Millimeter Wave Wireless Communication: Initial Acquisition, Data Communication and Relay Network Investigation

vorgelegt von
M.Eng.

Xiaoshen Song

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
-Dr.-Ing.-

genehmigte Dissertation

Promotionsausschuss:
Vorsitzender: Prof. Rafael Schaefer
Gutachter: Prof. Giuseppe Caire
Gutachter: Prof. Robert W. Heath Jr.
Gutachter: Prof. Joerg Widmer

Tag der wissenschaftlichen Aussprache: 18. September 2020

Berlin 2020

Abstract

Wireless communication has become an important part of our daily lives. In the past decades, the phenomenal increasing demand for mobile wireless data services has been pushing both industry and academia to move to millimeter wave (mmWave) frequencies (30-300 GHz) for the next generation (5G) mobile communication. The main motivation for mmWave communication is the unprecedented massive bandwidth (multi-GHz) which can offer multi-Gbps data rates for each mobile devices. However, mmWave signals experience high path loss, directivity and blockages, which severely limits the network performance.

To overcome the aforementioned challenges in mmWave communication, large antenna arrays are used on the transceivers aiming for a large beamforming gain to compensate the severe path loss. In addition, hybrid digital analog (HDA) architecture, with much smaller number of radio frequency (RF) chains in comparison with the number of antennas, are commonly implemented at the transceivers in order to reduce the hardware complexity and power consumption. All these new features rise a big challenge for signaling and networking for mmWave wireless systems.

The goal of this thesis is to clearly incorporate all the new features at mmWave frequencies, on top of which to provide new state of the art schemes regarding different communication phases. Specifically, this thesis contains four main contributions.

First, we propose an efficient beam alignment (BA) scheme for mmWave OFDM (orthogonal frequency division multiplexing) systems. In this scheme, we use pseudo-random multi-finger beam patterns in the downlink to explore the beam-domain channel, and then construct an estimate of the channel second-order statistics. By using non-negative least-squares (NNLS) technique, the resulting under-determined equations can be efficiently solved. Accordingly, the proposed BA scheme is very robust to channel time-dynamics and is strongly scalable to multi-user scenarios.

Second, we further explore single-carrier (SC) operation mode at mmWave frequencies and propose a new BA scheme for mmWave SC systems. In this scheme, the BS periodically probes the channel via a pre-specified pseudo-random beamforming codebook and pseudo-random spreading codes. Each UE formulates the BA problem as the estimation of a sparse non-negative second-order statistic channel vector, which can be efficiently solved by using NNLS technique. In addition to the advantage of multi-user

scalability, this proposed scheme is purely in time domain and is highly robust to fast channel variations caused by the large Doppler spread between multipath components.

Third, we define two HDA antenna architectures which can be regarded as two "extreme" cases, i.e., the fully-connected (FC) architecture and the one-stream-per-subarray (OSPS) architecture. We propose a joint performance evaluation of the initial BA, the consequent data communication as well as the hardware impairments, where we consider, from a realistic point of view, only a limited channel state information (CSI) that obtained from the BA phase. Also, a family of multi-user MIMO (MU-MIMO) precoding schemes are investigated to well adapt to the hybrid architectures and the beam information extracted from the BA phase. An interesting observation from this work is that the two aforementioned architectures achieve similar sum spectral efficiency, while the OSPS architecture is advantageous with respect to the FC case in terms of hardware complexity and power efficiency, only at the cost of a slightly longer BA time-to-acquisition due to its reduced beam angle resolution.

Fourth, we extend our work to relay networking to further increase the communication range at mmWave frequencies. For a general mmWave half-duplex (HD) relay network with arbitrary relay connections, we introduce the information theoretically optimal schedule to firstly do a topology simplification procedure, on top of which we propose two practical beam scheduling schemes, i.e., the deterministic edge coloring (EC) scheduler and the adaptive backpressure (BP) scheduler. The former is more suitable for static scenarios while the later is more favorable for time-varying scenarios. Both the proposed schedulers can effectively stabilize the network within its capacity range, meanwhile achieve much smaller queuing backlogs, much smaller backlog fluctuations, and much lower packet end-to-end delays in comparison with the reference baseline scheme.

Zusammenfassung

Drahtlose Kommunikation ist zu einem wichtigen Bestandteil unseres täglichen Lebens geworden. Die Nachfrage nach mobilen Datendiensten ist in den letzten Jahren massiv gestiegen. Dies hat sowohl die Industrie als auch die Wissenschaft dazu veranlasst, für die Mobilkommunikation der nächsten Generation (5G) auf Frequenzen im Bereich der Millimeterwellen (mmWave, 30-300GHz) umzusteigen. Die Hauptmotivation für die mmWave-Kommunikation ist die Verfügbarkeit einer enormen Bandbreite (mehrere GHz). Diese ermöglicht Datenraten von mehreren Gbps für mehrere Mobilfunkendgeräte. Bei mmWave-Signalen treten jedoch hohe Pfadverluste, sehr spezifische Richtcharakteristiken und Blockierungen auf, was die Netzwerkleistung stark einschränkt.

Um die oben genannten Herausforderungen für die mmWave-Kommunikation zu bewältigen, werden an den Transceivern große Antennenarrays verwendet. Diese ermöglichen einen großen Beamforminggewinn, um den hohen Pfadverlust zu kompensieren. Darüber hinaus wird an den Transceivern üblicherweise eine hybrid digital analog -Architektur (HDA-Architektur) mit einer im Vergleich zur Anzahl der Antennen viel geringeren Anzahl von Basisbandsignalpfaden eingesetzt. Dadurch kann die Hardwarekomplexität und der Stromverbrauch verringert werden. All diese neuen Funktionen stellen die Entwickler der physikalische Schicht von mmWave-Funksystemen und deren Netzwerken vor große Herausforderungen.

Ziel dieser Arbeit ist es, alle neuen Eigenschaften bei mmWave-Frequenzen klar einzubeziehen und darüber hinaus neue hochmoderne Algorithmen für verschiedene Phasen der Kommunikation bereitzustellen. Insbesondere enthält diese Arbeit vier Hauptbeiträge.

Zunächst stellen wir einen effizienten Algorithmus für die initiale Ausrichtung von Antennencharakteristiken (im Englischen beam alignment - BA) für mmWave orthogonal frequency division multiplexing-Systeme (OFDM-Systeme) vor. In diesem Algorithmus verwenden wir pseudozufällige Mehrfinger-Abstrahlcharakteristiken im Downlink, um den Mobilfunkkanal zu untersuchen, und schätzen dann die Statistik zweiter Ordnung des Kanals. Durch die Verwendung der Non-Negative Least Squares-Technik (NNLS-Technik) können die resultierenden unterbestimmten Gleichungen effizient gelöst werden. Der vorgeschlagene BA-Algorithmus ist sehr robust gegenüber der Zeitdynamik des Kanals und skaliert sehr gut für Mehrbenutzerszenarien.

Zweitens untersuchen wir den Einträger-Betriebsmodus (single carrier - SC) bei mmWave-Frequenzen. Wir schlagen einen neuen BA-Algorithmus für mmWave-SC-Systeme vor. In diesem Schema prüft die Basistation den Kanal periodisch über ein vorbestimmtes Codebuch aus pseudozufälligen Mehrfinger-Abstrahlcharakteristiken und Pseudozufalls-Spreizcodes. Jedes Endgerät formuliert das BA-Problem als Schätzung eines dünnbesetzten nicht negativen statistischen Kanalvektors zweiter Ordnung. Dieser kann unter Verwendung der NNLS-Technik effizient gelöst werden. Zusätzlich zu dem Vorteil der Mehrbenutzerskalierbarkeit arbeitet der vorgeschlagene Algorithmus nur im Zeitbereich und ist damit äußerst robust gegenüber schnellen Kanalschwankungen, die durch eine große Doppler-Verschiebung zwischen Pfadkomponenten des Funkkanals verursacht werden.

Drittens definieren wir zwei spezifische Strukturen der HDA-Architektur, die als zwei Extremfälle angesehen werden können. Diese sind die vollständig verbundene (fully connected - FC) Architektur und die Ein-Signalpfad-pro-Subarray-Architektur (one stream per subarray - OSPS). Wir betrachten zusammenhängend die Leistung des initialen BA, der daraus resultierenden Datenkommunikation sowie der Hardware-Beeinträchtigungen. Dabei nutzen wir realitätsnah nur eine begrenzte Zustandsinformation des Funkkanals (channel state information – CSI). Diese Information kann direkt aus der BA-Phase erhalten werden. Außerdem untersuchen wir verschiedene Multi-User-MIMO-Codierungen (MU-MIMO precoding), welche an die vorgeschlagenen HDA-Architekturen und die aus der BA-Phase extrahierten Kanalinformationen angepasst sind. Eine interessante Beobachtung aus dieser Arbeit ist, dass die beiden oben genannten Architekturen eine ähnliche spektrale Effizienz erzielen. Während die OSPS-Architektur im Vergleich zum FC-Fall in Bezug auf Hardwarekomplexität und Energieeffizienz vorteilhaft ist, zeigt die FC-Architektur eine etwas bessere Leistung im BA. Dies liegt an der geringeren Winkelauflösung der OSPS-Struktur.

Viertens erweitern wir unsere Arbeit auf Relay-Netzwerke, um die Reichweite bei mmWave-Frequenzen weiter zu vergrößern. Für ein allgemeines Relay-Netzwerk im mmWave-Bereich im Halbduplexbetrieb mit beliebigen Relayverbindungen führen wir das informationstheoretisch optimale Schema (im Englischen schedule) ein. Dadurch können wir zunächst ein Verfahren zur Vereinfachung der Topologie durchführen. Anschließend schlagen wir dann zwei praktische Schemata zur Strahlsteuerung vor. Zum einen das deterministic edge coloring (EC) Schema und zum anderen das adaptive backpressure (BP) Schema. Ersteres eignet sich besser für statische Szenarien, während letzteres für zeitlich variierende Szenarien günstiger ist. Beide vorgeschlagenen Schemata können das Netzwerk innerhalb seines Kapazitätsbereichs effektiv betreiben. Im Vergleich zum Referenzschema haben beide viel kleinere Warteschlangen, viel kleinere Schwankungen in der Warteschlangenauslastung und viel geringere Ende-zu-Ende-Verzögerungen.

To My Youth and My Beloved Husband.

— Xiaoshen Song

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my PhD supervisor Prof. Giuseppe Caire. Without him this research would have not been possible. He has always been there for me, providing unending support and motivation. He tolerates my shortcomings and helps me to overcome my weakness. Besides, he is passionate about new technologies and exciting ideas. He trained me in building efficient research skills which led me to finish my PhD research effectively and in time. I consider myself very lucky to have joined his group.

A very special thanks to Dr. Saeid Haghighatshoar. “Research life is tough, find a mentor!”. So yes, he is my best mentor. Without him I would never have got started in my PhD research. From academic writing to technical skill, he unconditionally passes on his valuable experience to me without any reservation. His critical advice, wide perspective and methodological precision have substantially shaped my scientific thinking.

I would also like to thank Mozhgan Bayat and Thomas Kühne. Adapting to a new country and a new culture is not easy. They have provided a listening ear, encouraged me and unconditionally helped me in private. They have made me feel that I am not alone in this country. I will always be grateful to them for their valuable friendship.

Many thanks to the doctoral committee experts Prof. Robert W. Heath Jr., Prof. Joerg Widmer and Prof. Rafael Schaefer for providing insightful comments to my research and thesis. I have learned a lot from their publications as well as from the long discussion with them over the defense.

In addition, I would like to express my gratitude and appreciation to my colleagues. They made my PhD journey funny and interesting. I will always hold dear the days and nights spent in the office.

I also thank my parents and my sisters for their inspiration and unequivocal support during my PhD journey. The journey from China to TU Berlin would not have been possible without their support.

Last but not least, I owe a thanks to my beloved husband Youjiang. I find it difficult to express my appreciation because it is so boundless. He has seen me through the ups and downs of the entire PhD journey. He is my most enthusiastic cheerleader; he is my best friend; and he is an amazing husband. Without his sunny optimism, I would be a much grumpier person; without his love and support, I would be lost. Fate brought us

together in the college. The past 9 years for our meeting and getting along with each other are recalled, being still happy and romantic. I have always been firm to spend my future with him together, pursuing our dreams and facing life's challenges hand in hand.

I would like to thank all my friends from different countries and cultures for long lasting friendships and enjoyable experiences during my stay in Berlin.

List of Publications

Below is a selection of publications that I authored / co-authored during my PhD candidate duration.

Journal Papers:

1. X. Song, S. Haghighatshoar, and G. Caire,“A scalable and statistically robust beam alignment technique for mm-wave systems,” IEEE Transactions on Wireless Communications, 2018. (page 20)
2. X. Song, S. Haghighatshoar, and G. Caire,“Efficient beam alignment for mmWave single-carrier systems with hybrid MIMO transceivers,” IEEE Transactions on Wireless Communications, 2019. (page 38)
3. X. Song, T. Kühne, and G. Caire, “Fully-/Partially-Connected Hybrid Beamforming Architectures for mmWave MU-MIMO,” IEEE Transactions on Wireless Communications, 2019. (page 58)
4. X. Song, Yahya H. Ezzeldin, Giuseppe Caire, Christina Fragouli, “Efficient Beam Scheduling for Half-Duplex mmWave Relay Networks,” IEEE Transactions on Wireless Communications, 2020. (to be submitted) (page 78)

Conference Papers:

1. X. Song, S. Haghighatshoar, and G. Caire, “A robust time-domain beam alignment scheme for multi-user wideband mmwave systems,” in WSA 2018; 22nd International ITG Workshop on Smart Antennas, 2018, pp. 1-7.
2. X. Song, S. Haghighatshoar, and G. Caire, “An Efficient CS-Based and Statistically Robust Beam Alignment Scheme for mmWave Systems,” in 2018 IEEE International Conference on Communications (ICC), 2018, pp. 1-6.
3. X. Song, T. Kühne, and G. Caire, “Fully-Connected vs. Sub-Connected Hybrid Precoding Architectures for mmWave MU-MIMO,” in ICC 2019-2019 IEEE International Conference on Communications (ICC), 2019, pp. 1-7.
4. X. Song, and G. Caire, “Queue-Aware Beam Scheduling for Half-Duplex mmWave Relay Networks,” in 2020 IEEE International Symposium on Information Theory (ISIT). (accepted)

5. T. Kühne, X. Song, G. Caire, K. Rasilainen, T. H. Le, M. Rossi, et al., "Performance Simulation of a 5G Hybrid Beamforming Millimeter-Wave System," in WSA 2020; 24th International ITG Workshop on Smart Antennas, 2020, pp. 1-6.

This thesis is an accumulation of publications. It is based on the above selected journal papers (three published papers after peer-reviewing and one to-be-submitted journal manuscript), which I wrote as first author. These four journal papers constitute the four main chapters (Chapter 3 - Chapter 6) in this thesis. At the beginning of the corresponding chapters, an introductory section with supplementary background information as well as the clarification of each authors' contributions are provided.

Table of Contents

1 Introduction

1.1 Background for mmWave communication

Wireless communication has become an integral part of our lives today. As data-hungry mobile devices and applications become increasingly prevalent, the mobile communication infrastructure needs to evolve dramatically to support the exploding demand for wireless data. Ericsson has predicted that the volume of mobile data traffic will increase five folds from 2018 to 2024, reaching 136 exabytes (EBs) per month (as illustrated in Figure. 1.1), equivalent to a compound annual growth rate of 31% [1]. This ever growing trend is expected to continue mainly due to the services that require massive data, such as high definition (HD) video streaming, online gaming, virtual reality applications and so on [2, 3, 4]. For example (e.g.), video streaming contributed 60% ($\sim$ 16 EB/month) of total mobile traffic in 2018, which is expected to reach 74% ($\sim$ 100 EB/month) by 2024 [1]. In addition, billions of new devices envisioned to be connected in the future generation of wireless networks to provide massive connectivity are also expected to contribute to the increase in data consumption [1].

The motivation for the evolution from 1G to 2G, 3G, and 4G (the first, second, third and fourth generation mobile communication, respectively) was to improve a particular aspect of mobile communication. For example, 1G to 2G transition improved voice service and increased network capacity by using digital communications; 3G and 4G were developed to improve the data rates. However, the evolution of the next generation mobile communication (5G) features concurrent improvements in many areas. Specifically, they include high data rates ($1 - 10$ Gbps), low latency (less than 1 milliseconds), massive connectivity ($\sim$ tens of billions new devices), and better quality of service [1, 5, 6].

The international telecommunication union (ITU) has classified the 5G usage scenarios (i.e., the international mobile telecommunications (IMT) for 2020 and beyond)

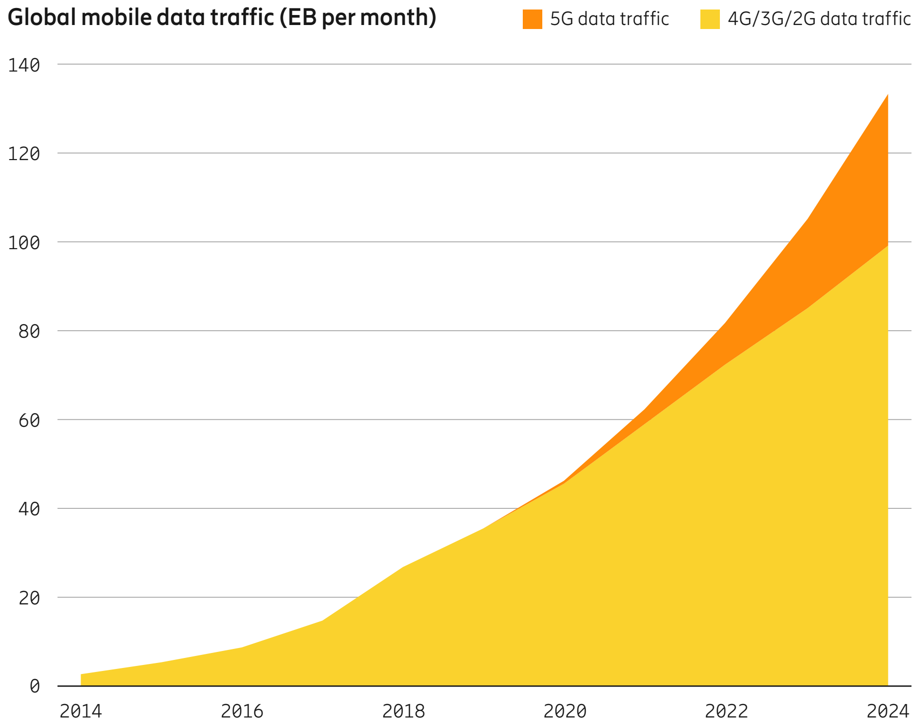


Figure 1.1: Expected worldwide mobile data demand [1].

into three broad categories [7]: Enhanced mobile broadband (eMBB), ultra-reliable and low-latency communications (uRLLC), and massive machine type communications (mMTC). Specifically, eMBB aims to meet the customers' demand for an increasingly digital lifestyle, and focuses on services that have high requirements for bandwidth, such as HD videos, virtual reality (VR), and augmented reality (AR). uRLLC aims to meet expectations for the demanding digital industry and focuses on latency-sensitive services, such as assisted and automated driving, and remote management. And mMTC aims to meet demands for a further developed digital society and focuses on services that include high requirements for connection density, such as smart city and smart agriculture. Figure. 1.2 illustrates some examples for the envisioned 5G usage scenarios in IMT 2020 and beyond [8].

Refer to [7], the IMT 2020 requirements for 5G data rates are: peak data rate of 20 Gbps in downlink, 10 Gbps in uplink, user perceived data rate of 100 Mbps in downlink and 50 Mbps in uplink, spectral efficiency of 30 bps/Hz in downlink and 15 bps/Hz in uplink. In order to achieve such high data rates, some key technologies are expected to be the integral parts of 5G, e.g., network densification with various small cells for interference management [9, 10, 11], massive multiple input multiple output (MIMO) for spatial multiplexing [12, 13], and shifting to higher frequency for larger bandwidth [14, 15]. In particular, due to the congestion of sub-6 GHz bands used by current cellular networks, immigration towards millimeter wave (mmWave) frequency bands (30-300 GHz) is considered as the most attractive enabler for 5G and

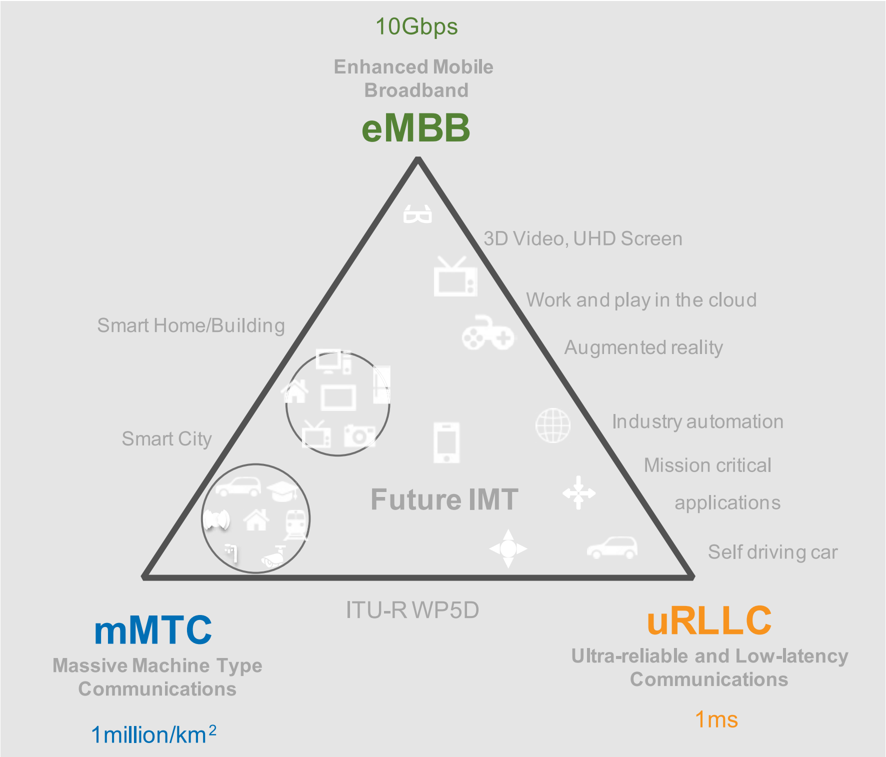


Figure 1.2: The envisioned three 5G usage scenarios for IMT 2020 and beyond [8].

beyond [16, 17, 18, 19]. The main reason is that in mmWave range, there do exist large amounts of relatively idle spectrum as shown in Figure. 1.3. Specifically, the large bandwidths in Ka-band (26.5-40 GHz), V-band (57-71 GHz), and E-band (71-76 GHz and 81-86 GHz), can significantly exceed all the allocations in contemporary cellular networks [20]. Therefore, world-wide academia and industry have intensively collaborated in establishing the foundation of mmWave 5G systems, where novel services and system efficiency are among the most important objectives.

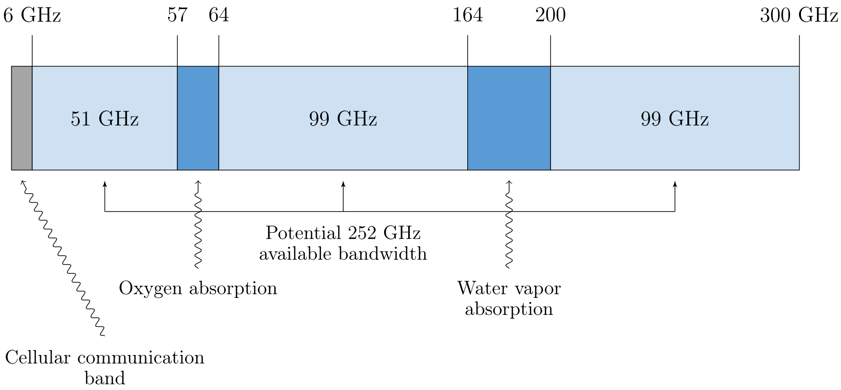


Figure 1.3: The spectrum availability in mmWave frequency range (30-300 GHz) [21].

1.1.1 Distinctive mmWave characteristics

Although the available bandwidth of mmWave frequencies is very large, the propagation characteristics are significantly different from that of the microwave frequency bands, which can be briefly summarized as follows [22, 16]:

- *Path loss.* From Friis's law, the isotropic path loss increases with the carrier frequency. As an example, the free-space path loss decays with the square of carrier frequency. Thus, in a point-to-point communication, one may expect significant path loss when we move from sub-6 GHz to great than 30 GHz carrier frequency [23].

- *Diffraction and blockage.* Diffraction leads to wave propagation in the geometrical shadow region behind obstacles. Diffraction may cause a non-negligible multipath propagation under both line of sight (LOS) and non-LOS conditions [24]. From electromagnetic theory, it is well understood that electromagnetic waves experience a difficulty to diffract when they propagate at obstacles with physical dimensions significantly larger than the wavelength [25]. Furthermore, signals at microwave frequencies can penetrate more easily through solid materials and buildings than mmWaves. For these reasons, mmWave signals are influenced by the effect of shadowing and diffraction to a much greater extent than microwave signals. For instance, one can observe more than 35 dB blockage losses due to bricks, concretes, etc., and around 35 dB due to human body, where these losses are negligible at microwave frequency bands [14].

- *Rain attenuation.* In general, the losses due to a rain attenuation at mmWave frequency bands are much larger than those of microwave bands. If we consider a typical mmWave frequency of 73 GHz, one can observe a rain attenuation of roughly 10 dB/km, which is quite large [26].

- *Atmospheric absorption.* Field measurement results have shown that mmWave signals are more susceptible to oxygen absorption than that of microwave signals. For instance, one can observe roughly 20 dB loss around 60 GHz mmWave signal (see Fig. 1 of [27]).

- *Foliage loss.* The attenuation of radio signals caused due to the presence of trees obstructing the radio link is termed as foliage loss. Foliage losses for mmWaves are significant and can be a limiting factor for some propagation environments. Empirical results demonstrate that at 10 m foliage penetration, the loss at 80 GHz mmWave frequency reaches around 23.5 dB, which is about 15 dB higher compared to that of the 3 GHz microwave frequency [28].

1.1.2 Potentials and challenges for mmWave communication

Having understood the distinctive characteristics of mmWave propagation, the advantages of mmWave communication are self-evident. As we have mentioned before, mmWave frequencies allow for larger channel bandwidth allocations which directly result in higher data rates and indirectly to reduced latencies of the network. Namely, service providers will be able to support user-data-hungry applications with minimal latency. Also, mmWave communication can be used in small cell setting with reduced coverage area, i.e., to establish more densely packed communication links and exploit spatial reuse to provide increased capacity gains.

In addition to the large bandwidth and capacity gain, the utilizing of massive MIMO techniques can guarantee many extra performance gains for mmWave communication:

- *Beamforming gain.* The small wavelength at mmWave frequencies enables to pack a large number of antenna elements in a small form factor, which offers high beamforming gain to compensate for the excessive free space propagation path loss.

- *Interference suppression.* In multi-user systems, the use of multiple antennas at the transmitter and receiver can significantly increase the transmission directivity, which accordingly can increase the potential to alleviate intra-channel interference. This is achieved via precoding at the transmitter or combining at the receiver, or a combination of both.

- *Diversity gain.* Spatial diversity can be exploited in mmWave systems with multiple antennas at both ends to mitigate the impact of fluctuations and loss of signals in the channel.

- *Multiplexing gain.* With multiple antennas at the transmitter, parallel streams can be transmitted to the users without using additional bandwidth or power. This increases the number of spatial dimensions for communication.

However, despite the great potential associated with mmWave communication, a number of challenges need to be addressed so as to be able to exploit these benefits.

- *Power consumption.* In a conventional fully-digital massive MIMO structure, a radio frequency (RF) chain is dedicated to each antenna element, which would result in using a large number of RF chains at mmWave frequencies, and consequently imposes prohibitive power consumption.

- *Hybrid transceiver architecture.* A common solution to reduce the power consumption at mmWave systems is to utilize a hybrid transceiver architecture. Namely, each RF chain is cascaded with a digital baseband unit, which leads to a much lower number of RF chains in comparison with the number of antennas. Accordingly, traditional channel estimation / precoding / combining approaches devised for fully-digital MIMO architectures are not applicable to mmWave hybrid MIMO architectures. This is because the transceivers lack straight access to each antenna element owing to the limited number of RF chains. In addition, the channel matrix is large due to employing large arrays in mmWave systems, and the signal to noise ratio (SNR) is fairly low as a result of severe path loss before beamforming.

- *User Mobility.* A major challenge that comes with user mobility in mmWave transmissions is the significant fluctuations of the channel coefficients since channel coherence time in the mmWave range is very small resulting in a large Doppler spreading. Thus, the signaling schemes used in mobile mmWave communication must take into account the fast time-varying channel states.

- *Integrated circuit (IC) design.* Additional factors that need to be considered when designing ICs for mmWave systems with high carrier frequencies and wide bandwidth include non-linear distortions in the power amplifiers (PAs), phase noise and IQ (in-phase and quadrature) imbalance because the severity of these errors scales up with high frequency transmissions.

- *mmWave relaying.* With the increasing interest in developing small cells for mmWave communication, how to use relays to increase coverage and to support mmWave wireless backhaul for dense small cell deployments remain a big challenge. The main concern thereby includes how to guarantee efficient beam scheduling, high data rate, network stability, low latency, etc..

1.2 Related works in the state of the art

Due to the great potential of mmWave communications, multiple international organizations have emerged for the standardization. In particular, IEEE 802.15.3c specifies the physical layer and MAC (Media Access Control) layer for indoor wireless personal area network (WPAN, also referred to as the piconet) at unlicensed 60 GHz band, which is composed of several wireless nodes and a single piconet controller [29]. IEEE 802.11ad specifies the physical layer and MAC layer for local area network (LAN) at 60 GHz band to support multi-Gbps wireless applications [30]. In particular for the physical layer, two operating modes are defined, the orthogonal frequency division multiplexing (OFDM) mode for high performance applications (e.g., high data rate), and the single carrier (SC) mode for low power and low complexity implementation [29, 6]. Additionally, 5G NR (new radio), which is designed to be the global standard for a unified and more capable 5G wireless air interface, specifies the capability to use mmWave bands to achieve high data rates, enhanced network energy performance, forward compatibility, low latency, and beam-centric design to allow for massive number of antennas [3]. I would refer to [3, 6, 29, 30] for a more detailed review of the corresponding standards.

In addition to the aforementioned standardization activities, many research efforts have also been put into mmWave system design. In terms of the different communication phases in a wireless system, we classify the related literature into four categories, i.e., the initial access phase, the data communication phase, the relay networking and finally the hardware aspect.

An essential component to obtain large antenna gains at mmWave frequencies consists of identifying suitable narrow beam combinations in the initial access phase. The problem of finding an AoA-AoD (angle of arrival, angle of departure) pair is referred to as beam alignment (BA). The inefficiency of naive exhaustive search and the spars characteristic of mmWave channels have motivated a large variety of BA approaches in the literature, e.g., the multi-level hierarchical schemes [31, 32, 33, 34, 35], the compressed sensing schemes [36, 37, 38, 39], etc.. All these algorithms, in some way, suffer from some limitations, e.g., non-scalable for multi-user scenarios, long-time invariant channel assumptions, limited to single-side training, etc..

A large number of works on hybrid architectures have investigated the data communication phase for mmWave systems with an assumption of full channel state information (CSI) [40, 41, 42, 43, 44, 45, 46, 47], namely, the vectors of baseband complex channel coefficients at each array element are known. These works focus on the optimization of the hybrid precoder using the full CSI knowledge. Unfortunately, this assumption is obviously not feasible in a realistic system, since in order to acquire such coefficients, one should be able to sample each antenna element, i.e., one would need an RF chain per antenna element or exhaustively measure all elements successively. Obviously, the former is prohibitively power consuming and the latter is prohibitively time consuming.

While relays on sub-6 GHz bands suffer from severe interference due to their ominidirectional transmissions, the directivity of mmWave antennas significantly mitigates interference, especially in backhaul systems [19, 48]. Recently many efforts have also been made to study the mmWave relay network regime with an emphasis on one or several aspects, such as relay selection, congestion control, routing, scheduling and so on [17, 49, 19, 50, 48, 51]. However, we observe that the existing works more or less encounter some limitations, e.g., the limitation of single path streaming, the ignorance of source admission control, etc.. Particularly, a fundamental information theoretical understanding of mmWave relay networks in terms of its potential at maximum is rather unexplored.

For the hardware aspects, a common theme that underlies most of the hybrid mmWave works is that the fully-connected (FC) architecture outperforms the subarray architecture only at the cost of a higher hardware complexity. However, many reference works [52, 41, 40, 46, 43] have ignored hardware impairments [42], such as the power dissipation, the PA nonlinear distortion and so on [53]. In particular, the nonlinear PAs employed at the BS can drastically distort the transmit signal when operated close to saturation [54]. To this end, a certain power backoff from the saturation power of a PA should be considered accordingly for different signaling schemes and transceiver architectures, such that the PAs can always work in their linear operating region.

As we can see, although many studies have been dedicated to mmWave communication in the last decade, there are still many research gaps regarding to a practical mmWave implementation.

1.3 Contributions and structure of this thesis

This thesis is an accumulation of publications. It is based on four selected journal papers (three published papers after peer-reviewing [55, 56, 57] and one to-be-submitted journal manuscript [58]), which I wrote as first author. These four journal papers constitute the four main chapters (Chapter 3 - Chapter 6) of this thesis.

An overview of the thesis structure and the contributions of each chapters is given in below.

- Chapter 1 is the introduction, which provides the background of mmWave communication as well as its distinctive characteristics, potential, challenges and the state of the art.

- Chapter 2 provides an description of mmWave wireless communication systems as well as the relevant concepts. The mathematical channel and signaling models for mmWave multi-user MIMO (MU-MIMO) are also provided in this chapter so as to prepare the reader for the technical subjects covered in this thesis.

- Chapter 3 studies the initial beam alignment (BA) problem for OFDM mmWave systems. This chapter presents an efficient BA scheme, which explores the AoA-AoD channel domain through pseudo-random multi-finger beam patterns, and then constructs an estimate of the resulting channel second-order statistics. The resulting under-determined system of equations is efficiently solved by using the technique of non-negative least-squares (NNLS). As a result of quadratic channel measuring, the proposed scheme is highly robust to variations of the channel time-dynamics compared with the concurrent approaches in the literature. Also, since all the estimations take place in the downlink, the proposed approach has a strong scalability for multi-user scenarios.

