," *Nature communications*, vol. 11, no. 1, pp. 1–6, 2020.

[5] E. Ghaderi, A. S. Ramani, A. A. Rahimi, D. Heo, S. Shekhar, and S. Gupta, "An integrated discrete-time delay-compensating technique for large-array beamformers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 9, pp. 3296–3306, 2019.

[6] V. Boljanovic, H. Yan, C.-C. Lin, S. Mohapatra, D. Heo, S. Gupta, and D. Cabric, "True-time-delay arrays for fast beam training in wideband millimeter-wave systems," *arXiv preprint arXiv:2007.08713*, 2020.

[7] A. R. Qualcomm, "Augmented and virtual reality: the first wave of 5g killer apps," https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/cr-qcom-140.pdf, 2017.

[8] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Network*, vol. 32, no. 2, pp. 8–15, 2018.

[9] "Speech and multimedia Transmission Quality (STQ); QoS parameters and test scenarios for assessing network capabilities in 5G performance measurements," 3rd Generation Partnership Project (3GPP), TR ETSI TR 103 702 V1.1.1 , 2020-11. [Online]. Available: http://www.etsi.org/deliver/etsi_tr/138900_138999/138901/14.00.00_60/tr_138901v140000p.pdf

[10] S. Mangiante, G. Klas, A. Navon, Z. GuanHua, J. Ran, and M. D. Silva, "Vr is on the edge: How to deliver 360 videos in mobile networks," in *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, 2017, pp. 30–35.

[11] J. Benesty, I. Cohen, and J. Chen, *Array Processing*. Springer, 2019.

[12] A. Nagulu, A. Gaonkar, S. Ahasan, S. Garikapati, T. Chen, G. Zussman, and H. Krishnaswamy, "A full-duplex receiver with true-time-delay cancelers based on switched-capacitor-networks operating beyond the delay–bandwidth limit," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 5, pp. 1398–1411, 2021.

[13] E. Ghaderi and S. Gupta, "A four-element 500-mhz 40-mw 6-bit adc-enabled time-domain spatial signal processor," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 6, pp. 1784–1794, 2020.

[14] I. K. Jain, R. Subbaraman, T. H. Sadarahalli, X. Shao, H.-W. Lin, and D. Bharadia, "mmobile: Building a mmwave testbed to evaluate and address mobility effects," in *Proceedings of the 4th ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, 2020, pp. 1–6.

[15] D. Caudill, J. Chuang, S. Y. Jun, C. Gentile, and N. Golmie, "Real-time mmwave channel sounding through switched beamforming with 3-d dual-polarized phased-array antennas," *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 11, pp. 5021–5032, 2021.

[16] J. Palacios, "Adaptive Codebook Optimization for Beam Training on Off-the-Shelf IEEE 802.11ad Devices," in *Proceedings of the 24rd Annual International Conference on Mobile Computing and Networking*. ACM, 2018.

[17] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2018.

[18] alvarovalcarce, "Nokia wireless suite - github reference," https://github.com/nokia/wireless-suite (2020/11/12).

[19] I. K. Jain, R. Kumar, and S. S. Panwar, "The impact of mobile blockers on millimeter wave cellular systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 854–868, 2019.

[20] "5G; Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), TR 138 901 V14.0.0, 2017-05. [Online]. Available: http://www.etsi.org/deliver/etsi_tr/138900_138999/138901/14.00.00_60/tr_138901v140000p.pdf

[21] I. K. Jain, R. Subbaraman, and D. Bharadia, "Two beams are better than one: towards reliable and high throughput mmwave links," in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, 2021, pp. 488–502.

[22] I. Aykin, B. Akgun, and M. Krunz, "Multi-beam transmissions for blockage resilience and reliability in millimeter-wave systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 12, pp. 2772–2785, 2019.

[23] J. A. Zhang, X. Huang, Y. J. Guo, J. Yuan, and R. W. Heath, "Multibeam for joint communication and radar sensing using steerable analog antenna arrays," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 671–685, 2018.

[24] H. Hassanieh, O. Abari, M. Rodriguez, M. Abdelghany, D. Katabi, and P. Indyk, "Fast millimeter wave beam alignment," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. ACM, 2018, pp. 432–445.

[25] D. Zhu, J. Choi, Q. Cheng, W. Xiao, and R. W. Heath, "High-resolution angle tracking for mobile wideband millimeter-wave systems with antenna array calibration," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7173–7189, 2018.

[26] R. Rotman, M. Tur, and L. Yaron, "True time delay in phased arrays," *Proceedings of the IEEE*, vol. 104, no. 3, pp. 504–518, 2016.

[27] S. K. Garakoui, E. A. Klumperink, B. Nauta, and F. E. van Vliet, "Compact cascadable gm-c all-pass true time delay cell with reduced delay variation over frequency," *IEEE journal of solid-state circuits*, vol. 50, no. 3, pp. 693–703, 2015.

[28] A. Wadaskar, V. Boljanovic, H. Yan, and D. Cabric, "3D Rainbow Beam Design for Fast Beam Training with True-Time-Delay Arrays in Wideband Millimeter-Wave Systems," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2021, pp. 85–92.

[29] V. Boljanovic and D. Cabric, "Compressive estimation of wideband mmw channel using analog true-time-delay array," in *2021 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, 2021, pp. 170–175.

[30] J. Tan and L. Dai, "Wideband beam tracking in thz massive mimo systems," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1693–1710, 2021.

[31] ——, "Delay-phase precoding for thz massive mimo with beam split," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[32] R. Garg, G. Sharma, A. Binaie, S. Jain, S. Ahasan, A. Dascurcu, H. Krishnaswamy, and A. S. Natarajan, "A 28-ghz beam-space mimo rx with spatial filtering and frequency-division multiplexing-based single-wire if interface," *IEEE Journal of Solid-State Circuits*, 2020.

[33] M.-K. Cho, I. Song, and J. D. Cressler, "A true time delay-based sige bi-directional t/r chipset for large-scale wideband timed array antennas," in *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*. IEEE, 2018, pp. 272–275.

[34] F. Hu and K. Mouthaan, "A 1–20 ghz 400 ps true-time delay with small delay error in 0.13 μm cmos for broadband phased array antennas," in *2015 IEEE MTT-S International Microwave Symposium*. IEEE, 2015, pp. 1–3.

[35] L.-H. Shen and K.-T. Feng, "Mobility-aware subband and beam resource allocation schemes for millimeter wave wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 893–11 908, 2020.

[36] W. Zhang, Y. Wei, S. Wu, W. Meng, and W. Xiang, "Joint beam and resource allocation in 5g mmwave small cell systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10 272–10 277, 2019.