- Chapter 4 is a horizontal extension of Chapter 3, which studies the BA problem for single-carrier (SC) mmWave systems. In this Chapter, we propose a new BA scheme where the base station (BS) periodically probes the channel in the downlink via a pre-specified pseudo-random beamforming codebook and pseudo-random spreading codes, letting each user equipment (UE) estimate its strongest path direction. This scheme again formulates the BA problem as the estimation of a sparse non-negative second-order statistic channel vector and then uses NNLS technique to efficiently find the strongest AoA-AoD pair connecting each UE to the BS. The proposed scheme is completely done in time domain and is highly robust to fast channel variations caused by the large Doppler spread between the multipath components. Furthermore, this chapter will show that after achieving BA, the beamformed channel is essentially frequency-flat, such that SC communication needs no equalization in the time domain.

- Chapter 5 focuses on data communication after BA is achieved. This chapter presents two typical hybrid digital analog (HDA) mmWave antenna architectures that can be regarded as two extreme cases, namely, the fully-connected (FC) and the one-stream-per-subarray (OSPS) architectures. A joint evaluation of the initial BA and the consequent data communication is considered, where the latter takes place by using the beam direction information obtained by the former. A family of MU-MIMO precoding schemes are investigated to well adapt to the hybrid architectures and the beam information extracted from the BA phase. In addition, the power efficiency of the two hybrid architectures are also evaluated by taking into account the power dissipation at different hardware components as well as the power backoff under typical power amplifier constraints. A small conclusion from this chapter is that the two architectures achieve similar sum spectral efficiency, while the OSPS architecture is advantageous with respect to the FC case in terms of hardware complexity and power efficiency, at the sole cost of a slightly longer BA time-to-acquisition due to its reduced beam angle resolution.

- Chapter 6 studies the relay networking for mmWave wireless systems. Although the optimal beam directions for each node pair can be obtained through an BA phase, how to efficiently schedule the beams, in terms of avoiding the queuing explosion as well as assuring large data rates and small end-to-end delays is the main focus of this chapter. More precisely, this chapter studies the beam scheduling problem for mmWave half-duplex (HD) relay networks, where the relay topology can be arbitrary and a link is active only if both nodes focus their beams to face each other. The approximate information theoretical Shannon capacity is introduced to help understand at maximum the potential of the underlying networks. Based on the theoretically optimal schedule results, a prior network simplification procedure is implemented to reduce the network topology complexity, on top of which two practical beam scheduling schemes, i.e., the deterministic edge coloring (EC) scheduler and the adaptive backpressure (BP) scheduler are presented. The former is a very simple one-time computation and then periodical state repetition, hence is more suitable for static scenarios. The later is an “online” approach which will update in every time slots, thus is more favorable for time-varying scenarios. Both of the proposed schedulers can achieve much smaller queuing backlogs, much smaller backlog fluctuations, and much lower packet end-to-end delays in comparison with the reference baseline scheme.

- Chapter 7 finally concludes the thesis and provides suggestions for future work.

1.4 Notations

Vectors, matrices and scalars are denoted by boldface small letters(e.g., $\mathbf{a}$) , boldface capital letters (e.g., $\mathbf{A}$) and non-boldface letters (e.g., a, A), respectively. Sets are denoted by calligraphic letter $\mathcal{A}$ with its cardinality denoted by $|\mathcal{A}|$. The empty set is denoted by $\emptyset$. $\mathbb{E}$ is for the expectation, $\otimes$ is for Kronecker product, $\odot$ is for Hadamard product, $\circledast$ is for continuous-time convolution. $\mathbf{A}^{\mathsf{T}}$ denotes transpose, $\mathbf{A}^{*}$ denotes conjugate, and $\mathbf{A}^{\mathsf{H}}$ denotes conjugate transpose of a matrix $\mathbf{A}$, respectively. The complex circularly symmetric Gaussian distribution with a mean μ and a variance γ is denoted by $\mathcal{CN}(\mu, \gamma)$. For an integer $K \in \mathbb{Z}_{+}$, the shorthand notation $[K]$ is used to represent the set of non-negative integers $\{1, ..., K\}$.

2

System Model for mmWave MU-MIMO

5G promises great flexibility to support a myriad of Internet Protocol (IP) devices, small cell architectures, and dense coverage areas. Applications envisioned for 5G include the Tactile Internet, vehicle-to-vehicle communication, vehicle-to-infrastructure communication, as well as peer-to-peer and machine-to-machine communication, all of which will require extremely low network latency and on-call demand for large bursts of data over minuscule time epochs. Figure. 2.1 shows how backhaul connects the fixed cellular infrastructure (e.g., BS) to the core telephone network and the Internet [5].

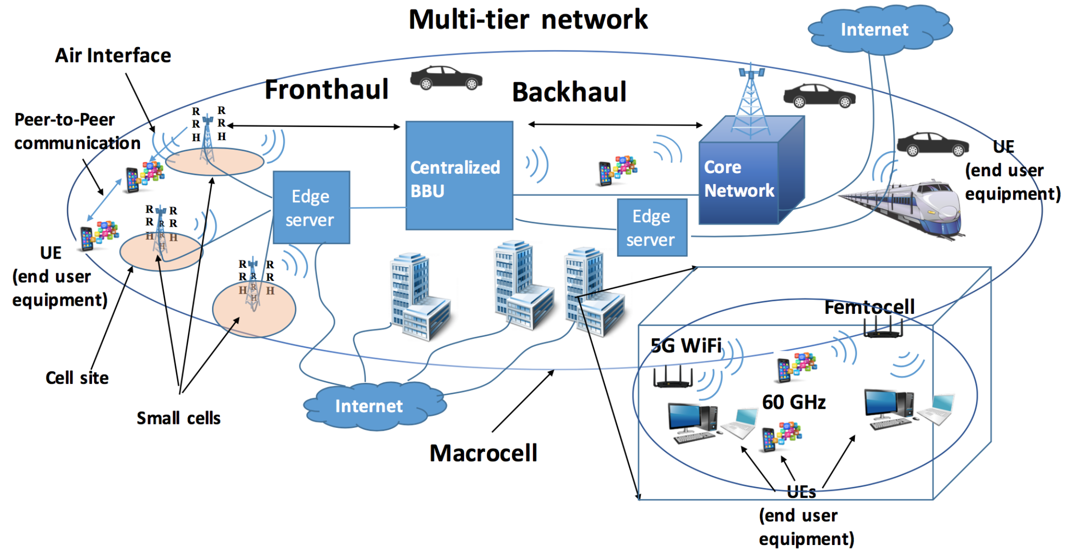
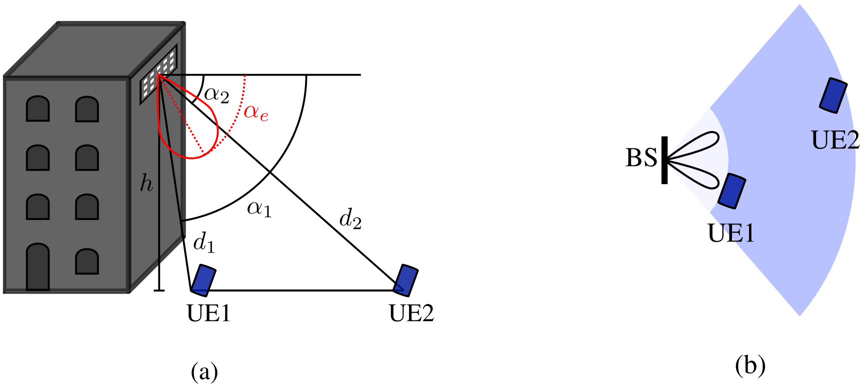


Figure 2.1: An illustration of 5G small cells, edge servers, wireless backhaul, and multi-tier architecture model [5].

As we have discussed before, to address the ever increasing data demand, the wireless industry for 5G is moving to mmWave frequencies, since for the backhaul/fronthaul, mmWave will offer fiber-like data rates and bandwidths to infrastructure without the expense of deploying wired backhaul networks or long-range digital radio-over-fiber (D-RoF). Also, for the small cell mobile users, mmWave will offer unprecedented spectrum and multi-Gbps data rates. However, since the wavelengths shrink by an order of magnitude at mmWave when compared to today's 4G microwave frequencies, diffraction and material penetration will incur greater attenuation, thus elevating the importance of LOS propagation, reflection, and scattering. Therefore, accurate propagation models are vital for the design of new mmWave signaling protocols (e.g., air interfaces). To this end, this chapter will provide the concept model of typical mmWave transceiver architectures, on top of which the mathematical mmWave channel model as well as the multi-user signaling model are also provided.

2.1 Hybrid mmWave transceiver architectures

An effective way to increase area spectral efficiency is to shrink cell size [5, 11], where the reduced number of users per cell, caused by cell shrinking, provides more spectrum to each user. Total network capacity would vastly increase by shrinking cells and reusing the spectrum. Without loss of generality, this chapter consider a small cell configurations as illustrated in Figure. 2.2 (a), where the BS creates a fixed arc-like sectorized beam in the elevation direction. The orientation of the BS beam center in the elevation direction tends to be fixed with an elevation angle α_e [11]. It follows that the probing area in the range direction is restricted and the intensive initial beam searching takes place mainly in the azimuth direction as shown in Figure. 2.2 (b). For notation simplicity, this thesis only focus on the 2D azimuth plane. Extension to the 3D geometry is conceptually straightforward although may lead to a rather high dimensional search for the initial beam acquisition phase.



(a)

(b)

Figure 2.2: Illustration of a small cell scenario with (a) 3D side view and (b) 2D top view.

Two "extreme" HDA architectures are depicted in Figure. 2.3. Figure. 2.3 (a) shows a fully-connected (FC) architecture, where each RF antenna port is connected to all antenna elements of the array. At the other extreme, Figure. 2.3 (b) shows what we refer to as the one-stream-per-subarray (OSPS) architecture, where each RF antenna port is connected to a disjoint subarray.

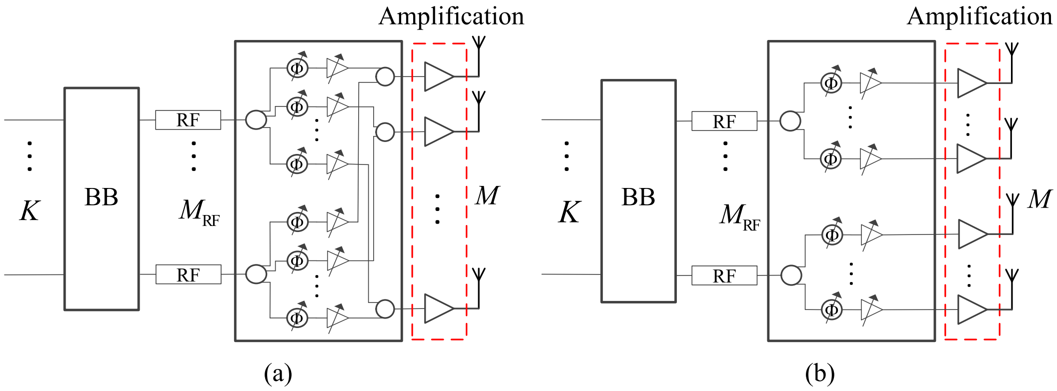


Figure 2.3: Hybrid digital analog (HDA) transmitter architectures: (a) fully-connected (FC), (b) partially-connected with one-stream-per-subarray (OSPS). The "BB" block denotes digital baseband beamforming, K is the number of data streams, $M_{\rm RF}$ is the number of RF chains, and M is the number of antennas.

2.2 Channel model

Extensive measurements have shown that mmWave channels typically exhibit a small number of multipath components (on average of up to 3 strong components), each corresponding to a scattering cluster with small delay / angle spreading [14, 59, 16]. Assume that the BS serves simultaneously K UEs. The BS is equipped with a uniform linear array (ULA) of M antennas and $M_{\rm RF}$ RF chains, where $K \leq M_{\rm RF} \ll M$. Each UE is equipped with a ULA of N antennas and $N_{\rm RF} \ll N$ RF chains. The propagation channel between the BS and the k-th UE, $k \in [K]$, consists of $L_k \ll \max\{M, N\}$ *significant* multipath components. As a result, the $N \times M$ baseband equivalent impulse response of the channel at time slot s can be written as

$$\begin{aligned} \mathsf{H}_{s,k}(t,\tau) &= \sum_{l=1}^{L_k} \rho_{s,k,l} e^{j2\pi\nu_{k,l}t} \mathbf{a}_{\rm R}(\phi_{k,l}) \mathbf{a}_{\rm T}(\theta_{k,l})^{\mathsf{H}} \delta(\tau - \tau_{k,l}) \\ &= \sum_{l=1}^{L_k} \mathsf{H}_{s,k,l}(t) \delta(\tau - \tau_{k,l}), \end{aligned} \tag{2.1}$$

where $\mathsf{H}_{s,k,l}(t) := \rho_{s,k,l} e^{j2\pi\nu_{k,l}t} \mathbf{a}_{\rm R}(\phi_{k,l}) \mathbf{a}_{\rm T}(\theta_{k,l})^{\mathsf{H}}$ and $\delta(\cdot)$ denotes the Dirac delta function. Each l-th multipath component is identified by the tuple $(\phi_{k,l}, \theta_{k,l}, \tau_{k,l}, \nu_{k,l})$ of angle of arrival (AoA), angle of departure (AoD), delay, and Doppler shift, respectively. The vectors $\mathbf{a}_{\rm T}(\theta_{k,l}) \in \mathbb{C}^M$ and $\mathbf{a}_{\rm R}(\phi_{k,l}) \in \mathbb{C}^N$ are the array response vectors of the BS and the k-th UE at the AoD $\theta_{k,l}$ and the AoA $\phi_{k,l}$, respectively. With the ULA configuration and the assumption that the spacing of the ULA antennas in each array / subarray equals to a half-wavelength $\lambda/2$, the elements of $\mathbf{a}_{\mathrm{T}}(\theta_{k,l})$ and $\mathbf{a}_{\mathrm{R}}(\phi_{k,l})$ are given by

$$[\mathbf{a}_{\mathrm{T}}(\theta)]_{(i'-1)\cdot\hat{M}+d} = e^{j(d-1)\pi\sin(\theta)} \cdot e^{j\Psi(i',\theta)}, d \in [\hat{M}] \tag{2.2a}$$

$$[\mathbf{a}_{\mathrm{R}}(\phi)]_n = e^{j(n-1)\pi\sin(\phi)}, n \in [N], \tag{2.2b}$$

where in (2.2a) we assume that $(i' \equiv 1, \hat{M} = M)$ for the FC architecture as shown in Figure. 2.3(a), and $(i' \in [M_{\mathrm{RF}}], \hat{M} = \frac{M}{M_{\mathrm{RF}}})$ for the OSPS architecture as shown in Figure. 2.3(b). The additional term $\Psi(i',\theta)$ in (2.2a) takes into account the phase shifts among different subarrays, given by

$$\Psi(i',\theta) = \frac{2\pi}{\lambda}(i'-1)\cdot D_x \cdot \sin(\theta), \tag{2.3}$$

where i' indicates the index of the subarrays and $D_x \geq 0$ denotes the subarray center-to-center spacing in the scan direction. Hence, in the special case with $D_x = 0$, all the subarrays are co-located; while with $D_x = \frac{M}{M_{\mathrm{RF}}} \cdot \frac{\lambda}{2}$, the antenna element layout in the scan direction for the OSPS architecture is exactly the same as for the FC architecture.

We adopt a block fading model, where the coefficient of the l-th multipath component $\rho_{s,k,l}$ is constant over a short interval (within one slot) and changes from slot to slot according to a wide-sense stationary process statistics characterized by its power spectral density (Doppler spectrum) [60]. When the channel coherence time (related to the inverse of the bandwidth of the Doppler spectrum, see [60]) is significantly larger than the slot duration but equal or smaller than the (non-consecutive) slot separation in time, a convenient model is to consider the coefficients as i.i.d. across different slots. Moreover, the Doppler shift $\nu_{k,l}$ as defined in (2.1) introduces a continuous phase rotation for each channel sample. Each multipath component (channel tap coefficient) is formed by the superposition of a large number of micro-scattering components (e.g., due to rough surfaces) having (approximately) the same AoA-AoD and delay. By the central limit theorem, it is customary to model the superposition of these many small effects as Gaussian [61, 62]. Hence, the multipath component coefficients can be modeled as Rice fading given by

$$\rho_{s,k,l} \sim \sqrt{\gamma_{k,l}} \left(\sqrt{\frac{\eta_{k,l}}{1+\eta_{k,l}}} + \frac{1}{\sqrt{1+\eta_{k,l}}} \check{\rho}_{s,k,l} \right), \tag{2.4}$$

where $\gamma_{k,l}$ denotes the overall multipath component strength, $\eta_{k,l} \in [0,\infty)$ indicates the strength ratio between the specular reflection (or LOS) and the scattered components, and $\check{\rho}_{s,k,l} \sim \mathcal{CN}(0,1)$ is a zero-mean unit-variance complex Gaussian random variable whose value changes in an i.i.d. fashion across different slots. In particular, $\eta_{k,l} \to \infty$ indicates a pure LOS path while $\eta_{k,l} = 0$ indicates a pure scattered path, affected by Rayleigh fading.

The AoA-AoDs $(\phi_{k,l}, \theta_{k,l})$ in (2.1) can take on arbitrary values in the continuous AoA-AoD domain. Following the widely used approach of [63], known as *beam-domain representation*, we obtain a finite-dimensional representation of the channel response (2.1). More precisely, we consider the discrete set of AoA-AoDs

$$\Phi := \left\{ \check{\phi} : (1 + \sin(\check{\phi}))/2 = \frac{n-1}{N}, \, n \in [N] \right\}, \tag{2.5a}$$

$$\Theta := \left\{ \check{\theta} : (1 + \sin(\check{\theta}))/2 = \frac{m-1}{M}, m \in [M] \right\}. \tag{2.5b}$$

It follows that the corresponding sets $\mathcal{A}_{\rm R} := \{\mathbf{a}_{\rm R}(\check{\phi}) : \check{\phi} \in \Phi\}$ and $\mathcal{A}_{\rm T} := \{\mathbf{a}_{\rm T}(\check{\theta}) : \check{\theta} \in \Theta\}$ form discrete dictionaries to represent the channel response. For the ULAs considered in this paper, the dictionaries $\mathcal{A}_{\rm R}$ and $\mathcal{A}_{\rm T}$, after suitable normalization, reduce to the columns of unitary discrete Fourier transform (DFT) matrices $\mathbf{F}_N \in \mathbb{C}^{N \times N}$ and $\mathbf{F}_M \in \mathbb{C}^{M \times M}$, with elements

$$[\mathbf{F}_N]_{n,n'} = \frac{1}{\sqrt{N}} e^{j2\pi(n-1)(\frac{n'-1}{N} - \frac{1}{2})}, n, n' \in [N], \tag{2.6a}$$

$$[\mathbf{F}_M]_{m,m'} = \frac{1}{\sqrt{M}} e^{j2\pi(m-1)(\frac{m'-1}{M} - \frac{1}{2})}, m, m' \in [M]. \tag{2.6b}$$

Consequently, based on a subarray basis indexed by i', the beam-domain representation of the channel response (2.1) is given by [63, 15]

$$\begin{aligned} \check{\mathsf{H}}^{i'}_{s,k}(t,\tau) &= \mathbf{F}_N^{\mathsf{H}} \mathsf{H}_{s,k}(t,\tau) \cdot \left(\mathbf{F}_M \odot \mathbb{1}_{\{(i'-1)\hat{M}+1:i'\hat{M},1:M\}} \right) \\ &= \sum_{l=1}^{L_k} \check{\mathsf{H}}^{i'}_{s,k,l}(t) \delta(\tau - \tau_l), \end{aligned} \tag{2.7}$$

where $(i' \equiv 1, \hat{M} = M)$ for the FC architecture, and $(i' \in [M_{\rm RF}], \hat{M} = \frac{M}{M_{\rm RF}})$ for the OSPS architecture. Here we define $\check{\mathsf{H}}^{i'}_{s,k,l}(t) := \mathbf{F}_N^{\mathsf{H}} \mathsf{H}_{s,k,l}(t) \cdot \left(\mathbf{F}_M \odot \mathbb{1}_{\{(i'-1)\hat{M}+1:i'\hat{M},1:M\}} \right)$ as the beam-domain l-th multipath component between the k-th UE and the BS, where $\mathbb{1}_{\{a_1:a_2,b_1:b_2\}} \in \mathbb{C}^{M \times M}$ is an indicator matrix, with 1 at the components indexed by rows from a_1 to a_2 and by columns from b_1 to b_2, otherwise zero. The indicator matrix takes into account the fact that the number of antenna elements for each subarray in the OSPS architecture is $M_{\rm RF}$ times less than that in the FC architecture.

2.3 Signaling model

Let $\mathbf{x}_s(t) = [x_{s,1}(t), x_{s,2}(t), ..., x_{s,K}(t)]^{\mathsf{T}}$ denote the continuous-time baseband equivalent signal (either pilot or data signal), transmitted from the BS over the s-th slot. With HDA beamforming, the beamformed signal at the output of the transmitter over the

s-th slot is generally given by

$$\hat{\mathbf{x}}_s(t) = \sqrt{E_0} \cdot \mathbf{U}_s^{\mathrm{RF}} \cdot \mathbf{W}_s^{\mathrm{BB}} \cdot \mathbf{x}_s(t), \tag{2.8}$$

where for simplicity of exposition we restrict to the case of uniform power allocation, with $E_0 = \frac{P_{\texttt{tot}} T_c}{K}$ indicating the per-chip energy of each signal stream, where $P_{\texttt{tot}}$ denotes the total radiated power at the BS and $T_c = \frac{1}{B}$ denotes the chip duration with B indicating the signaling bandwidth. In (2.8), we define $\mathbf{W}_s^{\mathrm{BB}} \in \mathbb{C}^{M_{\mathrm{RF}} \times K}$ and $\mathbf{U}_s^{\mathrm{RF}} \in \mathbb{C}^{M \times M_{\mathrm{RF}}}$ as the baseband (digital) and the RF analog beamforming matrices, respectively. Note that, depending on the transmitter architecture, the analog beamforming matrix $\mathbf{U}_s^{\mathrm{RF}}$ takes on the form

$$[\tilde{\mathbf{u}}_{s,1}, \tilde{\mathbf{u}}_{s,2}, \cdots, \tilde{\mathbf{u}}_{s,M_{\mathrm{RF}}}] \text{ and } \begin{bmatrix} \tilde{\mathbf{u}}_{s,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{u}}_{s,2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{u}}_{s,M_{\mathrm{RF}}} \end{bmatrix} \tag{2.9}$$

for the FC and the OSPS architectures, respectively, where $\tilde{\mathbf{u}}_{s,i} \in \mathbb{C}^{\hat{M}}$, $i \in [M_{\mathrm{RF}}]$, with $\hat{M} = M$ for the FC architecture and $\hat{M} = \frac{M}{M_{\mathrm{RF}}}$ for the OSPS architecture. Hence, in both cases $\mathbf{U}_s^{\mathrm{RF}}$ has dimension $M \times M_{\mathrm{RF}}$, but FC has a full matrix, while OSPS has a block-diagonal matrix, due to the constrained connectivity. Without loss of generality, the beamforming vectors are normalized as $\sum_{i=1}^{M_{\mathrm{RF}}} \|\mathbf{u}_{s,i}\|^2 = M_{\mathrm{RF}}$.

The beamformed signal (2.8) goes through the channel as defined in (2.1). At the UE side, because of the HDA architecture, the UE does not have direct access to each antenna element. Instead, at each slot s, the UE obtains only a projection of the received signal by applying some beamforming vector in the analog domain. For notation simplicity, let's consider a single RF chain at each UE with $N_{\mathrm{RF}} = 1$. The extension to $N_{\mathrm{RF}} > 1$ is straightforward and will be considered in later sections. Thus, the received signal at the k-th UE side is given by

$$\begin{aligned} \hat{y}_{s,k}(t) =& \mathbf{v}_{s,k}^{\mathsf{H}} \mathsf{H}_{s,k}(t,\tau) \circledast \hat{\mathbf{x}}_s(t) + z_{s,k}(t) \\ =& \sqrt{E_0} \mathbf{v}_{s,k}^{\mathsf{H}} \mathsf{H}_{s,k}(t,\tau) \circledast \left(\mathbf{U}_s^{\mathrm{RF}} \cdot \mathbf{W}_s^{\mathrm{BB}} \cdot \mathbf{x}_s(t) \right) \\ & + z_{s,k}(t), \end{aligned} \tag{2.10}$$

where $\mathbf{v}_{s,k} \in \mathbb{C}^N$ denotes the normalized beamforming vector with $\|\mathbf{v}_{s,k}\| = 1$ at the k-th UE, and $z_{s,k}(t)$ is the continuous-time complex additive white Gaussian noise (AWGN) at the output of the UE RF chain, with a power spectral density (PSD) of N_0 Watt/Hz.

In order to clearly describe the channel condition between the BS and a generic UE, it is useful to first define the channel SNR before beamforming (BBF) $\mathsf{SNR}_{\mathrm{BBF}}$, given by

$$\mathsf{SNR}_{\mathrm{BBF},\,k} = \frac{P_{\texttt{tot}} \sum_{l=1}^{L_k} \gamma_{k,l}}{N_0 B}. \tag{2.11}$$

where k is the index of the UE and $\gamma_{k,l}$ denotes the strength of the l-th multipath component. The SNR in (2.11) indicates the ratio of the total received signal power (summing over all the multipath components) over the total noise power at the receiver baseband processor input, assuming that the signal is isotropically transmitted by the BS and isotropically received at the k-th UE over the total bandwidth B. As mentioned before, one of the challenges of mmWaves communication is that the SNR before beamforming $\mathsf{SNR}_{\mathrm{BBF}}$ in (2.11) may be very low.

2.4 Summary

This chapter presents two “extreme” hybrid mmWave antenna architectures, on top of which the mathematical channel and signaling models are also provided. The main object of this chapter is to prepare the reader for the basic mmWave channel mathematics covered in this thesis.

3

Initial Beam Alignment for mmWave OFDM Systems

3.1 Introduction

To cope with the severe path loss at mmWave frequencies, directional beamforming both at the BS side and the UE side is necessary in order to establish a strong path conveying enough signal power. Finding such beamforming directions is referred to as beam alignment (BA). This chapter presents an efficient BA scheme which can be used in the initial access phase for mmWave OFDM systems.

3.2 Clarification of each authors' contributions

This chapter is a journal publication, which is a joint work with Saeid Haghighatshoar and Giuseppe Caire. I wrote this journal as the first author. The citation information is in below:

X. Song, S. Haghighatshoar, and G. Caire,"A scalable and statistically robust beam alignment technique for mm-wave systems," IEEE Transactions on Wireless Communications, 2018. DOI: 10.1109/TWC.2018.2831697

All the authors contributed to this paper, but I have implemented all the experiments and simulations. I also wrote the complete first draft (including all sections) of this paper.

Saeid Haghighatshoar provided valuable ideas for the channel and signaling model as well as the mathematical techniques for the channel estimation. He also modified my first draft in terms of its English expressions.

Giuseppe Caire, who is my PhD supervisor, provided valuable discussions in each meeting of this work. He also did a final modification of the overall draft.

3.3 Original journal article

The following article is a reprint of the original journal paper. It is the accepted version of the paper. The copyright information is given in page xii of this thesis as well as in the first page of the reprinted paper.

A Scalable and Statistically Robust Beam Alignment Technique for mm-Wave Systems

Xiaoshen Song, *Student Member, IEEE,* Saeid Haghighatshoar, *Member, IEEE,* Giuseppe Caire, *Fellow, IEEE*

***Abstract*—Millimeter-Wave (mm-Wave) frequency bands provide an opportunity for much wider channel bandwidth compared with the traditional sub-6 GHz band. Communication at mm-Waves is, however, quite challenging due to the severe propagation pathloss incurred by conventional isotropic antennas. To cope with this problem, directional beamforming both at the *Base Station* (BS) side and at the *User Equipment* (UE) side is necessary in order to establish a strong path conveying enough signal power. Finding such beamforming directions is referred to as *Beam Alignment* (BA). This paper presents a new scheme for efficient BA. Our scheme finds a strong propagation path identified by an Angle-of-Arrival (AoA) and Angle-of-Departure (AoD) pair, by exploring the AoA-AoD domain through pseudo-random multi-finger beam patterns, and constructing an estimate of the resulting second-order statistics (namely, the average received power for each pseudo-random beam configuration). The resulting under-determined system of equations is efficiently solved using non-negative constrained Least-Squares, yielding naturally a sparse non-negative vector solution whose maximum component identifies the optimal path. As a result, our scheme is highly robust to variations of the channel time-dynamics compared with alternative concurrent approaches based on the estimation of the instantaneous channel coefficients, rather than of their second-order statistics. In the proposed scheme, the BS probes the channel in the *Downlink* (DL) and trains simultaneously an arbitrarily large number of UEs. Thus, "beam refinement", with multiple interactive rounds of *Downlink/Uplink* (DL/UL) transmissions, is not needed. This results in a scalable BA protocol, where the protocol overhead is virtually independent of the number of UEs since all the UEs run the BA procedure at the same time. Extensive simulation results illustrate that our approach is superior to the state-of-the-art BA schemes proposed in the literature in terms of training overhead in multi-user scenarios and robustness to variations in the channel dynamics.**

***Index Terms*—Millimeter-Wave, Beam Alignment, Compressed Sensing, Non-Negative Least-Squares (NNLS).**

I. Introduction

Communication at millimeter-waves (mm-Waves) provides an opportunity to fulfill the demand for high data rates in the next generation communication networks because of the large available bandwidth [1]. A critical challenge to signaling at mm-Waves compared with sub-6 GHz spectrum is the severe propagation loss when conventional isotropic antennas are used [2]. The standard way to counter the isotropic pathloss consists of using antenna gain at both the transmitter and the receiver sides. In a mobile environment, such antenna gain is achieved by electronically steerable antenna arrays, in order to cope with beam direction changes due to the relative motion of transmitter and receiver. Fortunately, due to the small wavelength, it is possible to package a large number of antenna elements in a small form factor, such that large antenna arrays can be implemented at both the *Base Station* (BS) side and the *User Equipment* (UE) side. Moreover, it has been observed experimentally and modeled mathematically that the propagation channel at mm-Waves is formed by a very sparse collection of scatterers in the angle domain [3–6]. This implies that, to establish reliable communication, the BS and the UE need to focus their beams in the direction of a strong path. For example, in the case of *Line-of-Sight* (LoS) propagation, the beams must point at each other since the LoS path is typically the strongest one.

More in general, we refer to the problem of finding a narrow beam direction at both the BS and the user sides yielding a SNR *after beamforming* above a desired threshold as the *Beam Alignment* (BA) problem. This problem is quite well studied in the literature [3–16]. In particular, it is known to be a challenging problem since in mm-Waves the SNR *before beamforming* (i.e., in isotropic propagation conditions) is typically very low, especially in outdoor non-LoS conditions. Moreover, although the number of array antennas may be very large, the number of *Radio Frequency* (RF) chains is limited, due to the difficulty of implementing a full RF chain (including A/D conversion, modulation, and PA/LNA amplification) for each array element in a very small form factor and for a very large bandwidth. The small number of RF chains prevents the implementation of classical digital beamforming schemes in the baseband domain. Hence, a widely studied approach consists of *Hybrid-Digital-Analog* (HDA) beamforming [7, 17]. In this case, a naive sequential scanning of the *Angle-of-Departure* (AoD) and *Angle-of-Arrival* (AoA) domains with narrow beams in order to find an alignment to a strongly connected propagation path is very time-consuming and would incur a large initial acquisition protocol overhead, not suited for outdoor mobile applications [11–18].

The authors are with the Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: xiaoshen.song@campus.tu-berlin.de).

X. Song is sponsored by the China Scholarship Council (201604910530).

A. Related State-of-the-Art

The inefficiency of naive alignment search has motivated BA algorithms based on hierarchical adaptive search, interactive search, and *Compressed Sensing* (CS) techniques [8–16].

The fundamental idea of hierarchical methods is to use wider beam patterns at the start of the search and to refine them in several consecutive stages. In [11], for example, the authors develop a bisection algorithm in which the range of AoDs and AoAs are divided by a factor of 2 at each step and is refined by probing the resulting 2×2 sections and identifying the section with the maximum received power. A similar idea using overlapped beam patterns is used in [12]. Such hierarchical techniques, however, require the interaction of the BS with each individual UE, since the training is bi-directional and involves both *Downlink* (DL) probing and *Uplink* (UL) feedback for each iterative round.

In [13], a method is proposed where the BS and the UE iteratively and collaboratively identify the dominant eigenvector of their channel matrix via the well-known *power method*. However, this approach requires to demodulate the signal at each antenna both at the BS and at the UE sides. Therefore, this method is essentially incompatible with the HDA beamforming structure.

More recently, considering the natural channel sparsity in the AoA-AoD domain [3–6], CS-based algorithms have been proposed for BA in mm-Waves [14–16, 19–21]. These algorithms are efficient and particularly attractive for multi-user scenarios, but they are based on the assumption that the instantaneous channel remains invariant during the whole probing/measuring stage (the same assumption is also adopted in [11, 12]). This assumption is typically not satisfied in practice due to the large Doppler spread at mm-Waves, implying significant time-variations of the channel coefficients even in conditions of moderate mobility [22, 23].[1]

B. Contributions

In this paper, we propose a novel BA scheme that has the following advantages compared with the existing works in the literature:

1) *Low-Complexity Beam Direction Estimation:* Our scheme finds a strong propagation path identified by an AoA-AoD pair, by exploring the AoA-AoD domain through pseudo-random multi-finger beam patterns, and constructing an estimate of the resulting second-order statistics (namely, the average received power for each pseudo-random beam configuration). The resulting underdetermined system of equations is efficiently solved using *Non-Negative Least-Squares* (NNLS), yielding naturally a sparse non-negative vector solution whose maximum component identifies the optimal path.

2) *System-Level Scalability:* In our approach, the BS actively probes the channel by periodically broadcasting a beamforming codebook (consists of a sequence of *pseudo-random beamforming patterns*) over reserved beacon slots in the DL, while all UEs stay in listening mode. Measurements are collected by the UEs, which locally and independently identify the AoA-AoD of a strong multipath component. Since there is no need for interaction between the BS and each UE, the proposed BA scheme is highly scalable and its overhead and complexity do not grow with the number of active users in the system.

3) *User-Specific Beamforming Codebook:* During the beacon slots, each UE makes use of its own *receive beamforming codebook*. The BS needs no knowledge of such codebook, which can be locally generated by each UE. We shall show that the optimal *angular spreading factor* of the receiver beamforming patterns yielding the fastest BA acquisition time depends on the pre-beamforming SNR. Hence, our method has the advantage that beamforming codebook of each UE can be individually and locally tailored, depending on hardware constraints (number of RF chains) and SNR conditions, without impacting the overall system functions.

4) *Robustness to Variations in Channel Statistics:* Our scheme is based on quadratic measurements (i.e., averaged received power, yielding estimates of the channel second order statistics), rather than linear measurements of the channel coefficient vectors. As such, our scheme is highly robust to variations in the channel time-dynamics. We also illustrate via numerical simulations that existing CS-based algorithms fail to estimate the channel strong path direction when the channel is significantly time-varying, i.e., it undergoes several fading cycles during the estimation period, whereas our scheme performs well for a wide range of channel dynamics. Using channel second order statistics for BA is also considered in [24] via the *Maximum Likelihood* (ML) estimation of the channel covariance matrix. However, in [24] the channel probing signals are transmitted isotropically through a single antenna or via a fixed beamforming pattern from the BS side. The drawback is that with isotropic transmission the received SNR at the UE side might be very low whereas with fixed beamforming the transmit pattern might not hit any strong multipath component. Moreover, in [24] the UEs can estimate only their corresponding AoAs rather than the joint AoA-AoD pairs of the strong paths. In contrast, our scheme yields the joint AoA-AoD pairs, allowing full BA at both the BS and the UE sides.

Notation: We denote vectors by boldface small (e.g., $\mathbf{a}$) and matrices by boldface capital (e.g., $\mathbf{A}$) letters. Scalars are denoted by non-boldface letters (e.g., a, A). We represent sets by calligraphic letter $\mathcal{A}$ and their cardinality with $|\mathcal{A}|$. We denote the empty set by $\emptyset$. We use $\mathbb{E}$ for the expectation, $\otimes$ for the Kronecker product of two matrices,

[1] Notice that the channel time-variations are greatly reduced *after BA is achieved*, since once the beams are aligned, the effective channel angular spread is very small [23]. However, *before BA is achieved*, the channel variability over time can be large, since even a small motion of a few centimeters traverses several wavelengths, potentially producing multiple deep fades [22].

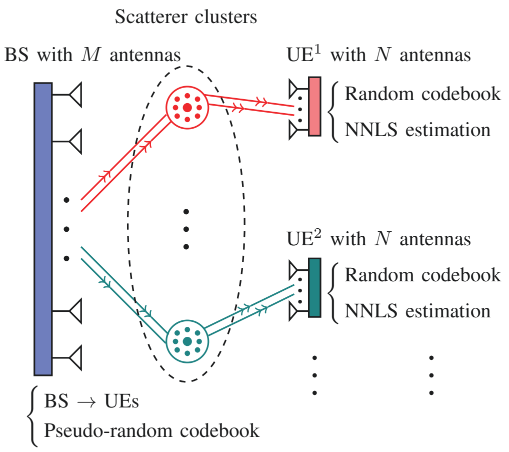


Fig. 1: *Illustration of the physical channel model and our proposed Beam Alignment (BA) scheme.*

$\mathbf{A}^{\mathsf{T}}$ for transpose, $\mathbf{A}^*$ for conjugate, and $\mathbf{A}^{\mathsf{H}}$ for conjugate transpose of a matrix $\mathbf{A}$. The output of an optimization problem such as $\arg\min_{x\in\mathcal{X}} f(x)$ is denoted by x^*. The complex circularly symmetric Gaussian distribution with a mean μ and a variance γ is denoted by $\mathcal{CN}(\mu,\gamma)$. For an integer $k \in \mathbb{Z}$, we use the shorthand notation $[k]$ for the set of non-negative integers $\{1,...,k\}$.

II. Basic Setup

A. Channel Model

We consider a mm-Wave system including a BS equipped with a *Uniform Linear Array* (ULA) with M antennas and $m \ll M$ RF chains. We consider a generic UE, also equipped with a ULA with N antennas and $n \ll N$ RF chains. We assume that both the BS and UE arrays have the antenna spacing $d = \frac{\lambda}{2}$, where λ is the wavelength given by $\lambda = \frac{c_0}{f_0}$, where c_0 is the speed of the light and f_0 is the carrier frequency. We denote by $\theta, \phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ the steering angles with respect to the BS and UE arrays. We represent the array responses of the BS and UE to a planar wave coming from the angles θ and ϕ by the M-dim and N-dim array vectors $\mathbf{a}(\theta) \in \mathbb{C}^M$ and $\mathbf{b}(\phi) \in \mathbb{C}^N$ respectively, with elements

$$[\mathbf{a}(\theta)]_k = e^{j(k-1)\pi\sin(\theta)}, k \in [M], \tag{1a}$$

$$[\mathbf{b}(\phi)]_l = e^{j(l-1)\pi\sin(\phi)},\ l \in [N]. \tag{1b}$$

We assume that the communication between the BS and the UE occurs via a collection of sparse *multi-path components* (MPCs) in the AoA-AoD-delay domain [1], where the $N \times M$ low-pass equivalent impulse response of the channel at a symbol time s is given by[2]

$$\mathsf{H}_s(\tau) = \sum_{l=1}^{L} \rho_{s,l}\mathbf{b}(\phi_l)\mathbf{a}(\theta_l)^{\mathsf{H}}\delta(\tau-\tau_l), \tag{2}$$

where $\rho_{s,l}$ is the random channel gain of the l-th MPC at AoA-AoD-delay $(\theta_l, \phi_l, \tau_l)$, $l \in [L]$. Typically the number of *significant* MPCs satisfies $L \ll \max\{M,N\}$ [2]. In practice there may be a large number of MPCs that convey such a small amount of signal power that can be simply neglected since in any case they will not be useful for signal transmission even after the BA is achieved. Note that in the channel model, we made the implicit assumption (very common in most beamforming and array processing literature) that the communication bandwidth B is much smaller than the carrier frequency f_0, such that the array responses in (1) are essentially constant with $f \in [f_0 - B/2, f_0 + B/2]$. We adopt a block fading model, where the channel gains $\rho_{s,l}$, $l \in [L]$, remain invariant over the channel coherence time Δt_c but change randomly across different coherence times according to a given wide-sense stationary process with given Doppler power spectral density [25]. We also assume that each MPC is formed by a cluster of micro-scatterers corresponding (roughly) to the same delay and AoA-AoD (see Fig. 1), such that the channel gains $\rho_{s,l} \sim \mathcal{CN}(0,\gamma_l)$ have a zero-mean complex Gaussian distribution.

We also assume that the angle coherence time, i.e., the time scale over which the AoA-AoDs of the scatterers $\{(\theta_l,\phi_l)\}_{l=1}^L$ change significantly, is much longer than the channel coherence time Δt_c. Hence, the angles can be treated as locally constant (but unknown) during the BA phase. This *local stationarity* of the scattering geometry is widely used in the literature and confirmed by channel sounding measurements (e.g., see [23, 26]).

B. Signaling Model

Consider the communication between the BS and a generic UE. Since the BS has m RF chains, it can transmit up to m different data streams. For a given signaling interval t_0, let $x_{s,i}(t)$, $t \in [st_0, (s+1)t_0)$, be the continuous-time baseband equivalent signal corresponding to the i-th data stream. We assume that the channel is time-invariant over each symbol, i.e., $t_0 < \Delta t_c$. To transmit the i-th data stream, the BS applies a beamforming vector $\mathbf{u}_{s,i} \in \mathbb{C}^M$. Without loss of generality, the beamforming vectors are normalized such that $\|\mathbf{u}_{s,i}\| = 1$.[3] The (baseband equivalent) transmitted signal at symbol time s is given by

$$\mathbf{x}_s(t) = \sum_{i=1}^{m} x_{s,i}(t)\mathbf{u}_{s,i}. \tag{3}$$

The corresponding received signal at the UE array is

$$\begin{aligned}\mathbf{r}_s(t) &= \int \mathsf{H}_s(\tau)\mathbf{x}_s(t-\tau)d\tau \\ &= \sum_{l=1}^{L}\sum_{i=1}^{m} \rho_{s,l}x_{s,i}(t-\tau_l)\mathbf{b}(\phi_l)\mathbf{a}(\theta_l)^{\mathsf{H}}\mathbf{u}_{s,i}\end{aligned}$$

[2] Consistently with the current technology trend in mm-Wave systems, in this paper we focus on a *Time-Division Duplexing* (TDD), where the UL and the DL communication occur over the same frequency band.

[3] Also, note that here we are assuming that the beamforming vectors $\mathbf{u}_{s,i}$, $i \in [m]$ are implemented in the RF domain via an analog beamforming network and therefore they are frequency flat, i.e., they are constant over the whole signal bandwidth.

$$= \sum_{l=1}^{L} \sum_{i=1}^{m} \rho_{s,l} g^{\mathrm{BS}}_{s,l,i} x_{s,i}(t - \tau_l) \mathbf{b}(\phi_l) \tag{4}$$

where $g^{\mathrm{BS}}_{s,l,i} := \mathbf{a}(\theta_l)^{\mathsf{H}} \mathbf{u}_{s,i}$ denotes the beamforming gain along the l-th MPC at the BS side for the i-th RF chain. As stated before, we assume that the UE is also equipped with n RF chains and the analog RF signal received at the UE antenna array is distributed into these chains for demodulation. This is achieved by signal splitters that divide the signal power by a factor of n. The noise in the receiver is mainly introduced by the RF chain electronics (filter, mixer, and A/D conversion). It follows that the noisy received signal at the output of the j-th RF chain at the UE side is given by

$$\begin{aligned} y_{s,j}(t) &= \frac{1}{\sqrt{n}} \mathbf{v}^{\mathsf{H}}_{s,j} \mathbf{r}_s(t) + z_{s,j}(t) \\ &= \frac{1}{\sqrt{n}} \sum_{l=1}^{L} \sum_{i=1}^{m} \rho_{s,l} g^{\mathrm{BS}}_{s,l,i} x_{s,i}(t-\tau_l) \mathbf{v}^{\mathsf{H}}_{s,j} \mathbf{b}(\phi_l) + z_{s,j}(t) \\ &= \sum_{i=1}^{m} \frac{1}{\sqrt{n}} \sum_{l=1}^{L} \rho_{s,l} g^{\mathrm{BS}}_{s,l,i} g^{\mathrm{UE}}_{s,l,j} x_{s,i}(t - \tau_l) + z_{s,j}(t) \end{aligned} \tag{5}$$

where $\mathbf{v}_{s,j} \in \mathbb{C}^N$ denotes the normalized beamforming vector of the j-th RF chain at the UE side, where $g^{\mathrm{UE}}_{s,l,j} := \mathbf{v}^{\mathsf{H}}_{s,j} \mathbf{b}(\phi_l)$ denotes the array gain of the j-th RF chain along the l-th MPC, and where $z_{s,j}(t)$ is the continuous-time complex *Additive White Gaussian Noise* (AWGN) at the output of the j-th RF chain, with *Power Spectral Density* (PSD) of N_0 Watt/Hz. The factor $1/\sqrt{n}$ in (5) takes into account the power split said above.

In this paper, we consider OFDM signaling with given subcarrier separation Δf, hence, each symbol $x_{s,i}(t)$ in the general model defined before corresponds here to an OFDM symbol. The number of subcarriers is given by $F := B/\Delta f$, where B denotes the channel bandwidth as defined before. We make the standard assumption that the duration τ_{cp} of the *Cyclic Prefix* (CP) of the OFDM modulation is longer than the channel delay spread, implying $t_0 = 1/\Delta f + \tau_{\mathrm{cp}}$ with $\tau_{\mathrm{cp}} \geq \max\{\tau_l\} - \min\{\tau_l\}$. Hence, after OFDM demodulation, the *Inter-Block Interference* is completely removed and we can focus on a per-symbol model in the frequency domain [25]. Also, for simplicity, we neglect the effect of pulse-shaping in the OFDM signaling and assume a frequency-flat pulse response. Applying the Fourier transform to the matrix-valued channel impulse response (2), the frequency-domain channel matrix at symbol interval s is given by

$$\check{\mathsf{H}}_s(f) = \sum_{l=1}^{L} \rho_{s,l} \mathbf{b}(\phi_l) \mathbf{a}(\theta_l)^{\mathsf{H}} e^{-j2\pi f \tau_l}. \tag{6}$$

We denote the OFDM subcarriers as $\{f_\omega = \frac{\omega}{t_0} : \omega \in [F]\}$. The channel matrix at subcarrier ω is given by $\mathbf{H}_s[\omega] := \check{\mathsf{H}}_s(f_\omega)$. Let $\check{x}_{s,i}[\omega]$ denote the frequency-domain data symbol for the i-th stream. Applying OFDM demodulation to the received signal (5), we obtain the corresponding frequency-domain received signal at the j-th receiver RF chain, with transmit beamforming vector $\mathbf{u}_{s,i}$ and receive beamforming vector $\mathbf{v}_{s,j}$ in the form

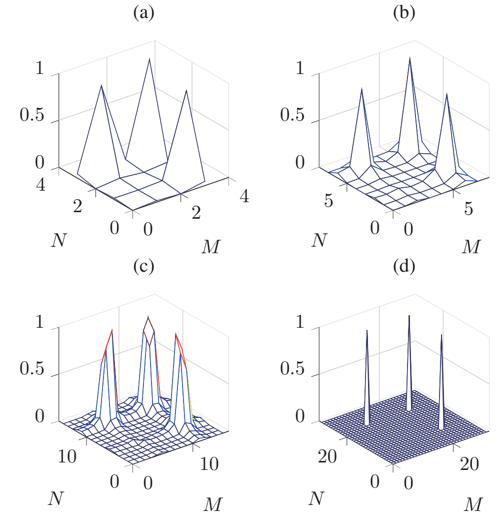


Fig. 2: *Illustration of the sparsity of the channel matrix* $\check{\mathbf{H}}_s[\omega]$ *at an arbitrary subcarrier* ω *consisting of* 3 *off-grid AoA-AoDs with increasing number of antennas for* $M = N = 4$ *(a),* $M = N = 8$ *(b),* $M = N = 16$ *(c),* $M = N = 32$ *(d).*

$$\begin{aligned} \check{y}_{s,i,j}[\omega] &= \frac{1}{\sqrt{n}} \mathbf{v}^{\mathsf{H}}_{s,j} \mathbf{H}_s[\omega] \mathbf{u}_{s,i} \check{x}_{s,i}[\omega] + \check{z}_{s,j}[\omega] \\ &= \frac{1}{\sqrt{n}} \sum_{l=1}^{L} \rho_{s,l} e^{-j2\pi \frac{\omega}{t_0} \tau_l} g^{\mathrm{BS}}_{s,l,i} g^{\mathrm{UE}}_{s,l,j} \check{x}_{s,i}[\omega] + \check{z}_{s,j}[\omega], \end{aligned} \tag{7}$$

where $\check{z}_{s,j}[\omega] \sim \mathcal{CN}(0, \sigma^2)$ denotes the noise at j-th RF chain of UE at subcarrier ω, with variance $\sigma^2 = \Delta f N_0$ which we assume is known for each UE [6].

C. Beam Alignment

During the DL probing slots (see frame structure discussed in Section III), we assume that the signal corresponding to different transmitted streams $x_{s,i}(t)$ are orthogonal, i.e.,

$$\langle x_{s,i}, x_{s,i'} \rangle := \int_{st_0}^{(s+1)t_0} x_{s,i}(t)^* x_{s,i'}(t) dt = E_i \delta_{i,i'}, \tag{8}$$

where E_i is the energy per symbol for the i-th data stream and $\delta_{i,i'}$ is the Kronecker delta symbol (equal to 1 for $i = i'$ and 0 otherwise). For example, this can be obtained in the frequency domain by using OFDM and mapping the different streams onto sets of non-overlapping subcarriers. Letting $r_{s,i,j}(t) := \sum_{l=1}^{L} \rho_{s,l} g^{\mathrm{BS}}_{s,l,i} g^{\mathrm{UE}}_{s,l,j} x_{s,i}(t - \tau_l)$ denote the signal contribution relative to the i-th transmitted data stream of the BS received at the output of the j-th RF chain of the UE (see (5)), defining B_i and $P_i = E_i/t_0$ to be the bandwidth and the average power of $x_{s,i}(t)$, respectively, and recalling that $\gamma_l = \mathbb{E}[|\rho_{s,l}|^2]$, the SNR *after beamforming* (ABF) for the i-th data stream received at the j-th RF chain at the UE is given by

$$\mathsf{SNR_{ABF}}^{i,j} := \frac{\frac{1}{t_0}\mathbb{E}\left[\int_{st_0}^{(s+1)t_0} |r_{s,i,j}(t)|^2 dt\right]}{nN_0B_i} = \frac{P_i \sum_{l=1}^{L} \gamma_l |g^{\mathrm{UE}}_{s,l,j}|^2 |g^{\mathrm{BS}}_{s,l,i}|^2}{nN_0B_i}. \tag{9}$$

We define the total transmit power of the BS as $P_{\mathtt{tot}} = \sum_{i=1}^{m} P_i$. In particular, for equal power allocation ($P_i = P_{\mathtt{tot}}/m$) over the streams, we have

$$\mathsf{SNR_{ABF}}^{i,j} = \frac{P_{\mathtt{tot}} \sum_{l=1}^{L} \gamma_l |g^{\mathrm{UE}}_{s,l,j}|^2 |g^{\mathrm{BS}}_{s,l,i}|^2}{mnN_0B_i}, \tag{10}$$

For later use, we also define the SNR before beamforming (BBF) by

$$\mathsf{SNR_{BBF}} := \frac{P_{\mathtt{tot}} \sum_{l=1}^{L} \gamma_l}{N_0 B}. \tag{11}$$

This is the SNR obtained when a single data stream ($m = 1$) is transmitted through a single BS antenna and is received in a single UE antenna (isotropic transmission) over a single RF chain ($n = 1$) with full-band spreading.

A challenge in mm-Wave communication is that the SNR before beamforming $\mathsf{SNR_{BBF}}$ in (11) is typically very low. This cannot be increased by simply boosting the transmit power $P_{\mathtt{tot}}$ because of hardware and regulation limitations, also because, in general, we would like to design energy-efficient systems. Another option consists of communicating over a small bandwidth $B' < B$. However, it is well-known that this strategy is suboptimal.[4] In fact, assuming a Gaussian channel with SNR equal to $\mathsf{SNR_{BBF}}$, Shannon's capacity formula yields that the achievable rate in bit/s when communicating over a bandwidth B' is given by $R = B' \log(1 + (B/B')\mathsf{SNR_{BBF}})$, which is increasing for $0 < B' \leq B$. Hence, by using a bandwidth smaller than the available channel bandwidth B, the achievable rate is reduced. It follows that the only viable alternative to achieve a reasonable SNR consists in using antenna arrays with a large number of antennas both at the BS and at the UE. The goal of BA is to find *good* beamforming vectors $\mathbf{u}_s$ and $\mathbf{v}_s$ at the BS and the UE, respectively, in order to boost the SNR by a factor $\approx M$ at the BS side and a factor $\approx N$ at the UE side. This is achieved by aligning the beamforming vectors along the AoA-AoD of a strong MPC of the channel.

D. Sparse Beamspace Representation

The AoA-AoDs (θ_l, ϕ_l) in (6) take continuous values. In this paper we adopt the approximate finite-dimensional (discrete) beamspace representation following the well-known approach of [1, 3, 31]. We consider the discrete set of AoA-AoDs

$$\Theta := \{\check{\theta} : (1 + \sin(\check{\theta}))/2 = \frac{k-1}{M}, k \in [M]\}, \tag{12a}$$

$$\Phi := \{\check{\phi} : (1 + \sin(\check{\phi}))/2 = \frac{k'-1}{N}, k' \in [N]\}, \tag{12b}$$

and use the corresponding array responses $\mathcal{A} := \{\mathbf{a}(\check{\theta}) : \check{\theta} \in \Theta\}$ and $\mathcal{B} := \{\mathbf{b}(\check{\phi}) : \check{\phi} \in \Phi\}$ as a discrete dictionary to represent the channel response. For the ULAs considered in this paper, the dictionary $\mathcal{A}$ and $\mathcal{B}$, after suitable normalization, yield orthonormal bases corresponding to the columns of the unitary DFT matrices $\mathbf{F}_M$ and $\mathbf{F}_N$ [5], where we define the D-dimensional DFT matrix with elements

$$[\mathbf{F}_D]_{k,k'} = \frac{1}{\sqrt{D}} e^{j2\pi(k-1)(\frac{k'-1}{D} - \frac{1}{2})}, k, k' \in [D]. \tag{13}$$

Hence, we obtain the beamspace representation of the channel matrix as $\mathbf{H}_s[\omega] = \mathbf{F}_N \check{\mathbf{H}}_s[\omega] \mathbf{F}_M^{\mathsf{H}}$, where

$$\check{\mathbf{H}}_s[\omega] = \sum_{l=1}^{L} \rho_{s,l} e^{-j2\pi \frac{\omega}{t_0} \tau_l} \check{\mathbf{b}}(\phi_l) \check{\mathbf{a}}(\theta_l)^{\mathsf{H}}, \tag{14}$$

where $\check{\mathbf{a}}(\theta_l) := \mathbf{F}_M^{\mathsf{H}} \mathbf{a}(\theta_l)$ and $\check{\mathbf{b}}(\phi_l) := \mathbf{F}_N^{\mathsf{H}} \mathbf{b}(\phi_l)$ denote the coefficient vectors of the array responses $\mathbf{a}(\theta_l)$ and $\mathbf{b}(\phi_l)$ with respect to the DFT bases, respectively. The m'-th entry of $\check{\mathbf{a}}(\theta_l)$ is given by

$$[\,\check{\mathbf{a}}(\theta_l)\,]_{m'} = \frac{1}{\sqrt{M}} \sum_{i=0}^{M-1} e^{-j2\pi i(\frac{m'-1}{M} - \frac{1}{2})} e^{j\pi i \sin(\theta_l)} = \frac{1}{\sqrt{M}} \frac{\sin(\pi\psi_l M)}{\sin(\pi\psi_l)} e^{-j\pi\psi_l(M-1)}, \tag{15}$$

where $\psi_l = \frac{m'-1}{M} - \frac{1}{2}\sin(\theta_l) - \frac{1}{2}$. A similar expression holds for $\check{\mathbf{b}}(\phi_l)$. It is seen from (15) that $|[\,\check{\mathbf{a}}(\theta_l)\,]_{m'}| = \frac{1}{\sqrt{M}} \frac{|\sin(\pi\psi_l M)|}{|\sin(\pi\psi_l)|}$ is a localized kernel around $\theta_l = \sin^{-1}[\frac{2(m'-1)}{M} - 1]$ with a resolution of $\frac{1}{M}$. In general, the AoA-AoDs of the MPCs are not aligned with the discrete grid $\mathcal{G} = \Theta \times \Phi$. However, as the number of antennas M at the BS and N at the UE increases, the DFT basis provide good sparsification of the channel matrix $\check{\mathbf{H}}_s[\omega]$. This is qualitatively illustrated in Fig. 2 for a channel with $L = 3$ discrete off-grid MPCs. It is seen that, as M and N increase, the resulting representation $\check{\mathbf{H}}_s[\omega]$ is more and more sparse.

III. Proposed Beam-Alignment Algorithm

A. High-Level Overview

In the proposed scheme, the channel is periodically probed by the BS while the UEs remain in the listening mode. During the listening mode, each UE gathers

[4] This statement holds only in the case where the channel coefficients change sufficiently slowly in time. More in general, for time-varying wideband fading channels, it has been shown (e.g., see [27–30]) that spreading the transmit power over the entire bandwidth is suboptimal and drives the achievable rate to zero for $B \to \infty$. Intuitively, this is due to the inability of the receiver to estimate the fading channel coefficients, as explained in [27]. The issue of optimal signaling in the presence of time-varying fading is quite intricate and goes beyond the scope of this paper. As a matter of fact, when a large beamforming gain is available at both the BS and the UE side, the effective channel coefficients *after beam alignment* are slowly varying (see [23]) and the SNR after beamforming is large enough, such that the channel can be treated as a standard block-fading AWGN channel with fully known channel coefficients.

measurements of the channel, which is continued until the UE gathers a sufficient number of measurements such that the AoA-AoD of a strong MPC can be reliably identified. After this directional channel estimation is done, the UE tries to announce its identity (user ID) and its beam ID (i.e., the index of the discrete AoD corresponding to the estimated strong MPC) to the BS by sending a control packet. Such control packet is sent over the *Random Access Control CHannel* (RACCH), i.e., a dedicated slot in the frame used for random access, as in virtually all current cellular standard in use today. During the RACCH, the BS stays in listening mode. If the control packet is successfully decoded, the BS responds with a beamformed ACK using the AoD information extracted from the control packet, over a DL data slot. During the data slots, the UE stays in listening mode using its own estimated beam. It follows that the ACK enjoys the full (two-sided) beamforming gain. At this point, BA is achieved and high-SNR communication can take place over the data slots. An overview of the proposed initial acquisition and BA protocol is illustrated in Fig. 3.

Fig. 4 illustrates the proposed frame structure, consisting of three parts: the DL beacon slot, the RACCH slot, and Data Transmission slot. During the DL beacon slots (corresponding to Fig. 3 #1), the BS probing signal is formed by a sequence of pseudo-random beam patterns (referred to as the transmit beamforming codebook), repeated periodically, and priori known to all UEs. Each UE makes measurements of the beacon transmission by applying its own (individual) sequence of receive beam patterns (referred to as the receive beamforming codebook). The number of measurements may differ from user to user, depending on the individual pre-beamforming SNR and on the number of receiver RF chains. We will show in the simulation section that, when a UE is close to the BS, i.e., its received signal power (SNR) is sufficiently high, it can use wider beams and take less measurement rounds in time to speed up the estimation. In contrast, when a UE is far from the BS, i.e., the received signal power (SNR) is very low, it applies narrower receive beams to achieve sufficiently large SNR and takes more rounds in time to collect sufficient number of measurements. In general, a UE might not know the SNR of its channel and may need an adaptive strategy to find a suitable beamwidth for BA. Nevertheless, since the beacon signal is repeated periodically, all the users no matter whether they are weak or strong are able to gather as many measurements as they need.

During the RACCH slots (Fig. 3 #2), the BS stays in listening mode and uses its m RF chains to form m coarse beam patterns (sectors), covering the whole BS angle domain, in order to provide some receiver beamforming gain. Notice that the control packet in the RACCH may fail because of incorrect directional channel estimation (i.e., the UE beam points in a wrong direction), or because of a collision in the RACCH due to another user, or also simply because of statistical fluctuations of the noise, yielding a small but non-zero packet error probability. In all these cases, the BS will not respond with the ACK packet in the data field, and the UE will try again, after gathering more beacon measurements. It should be noticed that packet losses in the RACCH are handled in various ways in all cellular standards in operation today, and surely the RACCH can be dimensioned such that it does not represent a system bottleneck. Furthermore, collisions in the RACCH are not a specific problem of our scheme. In fact, they exist in some form in any scheme for initial acquisition operating in a multiuser environment. Actually, schemes based on interactive beam refinement, requiring multiple control packets and pilot signals to be sent in both UL and DL, are definitely more prone to such problems than the proposed scheme. Since the RACCH is not specific of the proposed scheme, in the following we shall assume that, when the UE has correctly estimated its best MPC, the control packet is received without errors. This allows to compare different systems in a simple and direct manner, and focus on the important and specific aspects of BA.

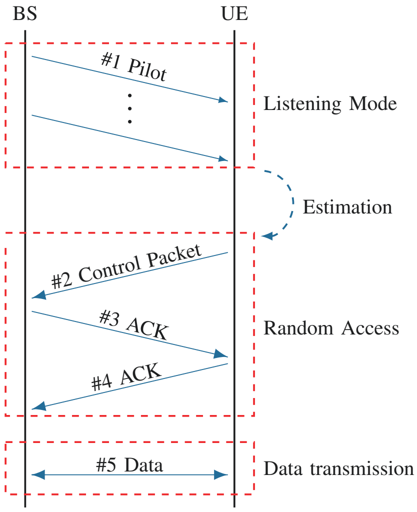


Fig. 3: *Illustration of the proposed Beam Alignment (BA) process between the BS and a generic UE. The procedures (#2∼#5) are independently done at each UE. All the UEs share the same BS beamforming codebook (#1).*

B. BS Channel Probing and UE Sensing

Without loss of generality, we focus on the BA procedure for a generic UE and omit the UE index. Consider the channel matrix $\mathbf{H}_s[\omega]$ between the BS and the UE arrays, as defined in Section II-D, and its beamspace representation $\check{\mathbf{H}}_s[\omega]$ at beacon slot $s \in [T]$ and subcarrier $\omega \in [F]$, where T is the effective period of beam training.

For simplicity of exposition, we assume that the beacon slot contains a single OFDM symbol interval.[5] At each beacon slot, the BS uses its m RF chains to probe the channel along m beamforming vectors $\mathbf{u}_{s,i}$, $i \in [m]$, by

[5] The generalization to multiple OFDM symbols per slot is immediate and slots of $S \geq 1$ OFDM symbols shall be used in the numerical results.

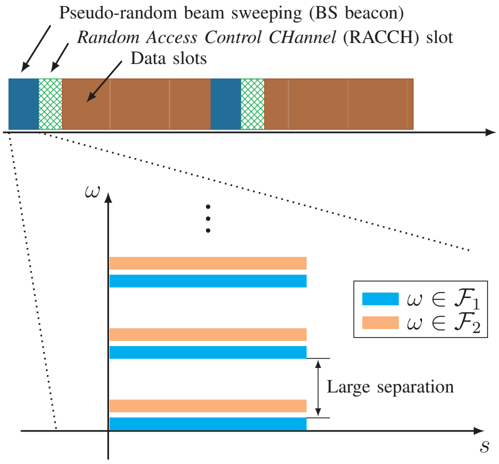

Pseudo-random beam sweeping (BS beacon)
Random Access Control CHannel (RACCH) slot
Data slots
ω
$\vdots$
$\omega \in \mathcal{F}_1$
$\omega \in \mathcal{F}_2$
Large separation
s

Fig. 4: *(Top) Frame structure of the proposed BA scheme. Notice that the beacon, RACCH, and data slots are multiplexed in time according to a TDD scheme. The data slot includes both DL and UL subslots. (Bottom) Different beacon signals are orthogonally multiplexed over disjoint sets of subcarriers $\omega \in \mathcal{F}_i$, $i \in [m]$. In the figure's example we have two orthogonal beacon signals on the "blue" and on the "orange" combs of subcarriers.*

transmitting an OFDM symbol $x_{s,i}(t)$ along each $\mathbf{u}_{s,i}$. We design the beacon OFDM symbols $x_{s,i}(t)$ such that they are mutually orthogonal in the frequency domain and can be separated at the UE side. Thanks to orthogonality, each beacon pilot stream provides the UE with different measurements from the underlying channel. In particular, for each $i \in [m]$ we define a subset $\mathcal{F}_i \subset [F]$ of size $|\mathcal{F}_i| \leq F$ such that $\mathcal{F}_i \cap \mathcal{F}_{i'} = \emptyset$ for $i \neq i'$ (see Fig. 4). We choose each subset $\mathcal{F}_i$ to form a "comb" of subcarriers of equal size $|\mathcal{F}_i| = F'$, with sufficient subcarrier separation such that the corresponding channel matrices $\{\mathbf{H}_s[\omega] : \omega \in \mathcal{F}_i\}$ are mutually uncorrelated.

A main ingredient of our proposed BA scheme is the pseudo-random beamforming codebook transmitted by the BS during the beacon slots, defined as the collection of sets $\mathcal{C}_{\text{BS}} := \{\mathcal{U}_{s,i} : s \in [T], i \in [m]\}$, where $\mathcal{U}_{s,i}$ is the angle-domain support (i.e., the subset of quantized angles in the virtual beamspace representation) defining the directions to which the transmit beam patterns $\mathbf{u}_{s,i}$ sends the signal power. We assume $|\mathcal{U}_{s,i}| = \kappa_u \leq M$ for all (s,i). The beamforming vectors are given by $\mathbf{u}_{s,i} = \mathbf{F}_M \check{\mathbf{u}}_{s,i}$, where $\check{\mathbf{u}}_{s,i} = \frac{\mathbf{1}_{\mathcal{U}_{s,i}}}{\sqrt{\kappa_u}}$, and where $\mathbf{1}_{\mathcal{U}_{s,i}}$ denotes a vector with 1 at components in the support set $\mathcal{U}_{s,i}$ and 0 elsewhere. An example of such patterns with the corresponding vector $\check{\mathbf{u}}_{s,i}$ is shown in Fig. 5 (a). The pseudo-random nature of the codebook is due to the fact that the sequences of angular support sets $\{\mathcal{U}_{s,i} : i \in [m], s \in [T]\}$ are generated in a pseudo-random manner.

The second ingredient of our proposed BA algorithm is a local receive codebook at each UE, through which the UE makes measurements in order to estimate the AoA-AoD information of its strong MPCs. Each UE can customize (locally) its own receive beamforming codebook defined by the collection of sets $\mathcal{C}_{\text{UE}} := \{\mathcal{V}_{s,j} : s \in [T], j \in [n]\}$, where $\mathcal{V}_{s,j}$ is the angle-domain support defining the directions from which the receiver beam patterns $\mathbf{v}_{s,i}$ collect signal power. We assume $|\mathcal{V}_{s,j}| = \kappa_v \leq N$ for all (s,j). The beamforming vectors are given by $\mathbf{v}_{s,j} = \mathbf{F}_N \check{\mathbf{v}}_{s,j}$, where $\check{\mathbf{v}}_{s,j} = \frac{\mathbf{1}_{\mathcal{V}_{s,j}}}{\sqrt{\kappa_v}}$. Similar to the parameter κ_u at the transmitter side, the parameter κ_v controls the spread of the sensing window at the UE side. This is illustrated again in Fig. 5 (a).

During the s-th beacon slot, the UE applies the receive beamforming vector $\mathbf{v}_{s,j}$ to its j-th RF chain, obtaining the frequency-domain received signal (after OFDM demodulation) given by (7) for $i \in [m]$ and $\omega \in \mathcal{F}_i$. Note that the m probing signals $x_{s,i}(t)$ are orthogonal in the frequency domain and therefore can be perfectly separated at the receiver. It is convenient to write (7) directly in terms of the beamspace representation as

$$\check{y}_{s,i,j}[\omega] = \frac{1}{\sqrt{n}} \check{\mathbf{v}}_{s,j}^{\mathsf{H}} \check{\mathbf{H}}_s[\omega] \check{\mathbf{u}}_{s,i} \check{x}_{s,i}[\omega] + \check{z}_{s,j}[\omega]. \tag{16}$$

The BS total transmit power $P_{\texttt{tot}}$ is allocated equally to all the probing streams $i \in [m]$, all the subcarriers in $\omega \in \mathcal{F}_i$, and all the κ_u beamspace directions. Hence, the symbols $\{\check{x}_{s,i}[\omega] : \omega \in \mathcal{F}_i\}$ have uniform power distribution with $\mathbb{E}[|\check{x}_{s,i}[\omega]|^2] = \frac{P_{\texttt{tot}}}{mF'} := P_{\texttt{dim}}$ (power per transmit signal dimension). In fact, without loss of generality, we choose the frequency-domain probing symbols to be constant and given by $\check{x}_{s,i}[\omega] = \sqrt{P_{\texttt{dim}}}$.

Considering the beamforming patterns defined by $\mathcal{C}_{\text{BS}}$ and $\mathcal{C}_{\text{UE}}$, it is not difficult to see that

$$|g_{s,l,i}^{\text{BS}}|^2 = |\mathbf{a}(\theta_l)^{\mathsf{H}} \mathbf{u}_{s,i}|^2 = |\mathbf{a}(\theta_l)^{\mathsf{H}} \mathbf{F}_M \check{\mathbf{u}}_{s,i}|^2 \leq \frac{M}{\kappa_u}, \tag{17a}$$

$$|g_{s,l,j}^{\text{UE}}|^2 = |\mathbf{v}_{s,j}^{\mathsf{H}} \mathbf{b}(\phi_l)|^2 = |\mathbf{b}(\phi_l)^{\mathsf{H}} \mathbf{F}_N \check{\mathbf{v}}_{s,j}|^2 \leq \frac{N}{\kappa_v}. \tag{17b}$$

Applying the upper bounds (17a) and (17b) in (10), we obtain the maximum possible SNR for channel estimation in the per-subcarrier observation (16), given by

$$\begin{aligned} \mathsf{SNR}_{\mathsf{ABF}}^{\mathsf{CE}} &:= \frac{P_{\texttt{dim}}}{n} \frac{MN \sum_{l=1}^{L} \gamma_l}{\kappa_u \kappa_v \sigma^2} \\ &= \frac{MN}{\kappa_u \kappa_v mn} \times \frac{B}{F' \Delta f} \times \mathsf{SNR}_{\mathsf{BBF}}. \end{aligned} \tag{18}$$

where F' denotes the adopted number of subcarriers and Δf denotes the subcarrier bandwidth.

By setting $\kappa_u = \kappa_v = 1$ in (18), we obtain the beamforming gain after BA, namely, aligning the beams along the strongest scatterer. Moreover, (18) puts in evidence the role of the different factors: the first term expresses the *power concentration in the spatial domain*, i.e., the ratio the maximal available beamforming gain MN, divided by the total signal dimensions in the spatial multiplexing domain $\kappa_u \kappa_v mn$ over which the signal is spread; the second term corresponds to the *power concentration in the frequency* *domain*; the third term is the SNR before beamforming, defined in (11).

The frequency spreading factor F' and angle spreading factors κ_u, κ_v can be optimized depending on the specific cell topology (e.g., on the size of the cell, which in turn determines the worst-case SNR before beamforming). Clearly, by making κ_u (resp., κ_v) larger, each beam pattern probes (resp., sense) simultaneously more directions, but the total power is spread over all such directions. In contrast, by making κ_u (resp., κ_v) smaller, the beam pattern explores less directions but obtains better power concentration in the angle domain. It is also important to notice the effect of F': as we shall see in Section III-C, the AoA-AoD estimator builds some sample-mean statistics by averaging over a sufficiently large number of uncorrelated channel fading realization over the frequency domain. Hence, larger F' provide better averaging at the cost of spreading the total power over more subcarriers.[6]

Remark *1:* Angular probing schemes via random, pseudo-random, or even adaptive codebooks can also be found in [11, 12, 15, 21]. Our proposed codebook in this paper can be seen as an improved version of those schemes where the width of the beam, i.e., κ_v can be individually selected by each UE to achieve an optimal tradeoff between angular exploration and the SNR obtained in each measurement. $\Diamond$

C. Channel Estimation at the UE Side

The strong MPCs of the channel correspond to the components (k, k') in the matrix $\check{\mathbf{H}}_s[\omega]$ with large second moment. An immediate consequence of the channel model definition and the standard assumption of uncorrelated MPCs is that the element second moments $\mathbb{E}[|(\check{\mathbf{H}}_s[\omega])_{k,k'}|^2]]$ are invariant both with respect to s (time) and with respect to ω (frequency) [32]. If we had direct access to measurements of the elements of $\check{\mathbf{H}}_s[\omega]$, a naive approach would build estimators for the second moments (sample covariance), and try to identify the largest element. However, this would require a number of RF chains equal to the number of antenna elements. In contrast, we have only access to the projections $\check{\mathbf{v}}_{s,j}^{\mathsf{H}} \check{\mathbf{H}}_s[\omega] \check{\mathbf{u}}_{s,i}$ from the observation in (16).

Using $\check{x}_{s,i}[\omega] = \sqrt{P_{\mathtt{dim}}}$ in (16), we can write the received beacon symbol observation at the UE as

$$\begin{aligned}\check{y}_{s,i,j}[\omega] &= \sqrt{\frac{P_{\mathtt{dim}}}{n}} \check{\mathbf{v}}_{s,j}^{\mathsf{H}} \check{\mathbf{H}}_s[\omega] \check{\mathbf{u}}_{s,i} + \check{z}_{s,j}[\omega] \\ &= \sqrt{\frac{P_{\mathtt{dim}}}{n}} (\check{\mathbf{u}}_{s,i}^{\mathsf{T}} \otimes \check{\mathbf{v}}_{s,j}^{\mathsf{H}}) \check{\mathbb{h}}_s[\omega] + \check{z}_{s,j}[\omega] \\ &= \sqrt{\frac{P_{\mathtt{dim}}}{n}}\, \mathbf{g}_{s,i,j}^{\mathsf{H}} \check{\mathbb{h}}_s[\omega] + \check{z}_{s,j}[\omega], \end{aligned} \tag{19}$$

where $\check{\mathbb{h}}_s[\omega] = \mathrm{vec}(\check{\mathbf{H}}_s[\omega])$ denotes the vectorized beamspace representation of the channel matrix at subcarrier $\omega \in \mathcal{F}_i$, where we used the well-known identity $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathsf{T}} \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$, and where we define the combined probing and sensing beamforming pattern as $\mathbf{g}_{s,i,j} = \check{\mathbf{u}}_{s,i}^* \otimes \check{\mathbf{v}}_{s,j} \in \mathbb{C}^{MN}$, which is common across all the subcarriers $\omega \in \mathcal{F}_i$ but differs for different pairs of BS and UE RF chains (i, j).

In practice, each beacon slot is formed by a block of $S \geq 1$ OFDM symbols. With a slight abuse of notation, we index the symbols belonging to the $(s+1)$-th slot as $sS+s'$, for $s' \in [S]$. In order to estimate the average received power at the UE j-th RF chain output due to the signal transmitted by the BS i-th RF chain in the s-th beacon slot, we form the averaged *quadratic measurement*

$$\begin{aligned}\check{q}_{s,i,j} &= \frac{1}{SF'} \sum_{s' \in [S]} \sum_{\omega \in \mathcal{F}_i} |\check{y}_{sS+s',i,j}[\omega]|^2 \\ &= \frac{P_{\mathtt{dim}}}{n} \mathbf{g}_{s,i,j}^{\mathsf{H}} \left(\frac{1}{SF'} \sum_{s' \in [S]} \sum_{\omega \in \mathcal{F}_i} \check{\mathbb{h}}_{sS+s'}[\omega] \check{\mathbb{h}}_{sS+s'}[\omega]^{\mathsf{H}} \right) \mathbf{g}_{s,i,j} \\ &\quad + \frac{1}{SF'} \sum_{s' \in [S]} \sum_{\omega \in \mathcal{F}_i} |\check{z}_{sS+s',j}[\omega]|^2 \\ &\quad + \frac{1}{SF'} \sum_{s' \in [S]} \sum_{\omega \in \mathcal{F}_i} \xi_{sS+s',j}[\omega], \end{aligned} \tag{20}$$

where the first and the second terms correspond to the signal contribution and to the noise contribution, and where

$$\xi_{sS+s',j}[\omega] = 2\sqrt{\frac{P_{\mathtt{dim}}}{n}} \mathrm{Re}\left\{ \mathbf{g}_{s,i,j}^{\mathsf{H}} \check{\mathbb{h}}_{sS+s'}[\omega] \check{z}_{sS+s',j}[\omega]^{\mathsf{H}} \right\} \tag{21}$$

denotes the signal-noise cross term. Note that since the AWGN noise $(\check{z}_{sS+s',j}[\omega])$ and the Gaussian channel coefficients $(\check{\mathbb{h}}_{sS+s'}[\omega])$ are independent, the cross term has a zero mean. Thus, when the number of dimensions $S \times F'$ (over which the instantaneous power $\check{q}_{s,i,j}$ is averaged) is large, it contributes negligibly to (20) and can be treated as a residual term in our formulation. Moreover, the empirical covariance matrix of the channel vector converges as

$$\begin{aligned} &\frac{1}{SF'} \sum_{s' \in [S]} \sum_{\omega \in \mathcal{F}_i} \check{\mathbb{h}}_{sS+s'}[\omega] \check{\mathbb{h}}_{sS+s'}[\omega]^{\mathsf{H}} \\ &\quad \to \mathbb{E}[\mathbb{h}_s[\omega] \mathbb{h}_s[\omega]^{\mathsf{H}}] =: \mathbf{\Sigma}_{\mathbb{h}}. \end{aligned} \tag{22}$$

Similarly, the noise term converges to

$$\frac{1}{SF'} \sum_{s' \in [S]} \sum_{\omega \in \mathcal{F}_i} |\check{z}_{sS+s',j}[\omega]|^2 \to \sigma^2. \tag{23}$$

Hence, the received power in (20) gives an approximate 1-dimensional noisy projection of the covariance matrix $\mathbf{\Sigma}_{\mathbb{h}}$ with respect to the combined probing and sensing vector $\mathbf{g}_{s,i,j}$. We include the signal-noise cross term in (21) and the difference between the empirical and statistical averages in (22) and (23) as a residual error. This results in

$$\check{q}_{s,i,j} = \frac{P_{\mathtt{dim}}}{n} \mathbf{g}_{s,i,j}^{\mathsf{H}} \mathbf{\Sigma}_{\mathbb{h}} \mathbf{g}_{s,i,j} + \sigma^2 + \check{w}_{s,i,j}, \tag{24}$$

[6] This tradeoff in the choice of the spreading parameters F' and κ_u, κ_v can be seen as an instance of the well-known *exploration-exploitation* tradeoff in statistics.

where $\breve{w}_{s,i,j}$ is the measurement error term. When all the AoA-AoDs lie on a discrete grid, $\check{\mathbb{h}}_s[\omega]$ is a sparse vector with i.i.d. components with only a few nonzero coefficients corresponding to the scatterers. In practice, for large M and N, $\check{\mathbb{h}}_s[\omega]$ is almost sparse with small clusters of non-zero coefficients concentrated around the AoA-AoD pairs of the strong MPCs (as illustrated in Fig. 2). Correspondingly, also $\boldsymbol{\Sigma}_{\mathbb{h}}$ is a very sparse matrix, with strong components localized on the main diagonal.

Next, we put (24) in the form suitable for the proposed AoA-AoD estimation algorithm. With reference to Fig. 5 (a), recall the beam probing and sensing vectors $\check{\mathbf{u}}_{s,i} = \frac{\mathbf{1}_{\mathcal{U}_{s,i}}}{\sqrt{\kappa_u}}$ and $\check{\mathbf{v}}_{s,j} = \frac{\mathbf{1}_{\mathcal{V}_{s,j}}}{\sqrt{\kappa_v}}$. With reference to Fig. 5 (b), let $\check{\boldsymbol{\Gamma}}$ denote the $N \times M$ matrix with elements $\frac{P_{\mathtt{dim}}}{n\kappa_u\kappa_v}\mathbb{E}[|[\check{\mathbf{H}}_s[\omega]]_{k,k'}|^2]$. Finally, define the $NM \times 1$ binary vectors $\mathbf{b}_{s,i,j} := \mathbf{1}_{\mathcal{U}_{s,i}} \otimes \mathbf{1}_{\mathcal{V}_{s,j}}$, with components equal to 0 or 1, where the 1's correspond to the positions in the set $\mathcal{V}_{s,j} \times \mathcal{U}_{s,i}$ of the quantized AoA-AoDs pairs probed/sensed by the beamforming vectors $\check{\mathbf{v}}_{s,j}$ and $\check{\mathbf{u}}_{s,i}$, respectively. With these definitions, after a little algebra, we can rewrite (24) as

$$\check{q}_{s,i,j} = \mathbf{b}_{s,i,j}^{\mathsf{T}} \mathrm{vec}(\check{\boldsymbol{\Gamma}}) + \sigma^2 + \breve{w}_{s,i,j}. \tag{25}$$

Since the BS transmits along m RF chain in each beacon slot and the UE has n RF chains to sense the channel, the UE obtains mn equations for the unknown vector $\mathrm{vec}(\check{\boldsymbol{\Gamma}})$ as in (25). Over T beacon slots the UE obtains mnT equations, which can be written in the form

$$\check{\mathbf{q}} = \mathbf{B} \cdot \mathrm{vec}(\check{\boldsymbol{\Gamma}}) + \sigma^2 \mathbf{1} + \check{\mathbf{w}}, \tag{26}$$

where the vector $\check{\mathbf{q}} = [\check{q}_{1,1,1}, \dots \check{q}_{1,m,n}, \dots, \check{q}_{T,m,n}]^{\mathsf{T}}$ consists of all the mnT measurements calculated as in (20), the $mnT \times MN$ matrix $\mathbf{B} = [\mathbf{b}_{1,1,1}, \dots, \mathbf{b}_{1,m,n}, \dots, \mathbf{b}_{T,m,n}]^{\mathsf{T}}$ is uniquely defined by the beamforming codebooks $\mathcal{C}_{\mathrm{BS}}$ and $\mathcal{C}_{\mathrm{UE}}$, and $\check{\mathbf{w}} \in \mathbb{R}^{mnT}$ is the residual noise and error in the measurements. At this point, some remarks are in order.

Remark *2:* An implicit assumption made here is that each UE is frame-synchronous with the BS, i.e., it knows, at each beacon slot s the subsets $\{\mathcal{U}_{s,i} : i = 1, \dots, m\}$ of beam directions of the BS. It is clear that a lack of frame synchronization between a UE and the BS would lead to a wrong construction of the measurement matrix $\mathbf{B}$ in (26). Notice however that, since the beacon patterns are repeated periodically with some period of T frames, this requires only to be aware of the start epoch of the period. This assumption is explicitly or implicitly made in virtually all works dealing with initial beam acquisition (aka, BA problem) [8–15], as reviewed in Section I. Therefore, this is not a particularly restrictive assumption specific to our approach. As in most works, we assume that such coarse frame information can be gathered from some external source. In practice, this may be either an overlay cell operating at some standard cellular frequency (e.g., typically in the range of sub-6 GHz) or, for a standalone small-cell mm-Wave system, by letting the cells be frame synchronous. In this way, the UE needs to acquire the frame period only once when it joins at first the system, at the cost of a small additional overhead.[7] Then, the frame synchronization is maintained while the UE roams from cell to cell. ◇

Remark *3:* As already remarked, there exist a fairly large number of existing/concurrent works that make use of pseudo-random beam patterns in order to gather linear (compressed) measurements of the channel matrix, with the goal of estimating the channel matrix coefficients. This is typically obtained by using some CS technique, leveraging the fact that the propagation channels are sparse in the angle/delay domain [14–16, 19–21]. It is important to note that our scheme differs from all these works in one key fact. Namely, our scheme gathers *quadratic* compressed measurements (see (24)) and not linear. While in this way we loose the ability of estimating the complex channel coefficients, we can estimate the channel second order statistics in the beamspace domain, and identify the strong MPCs in terms of the corresponding average received signal power. This information is much more stable and robust to variations in the channel time-dynamics than the channel coefficients themselves. In fact, it is easy to see that when the channel coefficients vary significantly over the measurement time, the system of linear equations in CS schemes tends to become unidentifiable. For example, in the limiting case of independent channel coefficients across the measurement slots, each new measurement depends on a new set of coefficients, such that the number of measurements is always less than the number of nonzero channel coefficients. In these conditions, fundamental information theoretic bounds show that stable reconstruction is not possible for any CS algorithm, no matter how sparse the channel is [33]. In contrast, focusing on the channel second-order statistics sampled by the quadratic measurements in (24)-(26), we can *always* gather a number of measurements mnT larger than the number of non-zero coefficients in $\mathrm{vec}(\check{\boldsymbol{\Gamma}})$, for sufficiently large T, such that the strong components in $\mathrm{vec}(\check{\boldsymbol{\Gamma}})$ can be identified with high probability. ◇

As a general comment, we notice here that it is more sensible and more robust to first estimate the beamforming direction (e.g., via the proposed scheme) and then estimate the beamformed channel in the regime of high SNR, rather than trying to first estimate the channel coefficients (in low SNR and at the mercy of the possibly large time-variations) and then computing the beamforming coefficients.

D. Non-Negative Least-Squares

In order to identify the AoA-AoD directions of the strong scatterers, we estimate the MN dimensional vector $\mathrm{vec}(\check{\boldsymbol{\Gamma}})$ from the mnT-dimensional observation given in (26). Because of the presence of the measurement noise $\check{\mathbf{w}}$, a standard approach consists of solving the Least-Squares

[7] For example, this can be obtained by trying sequentially the different cyclic shifts of the beacon sequence until successful alignment.

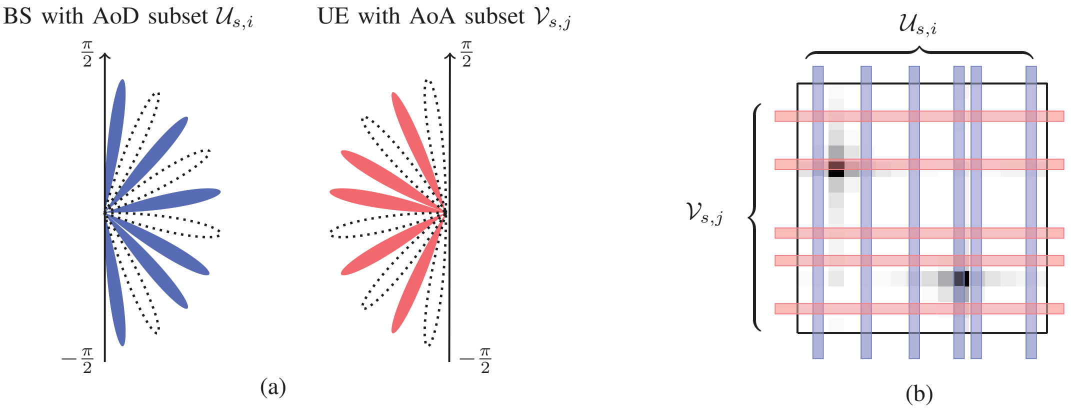


Fig. 5: *(a) Illustration of the subset of AoA-AoDs at time slot s probed by the i-th beacon stream transmitted by the BS and received by the j-th RF chain of the UE, for $M = N = 10$. The AoD subset is given by $\mathcal{U}_{s,i} = \{1,3,4,6,8,10\}$ (numbered counterclockwise) with beamforming vector $\breve{\mathbf{u}}_{s,i} = \frac{1}{\sqrt{6}}[1,0,1,0,1,0,1,1,0,1]^{\mathsf{T}}$. The AoA subset is given by $\mathcal{V}_{s,j} = \{2,4,5,7,9\}$ (numbered counterclockwise) with receive beamforming vector $\breve{\mathbf{v}}_{s,j} = \frac{1}{\sqrt{5}}[0,1,0,1,1,0,1,0,1,0]^{\mathsf{T}}$. (b) The channel gain matrix $\breve{\mathbf{\Gamma}}$ (with two strong MPCs indicated by the dark spots) measuring along $\mathcal{V}_{s,j} \times \mathcal{U}_{s,i}$.*

(LS) problem $\min_{\breve{\mathbf{\Gamma}}} \|\mathbf{B} \cdot \mathrm{vec}(\breve{\mathbf{\Gamma}}) + \sigma^2 \mathbf{1} - \breve{\mathbf{q}}\|^2$. However, in general MN is significantly larger than mnT, such that the system of equations is heavily underdetermined and the LS solution yields meaningless results. The key observation here is that $\breve{\mathbf{\Gamma}}$ is *sparse* (by assumption) and *non-negative* (by construction). Recent results in CS show that when the underlying parameter $\breve{\mathbf{\Gamma}}$ is non-negative, the simple non-negative constrained LS given by

$$\breve{\mathbf{\Gamma}}^* = \underset{\mathrm{vec}(\breve{\mathbf{\Gamma}}) \in \mathbb{R}_+^{MN}}{\arg\min} \|\mathbf{B} \cdot \mathrm{vec}(\breve{\mathbf{\Gamma}}) + \sigma^2 \mathbf{1} - \breve{\mathbf{q}}\|^2, \tag{27}$$

is still enough to impose sparsity of the solution $\breve{\mathbf{\Gamma}}^*$ [34, 35], with no need for an explicit sparsity-promoting regularization term in the objective function as in the classical LASSO algorithm [36]. The (convex) optimization problem (27) is generally referred to as *Non-Negative Least-Squares* (NNLS), and has been well investigated in the literature, starting with Donoho *et al.* in [37]. More recently, in the context of CS it has been shown that the non-negativity constraint alone might suffice to recover a sparse non-negative signal from under-determined linear measurements both in the noiseless case [38–41] and in the noisy case [34, 35]. Moreover, [34] demonstrates that NNLS has a noisy recovery performance comparable to that of LASSO. In [34] it is also shown that NNLS along with an appropriate thresholding provides state-of-the-art performance in terms of support estimation. This property is very relevant in the context of this paper, where the identification of the support of $\breve{\mathbf{\Gamma}}$ corresponds to finding the AoA-AoD directions strongly coupled by MPCs.

As discussed in [34], NNLS implicitly performs ℓ_1-regularization and promotes the sparsity of the resulting solution provided that the measurement matrix satisfies the $\mathcal{M}^+$-criterion [42]. This property is beneficial for our proposed BA scheme because of the natural sparsity of the mm-Wave channel in AoA-AoD domain. Posed in our framework, the measurement matrix $\mathbf{B}$ fulfills the $\mathcal{M}^+$-criterion if there is a vector $\mathbf{g}' \in \mathbb{R}_+^{mnT}$ such that $\mathbf{B}^{\mathsf{T}}\mathbf{g}' > 0$. It is not difficult to see that when $\mathbf{g}' = \mathbf{1}$ is an all-one vector of dimension mnT, the i-th component of $[\mathbf{B}^{\mathsf{T}}\mathbf{g}']_i$ corresponds to the number of measurement patterns that hit the AoA-AoD pair corresponding to $i \in [MN]$. Hence, the necessary condition $\mathbf{B}^{\mathsf{T}}\mathbf{g}' > 0$ can be simply interpreted as the fact that the set of mnT measurement patterns should hit all MN AoA-AoD pairs at least once. Also, as stated in [42], NNLS performs better when the condition number $\frac{\max_{i\in[MN]}[\mathbf{B}^{\mathsf{T}}\mathbf{g}']_i}{\min_{i\in[MN]}[\mathbf{B}^{\mathsf{T}}\mathbf{g}']_i}$ is close to 1, which is met when the measurement patterns (i.e., the rows of $\mathbf{B}$) cover the whole set of AoA-AoDs quite uniformly. This also provides a criterion to design good pseudo-random beamforming codebooks for the BA problem.

In terms of numerical implementations, the NNLS can be posed as an unconstrained LS problem over the positive orthant and can be solved by several efficient techniques such as Gradient Projection, Primal-Dual techniques, etc., with an affordable computational complexity [43], generally significantly less than CS algorithms for problems of the same size and sparsity level. We refer to [44, 45] for the recent progress on the numerical solution of NNLS and a discussion on other related work in the literature.

IV. Performance Evaluation

In this section we evaluate the performance of our proposed algorithm via numerical simulations. To run the NNLS optimization in (27), we use the implementation of NNLS in MATLAB© called `lsqnonneg.m`.

Channel and Signal Model. We assume $f_0 = 70$ GHz carrier frequency and bandwidth of $B = 1$ GHz. The

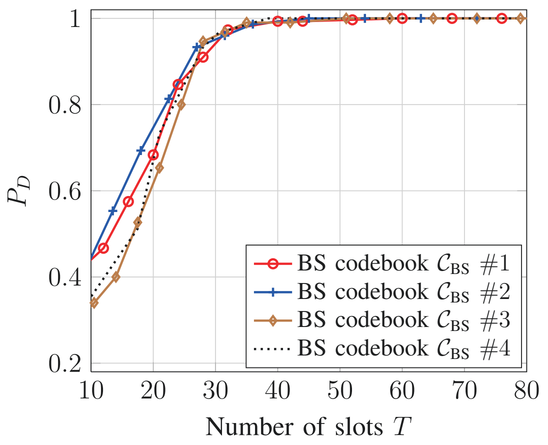


Fig. 6: *Detection probability P_D of the proposed scheme for different pseudo-random codebooks (denoted by $\mathcal{C}_{BS}$), where $M = N = 32$, $F' = 3$, $m = 3$, $n = 2$, $\kappa_u = \kappa_v = 8$, $\mathsf{SNR}_{BBF} = -33$ dB.*

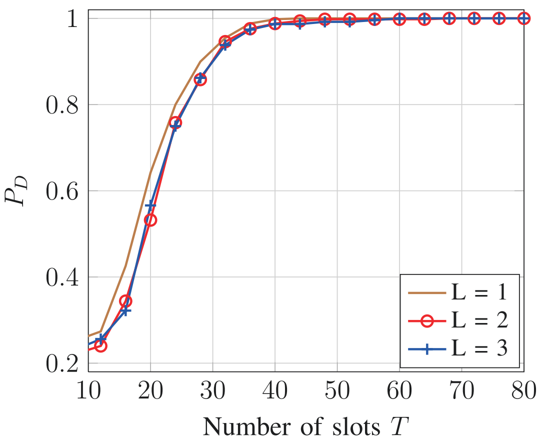


Fig. 7: *Detection probability P_D of the proposed scheme for different number of paths (scatterers) L, where $M = N = 32$, $F' = 3$, $m = 3$, $n = 2$, $\kappa_u = \kappa_v = 16$, $\mathsf{SNR}_{BBF} = -33$ dB ($L = 1$), -32 dB ($L = 2$), -31 dB ($L = 3$).*

OFDM subcarrier spacing is 480 kHz in compliance with recent 3GPP standard specifications [46, 47]. Assuming $\tau_{\rm cp}\Delta f = 0.07$ (i.e., the CP length is 7% of the OFDM duration), we obtain $t_0 = 2.23$ μs and around $F = 2048$ subcarriers (plus some guard band). We fix the frame duration of our scheme (i.e., the repetition interval of the beacon slot) to 1 ms, consists of 448 OFDM symbols (per subcarrier).

A *beacon slot* contains $S = 14$ OFDM symbols, the *random access slot* also contains 14 OFDM symbols, and the remaining 420 symbols are used for data transmission [46]. We assume that the BS has $M = 32$ antennas and $m = 3$ RF chains, and the UE has $N = 32$ antennas and $n = 2$ RF chains. We announce an individual experiment to be successful if the index of the strongest component in $\check{\mathbf{\Gamma}}$ is correctly estimated (i.e., it coincides with the actual strongest MPC AoA-AoD location, up to the discrete angle grid quantization).

Dependence on the Random BS Codebook. We generated 4 different beamforming codebooks at the BS side. Each codebook consists of a randomly generated sequence of patterns, identified by binary vectors of dimension M and Hamming weight κ_u, obtained by independently sampling the set of all possible $\binom{M}{\kappa_u}$ such equal-weight vectors. For simplicity, we consider a very sparse channel model with only one MPC. Fig. 6 illustrates the detection probability for the different pseudo-random codebooks, where the angle spreading factors at the BS and the user sides are set to $\kappa_u = \kappa_v = 8$, respectively. We repeated each experiment 200 times and plot the resulting detection probability versus training period length T. Notice that different codebooks have quite similar performances. This demonstrates the fact that, well-known in several CS contexts, that the scheme is quite insensitive to the specific measurement matrix, as long as it is sufficiently randomized.

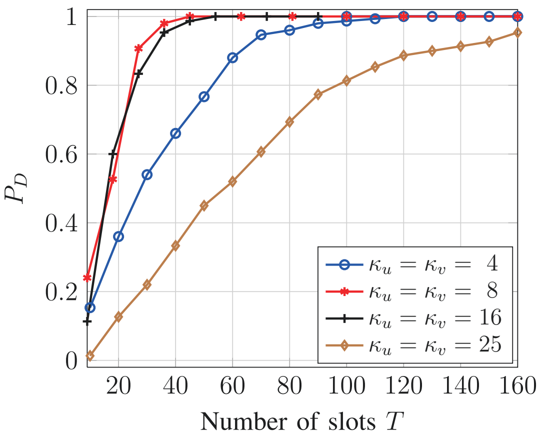


Fig. 8: *Detection probability P_D of the proposed scheme for different angle spreading factors (κ_u, κ_v), where $M = N = 32$, $F' = 3$, $m = 3$, $n = 2$, $\mathsf{SNR}_{BBF} = -33$ dB.*

Performance with Different Number of Paths (Scatterers) L. To illustrate that our proposed scheme works equally well for single-path and multi-path scenarios, we repeat the simulation with different number of MPCs $L = 1, 2, 3$, with different strengths $\gamma_1 > \gamma_2 = \gamma_3$. Fig. 7 shows the performance of the proposed scheme, where we announce an individual experiment to be successful if the strongest path (γ_1) is correctly identified. It is seen that the scheme performs equally well for a single MPC ($L = 1$) and multiple MPCs ($L = 2, 3$), where in both cases, at most $T = 40$ beacon slots are sufficient to ensure a successful BA with high probability.

Dependence on the Angle Spreading Factors κ_u and κ_v. The angle spreading factors κ_u and κ_v impose a tradeoff between the angle coverage of the probing/sensing matrix $\mathbf{B}$ (exploration) and its receive SNR at the user side (exploitation). By making κ_u (resp., κ_v) larger, each beam pattern probes simultaneously more directions, but the total power is spread over all such directions. In contrast, by making κ_u (resp., κ_v) smaller, the beam pattern explores less directions but obtains better power concentration in the angle domain. This is illustrated in Fig. 8. It is seen that increasing the spreading factor from $\kappa_u = \kappa_v = 4$ to $\kappa_u = \kappa_v = 8$ yields a performance improvement. However, a further increase to $\kappa_u = \kappa_v = 16$ slightly degrades the performance, and the degradation is severe for even larger values $\kappa_u = \kappa_v = 25$. As already remarked a few times in this paper, the choice of the spreading factor κ_v at the UE can be tailored to the individual SNR condition (e.g., these may depend on the distance between UE and BS). To pinpoint this, we repeated the simulation to find the best κ_v at the UE side as a function of its channel SNR. This is illustrated in Fig. 9, reporting the best κ_v and the best search time T as function of UE SNR (assuming a threshold of $P_D \geq 0.95$ and a spreading factor of $\kappa_u = 8$ at the BS side). It is seen that, as expected, when a UE enjoys a larger SNR (e.g., it is close to the BS), it should use a larger κ_v in order to better explore the channel thus reducing the search time T. In contrast, when a UE is in low SNR conditions (e.g., it is far from the BS), it should apply a smaller κ_v in order to gather measurements with sufficiently large SNR.

Dependence on the Number of Subcarriers F'. As explained in Section III-B and III-C, using a large number of subcarriers F' in the beacon signals ensures a reliable averaging of the received instantaneous power (see (24)) at the cost of a reduced SNR per subcarrier. As shown in Fig. 10, increasing the number of subcarriers from $F' = 1$ to $F' = 3$ improves the performance, but increasing further to $F' = 10, 30$ degrades the performance considerably.

Dependence on the Probing Dimensions ($\kappa_u\kappa_v mn$). Note that, for a certain pre-beamforming SNR, the output of the proposed BA scheme inherently depends on the probing dimensions, i.e., the product $\kappa_u\kappa_v mn$. This is illustrated by the curves marked as (#1, #1′) and (#2, #2′) in Fig. 11. For various different configurations of the parameters, if both the number of measurements (mn) and the probing dimensions ($\kappa_u\kappa_v mn$) in each slot are the same, a similar performance is achieved. This is useful in terms of system design where more complexity can be pushed towards the BS (e.g., more RF chains at the BS) while keeping the same system-level performance.

System-Level Scalability. We consider a multiuser scenario where K denotes the number of UEs in the system and $K(T)$ denotes the number of UEs that have achieved BA (i.e., that have successfully detected their strong MPC) after T frames. Fig. 12 compares the fraction $\frac{K(T)}{K}$ achieved by interactive bisection scheme [11] and our proposed scheme. In the simulations, we assume that for [11] the users are trained one by one with an ideal cost-free feedback in each round, whereas in our case, all the users are trained independently and simultaneously, i.e., all the users share the same BS pseudo-random probing codebook, while each user use its own sensing codebook which is also randomly generated. As we can see, the training overhead of interactive methods scales proportionally with the number of active users, whereas in our scheme the overhead does not grow with the number of users. Note that in practice, the feedback scheme for each iterative round in [11] costs UL transmissions and may not be ideal since the beamforming gains are very poor at the initial rounds. In contrast, the proposed scheme needs only one UL transmission of the RACCH packet, where the full beamforming gain at the UE side and the sectored beamforming gain at the BS side (as discussed in Section III-A) are available.

Robustness w.r.t. Variations in Channel Statistics. To investigate the sensitivity of the proposed scheme as well as competing CS-based schemes to channel time-variations, we consider a simple Gauss-Markov model for the channel correlation in time given by

$$\rho_{s,l} = \alpha\rho_{s-1,l} + \sqrt{1-|\alpha|^2}\,\nu_{s,l}, s \in \mathbb{Z}_+, \qquad (28)$$

where $\rho_{0,l} \sim \mathcal{CN}(0, \gamma_l)$, where $\nu_{s,l} \sim \mathcal{CN}(0, \gamma_l)$ is an i.i.d. sequence (innovation), and where $|\alpha| \in [0, 1]$ controls the channel correlation in time. This model is widely used as a simple and intuitive way to model correlated fading (see [48]). We assume that the channel is constant over each beacon slot of 14 OFDM symbols, and evolves in time according to (28) from slot to slot. More precisely, $|\alpha| = 1$ yields channel coefficients constant over the whole BA phase, while $|\alpha| = 0$ yields channel coefficient changing in an i.i.d. fashion over the beacon slots. In general, a full range of channel time variations can be obtained by varying $|\alpha|$ between 0 and 1. In Fig. 13, we compare the detection probability of the proposed scheme with that of other CS-based schemes presented in [15, 16]. In [15], the instantaneous channel coefficients are estimated using the *Orthogonal Matching Pursuit* (OMP) technique. In [16] an improvement is proposed where the congruence of the channel AoA/AoD components across a *Selected* comb of subcarriers is exploited by applying a *Simultaneous Orthogonal Matching Pursuit* (SS-OMP) technique. Fig. 13 illustrates the simulation results. It is seen that, the proposed NNLS scheme performs much better over a wide range of channel time-correlations whereas the OMP/SS-OMP schemes are quite fragile in the presence of channel time-variations.

V. Conclusion

In this paper, we proposed an efficient *Beam Alignment* (BA) scheme for mm-Wave multiuser MIMO systems. In the proposed scheme, the AoA/AoD of a strong MPC component is estimated by exploring the AoA-AoD domain through pseudo-random multi-finger beam patterns, and constructing an estimate of the resulting second-order statistics (namely, the average received power for each pseudo-random beam configuration). The resulting under-

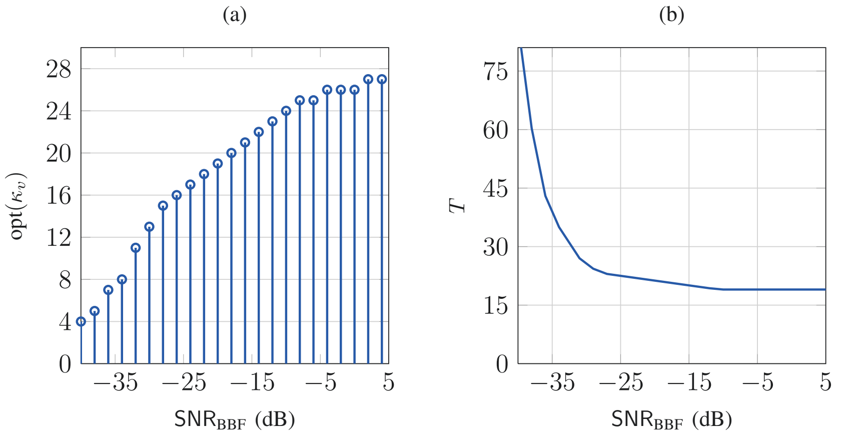


Fig. 9: *(a) The optimal spreading factor opt(κ_v) at the UE in terms of different* SNR$_{BBF}$ *when the BS spreading factor is fixed at* $\kappa_u = 8$. *(b) The average training slots T that ensures a high detection probability* $P_D \geq 0.95$.

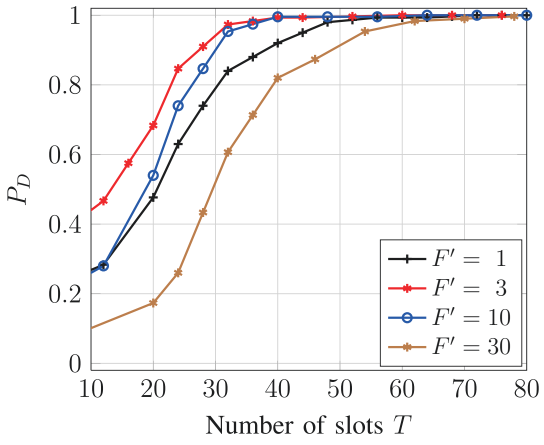


Fig. 10: *Detection probability P_D of the proposed scheme for different number of subcarriers F' deployed for each RF chain at the BS side, where $M = N = 32$, $m = 3$, $n = 2$,* SNR$_{BBF}$ $= -33$ *dB.*

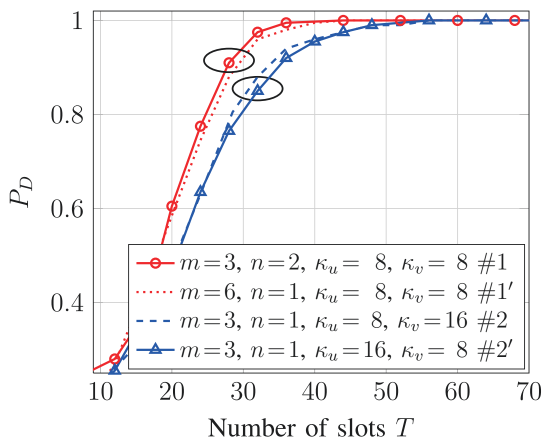


Fig. 11: *Detection probability P_D of the proposed scheme when the product of mn and $mn\kappa_u\kappa_v$, i.e., the number of measurements and the probing dimensions in the spatial multiplexing domain, are constant, and where $M = N = 32$, $F' = 3$,* SNR$_{BBF}$ $= -33$ *dB.*

determined system of equations is efficiently solved using NNLS, yielding naturally a sparse non-negative vector solution whose maximum component identifies the optimal path. In the proposed scheme, the channel is probed by the BS by sending (pseudo-random) beamformed beacon DL signals, and sensed by the UEs by applying (pseudo-random) receive beam patterns. The scheme can train simultaneously a large number of users, since it requires no interactive (multiple rounds) bi-directional transmission of pilots and/or control packets as in bisection methods. Also, the scheme is robust to the channel coefficient time-dynamics since it is based on the estimation of the channel second order statistics (received power) rather than on trying to estimate the complex channel coefficients, as done in other concurrent schemes also based on random beams and compressed sensing. Overall, the proposed scheme provides a very competitive performance both in terms of of scalability (with respect to the number of users) and robustness (with respect to the channel coefficient statistics), than the state-of-art algorithms for initial beam acquisition proposed so far.

Acknowledgments

X.S. is sponsored by the China Scholarship Council (201604910530). G.C. is supported by the Alexander von Humboldt Foundation through a Professorship Grant, and this work was also supported in part by a Collaborative Research Grant of Intel Research.

References

[1] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter

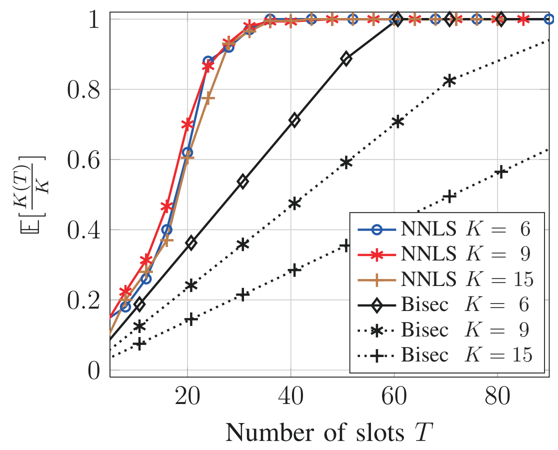


Fig. 12: *Comparison of the performance of our proposed scheme with that of interactive bisection method in [11] in terms of the fraction of users whose channel is estimated until a given time slot T given by $\frac{K(T)}{K}$. We take $M = N = 32$, $F' = 3$, $m = 3$, $n = 2$, $\kappa_u = \kappa_v = 8$, $\mathsf{SNR}_{BBF} = -33$ dB.*

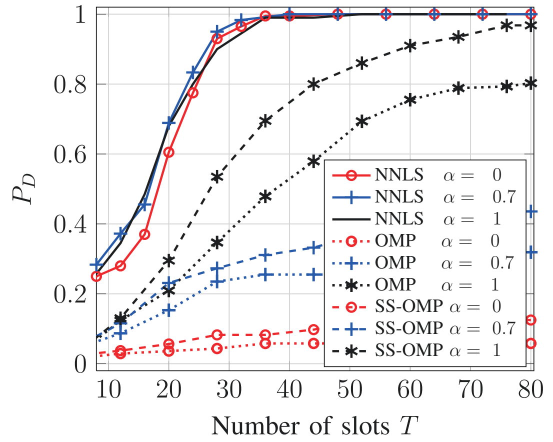


Fig. 13: *Comparison of detection probability P_D between proposed NNLS scheme ($m = 3, F' = 3, \kappa_u = \kappa_v = 8$), the OMP scheme in [15] ($m = 1, F' = 1, \kappa_u = \kappa_v = 16$), and the SS-OMP scheme in [16] ($m = 3, F' = 3, \kappa_u = \kappa_v = 8$) for $M = N = 32$, $n = 2$ and $\mathsf{SNR}_{BBF} = -33$ dB, when the path gains change from i.i.d. ($\alpha = 0$) to constant ($\alpha = 1$) over consecutive beacon slots.*

wave MIMO systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[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!" *Access, IEEE*, vol. 1, pp. 335–349, 2013.

[3] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2563–2579, 2002.

[4] T. Nitsche, C. Cordeiro, A. B. Flores, E. W. Knightly, E. Perahia, and J. C. Widmer, "IEEE 802.11 ad: directional 60 GHz communication for multi-Gigabit-per-second Wi-Fi," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 132–141, 2014.

[5] Z. Chen and C. Yang, "Pilot decontamination in wideband massive MIMO systems by exploiting channel sparsity," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 5087–5100, 2016.

[6] S. Haghighatshoar and G. Caire, "The beam alignment problem in mmwave wireless networks," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 741–745.

[7] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmwave communications," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 1926–1930.

[8] J. Wang, Z. Lan, C.-W. Pyo, T. Baykas, C.-S. Sum, M. A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada *et al.*, "Beam codebook based beamforming protocol for multi-gbps millimeter-wave wpan systems," *Selected Areas in Communications, IEEE Journal on*, vol. 27, no. 8, pp. 1390–1399, 2009.

[9] L. Chen, Y. Yang, X. Chen, and W. Wang, "Multi-stage beamforming codebook for 60ghz wpan," in *2011 6th International ICST Conference on Communications and Networking in China (CHINACOM)*, Aug 2011, pp. 361–365.

[10] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. Thomas, A. Ghosh *et al.*, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *Communications, IEEE Transactions on*, vol. 61, no. 10, pp. 4391–4403, 2013.

[11] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 5, pp. 831–846, 2014.

[12] M. Kokshoorn, H. Chen, P. Wang, Y. Li, and B. Vucetic, "Millimeter wave MIMO channel estimation using overlapped beam patterns and rate adaptation," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 601–616, 2017.

[13] P. Xia, R. W. Heath, and N. Gonzalez-Prelcic, "Robust analog precoding designs for millimeter wave MIMO transceivers with frequency and time division duplexing," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4622–4634, Nov 2016.

[14] D. E. Berraki, S. M. D. Armour, and A. R. Nix, "Application of compressive sensing in sparse spatial channel recovery for beamforming in mmwave outdoor systems," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, 2014, Conference Proceedings, pp. 887–892.

[15] A. Alkhateeb, G. Leusz, and R. W. Heath, "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, Conference Proceedings, pp. 2909–2913.

[16] J. Rodríguez-Fernández, N. González-Prelcic, K. Venugopal, and R. W. Heath Jr, "Frequency-domain compressive channel estimation for frequency-selective hybrid mmwave MIMO systems," *arXiv preprint arXiv:1704.08572*, 2017.

[17] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO-a survey," *arXiv preprint arXiv:1609.05078*, 2016.

[18] 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 Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6664–6678, 2015.

[19] J. Choi, "Beam selection in mm-wave multiuser MIMO systems using compressive sensing," *IEEE Transactions on Communications*, vol. 63, no. 8, pp. 2936–2947, 2015.

[20] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.

[21] K. Venugopal, A. Alkhateeb, N. G. Prelcic, and R. W. Heath, "Channel estimation for hybrid architecture based wideband millimeter wave systems," *IEEE Journal on Selected Areas in Communications*, 2017.

[22] R. J. Weiler, M. Peter, W. Keusgen, and M. Wisotzki, "Measuring the busy urban 60 ghz outdoor access radio channel," in *2014 IEEE International Conference on Ultra-WideBand (ICUWB)*, 2014, Conference Proceedings, pp. 166–170.

[23] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5014–5029, 2017.

[24] P. A. Eliasi, S. Rangan, and T. S. Rappaport, "Low-rank spatial channel estimation for millimeter wave cellular systems," *IEEE*

Transactions on Wireless Communications, vol. 16, no. 5, pp. 2748–2759, 2017.

[25] A. F. Molisch, *Wireless communications*. John Wiley & Sons, 2012, vol. 34.

[26] W. Shen, L. Dai, B. Shim, Z. Wang, and R. W. H. Jr., "Channel feedback based on aod-adaptive subspace codebook in FDD massive MIMO systems," *CoRR*, vol. abs/1704.00658, 2017. [Online]. Available: http://arxiv.org/abs/1704.00658

[27] M. Médard and R. G. Gallager, "Bandwidth scaling for fading multipath channels," *IEEE Transactions on Information Theory*, vol. 48, no. 4, pp. 840–852, 2002.

[28] A. Lozano and D. Porrat, "Non-peaky signals in wideband fading channels: Achievable bit rates and optimal bandwidth," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 246–257, 2012.

[29] F. Gómez-Cuba, J. Du, M. Médard, and E. Erkip, "Unified capacity limit of non-coherent wideband fading channels," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 43–57, 2017.

[30] G. Durisi, U. G. Schuster, H. Bolcskei, and S. Shamai, "Noncoherent capacity of underspread fading channels," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 367–395, 2010.

[31] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.

[32] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[33] Y. Wu and S. Verdú, "Optimal phase transitions in compressed sensing," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6241–6263, 2012.

[34] M. Slawski, M. Hein *et al.*, "Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization," *Electronic Journal of Statistics*, vol. 7, pp. 3004–3056, 2013.

[35] R. Kueng and P. Jung, "Robust nonnegative sparse recovery and the nullspace property of 0/1 measurements," *arXiv preprint arXiv:1603.07997*, 2016.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[37] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern, "Maximum entropy and the nearly black object," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 41–81, 1992.

[38] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of non-negative sparse & redundant representations," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 5145–5148.

[39] D. L. Donoho and J. Tanner, "Counting the faces of randomly-projected hypercubes and orthants, with applications," *Discrete & computational geometry*, vol. 43, no. 3, pp. 522–541, 2010.

[40] M. Wang and A. Tang, "Conditions for a unique non-negative solution to an underdetermined system," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2009, pp. 301–307.

[41] M. Wang, W. Xu, and A. Tang, "A unique "nonnegative" solution to an underdetermined system: From vectors to matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1007–1016, 2011.

[42] R. Kueng and P. Jung, "Robust nonnegative sparse recovery and the nullspace property of 0/1 measurements," *arXiv preprint arXiv:1603.07997*, 2016.

[43] D. P. Bertsekas and A. Scientific, *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

[44] D. Kim, S. Sra, and I. S. Dhillon, "Tackling box-constrained optimization via a new projected quasi-newton approach," *SIAM Journal on Scientific Computing*, vol. 32, no. 6, pp. 3548–3563, 2010.

[45] D. K. Nguyen and T. B. Ho, "Anti-lopsided algorithm for large-scale nonnegative least square problems," *arXiv preprint arXiv:1502.01645*, 2015.

[46] "3GPP TR 38.802 V2.0.0 (2017-03) - Study on New Radio (NR) Access Technology; Physical Layer Aspects (Release 14)," 2017.

[47] A. Ghosh. (2017) 5G mmWave Revolution & New Radio - IEEE 5G. [Online]. Available: {https://5g.ieee.org/images/files/pdf/5GmmWave_}\\ {Webinar_IEEE_Nokia_09_20_2017_final.pdf}

[48] C. C. Tan and N. C. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Transactions on Communications*, vol. 48, no. 12, pp. 2032–2040, 2000.

Xiaoshen Song (S'17) received the B.Sc. degree in Communication Engineering from Northwestern Polytechnical University, Xi'an, China, in 2013, and the M.Sc. degree in Communication and Information Systems from the Institute of Electronics, University of Chinese Academy of Sciences, Beijing, China, in 2016. Her master's thesis focuses on video synthetic aperture radar (VideoSAR) system design and imaging algorithms. She is currently pursuing the Ph.D. degree with the Communications and Information Theory (CommIT) group at Technische Universität Berlin, Berlin, Germany. Her research interests include wireless communication, mmWave massive MIMO, and compressed sensing.

Saeid Haghighatshoar (S'12–M'15) received the B.Sc. degree in Electrical Engineering (Electronics) in 2007 and the M.Sc. degree in Electrical Engineering (Communication Systems) in 2009, both from Sharif University of Technology, Tehran, Iran, and the Ph.D. degree in Computer and Communication Sciences from École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2014. Since 2015, he is a postdoctoral researcher with Communications and Information Theory (CommIT) group at Technische Universität Berlin, Berlin, Germany. His research interests lie in Information Theory, Communication Systems, Wireless Communication, Optimization Theory, and Compressed Sensing.

Giuseppe Caire (S'92 – M'94 – SM'03 – F'05) was born in Torino, Italy, in 1965. He received the B.Sc. in Electrical Engineering from Politecnico di Torino (Italy), in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992 and the Ph.D. from Politecnico di Torino in 1994. He has been a post-doctoral research fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands) in 1994-1995, Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, and he is currently an Alexander von Humboldt Professor with the Electrical Engineering and Computer Science Department of the Technical University of Berlin, Germany.

He served as Associate Editor for the IEEE Transactions on Communications in 1998-2001 and as Associate Editor for the IEEE Transactions on Information Theory in 2001-2003. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society & Information Theory Society Joint Paper Award in 2004 and in 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, and the Vodafone Innovation Prize in 2015. Giuseppe Caire is a Fellow of IEEE since 2005. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications theory, information theory, channel and source coding with particular focus on wireless communications.

4

Initial Beam Alignment for mmWave Single-Carrier Systems

4.1 Introduction

As discussed before, the IEEE 802.11.ad standard specifies two operating modes at 60 GHz bands, i.e, the OFDM mode for high performance applications (e.g., high data rate), and the single carrier (SC) mode for low power and low complexity implementation. On top of the efficient BA scheme for mmWave OFDM systems provided in the last chapter, this chapter focuses on developing a new BA scheme for mmWave SC systems.

4.2 Clarification of each authors' contributions

This chapter is a journal publication, which is a joint work with Saeid Haghighatshoar and Giuseppe Caire. I wrote this journal as the first author. The citation information is in below:

X. Song, S. Haghighatshoar, and G. Caire,"Efficient beam alignment for mmWave single-carrier systems with hybrid MIMO transceivers," IEEE Transactions on Wireless Communications, 2019. DOI: 10.1109/TWC.2019.2892043

All the authors contributed to this paper, but I have implemented all the experiments and simulations. I also wrote the complete first draft (including all sections) of this paper.

Saeid Haghighatshoar provided valuable ideas for the signaling model. He also modified my first draft in terms of its English expressions.

Giuseppe Caire, who is my PhD supervisor, provided valuable discussions in each meeting of this work. He also did a final modification of the overall draft.

4.3 Original journal article

The following article is a reprint of the original journal paper. It is the accepted version of the paper. The copyright information is given in page xii of this thesis as well as in the first page of the reprinted paper.

Efficient Beam Alignment for mmWave Single-Carrier Systems with Hybrid MIMO Transceivers

Xiaoshen Song, *Student Member, IEEE,* Saeid Haghighatshoar, *Member, IEEE,* Giuseppe Caire, *Fellow, IEEE*

***Abstract*—Communication at *millimeter wave* (mmWave) bands is expected to become a key ingredient of next generation (5G) wireless networks. Effective mmWave communications require fast and reliable methods for beamforming at both the *User Equipment* (UE) and the *Base Station* (BS) sides, in order to achieve a sufficiently large *Signal-to-Noise Ratio* (SNR) after beamforming. We refer to the problem of finding a pair of strongly coupled narrow beams at the transmitter and receiver as the *Beam Alignment* (BA) problem. In this paper, we propose an efficient BA scheme for single-carrier mmWave communications. In the proposed scheme, the BS periodically probes the channel in the downlink via a pre-specified pseudo-random beamforming codebook and pseudo-random spreading codes, letting each UE estimate the *Angle-of-Arrival / Angle-of-Departure* (AoA-AoD) pair of the multipath channel for which the energy transfer is maximum. We leverage the sparse nature of mmWave channels in the AoA-AoD domain to formulate the BA problem as the estimation of a sparse non-negative vector. Based on the recently developed *Non-Negative Least Squares* (NNLS) technique, we efficiently find the strongest AoA-AoD pair connecting each UE to the BS. We evaluate the performance of the proposed scheme under a realistic channel model, where the propagation channel consists of a few multipath components each having different delays, AoAs-AoDs, and Doppler shifts. The channel model parameters are consistent with experimental channel measurements. Simulation results indicate that the proposed method is highly robust to fast channel variations caused by the large Doppler spread between the multipath components. Furthermore, we also show that after achieving BA the beamformed channel is essentially frequency-flat, such that single-carrier communication needs no equalization in the time domain.**

***Index Terms*—mmWave, Beam Alignment, Single-Carrier, Compressed Sensing, Non-Negative Least Squares (NNLS).**

I. Introduction

The majority of existing wireless communication systems operate in the sub-6 GHz microwave spectrum, which has now become very crowded. As a result, *millimeter wave* (mmWave) spectrum ranging from 30 to 300 GHz has been considered as an alternative to achieve very high data rates in the next generation wireless systems. At these frequencies, a signal bandwidth of 1GHz with *Signal-to-Noise Ratio* (SNR) between 0 dB and 3 dB yields data rates $\sim$ 1 Gb/s per data stream. A mmWave *Base Station* (BS) supporting multiple data streams through the use of multiuser *Multiple-Input Multiple-Output* (MIMO) can achieve tens of Gb/s of aggregate rate, thus fulfilling the requirements of enhanced Mobile Broad Band (eMBB) in 5G [1, 2].

A main challenge of communication at mmWaves is the short range of isotropic propagation. According to Friis's Law [3], the effective area of an isotropic antenna decreases polynomially with frequency, therefore, the isotropic pathloss at mmWaves is considerably larger compared with sub-6 GHz counterpart. Moreover, signal propagation through scattering elements also suffers from a large attenuation at high frequencies. Fortunately, the small wavelength of mmWave signals enables to pack a large number of antenna elements in a small form factor, such that it is possible to cope with the severe isotropic pathloss by using large antenna arrays both at the BS side and the *User Equipment* (UE) side, providing an overall large beamforming gain. An essential component to obtain such large antenna gains consists of identifying suitable narrow beam combinations, i.e., a pair of *Angle of Departure* (AoD) at the BS and *Angle of Arrival* (AoA) at the UE, yielding a sufficiently large beamforming gain through the scatterers in the channel. [1] The problem of finding an AoA-AoD pair with a large channel gain is referred to as *Initial Beam Training, Acquisition, or Alignment* in the literature (see references in Section I-A). Consistently with our previous work [4], we shall refer to it simply as *Beam Alignment* (BA).

It is important to define the conditions under which the BA operation must be performed. In this work, we focus on MIMO devices with a *Hybrid Digital Analog* (HDA)

The authors are with the Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: xiaoshen.song@campus.tu-berlin.de).

The work of X. Song is sponsored by the China Scholarship Council (201604910530).

The work of G. Caire and S. Haghighatshoar is partially funded by a Professorship Grant of the Alexander von Humboldt Foundation and by the EU H2020 Project SERENA.

[1] We refer to AoD for the BS and AoA for the UE since the proposed scheme consists of downlink probing from the BS to the UEs. Of course, due to the propagation angle reciprocity, the role of AoA and AoD is referred in the uplink.

structure. HDA MIMO is widely proposed especially for mmWave systems, since the size and power consumption of all-digital architectures prevent the integration of many antenna elements on a small space. A HDA transceiver architecture consists of the concatenation of an analog part implementing the beamforming functions, and a digital part implementing the baseband processing [5, 6]. This poses some specific challenges: i) The signal received at the antennas passes through an analog beamforming network with only a limited number of *Radio Frequency* (RF) chains, much smaller than the number of antennas. Hence, the baseband vector of received signal samples at the output of the physical antenna array are not simultaneously available; ii) Due to the large isotropic pathloss, the received signal power is very low before beamforming, i.e., at every antenna port. Therefore, the BA scheme must be able to operate in very low SNR conditions; iii) Because of the large number of antennas at both sides, the size of the channel matrix between each UE and the BS is very large. However, extensive measurements have shown that mmWave channels typically exhibit a small number of multipath components (on average of up to 3 strong components), each corresponding to a scattering cluster with small delay / angle spreading [7, 8]. Considering the discretization of the AoA-AoD domain according to the UE and BS array resolution, a suitable BA scheme requires the identification of a very sparse set of AoA-AoD pairs coupled via strong propagation coefficient in the very high-dimensional matrix of all possible pairs of discrete beam directions [9, 10].

The other fundamental aspect to the BA problem is that this is the first operation that a UE must accomplish in order to communicate with the BS. Hence, while coarse frame and carrier frequency synchronization may be assumed (especially for the non-stand alone system, assisted by some other existing cell operating at lower frequencies), the fine timing and Doppler shift compensation cannot be assumed. It follows that the BA operation must cope with significant timing offsets and Doppler shifts. In addition, in a multipath propagation environment with paths coming from different directions, each path may be affected by a different Doppler shift. In multicarrier (OFDM-based) systems, this may lead to significant inter-carrier interference, which has been typically ignored in most of the current literature.

A. Related Work

The most straightforward BA method is an exhaustive search, where the BS and the UE scan all the AoA-AoD beam pairs until they find a strong one [7]. This is, however, prohibitively time-consuming, especially considering the very large dimension of the channel matrix due to very large number of antennas. Several BA algorithms have been recently proposed in the literature. All these algorithms, in some way, aim at achieving reliable BA while using less overhead than exhaustive search.

In [11], a two-stage pseudo-exhaustive BA scheme was proposed, where in the first stage, the BS isotropically probes the channel, while the UE scans its discrete beam directions (beam sweeping) to find the best AoA. In the second stage, the UE probes the channel along the AoA found in the first stage, while the BS performs beam sweeping to find the best AoD. A main limitation of [11] is that, due to the isotropic BS beamforming in the first stage, the scheme may suffer from a low pre-beamforming SNR [9, 12, 13], which may impair the whole BA performance.

Some mmWave standards such as IEEE 802.11ad [14] proposed to use multi-level hierarchical BA schemes (e.g., see also [15–18]). The underlying idea is to start with sectors of wide beams to do a coarse BA and then shrink the beamwidth adaptively and successively to obtain a more refined BA. The drawback of such schemes, however, is that each UE has its own specific AoA as seen from the BS side, thus, the BS needs to interact with each UE individually. As a result, all these hierarchical schemes require non-trivial coordination among the UEs and the BS, which is difficult to have at the initial channel acquisition stage. Moreover, since hierarchical schemes require interactive uplink-downlink communication between the BS and each individual UE, it is not clear how the overhead of such schemes scales in small cell scenarios with significant mobility of users across cells, where the BA procedure should be repeated at each handover.

The sparse nature of mmWave channels, i.e., large-dimension channel matrices along with very sparse scatterers in the AoA-AoD domain [7, 8], motivates the application of *Compressed Sensing* (CS) methods to speed up the BA. There are two groups of CS-based methods in the literature. The first group (e.g., see [9, 19–21]) applies CS to estimate the complex baseband channel coefficients. These algorithms are efficient and particularly attractive for multiuser scenarios, but they are based on the assumption that the instantaneous channel remains invariant during the whole probing/measuring stage. As anticipated before, this assumption is difficult to meet at mmWaves because of the large Doppler spread between the multipath components coming from different angles, implying significant time-variations of the channel coefficients even for UEs with small mobility [10, 22, 23].[2] The second group of CS-based schemes focuses on estimating the second-order statistics of the channel, i.e., the covariance of the channel matrix, which is very robust to channel variations. In [10] for example,

[2] Notice that the channel delay spread and time-variation are greatly reduced *after BA is achieved*, since once the beams are aligned, the communication occurs only through a single multipath component with small effective angular spread, whose delay and Doppler shift can be well compensated [23]. However, *before BA is achieved* the channel delay spread and time-variation can be large due to the presence of several mulipath components, each with its own delay and Doppler shift. In this case, even a small motion of a few centimeters traverses several wavelengths, potentially producing multiple deep fades [22].

a *Maximum Likelihood* (ML) method was proposed to estimate the covariance of the channel matrix. However, this scheme suffers from low SNR and the BA is achieved only at the UE side because of isotropic probing at the BS.

In our previous work [4], we proposed a novel efficient BA scheme that jointly estimates the two-sided AoA-AoD of the strongest path from the second-order statistics of the channel matrix. A limitation of [4], as well as most works based on OFDM signaling [9, 10], is the assumption of perfect OFDM frame synchronization and no inter-carrier interference. This is in fact difficult to achieve at mmWaves due to the potentially large multipath delay spread, Doppler shifts, and very low SNR before BA. These weaknesses, together with the fact that OFDM signaling suffers from large *Peak-to-Average Power Ratio* (PAPR), has motivated the proposal of single-carrier transmission [24, 25] as a more favorable option at mmWaves. Recently, [20, 21] proposed a time-domain BA approach based on CS techniques for single-carrier mmWave systems. However, as in [9, 19], this work focuses on estimating the instantaneous complex channel coefficients, with the assumption that these complex coefficients remain invariant over the whole training stage, which is an unrealistic assumption, as discussed above [4, 10, 22, 23].

B. Contributions

In this paper, we propose a novel efficient BA scheme for single-carrier mmWave communications with HDA transceivers and frequency-selective multipath channels. In the proposed scheme, each UE independently estimates its best AoA-AoD pair over the reserved beacon slots (see Section III), during which the BS periodically broadcasts its probing time-domain sequences. We exploit the sparsity of the mmWave channel in both angle and delay domains [26] to reduce the training overhead. We also pose the estimation of the strongest AoA-AoD pair as a *Non-Negative Least Squares* (NNLS) problem, which can be efficiently solved by standard techniques. Our main contributions can be summarized as follows:

1) *Pure time-domain operation.* Unlike our prior work in [4] and other works based on OFDM signaling [9, 10], the scheme proposed in this paper takes place completely in the time-domain and uses *Pseudo-Noise* (PN) sequences with *good* correlation properties that suits single-carrier mmWave systems.

2) *More general and realistic mmWave channel model.* We consider a quite general mmWave wireless channel model, taking into account the fundamental features of mmWave channels such as fast time-variation due to Doppler, frequency-selectivity, and the AoA-AoD sparsity [10, 22, 27].

3) *Tolerance to large Doppler shifts.* As in [4, 10], we design a signaling scheme to collect *quadratic measurements*, yielding estimates of the channel second-order statistics in the discretized AoA-AoD domain. Since quadratic measurements are related to the estimation of the received signal power, which is invariant with respect to the phase rotation of the channel taps, the proposed scheme is highly robust to the channel time-variations caused by the large Doppler spread between the multipath components.

4) *Impact of the PN sequence length.* Unlike our prior work in [28] and the work in [6], where Doppler is modeled as a phase rotation across different frames but the phase is kept constant over each beacon slot, here we consider a truly continuous linear (in time) phase rotation within the whole beacon slot. As a by-product of this realistic Doppler model, we notice that longer PN sequences do not necessarily exhibit better performance since they undergo larger phase rotations. We illustrate by numerical simulations that there is an optimal PN sequence length based on the given set of parameters, using which the proposed scheme achieves better performance in the presence of large Doppler shifts.

5) *System-level scalability and low-complexity beam direction estimation.* In our scheme, the BS actively probes the channel by periodically broadcasting a pseudo-random beamforming codebook over reserved beacon slots while all the UEs remain in the listening mode. Therefore, each UE is able to collect measurements from its channel locally and independently of all the other UEs. We pose the identification of the strongly coupled AoA-AoD pairs based on the measurements of each UE as an underdetermined system of noisy linear equations and solve it efficiently using *Non-Negative Least-Squares* (NNLS). Due to the properties of the NNLS, this yields a sparse estimate of the vector of non-negative channel gain coefficients in the discrete AoA-AoD domain. We illustrate via numerical simulations that the proposed scheme outperforms existing time-domain BA algorithms proposed in the literature in terms of training overhead. Moreover, in contrast with hierarchical algorithms, it does not require multiple rounds of uplink-downlink interaction between the BS and the UEs during the BA. Therefore, the proposed scheme is *scalable*, in the sense that its protocol overhead is essentially constant with the number of active UEs in the system.

6) *Effectiveness of single-carrier modulation.* Our proposed time-domain BA scheme is tailored to single-carrier mmWave systems. In particular, we show that, after achieving BA, the effective channel reduces essentially to a single path with a single delay and Doppler shift, with relatively large SNR due to the high beamforming gain. This means that single-carrier modulation needs no time-domain equalization and the baseband signal processing after BA is indeed very simple, requiring only standard timing and *carrier frequency offset* (CFO) recovery, operating in relatively large SNR conditions (after beamforming).

Notation: We denote vectors, matrices and scalars by $\mathbf{a}$, $\mathbf{A}$ and a (A) respectively. We represent sets by $\mathcal{A}$ and

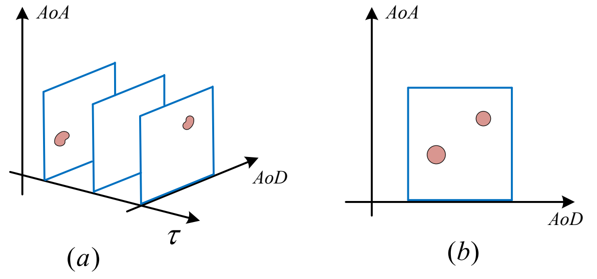


Fig. 1: *Illustration of the channel sparsity in the* Angle of Arrival *(AoA),* Angle of Departure *(AoD), and delay domains. (a) Slices of the channel power spread function over discrete delay taps, where only a few slices contain scattering components with large power. (b) Marginal power spread function of the channel in the AoA-AoD domain obtained from the integration of the power spread function over the delay domain.*

their cardinality with $|\mathcal{A}|$. We use $\mathbb{E}$ for the expectation, $\otimes$ for the Kronecker product, $\mathbf{A}^{\mathsf{T}}$ for transpose, $\mathbf{A}^*$ for conjugate, and $\mathbf{A}^{\mathsf{H}}$ for conjugate transpose. We define the vectorization operator as $\text{vec}(\cdot)$. For an integer $k \in \mathbb{Z}$, we use the shorthand notation $[k]$ for the index set $\{1, ..., k\}$.

II. Problem Statement

In this section, we provide a general overview of the BA problem based on the channel second-order statistics. Then, in Sections III and IV we provide the fully detailed system model and the proposed algorithm.

A. Channel Second-Order Statistics

We consider a widely used and well accepted mmWave scattering channel model (e.g., see [7, 8]), where the propagation between the BS and a generic UE occurs along a sparse collection of multipath components in the continuous AoA-AoD-delay (ϕ, θ, τ) domain, including a possible *Line-of-Sight* (LOS) component as well as some Non-Line-of-Sight (NLOS) reflected paths [25]. The channel follows *locally* the classical *Wide-Sense Stationary with Uncorrelated Scattering* (WSSUS) model [29, 30]. The average signal energy distribution over the AoA-AoD-Delay domain is described by the *Power Spread Function* (PSF) $f_p(\phi, \theta, \tau)$. In brief, $f_p(\phi, \theta, \tau) d\phi d\theta d\tau$ is the aggregate signal power transfer coefficient for the propagation paths in the AoA-AoD region $[\phi, \phi + d\phi) \times [\theta, \theta + d\theta)$ with path delays in $[\tau, \tau + d\tau)$. The PSF encodes the second-order statistics of the channel and it is locally time-invariant as long as the propagation geometry does not change significantly. The time scale over which the PSF is time-invariant is very large with respect to the inverse of the signaling bandwidth, justifying the local WSS assumption. Practical channel measurements have shown that only a few discrete delays carry significant signal energy, corresponding to the propagation delays of the LOS and some reflection NLOS paths [7, 8, 26]. This is illustrated in Fig. 1 (a), where only a few slices of the PSF with respect to the delay domain contain scattering components with large power. The marginal PSF of the channel in the AoA-AoD domain is obtained by integrating over the delay variable as

$$f_p(\phi, \theta) = \int_\tau f_p(\phi, \theta, \tau) d\tau, \tag{1}$$

and it is typically very sparse in the continuous angle domain (see, e.g., Fig. 1 (b)).

B. Beam-Alignment Using Second-order Statistics

In terms of BA, we are interested in finding an AoA-AoD pair corresponding to strong communication path between the UE and the BS. If the marginal PSF of the channel in the AoA-AoD domain $f_p(\phi, \theta)$ in (1) was *a-priori* known, the BA problem would simply boil down to finding the support of $f_p(\phi, \theta)$ (e.g., see the two bubbles in Fig. 1 (b)). In practice, however, $f_p(\phi, \theta)$ is not known and should be estimated via a suitable signaling scheme. With this in mind, we can pose the BA problem as follows.

BA Problem: Design a suitable signaling between the BS and the UE, find an estimate of the AoA-AoD PSF $f_p(\phi, \theta)$, and identify an AoA-AoD pair (ϕ_0, θ_0) with a sufficiently large strength $f_p(\phi_0, \theta_0)$.

In this paper, we use pseudo-random waveforms with nice auto-/cross-correlation properties as the probing signals. We will show that, using the proposed signaling, each UE is able to collect its own quadratic measurements, which yield noisy linear projections of a suitably discretized version of the marginal PSF $f_p(\phi, \theta)$. By expressing such linear projections as a matrix-vector product, we formulate the PSF estimation as the Least-Squares solution of an underdetermined system of linear equations. Imposing the non-negativity of the discretized PSF coefficients, we are in the presence of a NNLS problem, which naturally yields a sparse solution [31, 32].

Fig. 2 (a) illustrates the proposed frame structure which consists of three parts: the downlink beacon slot, the *Random Access Control CHannel* (RACCH) slot, and the data slot. An overview of the proposed initial acquisition and BA protocol is illustrated in Fig. 2 (b). As in [4, 28], the measurements are collected by the UEs from the sequence of downlink beacon slots broadcasted by the BS. By running the NNLS estimation algorithm mentioned above, each UE selects its strongest AoA-AoD pair, i.e., the discrete beam indices corresponding to the strongest path in the estimated discretized PSF. Then, the initial acquisition protocol proceeds as described in [4, 28]. Namely, the UE sends a beamformed packet to the BS in the RACCH slot, during which the BS stays in listening mode and uses its M_{RF} RF chains to form M_{RF} coarse beam patterns (sectors) covering the whole BS angle domain, in order to provide some receiver beamforming gain. The RACCH packet contains basic information such as user ID and the

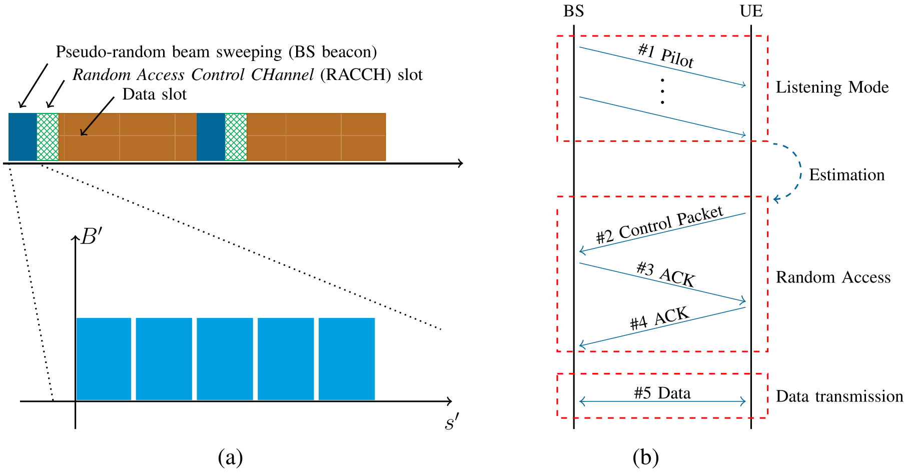


Fig. 2: *(a) (Top) Frame structure of the proposed Beam Alignment (BA) scheme. (Bottom) Each beacon slot consists of S PN sequences indexed by $s' \in [S]$, and all the PN sequences have access to the whole effective bandwidth $B' \leq B$. (b) Illustration of the proposed BA process between the BS and a generic UE, where the procedures (#2∼#5) are independently done at each UE, and all the UEs share the same BS beamforming codebook (#1).*

index of the beam corresponding to the selected AoD. The BS responds with a data packet with piggybacked acknowledgment in the data slot of a next frame, and from this moment on the BS and the UE are connected. Further beam refinement and tracking is possible in order to adapt to small variations of the propagation geometry. However, this can be achieved by rather standard array processing and goes outside the scope of this paper.

C. Equivalent Channel after Beam-Alignment

After completing a BA cycle as described above, the UE and the BS focus their beams on a specific AoA-AoD pair (ϕ_0, θ_0) to boost the SNR as much as possible.[3] As a result, the equivalent channel after beamforming along (ϕ_0, θ_0) can be represented by a *Single-Input Single-Output* (SISO) channel. The PSF of the resulting SISO channel can be well approximated by $f_p(\tau) = f_p(\phi_0, \theta_0, \tau)$ in the delay domain, which simply corresponds to the PSF obtained by focusing the transmitted signal power along the estimated AoA-AoD (ϕ_0, θ_0). Due to the underlying channel sparsity [7, 8, 26], we expect that, after BA, the PSF $f_p(\tau)$ consists of a nearly single-tap channel with delay τ_0 and Doppler shift ν_0, which can be estimated and compensated by a standard timing and frequency offset synchronization subsystem. Since these functions are performed after BA, the operating SNR is not at all critical. Therefore, standard techniques for single-carrier synchronization can be used. Furthermore, since the effective channel after BA reduces virtually to a single tap, it is essentially frequency-flat. This fact was observed experimentally in [22] and its impact on modulation and receiver schemes was discussed in [23]. In particular, for the class of sparse mmWave channels considered on this work, we argue that near-optimal performance can be achieved by single-carrier communication with no need of equalization. This is confirmed by the results in Section V, where we use the effective SISO channel to derive upper and lower bounds on the achievable ergodic rate after BA, showing that time-domain equalization is effectively not needed.

III. Mathematical Modeling

A. Channel Model

Consider a generic UE in a mmWave system served by a specific BS. Suppose that the BS is equipped with a *Uniform Linear Array* (ULA) having M antennas and $M_{\text{RF}} \ll M$ RF chains. The UE also has a ULA with N antennas and $N_{\text{RF}} \ll N$ RF chains. We assume that both the BS and the UE apply a hybrid beamforming consisting of an analog precoder/combiner and a digital precoder/combiner. In this paper, we will focus mainly on training the analog precoders/combiners in the initial BA phase. We assume that the propagation channel between the BS and the UE consists of $L \ll \max\{M, N\}$ multipath components, where the $N \times M$ baseband equivalent impulse response of the channel at time slot s is given by

$$\mathsf{H}_s(t, \tau) = \sum_{l=1}^{L} \rho_{s,l} e^{j2\pi\nu_l t} \mathbf{a}_{\text{R}}(\phi_l) \mathbf{a}_{\text{T}}(\theta_l)^{\mathsf{H}} \delta(\tau - \tau_l), \quad (2)$$

[3] In practice, beamforming patterns with some efficient shape and some assigned beamwidth can be formed in the selected directions (ϕ_0, θ_0), where the beamwidth can be optimized in order to trade-off beamforming gain for robustness to pointing errors due to small movements. The design of the beam shape used for data communication (namely, after the BA is achieved) is treated comprehensively in the literature and is out of the scope of this paper.

where $(\phi_l, \theta_l, \tau_l, \nu_l)$ denote the AoA, AoD, delay, and Doppler shift of the l-th component, and $\delta(\cdot)$ denotes the Dirac delta function. The vectors $\mathbf{a}_{\text{T}}(\theta_l) \in \mathbb{C}^M$ and $\mathbf{a}_{\text{R}}(\phi_l) \in \mathbb{C}^N$ are the array response vectors of the BS and UE at AoD θ_l and AoA ϕ_l respectively, with elements given by

$$[\mathbf{a}_{\text{T}}(\theta)]_m = e^{j(m-1)\pi \sin(\theta)}, m \in [M], \tag{3a}$$

$$[\mathbf{a}_{\text{R}}(\phi)]_n = e^{j(n-1)\pi \sin(\phi)}, \ n \in [N], \tag{3b}$$

where we assume that the spacing of the ULA antennas equals to half wavelength.

For the sake of modeling simplicity, we assume in (2) that each multipath component has a very narrow footprint over the AoA-AoD and delay domain. Extension to more widely spread multipath clusters is straightforward and will be applied in the numerical simulations. Moreover, we make the very standard assumption in array processing that the array response vectors are invariant with frequency over the signal bandwidth. More precisely, we assume that the wavelength λ over the frequency interval $f \in [f_0 - B/2, f_0 + B/2]$ can be approximated as $\lambda_0 = c/f_0$, where c denotes the speed of light. This is indeed well verified when B is less than $1/10$ of the carrier frequency (e.g., $B = 1$ GHz with carrier between 30 and 70 GHz). Each scatterer corresponding to a AoA-AoD-Delay $(\phi_l, \theta_l, \tau_l)$ has a Doppler shift $\nu_l = \frac{\Delta v_l f_0}{c}$ where Δv_l indicates the relative speed of the receiver, the l-th scatterer, and the transmitter [6]. We adopt a block fading model, where the coefficient of the l-th multipath component $\rho_{s,l}$ remains invariant over the channel *coherence time* Δt_c but change i.i.d. randomly across different *coherence times* [10]. Each scatterer is formed by the superposition of a large number of micro-scattering components (e.g., due to rough surfaces) having (approximately) the same AoA-AoD and delay. By the central limit theorem it is customary to model the superposition of these many small effects as Gaussian [29, 30]. Hence, the multipath component coefficients are modeled as Rice fading given by

$$\rho_{s,l} \sim \sqrt{\gamma_l} \left(\sqrt{\frac{\eta_l}{1+\eta_l}} + \frac{1}{\sqrt{1+\eta_l}} \check{\rho}_{s,l} \right), \tag{4}$$

where γ_l denotes the overall multipath component strength, $\eta_l \in [0, \infty)$ indicates the strength ratio between the LOS and the NLOS components, and $\check{\rho}_{s,l} \sim \mathcal{CN}(0,1)$ is a zero-mean unit-variance complex Gaussian random variable. In particular, $\eta_l \to \infty$ indicates a pure LOS path while $\eta_l = 0$ indicates a pure NLOS path, affected by standard Rayleigh fading.

B. Proposed Signaling Scheme

We assume that the BS can simultaneously transmit up to $M_{\text{RF}} \ll M$ different pilot streams. In our previous work [4], we considered OFDM signaling where the different pilot streams are assigned to non-overlapping sets of orthogonal subcarriers, such that (in the absence of inter-carrier interference) they can be perfectly separated by the UE in the frequency domain. However, such scheme may incur performance degradation in the presence of significant Doppler spread between the different multipath components and/or carrier frequency offset between the BS transmitter and UE receiver. Hence, in this work we consider single-carrier signaling where different PN sequences are assigned to each pilot stream, similar to *Code Division Multiple Access* (CDMA). In the proposed scheme, the different pilot streams are generally non-orthogonal but the cross-interference is very small if the assigned PN sequences are sufficiently long with good cross-correlation properties. As we shall see, this signaling scheme yields very good robustness to Doppler.

Let $x_{s,i}(t)$, $t \in [st_0, (s+1)t_0)$, be the continuous-time baseband equivalent PN signal corresponding to the i-th ($i \in [M_{\text{RF}}]$) pilot stream transmitted over s-th slot, given by

$$x_{s,i}(t) = \sum_{n=1}^{N_c} \varrho_{n,i} p_r(t - nT_c), \quad \varrho_{n,i} \in \{1, -1\}, \tag{5}$$

where t_0 denotes the duration of the PN sequence, $p_r(t)$ is a square-root Nyquist pulse[4] [33] with normalized energy $\int |p_r(t)|^2 dt = 1$, and $\{\varrho_{n,i} : n \in [N_c]\}$ is the n-th chip symbol. The PN sequence has a chip duration T_c, bandwidth $B' \approx 1/T_c \leq B$ (where B denotes the maximum available channel bandwidth), and a total of $N_c = t_0/T_c$ chips. We shall choose a suitable sequence length N_c, such that the time-domain signal (5) is transmitted in a sufficiently small time-interval t_0 over which the channel can be considered time-invariant, i.e., $t_0 \ll \Delta t_c$.

To transmit the i-th pilot stream, the BS applies a beamforming vector $\mathbf{u}_{s,i} \in \mathbb{C}^M$. Without loss of generality, the beamforming vectors are normalized such that $\|\mathbf{u}_{s,i}\| = 1$. As mentioned before, we consider a HDA beamforming architecture where the beamforming function is implemented in the analog RF domain. Hence, the beamforming vectors $\mathbf{u}_{s,i}$, $i \in [M_{\text{RF}}]$, are independent of frequency and constant over the whole bandwidth. The transmitted signal at slot s is given by

$$\begin{aligned} \mathbf{x}_s(t) &= \sum_{i=1}^{M_{\text{RF}}} \sqrt{\frac{P_{\texttt{tot}} T_c}{M_{\text{RF}}}} x_{s,i}(t) \mathbf{u}_{s,i} \\ &= \sum_{i=1}^{M_{\text{RF}}} \sum_{n=1}^{N_c} \sqrt{\frac{P_{\texttt{tot}} T_c}{M_{\text{RF}}}} \varrho_{n,i} p_r(t - nT_c) \mathbf{u}_{s,i}, \end{aligned} \tag{6}$$

where $P_{\texttt{tot}}$ is the total transmit power which is equally distributed into the M_{RF} RF chains from BS. The term $\frac{P_{\texttt{tot}} T_c}{M_{\text{RF}}}$ indicates the energy per chip of the transmitted PN sequences, where T_c denotes the chip duration.

[4] A square-root Nyquist pulse is a finite-energy waveform $p_r(t)$ such that the squared magnitude of its spectrum $|P_r(f)|^2$ satisfies the Nyquist criterion [33].

Consequently, the received baseband equivalent signal at the UE array is

$$\begin{aligned}\mathbf{r}_s(t) &= \int \mathsf{H}_s(t,\tau)\mathbf{x}_s(t-\tau)d\tau \\ &= \sum_{l=1}^{L} \mathsf{H}_{s,l}(t)\mathbf{x}_s(t-\tau_l) \\ &= \sum_{i=1}^{M_{\rm RF}}\sum_{l=1}^{L}\sqrt{\frac{P_{\tt tot}T_c}{M_{\rm RF}}}\mathsf{H}_{s,l}(t)x_{s,i}(t-\tau_l)\mathbf{u}_{s,i}, \qquad (7)\end{aligned}$$

where $\mathsf{H}_{s,l}(t) := \rho_{s,l}e^{j2\pi\nu_l t}\mathbf{a}_{\rm R}(\phi_l)\mathbf{a}_{\rm T}(\theta_l)^{\mathsf{H}}$, $l \in [L]$ are the time-varying MIMO channel taps corresponding to the L multipath components as in (2).

With a hybrid MIMO structure, the UE does not have direct access to (a sampled version of) the components of $\mathbf{r}_s(t)$. Instead, at each beacon slot s, the UE must apply some beamforming vector in the analog domain obtaining a projection of the received signal. Since the UE has $N_{\rm RF}$ RF chains, it can obtain up to $N_{\rm RF}$ such projections per slot. The analog RF signal received at the UE antenna array is distributed across the $N_{\rm RF}$ RF chains for demodulation. This is achieved by signal splitters that divide the signal power by a factor of $N_{\rm RF}$. Thus, the received signal at the output of the j-th RF chain at the UE side is given by

$$\begin{aligned}\hat{y}_{s,j}(t) &= \frac{1}{\sqrt{N_{\rm RF}}}\mathbf{v}_{s,j}^{\mathsf{H}}\mathbf{r}_s(t) + z_{s,j}(t) \\ &= \sum_{i=1}^{M_{\rm RF}}\sum_{l=1}^{L}\sqrt{E_{\tt dim}}\mathbf{v}_{s,j}^{\mathsf{H}}\mathsf{H}_{s,l}(t)\mathbf{u}_{s,i}x_{s,i}(t-\tau_l)+z_{s,j}(t), \qquad (8)\end{aligned}$$

where $E_{\tt dim} = \frac{P_{\tt tot}T_c}{M_{\rm RF}N_{\rm RF}}$ indicates the per-stream pilot chip energy distributed over the transmit and receive RF chains, $\mathbf{v}_{s,j} \in \mathbb{C}^N$ denotes the normalized beamforming vector of the j-th RF chain at the UE side with $\|\mathbf{v}_{s,j}\| = 1$, $z_{s,j}(t)$ is the continuous-time complex *Additive White Gaussian Noise* (AWGN) at the output of the j-th RF chain, with a *Power Spectral Density* (PSD) of N_0 Watt/Hz. The noise at the receiver is mainly introduced by the RF chain electronics, e.g., filter, mixer, and A/D conversion. The factor $\frac{1}{\sqrt{N_{\rm RF}}}$ in (8) takes into account the power split said above, assuming that this only applies to the useful signal and not to the thermal noise. Therefore, this received signal model is a conservative worst-case assumption.

In realistic conditions, we have $T_c\nu_l \ll 1$.[5] Hence, the phase time-variation over the duration of the chip pulse shape is negligible. It follows that we can replace the continuously time-varying matrix tap coefficient $\mathsf{H}_{s,l}(t)$ with its discrete approximation, which can be simply written in the form

$$\begin{aligned}\mathsf{H}_{s,l}(t)\Big|_{t\in[nT_c,(n+1)T_c)} &\approx \rho_{s,l}e^{j2\pi(\breve{\nu}_{s,l}+\nu_l nT_c)}\mathbf{a}_{\rm R}(\phi_l)\mathbf{a}_{\rm T}(\theta_l)^{\mathsf{H}} \\ &= \mathsf{H}_{s,l}e^{j2\pi\nu_l nT_c} \qquad (9)\end{aligned}$$

with $n \in [N_c]$, where $\mathsf{H}_{s,l} := \rho_{s,l}e^{j2\pi\breve{\nu}_{s,l}}\mathbf{a}_{\rm R}(\phi_l)\mathbf{a}_{\rm T}(\theta_l)^{\mathsf{H}}$, and where $\breve{\nu}_{s,l}$ represents a phase rotation at the beginning of the s-th beacon slot which is irrelevant since it can be incorporated in the Gaussian coefficient $\rho_{s,l}$. As a result, the product term $\mathsf{H}_{s,l}(t)x_{s,i}(t-\tau_l)$ in (8) can be written as

$$\begin{aligned}\mathsf{H}_{s,l}(t)x_{s,i}(t-\tau_l) &= \mathsf{H}_{s,l}\sum_{n=1}^{N_c}\varrho_{n,i}e^{j2\pi\nu_l nT_c}\, p_r(t-nT_c-\tau_l) \\ &:= \mathsf{H}_{s,l}x_{s,i}^l(t-\tau_l), \qquad (10)\end{aligned}$$

where $x_{s,i}^l(t)$ is given by

$$x_{s,i}^l(t) = \sum_{n=1}^{N_c}\varrho_{n,i}e^{j2\pi\nu_l nT_c}\, p_r(t-nT_c). \qquad (11)$$

Notice that $x_{s,i}^l(t)$ consists of a modified modulated PN sequence where the chip symbols $\varrho_{n,i}$ are rotated by the time-varying phase factor $e^{j2\pi\nu_l nT_c}$ due to the Doppler shift. Substituting (10) into (8), we can write the received signal $\hat{y}_{s,j}(t)$ in (8) as

$$\hat{y}_{s,j}(t)=\sum_{i=1}^{M_{\rm RF}}\sum_{l=1}^{L}\sqrt{E_{\tt dim}}\mathbf{v}_{s,j}^{\mathsf{H}}\mathsf{H}_{s,l}\mathbf{u}_{s,i}x_{s,i}^l(t-\tau_l)+z_{s,j}(t). \qquad (12)$$

Since the PN sequences assigned to the $M_{\rm RF}$ RF chains are mutually (roughly) orthogonal, the $M_{\rm RF}$ pilot streams transmitted from the BS side can be approximately separated at the UE by passing each j-th received signal (12) through a bank of matched filters where the i filter has impulse response $x_{s,i}^*(-t) = \sum_{n=1}^{N_c}\varrho_{n,i}p_r^*(-t+nT_c)$. Consequently, the i-th BS pilot stream received through the j-th RF chain at the UE is given by

$$\begin{aligned}y_{s,i,j}(t) &= \int \hat{y}_{s,j}(\tau)x_{s,i}^*(\tau-t)d\tau \\ &= \sum_{l=1}^{L}\sum_{i'=1}^{M_{\rm RF}}\sqrt{E_{\tt dim}}\mathbf{v}_{s,j}^{\mathsf{H}}\mathsf{H}_{s,l}\mathbf{u}_{s,i}R_{i',i}^{x^l}(t-\tau_l)+z_{s,j}^c(t) \\ &\overset{(a)}{\approx} \sum_{l=1}^{L}\sqrt{E_{\tt dim}}\mathbf{v}_{s,j}^{\mathsf{H}}\mathsf{H}_{s,l}\mathbf{u}_{s,i}R_{i,i}^{x^l}(t-\tau_l)+z_{s,j}^c(t) \qquad (13)\end{aligned}$$

where $\forall i, i' \in [M_{\rm RF}]$, $R_{i',i}^{x^l}(t) := \int x_{s,i'}^l(\tau)x_{s,i}^*(\tau-t)d\tau$ represents the correlation between the Doppler-rotated sequence $x_{s,i'}^l(t)$ given by (11) and the desired sequence $x_{s,i}(t)$, and $z_{s,j}^c(t) = \int z_{s,j}(\tau)x_{s,i}^*(\tau-t)d\tau$ denotes the noise at the output of the matched filter. The approximation (a) in (13) follows the fact that, the cross-correlations between different PN sequences are nearly zero, i.e., $R_{i',i}^{x}(t) = \int x_{s,i'}(\tau)x_{s,i}^*(\tau-t)d\tau \approx 0$, for $i' \neq i$. Since the phase rotation introduced by Doppler is very small ($\nu_l T_c \ll 1$), we can also safely assume that $R_{i',i}^{x^l}(t) = \int x_{s,i'}^l(\tau)x_{s,i}^*(\tau-t)d\tau \approx 0$, for $i' \neq i$. However, it is important to point out that these are only working assumptions in order to derive our algorithm. The actual

[5] For example, consider $T_c = 1$ ns, $\Delta v_l = 10$ m/s and $f_0 = 60$ GHz yielding $\nu_l = 2$ kHz and $T_c\nu_l = 2 \cdot 10^{-6}$.

performance of the scheme will of course depend also on the residual non-zero cross-interference between the PN sequences. Hence, in our numerical simulations, we made no such simplification and took into account all the cross terms arising from non-perfect orthogonality.

Consider (13) and suppose that the output signal at the UE side is sampled at the chip-rate. The resulting discrete-time signal can be written as

$$\begin{aligned} y_{s,i,j}[k] &= y_{s,i,j}(t)|_{t=kT_c} \\ &= \sum_{l=1}^{L} \sqrt{E_{\mathtt{dim}}} \mathbf{v}_{s,j}^{\mathsf{H}} \mathsf{H}_{s,l} \mathbf{u}_{s,i} R_{i,i}^{x^l}(kT_c - \tau_l) + z_{s,j}^{c}[k], \end{aligned} \tag{14}$$

where $k \in [\check{N}_c]$ indicates the sampling index, $\check{N}_c \geq N_c + \frac{\Delta\tau_{\max}}{T_c}$ denotes the total number of samples in each received PN sequence, and $\Delta\tau_{\max} = \max\{|\tau_l - \tau_{l'}| : l, l' \in [L]\}$ denotes the maximum delay spread of the channel. Note that for PN sequences, the sequence of samples $\{|R_{i,i}^{x^l}(kT_c - \tau_l)| : k \in [\check{N}_c]\}$ in (14) has sharp peaks at indices $k_l \approx \frac{\tau_l}{T_c}$, corresponding to the delays of the channel multipath components. Intuitively speaking, the output $y_{s,i,j}[k]$ at those indices k_l yields Gaussian variables whose power is obtained by projecting the AoA-AoD-Delay PSF $f_p(\phi, \theta, \tau)$ along beamforming vectors $(\mathbf{u}_{s,i}, \mathbf{v}_{s,j})$ in the angular domain and along the k_l-th delay slice $\tau \in [k_l T_c, (k_l + 1)T_c]$. The slicing in the delay domain results from the fact that, as said before, $|R_{i,i}^{x^l}(kT_c - \tau_l)|$ is well localized around k_l. We refer to Fig. 1 (a) for an illustration and will use this property later on in the paper to design our BA algorithm.

C. Sparse Beam-space Representation

The AoA-AoDs (ϕ_l, θ_l) in (2) take on arbitrary values in the continuous AoA-AoDs domain. Following the widely used approach of [34], known as *beam-space representation*, we obtain a finite-dimensional representation of the channel response (2) by discretizing the angle domain. Consider the discrete set of AoA-AoDs

$$\Phi := \{\check{\phi} : (1 + \sin(\check{\phi}))/2 = \frac{n-1}{N}, n \in [N]\}, \tag{15a}$$

$$\Theta := \{\check{\theta} : (1 + \sin(\check{\theta}))/2 = \frac{m-1}{M}, m \in [M]\}. \tag{15b}$$

The corresponding sets of array responses $\mathcal{A}_{\mathrm{R}} := \{\mathbf{a}_{\mathrm{R}}(\check{\phi}) : \check{\phi} \in \Phi\}$ and $\mathcal{A}_{\mathrm{T}} := \{\mathbf{a}_{\mathrm{T}}(\check{\theta}) : \check{\theta} \in \Theta\}$ form discrete dictionaries to represent the channel response. For the ULAs considered in this paper, the dictionaries $\mathcal{A}_{\mathrm{R}}$ and $\mathcal{A}_{\mathrm{T}}$, after suitable normalization, reduce to the columns of unitary *Discrete Fourier Transform* (DFT) matrices $\mathbf{F}_N \in \mathbb{C}^{N\times N}$ and $\mathbf{F}_M \in \mathbb{C}^{M\times M}$, with elements

$$[\mathbf{F}_N]_{n,n'} = \frac{1}{\sqrt{N}} e^{j2\pi(n-1)(\frac{n'-1}{N} - \frac{1}{2})}, n, n' \in [N], \tag{16a}$$

$$[\mathbf{F}_M]_{m,m'} = \frac{1}{\sqrt{M}} e^{j2\pi(m-1)(\frac{m'-1}{M} - \frac{1}{2})}, m, m' \in [M]. \tag{16b}$$

The channel beam-space representation consists of expressing the channel matrix as the linear combination of the outer product of rank-1 matrices of the form $\mathbf{f}_{N,n}\mathbf{f}_{M,m}^{\mathsf{H}}$ for all $n \in [N]$ and $m \in [M]$, where $\mathbf{f}_{N,n}$ and $\mathbf{f}_{M,m}$ denote the n-th and m-th columns of $\mathbf{F}_N$ and of $\mathbf{F}_M$, respectively. Explicitly, the beam-space representation expression is given by

$$\begin{aligned} \mathsf{H}_s(t,\tau) &= \sum_{n=1}^{N} \sum_{m=1}^{M} \left[\check{\mathsf{H}}_s(t,\tau)\right]_{n,m} \mathbf{f}_{N,n} \mathbf{f}_{M,m}^{\mathsf{H}} \\ &= \mathbf{F}_N \check{\mathsf{H}}_s(t,\tau) \mathbf{F}_M^{\mathsf{H}}, \end{aligned} \tag{17}$$

where the beam-space representation of the channel response is given by

$$\check{\mathsf{H}}_s(t,\tau) = \mathbf{F}_N^{\mathsf{H}} \mathsf{H}_s(t,\tau) \mathbf{F}_M = \sum_{l=1}^{L} \check{\mathsf{H}}_{s,l}(t) \delta(\tau - \tau_l), \tag{18}$$

where $\check{\mathsf{H}}_{s,l}(t) := \mathbf{F}_N^{\mathsf{H}} \mathsf{H}_{s,l}(t) \mathbf{F}_M$ corresponds to the beam-space l-th channel path.

As shown in our earlier work [4], as the number of antennas M at the BS and N at the UE increases, the DFT basis provides a good sparsification of the propagation channel. As a result, $\check{\mathsf{H}}_s(t,\tau)$ can be approximated as a sparse matrix, with non-zero elements in the locations corresponding to small clusters of discrete AoA-AoD pairs in the proximity of the (continuous) angle pairs of the L scatterers of the physical channel. We may encounter a grid error in (18) since the AoAs/AoDs do not necessarily fall into the uniform grid $\Phi \times \Theta$. Nevertheless, as shown in [4], the grid error becomes negligible by increasing the number of antennas (i.e., the grid resolution). We hasten to say that, in our simulations, we do not constrain the AoA-AoD pairs of the physical channel to take on values on the discrete grid; therefore, the grid discretization effects is fully taken into account in our numerical results.

IV. Proposed Beam Alignment Scheme

A. BS Channel Probing and UE Sensing

Consider the scattering channel model in (2) and its beam-space representation in (18). In our proposed scheme, at each beacon slot s, the BS probes the channel along M_{RF} beamforming vectors $\mathbf{u}_{s,i}$, $i \in [M_{\mathrm{RF}}]$, each of which is applied to a unique PN sequence signal $x_{s,i}(t)$. We select the beamforming vectors at the BS side according to a pre-defined pseudo-random codebook, which is a collection of the angle sets $\mathcal{C}_{\mathrm{T}} := \{\mathcal{U}_{s,i} : s \in [T], i \in [M_{\mathrm{RF}}]\}$, where $\mathcal{U}_{s,i}$ denotes the angle-domain support of the beamforming vector $\mathbf{u}_{s,i}$, i.e., the indices of the quantized angles in the beam-space representation of $\mathbf{u}_{s,i}$, and where T is the effective period of beam training. We assume that the beamforming vector $\mathbf{u}_{s,i}$ sends equal power along the directions in $\mathcal{U}_{s,i}$ with the number of active angles given by $|\mathcal{U}_{s,i}| =: \kappa_u \leq M$, which we assume to be the same for all (s,i). We call κ_u the *angle spreading factor* with respect to the transmit beamforming vectors. Consequently, we can write such beamforming vectors as $\mathbf{u}_{s,i} = \mathbf{F}_M \check{\mathbf{u}}_{s,i}$, where $\check{\mathbf{u}}_{s,i} = \frac{\mathbf{1}_{\mathcal{U}_{s,i}}}{\sqrt{\kappa_u}}$, and where $\mathbf{1}_{\mathcal{U}_{s,i}}$ denotes a vector with 1 at components in the support set $\mathcal{U}_{s,i}$ and 0 elsewhere. One can simply imagine the vector $\check{\mathbf{u}}_{s,i}$ as a multi-finger beam pattern in the angle-domain as illustrated in Fig. 3 (a).[6] We assume that the angle indices in $\mathcal{U}_{s,i}$ in the codebook $\mathcal{C}_{\text{T}}$ are a priori generated in a random manner and are a priori known to all UEs in the system. This is similar to the BS-dependent pseudo-random synchronization codes used in the 3G WCDMA standard [36]. Thus, we call $\mathcal{C}_{\text{T}}$ a pseudo-random codebook.

At the UE side, each UE can locally customize its own receive beamforming codebook defined as $\mathcal{C}_{\text{R}} := \{\mathcal{V}_{s,j} : s \in [T], j \in [N_{\text{RF}}]\}$, where $\mathcal{V}_{s,j}$, with $|\mathcal{V}_{s,j}| = \kappa_v \leq N$ for all (s, j), is the angle-domain support, defining the directions from which the receiver beam patterns collect the signal power. We define the beamforming vectors at the UE side by $\mathbf{v}_{s,j} = \mathbf{F}_N \check{\mathbf{v}}_{s,j}$, where $\check{\mathbf{v}}_{s,j} = \frac{\mathbf{1}_{\mathcal{V}_{s,j}}}{\sqrt{\kappa_v}}$ again defines the finger-shaped beam patterns as shown in Fig. 3 (a). Similar to the power spreading factor κ_u at the BS, the parameter κ_v controls the spread of the sensing beam patterns at the UE.

In our scheme, the UEs collect their measurements independently and simultaneously, without any influence or coordination to each other. Therefore, the scheme is quite scalable for multiuser scenarios, where the overhead of training all the UEs does not increase with the number of UEs. This represents a significant advantage with respect to traditional multi-level/interactive BA schemes, that require multiple beam-sweeping rounds and interactive data exchanges between the BS and each UE, such that the acquisition protocol overhead grows proportionally to the number of UEs being acquired.

B. UE Measurement Sparse Formulation

During the s-th beacon slot, the UE applies the receive beamforming vector $\mathbf{v}_{s,j}$ to its j-th RF chain. Assuming that the probing PN signals $x_{s,i}(t)$ are approximately orthogonal in the time domain as discussed before, each RF chain at the UE side can almost perfectly separate the transmitted M_{RF} pilot streams. Thus, using the beam-space representation of the channel in (18), we can write (14) as

$$y_{s,i,j}[k] = \sum_{l=1}^{L} \sqrt{E_{\texttt{dim}}} \check{\mathbf{v}}_{s,j}^{\mathsf{H}} \check{\mathsf{H}}_{s,l} \check{\mathbf{u}}_{s,i} R_{i,i}^{x^l}(kT_c - \tau_l) + z_{s,j}^c[k], \tag{19}$$

where $\check{\mathbf{u}}_{s,i} = \mathbf{F}_M^{\mathsf{H}} \mathbf{u}_{s,i}$ and $\check{\mathbf{v}}_{s,j} = \mathbf{F}_N^{\mathsf{H}} \mathbf{v}_{s,j}$ are the beamforming vectors in the beam-space domain. Here, we used the unitary property of the DFT matrices, i.e., $\mathbf{F}_M^{\mathsf{H}} \mathbf{F}_M = \mathbf{I}_M$ and $\mathbf{F}_N^{\mathsf{H}} \mathbf{F}_N = \mathbf{I}_N$, where $\mathbf{I}_M$ and $\mathbf{I}_N$ are identity matrices of dimension M and N respectively.

To formulate the sparse estimation problem, we define the vectors $\check{\mathbf{h}}_{s,l} = 1/\sqrt{N_{\text{RF}}} \cdot \text{vec}(\check{\mathsf{H}}_{s,l})$, $l \in [L]$, resulting in a reformulated channel matrix $\check{\mathbf{H}}_s = [\check{\mathbf{h}}_{s,1}, \cdots, \check{\mathbf{h}}_{s,L}]$ that collects all the channel coefficients in the beam-space domain. We also define a vector $\mathbf{c}_k^i = [R_{i,i}^{x^1}(kT_c - \tau_1), \cdots, R_{i,i}^{x^L}(kT_c - \tau_L)]^{\mathsf{T}} \cdot \sqrt{E_{\texttt{dim}}}$, which can be regarded as the *Power Delay Profile* (PDP) of the i-th pilot stream transmitted along the L paths and sampled at the k-th discrete delay tap kT_c. Consequently, we can express the received beacon signal (19) at the UE as

$$\begin{aligned} y_{s,i,j}[k] &= \sum_{l=1}^{L} \sqrt{E_{\texttt{dim}}} \check{\mathbf{v}}_{s,j}^{\mathsf{H}} \check{\mathsf{H}}_{s,l} \check{\mathbf{u}}_{s,i} R_{i,i}^{x^l}(kT_c - \tau_l) + z_{s,j}^c[k] \\ &= (\check{\mathbf{u}}_{s,i} \otimes \check{\mathbf{v}}_{s,j}^*)^{\mathsf{T}} \check{\mathbf{H}}_s \mathbf{c}_k^i + z_{s,j}^c[k] \\ &= \mathbf{g}_{s,i,j}^{\mathsf{T}} \check{\mathbf{H}}_s \mathbf{c}_k^i + z_{s,j}^c[k], \end{aligned} \tag{20}$$

where we used the well-known identity $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathsf{T}} \otimes \mathbf{A})\text{vec}(\mathbf{B})$, and where $\mathbf{g}_{s,i,j} := \check{\mathbf{u}}_{s,i} \otimes \check{\mathbf{v}}_{s,j}^*$ denotes the combined beam-space representation of the beamforming vectors corresponding to the i-th RF chain at the BS and the j-th RF chain at the UE.

Next, we introduce a slight generalization of the scheme illustrated so far, by allowing the repetition of the PN beacon sequences $S \geq 1$ times during each beacon slot (see Fig. 2 (a)). Hence, each beacon slot consists of S subslots, each of which contains a PN sequence transmission as explained above. Since beamforming is implemented in the analog RF domain, it is typically impractical to switch the beamforming pattern during the beacon slot. Hence, we assume that the combined beamforming vector $\mathbf{g}_{s,i,j}$ remains constant over the S subslots, whereas $\check{\mathbf{H}}_s$ changes because of the Doppler shifts ν_l. Over different beacon slots, in contrast, the beamforming vector $\mathbf{g}_{s,i,j}$ changes periodically according to the pre-defined pseudo-random beamforming codebook $\mathcal{U}_{s,i} \times \mathcal{V}_{s,j}$ as said before. In order to accommodate for this extension, with a slight abuse of notation, we index the received subslots belonging to the s-th beacon slot as $sS + s'$, $s' \in [S]$, where the index s labels the beacon slots and the index s' labels the subslots inside each beacon slot. It follows that the received signal through the i-th RF chain at the BS and the j-th RF chain at the UE after matched filtering (refer to (20)) can be written as

$$y_{sS+s',i,j}[k] = \mathbf{g}_{s,i,j}^{\mathsf{T}} \check{\mathbf{H}}_{sS+s'} \mathbf{c}_k^i + z_{sS+s',j}^c[k]. \tag{21}$$

As anticipated in Section I, in order to ensure a robust scheme with respect to fast channel variations [10], we focus on the second-order statistics of the channel coefficients. More specifically, we accumulate the energy at the output of the matched filter across all the $\tilde{N}_c$ discrete delay taps, by computing the following quadratic measurements in (22), where the first two terms correspond to the useful signal and noise contributions, respectively,

[6] Note that, in our scheme, the beamforming vectors result from a uniform linear combining of the DFT vectors. Further optimization of the beamforming vectors with non-uniform combining [35] is possible. However, this goes outside the scope of the present work and it is left for future investigation.

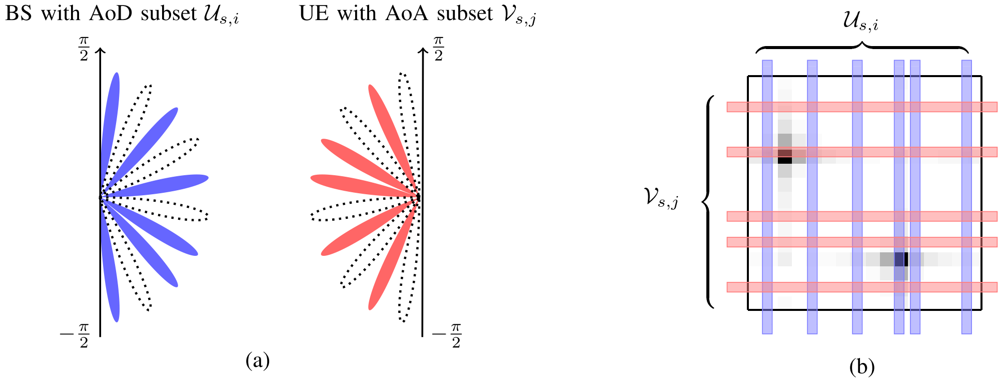


Fig. 3: *(a) Illustration of the subset of AoA-AoDs at time slot s probed by the i-th beacon stream transmitted by the BS and received by the j-th RF chain of the UE, for $M = N = 10$. The AoD subset is given by $\mathcal{U}_{s,i} = \{1, 3, 4, 6, 8, 10\}$ (numbered counterclockwise) with beamforming vector $\check{\mathbf{u}}_{s,i} = \frac{1}{\sqrt{6}}[1, 0, 1, 0, 1, 0, 1, 1, 0, 1]^{\mathsf{T}}$. The AoA subset is given by $\mathcal{V}_{s,j} = \{2, 4, 5, 7, 9\}$ (numbered counterclockwise) with receive beamforming vector $\check{\mathbf{v}}_{s,j} = \frac{1}{\sqrt{5}}[0, 1, 0, 1, 1, 0, 1, 0, 1, 0]^{\mathsf{T}}$. (b) The channel gain matrix $\check{\mathbf{\Gamma}}$ (with two strong MPCs indicated by the dark spots) measuring along $\mathcal{V}_{s,j} \times \mathcal{U}_{s,i}$.*

$$\begin{aligned}
\check{q}_{sS+s',i,j} &= \sum_{k=1}^{\check{N}_c} |y_{sS+s',i,j}[k]|^2 \\
&= \mathbf{g}^{\mathsf{T}}_{s,i,j} \left(\sum_{l=1}^{L} \check{\mathbf{h}}_{sS+s',l} \check{\mathbf{h}}^{\mathsf{H}}_{sS+s',l} \sum_{k=1}^{\check{N}_c} E_{\mathtt{dim}} |R^{x^l}_{i,i}(kT_c - \tau_l)|^2 \right) \mathbf{g}^*_{s,i,j} + \sum_{k=1}^{\check{N}_c} |z^c_{sS+s',j}[k]|^2 && \text{(22a)} \\
&\quad + 2\mathrm{Re}\left\{ \sum_{k=1}^{\check{N}_c} \mathbf{g}^{\mathsf{T}}_{s,i,j} \check{\mathbf{H}}_{sS+s'} \mathbf{c}^i_k (z^c_{sS+s',j}[k])^* \right\} && \text{(22b)} \\
&\quad + \mathbf{g}^{\mathsf{T}}_{s,i,j} \left(\sum_{l \neq l'}^{L} \check{\mathbf{h}}_{sS+s',l} \check{\mathbf{h}}^{\mathsf{H}}_{sS+s',l'} \sum_{k=1}^{\check{N}_c} E_{\mathtt{dim}} R^{x^l}_{i,i}(kT_c - \tau_l) R^{x^l}_{i,i}(kT_c - \tau_{l'})^* \right) \mathbf{g}^*_{s,i,j} && \text{(22c)}
\end{aligned}$$

the third term is the signal×noise contribution, typically arising in these forms of quadratic non-coherent detection [33], and the fourth term contains the self-interference due to non-perfect orthogonality of the transmitted PN sequences.

To obtain more reliable statistics, we average the terms as in (22) over the S subslots within each beacon slot. Note that the channel coefficients corresponding to different multipath components are mutually independent, as well as the channel coefficients and the noise samples. Consequently, the terms (22b) and (22c) have a zero mean. Thus, when the number of subslots S is large, these terms contribute negligibly to the observation. As a result, we can write the measurement obtained from each beacon slot s in the form of (23), where $w_{s,i,j}$ represents the contribution of the terms in (22b) and (22c).

In the proposed system all PN sequences have equal energy given by $R^x(0) = R^x_{i,i}(0) = N_c$, $\forall i \in [M_{\mathrm{RF}}]$. Since $\nu_l T_c \ll 1$, the energy loss due to mismatched filtering for the original PN signal $x_{s,i}(t)$ instead of the (unknown) Doppler-modulated signal $x^l_{s,i}(t)$ in (11) is also very small, i.e., $|R^{x^l}_{i,i}(0)| \approx R^x(0)$. Furthermore, due to the peakiness of the PN sequence autocorrelation function, the samples $|R^{x^l}_{i,i}(kT_c - \tau_l)|^2$ in (23a) are all almost zero except for those indices k for which $|kT_c - \tau_l| \leq T_c$. Hence, we can write

$$\sum_{k=1}^{\check{N}_c} |R^{x^l}_{i,i}(kT_c - \tau_l)|^2 \approx |R^{x^l}_{i,i}(0)|^2 \approx |R^x(0)|^2 \tag{24}$$

for all $i \in [M_{\mathrm{RF}}]$. Define the $N \times M$ matrix $\mathbf{\Gamma}$ with elements

$$[\mathbf{\Gamma}]_{n,m} = \sum_{l=1}^{L} \mathbb{E}\left[|[\check{\mathsf{H}}_{sS+s',l}]_{n,m}|^2 \right] \cdot \frac{E_{\mathtt{dim}}}{\kappa_u \kappa_v} \cdot |R^{x^l}(0)|^2, \tag{25}$$

and notice that, because of the WSS assumption, the second moments $\mathbb{E}\left[|[\check{\mathsf{H}}_{sS+s',l}]_{n,m}|^2 \right]$ are independent of the slot time index $sS + s'$. By construction, $[\mathbf{\Gamma}]_{n,m}$ yields the received signal energy from the virtual propagation paths corresponding to the quantized AoA-AoD pair in location (n, m) of the discretized angle domain $\Phi \times \Theta$. Also, it is well-known that the noise samples at the output of the matched filter with autocorrelation function $R^x(t)$

$$
\begin{aligned}
q_{s,i,j} &= \frac{1}{S}\sum_{s'=1}^{S} \check{q}_{sS+s',i,j} \\
&= \mathbf{g}_{s,i,j}^{\mathsf{T}} \left(\sum_{l=1}^{L}\left(\frac{1}{S}\sum_{s'=1}^{S} \check{\mathbf{h}}_{sS+s',l}\check{\mathbf{h}}_{sS+s',l}^{\mathsf{H}}\right)\sum_{k=1}^{\check{N}_c} E_{\mathtt{dim}}|R_{i,i}^{x^l}(kT_c-\tau_l)|^2\right)\mathbf{g}_{s,i,j}^{*} && \text{(23a)}\\
&+\sum_{k=1}^{\check{N}_c}\left(\frac{1}{S}\sum_{s'=1}^{S}|z_{sS+s',j}^{c}[k]|^2\right)+w_{s,i,j}, && \text{(23b)}
\end{aligned}
$$

have variance $N_0R^x(0)$ [33]. Hence, we can assume the approximation

$$\frac{1}{S}\sum_{s'=1}^{S}|z_{sS+s',j}^{c}[k]|^2 \approx \mathbb{E}[|z_{sS+s',j}^{c}[k]|^2] = N_0R^x(0), \tag{26}$$

which holds true in the limit of large S. Using the definition of $\mathbf{\Gamma}$ in (25), the approximations (24) and (26), and defining the NM-dimensional binary vectors

$$\mathbf{b}_{s,i,j} := \mathbf{g}_{s,i,j}\sqrt{\kappa_u\kappa_v} = \mathbf{1}_{\mathcal{U}_{s,i}} \otimes \mathbf{1}_{\mathcal{V}_{s,j}}, \tag{27}$$

we can eventually write (23) in the convenient compact form

$$q_{s,i,j} = \mathbf{b}_{s,i,j}^{\mathsf{T}}\mathrm{vec}(\mathbf{\Gamma}) + \check{N}_cN_0R^x(0) + \widetilde{w}_{s,i,j}, \tag{28}$$

where $\widetilde{w}_{s,i,j}$ collects the error term $w_{s,i,j}$ plus all the residual errors incurred by the above approximations.[7] Notice that $\mathbf{b}_{s,i,j}$ contains "ones" in the positions corresponding to the discrete angle support $\mathcal{U}_{s,i} \times \mathcal{V}_{s,j}$ from the beamforming codebook, while it contains "zeros" everywhere else. Hence, the inner product $\mathbf{b}_{s,i,j}^{\mathsf{T}}\mathrm{vec}(\mathbf{\Gamma})$ corresponds effectively to collecting all the signal energy received from the AoA-AoD pairs indexed by the angular support $\mathcal{U}_{s,i} \times \mathcal{V}_{s,j}$. An example of the probing geometry is illustrated in Fig. 3 (b).

In order to gain insight on the role of the algorithm parameters $\kappa_u, \kappa_v, M_{\mathrm{RF}}, N_{\mathrm{RF}}$, and T_c, it is useful to compare the SNR before beamforming (BBF) with the SNR associated to each of the measurements in (28). We define the SNR BBF as

$$\mathsf{SNR}_{\mathrm{BBF}} = \frac{P_{\mathtt{tot}}\sum_{l=1}^{L}\gamma_l}{N_0B}. \tag{29}$$

This is the ratio of the total received signal power (summing over all the multipath components) over the total noise power at the receiver baseband processor input, with total bandwidth B. As mentioned before, one of the challenges of BA and in general communication at mmWaves is that the SNR before beamforming $\mathsf{SNR}_{\mathrm{BBF}}$ in (29) is typically very low. The average SNR of the measurements in (28), with average taken over the randomness of the beam codebook and the channel, can be qualitatively quantified as

$$\mathsf{SNR}_{\mathrm{MEA}} = \frac{P_{\mathtt{tot}}T_c\sum_{l=1}^{L}\gamma_l \cdot MN}{\kappa_u\kappa_vM_{\mathrm{RF}}N_{\mathrm{RF}}N_0}. \tag{30}$$

This quantity is explained as follows: the energy per chip $P_{\mathtt{tot}}T_c$ is uniformly spread over the angular fraction $\kappa_u\kappa_v/(MN)$ and over the $M_{\mathrm{RF}}N_{\mathrm{RF}}$ measurements obtained in each beacon slot. Comparing (29) and (30) prompts to the following qualitative observations: i) by making the product $\kappa_u\kappa_v$ large, we explore simultaneously more angle directions, but the signal power is spread over a broader angle such that the SNR per measurement decreases. Therefore, we expect the existence of an exploration/exploitation trade-off with respect to the product $\kappa_u\kappa_v$ (as noticed in [4]). ii) The scheme gathers $M_{\mathrm{RF}}N_{\mathrm{RF}}$ new measurements for each beacon slot, but the SNR per measurements decreases with $M_{\mathrm{RF}}N_{\mathrm{RF}}$. Hence, a similar exploitation/exploration trade-off exists with respect to the number of RF chains used in the BA algorithm (see also [4]). iii) By making T_c larger than $1/B$, the signal power is effectively concentrated in a bandwidth $1/T_c < B$. This energy accumulation in the frequency domain improves the SNR per measurement. However, given a total pilot signal duration, increasing T_c decreases the number of chips of the PN sequence such that the cross-interference between PN sequences and their delayed versions increases. Therefore, there exists a trade-off between energy concentration in the frequency domain and self-interference in the system, reflected in the variance of the error term $\widetilde{w}_{s,i,j}$.

C. Path Strength Estimation via Non-Negative Least Squares

After T beacon slots, the UE obtains a total number of $M_{\mathrm{RF}}N_{\mathrm{RF}}T$ equations, given by

$$\mathbf{q} = \mathbf{B}\cdot\mathrm{vec}(\mathbf{\Gamma}) + \check{N}_cN_0R^x(0)\cdot\mathbf{1} + \widetilde{\mathbf{w}}, \tag{31}$$

where the vector $\mathbf{q} = [q_{1,1,1}, \dots q_{1,M_{\mathrm{RF}},N_{\mathrm{RF}}}, \dots, q_{T,M_{\mathrm{RF}},N_{\mathrm{RF}}}]^{\mathsf{T}} \in \mathbb{R}^{M_{\mathrm{RF}}N_{\mathrm{RF}}T}$ consists of all $M_{\mathrm{RF}}N_{\mathrm{RF}}T$ measurements achieved as in (28), $\mathbf{B} = [\mathbf{b}_{1,1,1}, \dots, \mathbf{b}_{1,M_{\mathrm{RF}},N_{\mathrm{RF}}}, \dots, \mathbf{b}_{T,M_{\mathrm{RF}},N_{\mathrm{RF}}}]^{\mathsf{T}} \in$

[7] We should point out here again that the goodness of such approximations reflects in the variance of the error term $\widetilde{w}_{s,i,j}$. The fact that our algorithm works very well in very low SNR conditions (see Section V) confirms that all the working assumptions made here are valid and justified.

$\mathbb{R}^{M_{\rm RF}N_{\rm RF}T\times MN}$ is uniquely defined by the pseudo-random beamforming codebook of the BS and the local beamforming codebook of the UE, and $\widetilde{\mathbf{w}} \in \mathbb{R}^{M_{\rm RF}N_{\rm RF}T}$ denotes the residual error.

In order to identify the strong AoA-AoD quantized directions, the UE needs to estimate the MN-dim vector $\text{vec}(\mathbf{\Gamma})$ from the $M_{\rm RF}N_{\rm RF}T$-dim observation (31) in presence of the measurement noise $\widetilde{\mathbf{w}}$, where in general, MN is significantly larger than $M_{\rm RF}N_{\rm RF}T$. There are a great variety of algorithms to solve (31) in the Least-Squares sense. The key observation here is that $\mathbf{\Gamma}$ is sparse (by the sparse nature of mmWave channels) and non-negative (by the second-order statistic construction of our scheme). As discussed in our previous work [4], recent results in CS show that when the underlying parameter $\mathbf{\Gamma}$ is non-negative, the simple non-negative constrained *Least Squares* (LS) given by

$$\mathbf{\Gamma}^\star = \mathop{\arg\min}_{\mathbf{\Gamma}\in\mathbb{R}_+^{N\times M}} \|\mathbf{B}\cdot\text{vec}(\mathbf{\Gamma}) + \check{N}_c N_0 R^x(0)\cdot\mathbf{1} - \mathbf{q}\|^2, \tag{32}$$

is sufficient to yield a sparse solution $\mathbf{\Gamma}^\star$ [31, 32], without the need for an explicit sparsity-promoting regularization term in the objective function as for example in the classical LASSO algorithm [37]. The (convex) optimization problem (32) is generally referred to as *Non-Negative Least Squares* (NNLS), and has been well investigated in the literature. As discussed in [31], NNLS implicitly performs ℓ_1-regularization and promotes the sparsity of the resulting solution provided that the measurement matrix $\mathbf{B}$ satisfies the $\mathcal{M}^+$-criterion [32], i.e., there exits a vector $\mathbf{d} \in \mathbb{R}_+^{M_{\rm RF}N_{\rm RF}T}$ such that $\mathbf{B}^{\sf T}\mathbf{d} > 0$. In our case, this criterion can be simply interpreted as the fact that the set of $M_{\rm RF}N_{\rm RF}T$ measurement beam patterns should hit all the MN AoA-AoD pairs at least once, which is almost fully satisfied in our scheme because of the random finger-shaped beam patterns, also because of the pseudo-random property of the designed beamforming codebook.

In terms of numerical implementation, the NNLS can be posed as an unconstrained LS problem over the positive orthant and can be solved by several efficient techniques such as Gradient Projection, Primal-Dual techniques, etc., with an affordable computational complexity [38], which is generally significantly less than conventional CS algorithms for problems of the same size and sparsity level. We refer to [39, 40] for the recent progress on the numerical solution of NNLS and a discussion on other related work in the literature.

V. PERFORMANCE EVALUATION

We consider a system with $M = 32$ antennas, $M_{\rm RF} = 3$ RF chains at the BS, and $N = 32$ antennas, $N_{\rm RF} = 2$ RF chains at a generic UE. We assume a short preamble structure used in IEEE 802.11ad [20, 41], where the beacon slot is of duration $t_0 S = 1.891\,\mu$s. The system is assumed to work at $f_0 = 70$ GHz, has a maximum available bandwidth of $B = 1.76$ GHz, namely, each beacon slot amounts to more than 3200 chips as in [20, 21]. We assume the channel contains $L = 3$ links given by $(\gamma_l = 1, \eta_1 = 100)$, $(\gamma_2 = 0.6, \eta_2 = 10)$ and $(\gamma_3 = 0.6, \eta_3 = 0)$, where γ_l denotes the scatterer strength, η_l indicates the strength ratio between the LOS and the NLOS propagation as in (4). Thus, the first scatterer can be roughly regarded as the LOS path, while the remaining scatterers represent the NLOS paths. This is consistent with the practical mmWave MIMO channel measurements in [27], where the relative power level of the NLOS path is around 10 dB lower than the desired LOS path. We assume that the relative speed Δv_l for each path is in the range $0 \sim 8$ m/s. We announce a success if the location of the strongest component in $\mathbf{\Gamma}^\star$ (see (32)) coincides with the LOS path.[8]

In the following simulations,[9] we evaluate the performance of our time-domain BA scheme according to three viewpoints: i) We study the effect of various scheme parameters on the achieved BA probability; ii) We show the superiority of our proposed scheme in comparison with other recently proposed time-domain BA schemes [20, 21]; iii) We consider the effectiveness of the BA scheme in the context of single-carrier modulation. To tackle the latter aspect we compute upper and lower bounds on the ergodic achievable rate for the effective SISO channel between the BS and the UE after BA. These bounds show that BA yields essentially a frequency-flat channel even when the original channel has multiple multipath components. Also, the effective SNR of the channel after BA is quite large. Therefore, single-carrier modulation with standard timing and carrier synchronization and without time-domain equalization works very well.

A. Success Probability of the Proposed BA Scheme

Dependence on the beam spreading factors (κ_u, κ_v). As discussed at the end of Section IV-B (see also our previous work [4, 28]), the trade-off between the angle exploration of the measuring matrix $\mathbf{B}$ and the SNR of received measurements is illustrated in Fig. 4 (a). Increasing the angular spreading factor from $\kappa_u = \kappa_v = 4$ to $\kappa_u = \kappa_v = 8$ improves the performance. However, the performance keeps degrading when (κ_u, κ_v) are increased to $\kappa_u = \kappa_v = 16, 22$.

Dependence on the PN sequence length N_c and robustness to Doppler shifts. In general, larger PN

[8] In the case that there is no LOS link, one can announce a success if the location of the strongest component in $\Gamma^\star$ coincides with the central AoA-AoD of the strongest scatterer cluster.

[9] We will use `lsqnonneg.m` in MATLAB©to solve the NNLS optimization problem in (32). Also, for simplicity, in our simulations, we assume that the sizes of the beamforming codebooks given by $\frac{1}{M_{\rm RF}}|\mathcal{C}_T|$ and $\frac{1}{N_{\rm RF}}|\mathcal{C}_R|$ on both sides, are the same as the number of effective beam training beacon slots T. In practical implementation, however, the BS codebook size should be fixed and used periodically, since it is shared to all UEs in advance; while the local beamforming codebook for each UE can be set to any size depending on the individual UE.

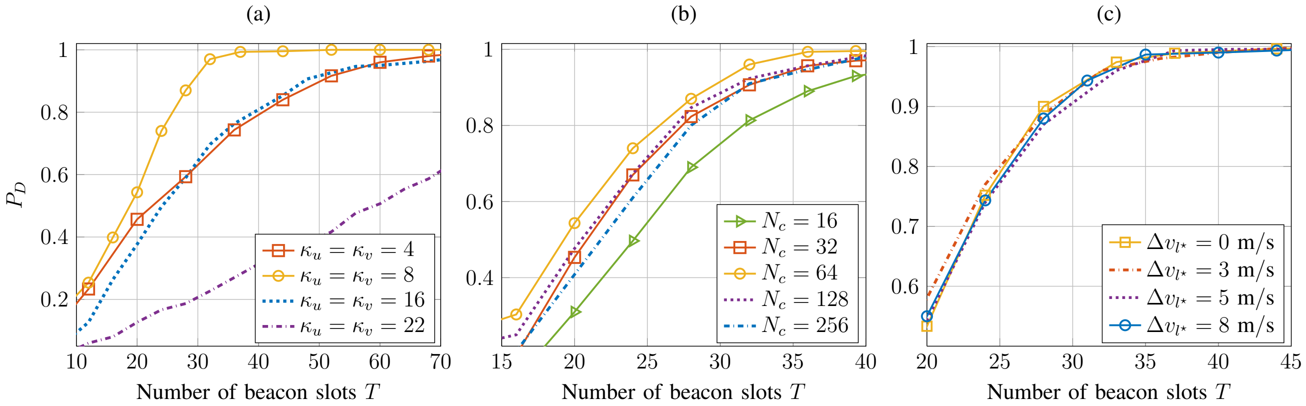


Fig. 4: *Detection probability P_D of the proposed time-domain scheme with respect to (a) different power spreading factors (κ_u, κ_v), where $N_c = 64$ and the relative speed of the strongest path $\Delta v_{l^\star} = 5$ m/s; (b) different PN sequence lengths N_c, where $\kappa_u = \kappa_v = 8$ and the relative speed of the strongest path $\Delta v_{l^\star} = 5$ m/s; (c) different relative speed values of the strongest path $\Delta v_{l^\star}$, where $\kappa_u = \kappa_v = 8$, $N_c = 64$. In all the above cases, $M = N = 32$, $M_{RF} = 3$, $N_{RF} = 2$, $B' = B$, $\mathsf{SNR}_{BBF} = -14$ dB.*

sequence length N_c provides better correlation properties, such that different pilot streams can be well separated at the UE. However, increasing N_c increases the whole duration $t_0 = N_c T_c$ of the transmitted signal. Thus, because of the Doppler shift, the received PN sequence undergoes larger phase rotation of the chips. This rotation degrades the PN sequence correlation property. This is illustrated in Fig. 4 (b). Increasing the PN sequence length N_c from $N_c = 16$ to $N_c = 32, 64$ improves the performance of the proposed scheme. However, the performance degrades slightly when N_c is increased to $N_c = 128, 256$. In general, our scheme is highly insensitive to the Doppler spread between different multipath components, as illustrated in Fig. 4 (c). For example, varying the speed difference between the paths from 0 to 8 m/s, the BA success probability remains virtually unchanged. This provides a significant advantage with respect to schemes based on OFDM signaling, which is known to be fragile to uncompensated Doppler shifts yielding inter-carrier interference.

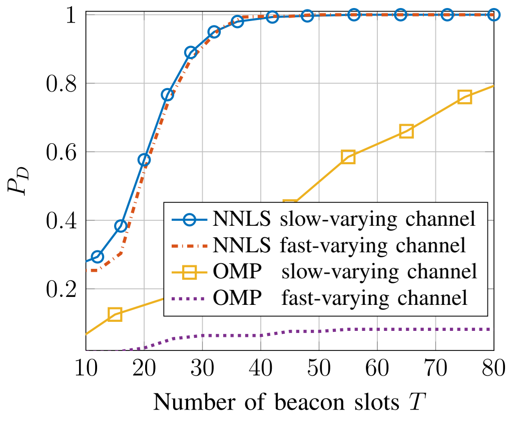


Fig. 5: *Comparison of the proposed scheme based on NNLS with that in [20, 21] based on OMP for both slow-varying channels (i.e., when the instantaneous channel coefficients are time-invariant) and fast-varying channels (i.e., when the instantaneous channel coefficients decorrelate almost completely from slot to slot due to the large Doppler spread), where $M = N = 32$, $M_{RF} = 3$, $N_{RF} = 2$, $\kappa_u = \kappa_v = 8$, $B' = B$, $N_c = 64$, $\mathsf{SNR}_{BBF} = -14$ dB.*

Comparison with other time-domain methods. Fig. 5 compares the performance of our proposed scheme with a recently proposed time-domain approach [20, 21] based on the *Orthogonal Matching Pursuit* (OMP) CS technique. The approach in [20, 21] assumes that the channel vector coefficients remain constant over the whole training stage (in other words, it assumes a completely stationary situation with zero Doppler shifts). It can be seen from Fig. 5 that the proposed scheme exhibits much more robust performance with respect to the channel time-variations whereas the approach in [20, 21] fails when the channel is fast time-varying.

Remark *1:* In all the simulations so far, for simplicity, we have considered path delays equal to integer multiples of the chip duration, namely, $\tau_l = G_l \cdot T_c$, for some integer G_l. In such cases, the chip pulse shape with any arbitrary square-root Nyquist pulse [33] yields the same performance since the samples of any Nyquist pulse at the output of the matched filter (see, e.g., (14)) are zero at all integer multiples of T_c except 0. In practice, however, the path delays are not integer multiple of T_c and can be generally written as $\tau_l = G_l \cdot T_c + \Delta\tau_l$ for some $0 < \Delta\tau_l < T_c$, referred to as the delay *fractional part*. In general, this is not an issue during the data communication phase since the delays are well compensated by suitable synchronization at the receiver, but it may affect the performance of our proposed BA since it is unrealistic to assume any proper synchronization during the BA. As a result, in the presence of non-null delay fractional parts, the performance of our scheme may depend on the specific

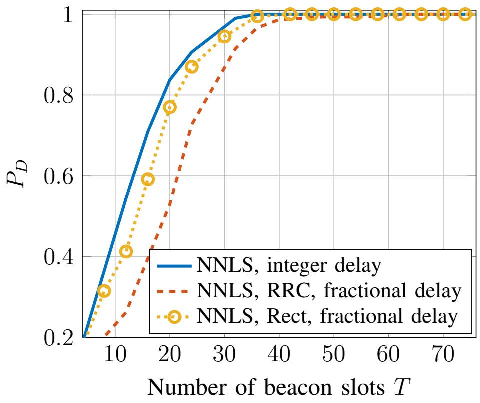


Fig. 6: *Illustration of pulse shaping effect, where $M = N = 32$, $M_{RF} = 3$, $N_{RF} = 2$, $\kappa_u = \kappa_v = 8$, $B' = B$, $N_c = 64$, $\mathsf{SNR}_{BBF} = -14$ dB, the relative speed of the strongest path $\Delta v_{l^\star} = 5$ m/s.*

chip pulse used. In order to investigate this effect, we perform numerical simulations with two different pulse shapes: root-raised-cosine (RRC) and rectangular (Rect) pulses [33]. Fig. 6 illustrates the simulation results for random fractional delays. As expected, we see a slight performance degradation compared with the ideal integer delay curve (an additional $5 \sim 10$ slots for $P_D \geq 0.95$). However, the effect of non-integer delays and of the specific chip pulse shape is rather small. Furthermore, it is observed that the Rect pulse yields less degradation than RRC.

B. Effectiveness of Single-Carrier Modulation

After running the BA protocol as described in Section IV, and assuming that the strongest multipath component is correctly identified, we denote such component as $l^\star$-th. Hence, the estimated beamforming vectors for the data transmission are given by $\mathbf{u}_{l^\star} = \mathbf{F}_M \check{\mathbf{u}}_{l^\star}$ at the BS and $\mathbf{v}_{l^\star} = \mathbf{F}_N \check{\mathbf{v}}_{l^\star}$ at the UE respectively, where $\check{\mathbf{u}}_{l^\star} \in \mathbb{C}^M$ is an all-zero vector with a 1 at the component corresponding to the AoD of the $l^\star$-th scatterer, and $\check{\mathbf{v}}_{l^\star} \in \mathbb{C}^N$ is an all-zero vector with a 1 at the component corresponding to the AoA of the $l^\star$ scatterer. In this section we focus on the data transmission phase under the beam alignment assumption. During the data transmission phase, a standard single-carrier linear modulated signal consisting of N_d information symbols is used. The complex baseband signal is given by $x(t) = \sum_{n=1}^{N_d} \sqrt{P_{\mathtt{tot}} T_d} \cdot d_n p_r(t - nT_d)$, where $p_r(t)$ denotes a unit-energy square-root Nyquist pulse shaping filter (e.g., a RRC pulse), $T_d = 1/B$ indicates the symbol interval, and $\{d_n\}$ is the sequence of unit-energy modulation symbols, belonging to a suitable modulation constellation [33]. From (7) and (8), the received signal including the transmit and receive beamforming vectors $(\mathbf{v}_{l^\star}, \mathbf{u}_{l^\star})$ is given by

$$\hat{y}(t) = \int \mathbf{v}_{l^\star}^{\mathsf{H}} \mathsf{H}(t,\tau) \mathbf{u}_{l^\star} x(t-\tau) d\tau + z(t)$$

$$= \sum_{l=1}^{L} \sum_{n=1}^{N_d} C_l d_n p_r(t - nT_d - \tau_l) e^{j2\pi(\check{\nu}_l + \nu_l n T_d)} + z(t), \tag{33}$$

where $C_l := \sqrt{P_{\mathtt{tot}} T_d} \rho_l \mathbf{v}_{l^\star}^{\mathsf{H}} \mathbf{a}_{\mathrm{R}}(\phi_l) \mathbf{a}_{\mathrm{T}}(\theta_l)^{\mathsf{H}} \mathbf{u}_{l^\star}$. The receiver uses standard timing and carrier synchronization with respect to the multipath component $l^\star$ selected by the BA algorithm. When BA is achieved, the SNR corresponding such $l^\star$ multipath component is quite large, since it is boosted by the combined beamforming gain of the UE and the BS. For example, in order to support a spectral 1 bit/s/Hz with practical coded modulation (e.g., using a QPSK constellation with binary coding rate $1/2$), the SNR after beamforming should be between 0 and 3 dB, depending on the coding scheme used. In these conditions, it is well-known that timing and carrier synchronization can be considered virtually ideal. Therefore, the receiver performs matched filtering with respect to the symbol pulse $p_r(t)$, sampling at epochs $t = kT_d + \tau_{l^\star}$, and symbol de-rotation by the factor $e^{-j2\pi(\check{\nu}_{l^\star} + \nu_{l^\star} nT_d)}$. It follows that the discrete-time baseband signal at the output of the matched filter and synchronizer takes on the form of (34), where $z_n^c[k]$ denotes the noise at the output of the matched filter with variance N_0,[10] and we define $\varphi(t) = \int p_r(\tau) p_r^*(\tau - t) d\tau$. In (34) (a) we used the fact that since $p_r(t)$ is a square-root Nyquist pulse, then $\varphi(\bar{k} \cdot T_d)$ is equal to 1 for $\bar{k} = 0$ and is zero otherwise. The first term in (34) corresponds to the desired symbol d_k multiplied by an overall channel coefficient $C_{l^\star}$ that contains the beamforming gain achieved by BA, whereas the last two terms correspond the inter-symbol interference and noise, respectively.

The resulting SNR after beamforming $\mathsf{SNR}_{\mathrm{ABF}}$ is given by (35), where in (35) (a) we used the fact that $\varphi(t) \approx 0$ for $|t| > T_d$, thus, $\sum_{n \in [N_d]} \mathbb{E}[|C_l \varphi((k-n)T_d + \tau_{l^\star} - \tau_l)|^2] \lesssim \mathbb{E}[|C_l|^2]$, in (35) (b) we used the fact that the interference caused by the other paths is negligible (compared with the noise floor of the receiver) since for the paths whose AoA-AoD is away from the beamforming directions the SNR is even lower than the isotropic SNR $\mathsf{SNR}_{\mathrm{BBF}}$ defined in (29). Finally, in (35) (c) we used the fact that the dominant path $l^\star$ has nearly full beamforming gain MN. It is seen from (35) that $\mathsf{SNR}_{\mathrm{ABF}}$ is around MN times larger than $\mathsf{SNR}_{\mathrm{BBF}}$. This justifies the assumption of nearly ideal timing and carrier recovery.

Consequently, the ergodic achievable rate in (34) can be upper and lower bounded as (36) and (37), respectively [42]. The upper bound (36) is obtained via the *Maximum Ratio Combining* for the case where all the delayed versions of the transmitted signal are separately observable (this is sometimes referred to as "matched filter upper bound"). The lower bound is actually achieved by a simple

[10] As usual, we assume that the symbol pulse has unit energy, i.e., $\int |p_r(t)|^2 dt = 1$, therefore the noise sample has the same variance of the noise power spectral density N_0 [33].

$$\begin{aligned} y(t)|_{t=kT_d+\tau_{l^\star}} = & \underbrace{\sum_{n=1}^{N_d} d_n C_{l^\star} \varphi\left[(k-n)T_d\right]}_{\overset{(a)}{=} d_k C_{l^\star}} \\ & + \sum_{n=1}^{N_d} d_n \sum_{l \neq l^\star} C_l \varphi\left[(k-n)T_d + \tau_{l^\star} - \tau_l\right] \cdot e^{j2\pi(\check{\nu}_l - \check{\nu}_{l^\star} + (\nu_l - \nu_{l^\star})nT_d)} + \sum_{n=1}^{N_d} z_n^c[k], \end{aligned} \tag{34}$$

$$\begin{aligned} \mathsf{SNR}_{\text{ABF}} = & \frac{\mathbb{E}[|C_{l^\star} d_k|^2]}{\sum_{l \neq l^\star} \sum_{n \in [N_d]} \mathbb{E}[|d_n|^2] \mathbb{E}[|C_l \varphi((k-n)T_d + \tau_{l^\star} - \tau_l)|^2] + \mathbb{E}[|z_n^c[k]|^2]} \\ \overset{(a)}{\approx} & \frac{\mathbb{E}[|d_k|^2] \times \mathbb{E}[|C_{l^\star}|^2]}{\mathbb{E}[|d_n|^2] \times \sum_{l \neq l^\star} \mathbb{E}[|C_l|^2] + \mathbb{E}[|z_n^c[k]|^2]} \\ \overset{(b)}{\approx} & \frac{\mathbb{E}[|d_k|^2] \times \mathbb{E}[|C_{l^\star}|^2]}{\mathbb{E}[|z_n^c[k]|^2]} \overset{(c)}{=} \frac{P_{\mathtt{tot}} \cdot \gamma_{l^\star} \cdot MN}{N_0 B}, \end{aligned} \tag{35}$$

$$R^{ub^\star} = \mathbb{E}\left[\log_2\left(1 + \frac{\sum_{l=1}^{L} |C_l \varphi(\tau_{l^\star} - \tau_l)|^2}{N_0}\right)\right], \tag{36}$$

$$R^{lb^\star} = \log_2\left(1 + \frac{|\mathbb{E}[C_{l^\star} \varphi(0)]|^2}{N_0 + \mathbb{V}\mathsf{ar}(C_{l^\star} \varphi(0)) + \sum_{m \in [N_d]} \sum_{l \neq l^\star} \mathbb{E}[|C_l \varphi(mT_d + \tau_{l^\star} - \tau_l)|^2]}\right). \tag{37}$$

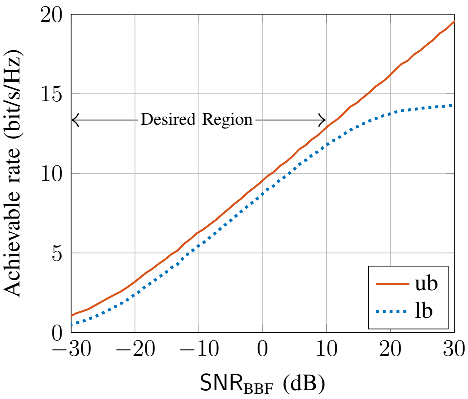


Fig. 7: *The ergodic achievable rate after a successful* Beam Alignment *using the proposed time-domain scheme, where $M = N = 32$, $B' = B$, $N_c = 64$, the relative speed of the strongest path $\Delta v_{l^\star} = 5$ m/s.*

receiver that treats all the *Inter-symbol Interference* (ISI) as a Gaussian noise.

Ergodic achievable rate bounds. In Fig. 7, we illustrate the lower and upper bounds on the achievable ergodic rate (36) (37) as a function of $\mathsf{SNR}_{\text{BBF}}$, under the assumption of successful BA, i.e., that the BA algorithm found beam indices of the strongest path.[11] It is clear that the lower bound is self-interference limited while the upper bound is not. However, the gap between the bounds is quite small in the regime of low pre-beamforming SNR ($\mathsf{SNR}_{\text{BBF}} <$ 10 dB), while the achievable ergodic spectral efficiency in this regime can be quite high, which is relevant in mmWave applications. In particular, it is important to recall here that the lower bound refers to the case of single-carrier transmission without any equalization. For example, focusing on a realistic spectral efficiency between $1 \sim 2$ bit/s/Hz, we notice that single-carrier with the proposed BA scheme and no equalization (just standard post-beamforming timing and frequency synchronization) achieves the relevant spectral efficiency in the range of $\mathsf{SNR}_{\text{BBF}}$ between -30 and -20 dB, and suffers from a very small gap with respect to the best possible equalization (given by the upper bound).

PDP before and after Beam Alignment (BA). Fig. 8 compares the average PDP of the mmWave channel with $L = 3$ multipath components before and after BA. It can be seen from Fig. 8 (a) that, before BA, the channel has a relatively large delay spread and is highly frequency selective. Moreover, since different multipath components are mixed with each other and since each one has its own delay and Doppler shift, the time-domain channel is highly time-varying. In contrast, as seen from Fig. 8 (b), after BA, the channel effectively consists of a single multipath component, thus, it is almost flat in frequency. Also, note that in contrast with the former case where different

[11] As seen before, this happens with probability ≈ 1 after a few tens of beacon slots.

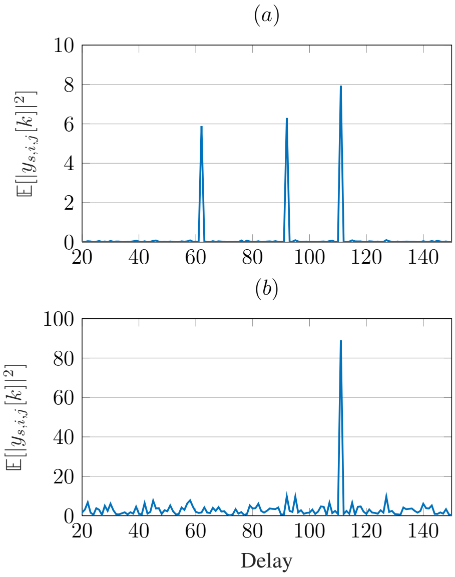


Fig. 8: *Illustration of the PDP with multipath* ($L = 3$) *channel in* (14). (a) *Before* Beam Alignment. (b) *After* Beam Alignment

multipath components were mixed with different Doppler frequencies, in the latter case the Doppler frequency of the single multipath component can be easily compensated by standard timing, frequency, and phase synchronization techniques at the receiver.

VI. Conclusion

In this paper, we proposed a novel time-domain *Beam Alignment* (BA) scheme for mmWave MIMO systems with a HDA architecture. The proposed scheme is particularly suited for single-carrier multiuser mmWave communication, where each user has access to the whole bandwidth, and all the users within the BS coverage can be trained simultaneously. We focused on the channel second-order statistics, incorporating both the random channel gains and Doppler shifts into the channel matrix to further capture the realistic features of mmWave channels. We applied the recently developed *Non-Negative Least Squares* (NNLS) technique to efficiently find the strongest path for each user. Simulation results showed that the proposed scheme incurs moderately low training overhead, achieves very good robustness to fast time-varying channels, and it is very robust to large Doppler shifts among different multipath components. Furthermore, we have shown that the multipath channel after BA reduces essentially to a single giant tap. Hence, single-carrier signaling can perform very efficiently and requires just standard timing and frequency synchronization (that works well at high SNR after beamforming) while it requires no time-domain equalization. This makes the proposed BA scheme together with single-carrier signaling a strong contender for future mmWave systems, especially in outdoor mobile scenarios.

References

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] J. Zhao, X. Wang, and H. Viswanathan, "Directional beam alignment for millimeter wave cellular systems," in *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, June 2016, pp. 619–628.

[4] X. Song, S. Haghighatshoar, and G. Caire, "A scalable and statistically robust beam alignment technique for mm-Wave systems," *IEEE Trans. on Wireless Comm.*, vol. PP, pp. 1–1, 2018.

[5] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 134–141, 2017.

[6] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[7] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[8] P. Schniter and A. Sayeed, "Channel estimation and precoder design for millimeter-wave communications: The sparse way," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 273–277.

[9] J. Rodríguez-Fernández, N. González-Prelcic, K. Venugopal, and R. W. Heath Jr, "Frequency-domain compressive channel estimation for frequency-selective hybrid mmWave MIMO systems," *arXiv preprint arXiv:1704.08572*, 2017.

[10] P. A. Eliasi, S. Rangan, and T. S. Rappaport, "Low-rank spatial channel estimation for millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2748–2759, 2017.

[11] J. Palacios, D. D. Donno, and J. Widmer, "Tracking mm-Wave channel dynamics: Fast beam training strategies under mobility," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, May 2017, pp. 1–9.

[12] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4391–4403, October 2013.

[13] S. Haghighatshoar and G. Caire, "The beam alignment problem in mmWave wireless networks," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 741–745.

[14] IEEE P802.11ad, Part 11, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 3: enhancements for very high throughput in the 60 GHz band," *IEEE Computer Society*, 2012.

[15] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 5, pp. 831–846, 2014.

[16] M. Kokshoorn, H. Chen, P. Wang, Y. Li, and B. Vucetic, "Millimeter wave MIMO channel estimation using overlapped beam patterns and rate adaptation," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 601–616, 2016.

[17] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5689–5701, Sept 2017.

[18] M. Hussain and N. Michelusi, "Throughput optimal beam alignment in millimeter wave networks," in *2017 Information Theory and Applications Workshop (ITA)*, Feb 2017, pp. 1–6.

[19] A. Alkhateeb, G. Leus, and R. W. Heath, "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp.

2909–2913.

[20] K. Venugopal, A. Alkhateeb, N. G. Prelcic, and R. W. Heath, “Channel estimation for hybrid architecture-based wideband millimeter wave systems,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1996–2009, 2017.

[21] K. Venugopal, A. Alkhateeb, R. W. Heath, and N. G. Prelcic, “Time-domain channel estimation for wideband millimeter wave systems with hybrid architecture,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, Conference Proceedings, pp. 6493–6497.

[22] R. J. Weiler, M. Peter, W. Keusgen, and M. Wisotzki, “Measuring the busy urban 60 GHz outdoor access radio channel,” in *2014 IEEE International Conference on Ultra-WideBand (ICUWB)*, Sept 2014, pp. 166–170.

[23] V. Va, J. Choi, and R. W. Heath, “The impact of beamwidth on temporal channel variation in vehicular channels and its implications,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5014–5029, 2017.

[24] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, “Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152–1163, June 2014.

[25] S. Buzzi, C. D’Andrea, T. Foggi, A. Ugolini, and G. Colavolpe, “Single-carrier modulation versus OFDM for millimeter-wave wireless MIMO,” *IEEE Transactions on Communications*, vol. PP, no. 99, pp. 1–1, 2017.

[26] A. Nasser and M. Elsabrouty, “Frequency-selective massive MIMO channel estimation and feedback in angle-time domain,” in *2016 IEEE Symposium on Computers and Communication (ISCC)*, June 2016, pp. 1018–1023.

[27] T. Hälsig, D. Cvetkovski, E. Grass, and B. Lankl, “Statistical properties and variations of LOS MIMO channels at millimeter wave frequencies,” *arXiv preprint arXiv:1803.07768*, 2018.

[28] X. Song, S. Haghighatshoar, and G. Caire, “A robust time-domain beam alignment scheme for multi-user wideband mmWave systems,” in *WSA 2018; 22th International ITG Workshop on Smart Antennas (to be published)*, March 2018, pp. 1–7.

[29] P. Bello, “Characterization of randomly time-variant linear channels,” *IEEE Transactions on Communications Systems*, vol. 11, no. 4, pp. 360–393, 1963.

[30] A. Goldsmith, *Wireless communications*. Cambridge University Press, 2005.

[31] M. Slawski, M. Hein *et al.*, “Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization,” *Electronic Journal of Statistics*, vol. 7, pp. 3004–3056, 2013.

[32] R. Kueng and P. Jung, “Robust nonnegative sparse recovery and the nullspace property of 0/1 measurements,” *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 689–703, Feb 2018.

[33] J. G. Proakis and M. Salehi, *Digital communications*. McGraw-Hill, 2008.

[34] A. M. Sayeed, “Deconstructing multiantenna fading channels,” *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2563–2579, 2002.

[35] J. Song, J. Choi, T. Kim, and D. J. Love, “Advanced quantizer designs for FDD-based FD-MIMO systems using uniform planar arrays,” *IEEE Transactions on Signal Processing*, vol. 66, no. 14, pp. 3891–3905, July 2018.

[36] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson, and C. Roobol, “WCDMA – the radio interface for future mobile multimedia communications,” *IEEE Transactions on Vehicular Technology*, vol. 47, no. 4, pp. 1105–1118, 1998.

[37] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[38] D. P. Bertsekas and A. Scientific, *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

[39] D. Kim, S. Sra, and I. S. Dhillon, “Tackling box-constrained optimization via a new projected quasi-Newton approach,” *SIAM Journal on Scientific Computing*, vol. 32, no. 6, pp. 3548–3563, 2010.

[40] D. K. Nguyen and T. B. Ho, “Anti-lopsided algorithm for large-scale nonnegative least square problems,” *arXiv preprint arXiv:1502.01645*, 2015.

[41] E. Perahia, C. Cordeiro, M. Park, and L. L. Yang, “IEEE 802.11 ad: Defining the next generation multi-Gbps Wi-Fi,” in *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE*. IEEE, Conference Proceedings, pp. 1–5.

[42] G. Caire, “On the ergodic rate lower bounds with applications to massive MIMO,” *arXiv preprint arXiv:1705.03577*, 2017.

Xiaoshen Song (S’17) received the B.Sc. degree in Communication Engineering from Northwestern Polytechnical University, Xi’an, China, in 2013, and the M.Sc. degree in Communication and Information Systems from the Institute of Electronics, University of Chinese Academy of Sciences, Beijing, China, in 2016. Her master’s thesis focuses on video synthetic aperture radar (VideoSAR) system design and imaging algorithms. She is currently pursuing the Ph.D. degree with the Communications and Information Theory (CommIT) group at Technische Universität Berlin, Berlin, Germany. Her research interests include wireless communication, mmWave MIMO, and compressed sensing.

Saeid Haghighatshoar (S’12–M’15) received the B.Sc. degree in Electrical Engineering (Electronics) in 2007 and the M.Sc. degree in Electrical Engineering (Communication Systems) in 2009, both from Sharif University of Technology, Tehran, Iran, and the Ph.D. degree in Computer and Communication Sciences from École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2014. Since 2015, he is a postdoctoral researcher with Communications and Information Theory (CommIT) group at Technische Universität Berlin, Berlin, Germany. His research interests lie in Information Theory, Communication Systems, Wireless Communication, Optimization Theory, and Compressed Sensing.

Giuseppe Caire (S'92 – M'94 – SM'03 – F'05) was born in Torino, Italy, in 1965. He received the B.Sc. in Electrical Engineering from Politecnico di Torino (Italy), in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992 and the Ph.D. from Politecnico di Torino in 1994. He has been a post-doctoral research fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands) in 1994-1995, Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, and he is currently an Alexander von Humboldt Professor with the Electrical Engineering and Computer Science Department of the Technical University of Berlin, Germany.

He served as Associate Editor for the IEEE Transactions on Communications in 1998-2001 and as Associate Editor for the IEEE Transactions on Information Theory in 2001-2003. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society & Information Theory Society Joint Paper Award in 2004 and in 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, and the Vodafone Innovation Prize in 2015. Giuseppe Caire is a Fellow of IEEE since 2005. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications theory, information theory, channel and source coding with particular focus on wireless communications.

5

Data Communication for mmWave Multi-User MIMO

5.1 Introduction

Hybrid digital analog (HDA) beamforming is the most practical solution for mmWave communication regarding the implementation cost, performance and power efficiency. This chapter presents two HDA mmWave antenna architectures that can be regarded as two extreme cases, namely, the fully-connected (FC) architecture and the one-stream-per-subarray (OSPS) architecture. A joint performance evaluation of the initial beam alignment and the consequent data communication will be provided, such that the latter takes place by using the beam direction information obtained by the former. In addition, the power efficiency of the the two architectures will also be evaluated, which takes into account the hardware impairments, e.g., power dissipation, power backoff, etc..

5.2 Clarification of each authors' contributions

This chapter is a journal publication, which is a joint work with Thomas Kühne and Giuseppe Caire. I wrote this journal as the first author. The citation information is in below:

X. Song, T. Kühne, and G. Caire, "Fully-/Partially-Connected Hybrid Beamforming Architectures for mmWave MU-MIMO," IEEE Transactions on Wireless Communications, 2019. DOI: 10.1109/TWC.2019.2957227

All the authors contributed to this paper. I authored the channel and signaling models. I implemented the simulations for beam alignment section and data communication section. I also wrote the complete first draft (including all sections) of this paper.

Thomas Kühne authored all the hardware sections. He proposed the HDA antenna architecture model and the hardware impairment model.

Giuseppe Caire, who is my PhD supervisor, provided valuable discussions in each meeting of this work. He also did a final modification of the overall draft.

5.3 Original journal article

The following article is a reprint of the original journal paper. It is the accepted version of the paper. The copyright information is given in page xii of this thesis as well as in the first page of the reprinted paper.

Fully- / Partially-Connected Hybrid Beamforming Architectures for mmWave MU-MIMO

Xiaoshen Song, *Student Member, IEEE,* Thomas Kühne, and Giuseppe Caire *Fellow, IEEE*

Abstract—**Hybrid digital analog (HDA) beamforming has attracted considerable attention in practical implementation of millimeter wave (mmWave) multiuser multiple-input multiple-output (MU-MIMO) systems due to the low power consumption with respect to its fully digital baseband counterpart. The implementation cost, performance, and power efficiency of HDA beamforming depends on the level of connectivity and reconfigurability of the analog beamforming network. In this paper, we investigate the performance of two typical architectures that can be regarded as extreme cases, namely, the fully-connected (FC) and the one-stream-per-subarray (OSPS) architectures. In the FC architecture each RF antenna port is connected to all antenna elements of the array, while in the OSPS architecture the RF antenna ports are connected to disjoint subarrays. We jointly consider the initial beam acquisition and data communication phases, such that the latter takes place by using the beam direction information obtained by the former. We use the state-of-the-art beam alignment (BA) scheme previously proposed by the authors and consider a family of MU-MIMO precoding schemes well adapted to the beam information extracted from the BA phase. We also evaluate the power efficiency of the two HDA architectures taking into account the power dissipation at different hardware components as well as the power backoff under typical power amplifier constraints. Numerical results show that the two architectures achieve similar sum spectral efficiency, while the OSPS architecture is advantageous with respect to the FC case in terms of hardware complexity and power efficiency, at the sole cost of a slightly longer BA time-to-acquisition due to its reduced beam angle resolution.**

Index Terms—**Millimeter Waves, MU-MIMO, HDA Beamforming, Beam Acquisition, Spectral Efficiency, Power Efficiency.**

I. Introduction

Millimeter wave (mmWave) multiuser multiple-input multiple-output (MU-MIMO) communications have emerged as one of the most promising techniques for the second phase of 5G wireless systems, aimed at achieving broadband data communications at unprecedented high rates ($\geq$ 1 Gb/s per user) in very dense urban small-cell environments. The relatively underutilized mmWave spectrum (30-300 GHz) allows to achieve a target $\sim$ 1 Gb/s per data stream with $\sim$ 1 GHz signal bandwidth, provided that the system can support a spectral efficiency of about 1 bit/s/Hz. Such relatively low spectral efficiency per stream can be achieved with rather standard modulation and coding techniques (e.g., binary codes of rate $1/2$ mapped onto a QPSK constellation), when that the signal to interference plus noise ratio (SINR) at the receiver is between 0 and 3 dB (depending on the gap to capacity of the underlying code).[1]

Due to the severe isotropic pathloss incurred by mmWave frequencies, large antenna gains are required both at the base station (BS) side and the user equipment (UE) side. Fortunately, the small carrier wavelength associated with mmWave frequencies enables large antenna arrays to be packed in a small form factor, such that the required large antenna gain can be obtained using beamforming. For example, in a single-user scenario where the signal-to-noise ratio (SNR) at the receiver in isotropic propagation conditions[2] is between -30 dB ~ -20 dB (a quite realistic situation for outdoor mmWave channels), a combined Tx and Rx beamforming gain of 30 dB is needed such that, when the Tx and the Rx beams are well aligned, the resulting SNR *after beamforming* reaches the desired target (a bit above 0 dB, as argued before).

Realizing fast and accurate digitally steerable beamforming at mmWave, however, is not a trivial task. One main challenge is that the conventional full digital transceiver architecture (with one radio frequency (RF) chain per antenna element) is infeasible due to hardware cost, power consumption, and above all power dissipation in the small integrated array form factor. Each RF chain consists of (roughly speaking) analog-to-digital / digital-to-analog (A/D, D/A) converters, up / down-conversion mixers, filters, power amplifiers (PAs), and low-noise amplifiers (LNAs). It follows that a design goal for mmWave transceivers is to reduce the number of RF chains to be significantly smaller than the number of antenna array elements. For this reason, the

The authors are with the Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: xiaoshen.song@campus.tu-berlin.de).

X. Song is sponsored by the China Scholarship Council (201604910530). This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 779305 (SERENA).

[1] With ideal single-user capacity achieving codes for the Gaussian channel, we have that $\log(1 + \mathsf{SINR}) = 1$ bit/s/Hz is achieved for $\mathsf{SINR} = 1$ (i.e., 0 dB). In practice, gaps of a fraction of a dB to 3-4 dB are obtained by actual coding schemes adopted in current standards.

[2] Here the isotropic propagation conditions correspond to one active antenna at the transmitter (Tx) and one active antenna at the receiver (Rx), respectively.

concatenation of digital and analog beamforming, known as hybrid digital analog (HDA) beamforming architecture, has been widely considered. In such a context, the limited number of RF chains are used to enable the multistream baseband processing, while an analog processing is used to realize the antenna beamforming gain. A primary objective of HDA beamforming is to maximize the multiuser sum rate, while keeping the hardware costs, complexity, and power efficiency, within some desirable targets.

A. *Related Work*

A large number of works have addressed HDA beamforming for mmWave communication systems. Rather than giving a complete account of such considerable body of literature (out of scope of the present non-tutorial paper), we consider a few significant representatives and examine their proposed approaches in a critical manner. A common assumption in most of existing works is that the analog part of the HDA precoder can only utilize phase control. This phase control can be realized through either phase shifters [1–6] or lenses [7, 8]. Consequently, the problem of finding the (sub-) optimal analog and digital precoding matrices is transformed into a series of relatively complicated decomposition steps [2–6], since the underlying optimization problem is non-convex. This phase-only control assumption may somewhat reduce the hardware complexity. However, the signaling freedom is also drastically reduced and the corresponding optimization computational complexity is typically prohibitive for practical real-time implementations. These drawbacks motivate the exploration of an analog precoding architecture with both phase and amplitude controls [9, 10]. In fact, it has been demonstrated in practice that simultaneous phase and amplitude control is fully feasible at mmWaves with good accuracy, low complexity, and low cost [11, 12].

Another severe limitation appearing in several HDA beamforming works is the assumption of invariant instantaneous channel coefficients over a large time duration [1, 13, 14]. It is known that, in order to overcome the heavy signal attenuation, communication at mmWaves requires an initial beam acquisition (which we refer as beam alignment (BA)) [7, 15, 16]. The goal of BA is to find a pair of narrow beams connecting each UE with the BS.[3] Thus, the nearly invariant channel assumption only makes sense *after BA is achieved*, since once the beams are aligned, the communication occurs only through a single narrow path with small effective angular spread, whose delay and Doppler shift can be easily compensated using standard synchronization techniques [17–19]. However, before BA is achieved, the channel delay spread and time-variation can be large due to the presence of several multipath components (both the LOS and the non-LOS (NLOS) paths), each with its own delay and Doppler shift. In this case, the instantaneous channel coefficients change very fast. Any BA algorithms relying on an invariant instantaneous channel assumption are no-longer feasible, since for example, even a small motion of a few centimeters traverses several wavelengths, potentially producing multiple deep fades [20, 21].

In addition, a large number of works on HDA architectures investigated only the data communication phase and assume full channel state information (CSI) [2–6, 10, 22, 23], i.e., that the vectors of baseband complex channel coefficients at each array element are known. These works focus on the optimization of the HDA precoder using the full CSI knowledge. Unfortunately, this assumption is obviously not feasible in a realistic system. In order to acquire such coefficients, one should be able to sample each antenna element, i.e., one would need an RF chain per antenna element or exhaustively measure all elements successively. Hence, if full CSI knowledge was possible, no HDA beamforming would be needed, since we could simply implement baseband digital beamforming/multiuser precoding, which is performance-wise more efficient. As a matter of fact, it makes sense to study HDA architectures under the assumption that only a low-dimensional projection of the channel vectors can be measured by the limited number of RF chains. To this end, a hybrid precoding scheme exploiting implicit CSI (i.e., the couplings of all possible pairs of analog beamforming vectors) was proposed in [24]. However, the work in [24] (as well as in [4, 6, 10, 22, 23]) is limited to a single-user configuration and does not treat the MU-MIMO case.

It is known that MU-MIMO is superior to single-user beamforming from a network spectral efficiency perspective even under HDA, provided that the user density is rich enough such that the BS can schedule subsets of UEs to be served by spatial multiplexing with sufficient angular separation [25, 26]. Hence, this motivates us to consider the implementation of MU-MIMO schemes under realistic HDA architecture constraints. Two "extreme" HDA architectures are depicted in Fig. 1 [27]. Fig. 1 (a) shows a fully-connected (FC) architecture, where each RF antenna port is connected to all antenna elements of the array. At the other extreme, Fig. 1 (b) shows what we refer to as the one-stream-per-subarray (OSPS) architecture, where each RF antenna port is connected to a disjoint subarray. A common theme that underlies most of the HDA works is that the FC architecture outperforms the OSPS architecture only at the cost of higher hardware complexity. However, many reference works [3, 8, 10, 22, 23] ignore hardware impairments [6], such as the power dissipation and the PA nonlinear distortion. In particular, the nonlinear PAs employed at the BS can drastically distort the transmit signal when operated close to saturation [28]. To this end, a certain power backoff from the saturation power of a PA should be considered accordingly for different signaling schemes

[3] E.g., in line-of-sight (LOS) propagation, the aligned directions typically coincide with the AoA and AoD of the LoS path.

and transceiver architectures, such that the PAs can always work in their linear operating region.

B. Contributions

In this paper we overcome the shortcomings of the present literature outlined before, and comprehensively evaluate the performance of HDA architectures (in particular, as shown in Fig. 1), where we assume both amplitude and phase control for each analog path. Our main focus is on the MU-MIMO downlink, but similar and symmetric conclusions can be reached for the uplink as well. Our main contributions are summarized as follows:

1) *More general and realistic mmWave channel model.* We consider a quite general mmWave wireless channel model, taking into account the fundamental features of mmWave channels such as fast time-variation due to Doppler, frequency-selectivity, and the AoA-AoD sparsity [20, 21, 29]. The numerical results based on our proposed channel model are further verified on the 3D geometry based channel generator QuaDRiGa [30], which has become a standard tool in industrial R&D as well as in 3GPP standardization.

2) *More practical hardware impairments and power efficiency analysis.* When comparing the HDA beamforming performance of different transmitter architectures, we take into account the practical hardware impairments, particularly, the potential power dissipation of the underlying analog network components, as well as the unavoidable power backoff for the nonlinear PAs. While the former is not difficult to be compensated, the latter is highly dependent on the peak-to-average power ratio (PAPR) of the input signal, which (as illustrated in Section V) should be carefully investigated in terms of different signaling and modulating schemes. On top of the potential hardware impairments, we also evaluate the power efficiency of the most power consuming PAs with respect to different transmitter architectures. Numerical results show that the OSPS architecture with single-carrier (SC) modulation achieves the highest power efficiency.

3) *A joint evaluation of initial BA and data communication.* As mentioned before, a main limitation in most hybrid beamforming works is that they only focus on the data communication and assume full CSI. To address this issue, we consider both initial BA and consecutive data communication in this paper. We assume that the precoder in the data communication phase can only exploit a limited amount of CSI, which is obtained along the beams acquired in the BA phase. Hence, the signaling and communication procedure in our paper captures the fundamental features of practical mmWave communication.

4) *Low-complexity data transmit precoding.* In the BA phase, we use our previously proposed BA scheme [16, 18, 19], after which each UE obtains a sparse estimate of the channel gains associated to all pairs of AoA-AoD on a finely spaced discrete grid, corresponding to the Tx and Rx beamforming codebooks. For the data communication phase, we consider three alternative precoding options on top of the effective channel after the BA phase. These are referred to as beam steering (BST), analog maximum ratio transmission (MRT), and joint analog maximum ratio and baseband zeroforcing (MR-ZF), respectively. The proposed schemes are very suitable for practical implementations due to the low-time-overhead and low-complexity. In particular, the MR-ZF precoding scheme proposed in this paper outperforms the state-of-the-art counterparts in the literature.

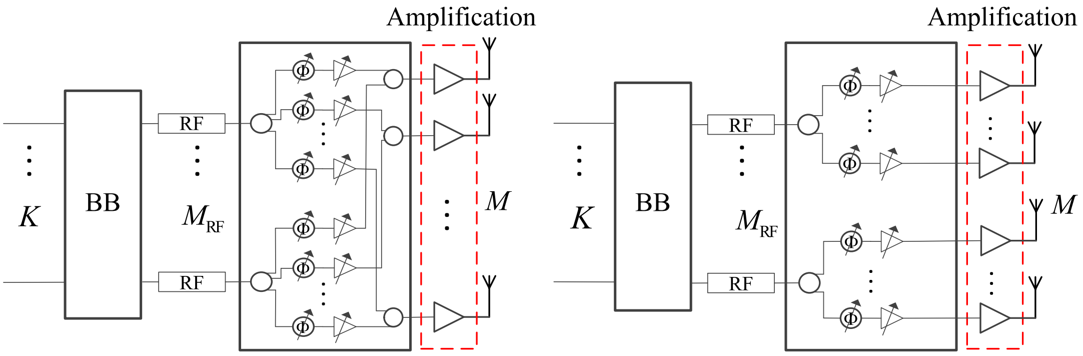
Notation: We denote vectors, matrices, and scalars by $\mathbf{a}$, $\mathbf{A}$, and a (A), respectively. For an integer $K \in \mathbb{Z}$, $[K]$ denotes the index set $\{1, ..., K\}$. We represent sets by calligraphic $\mathcal{A}$ and their cardinality with $|\mathcal{A}|$. We use $\mathbb{E}[\cdot]$ for the expectation, $\|\cdot\|$ for l_2-norm, $\circledast$ for continuous-time convolution, $\otimes$ for the Kronecker product, $\odot$ for Hadamard product.

II. Channel and Signal Models

A. Channel Model

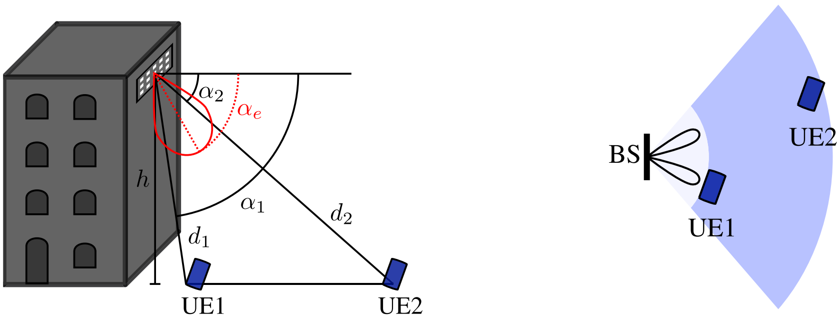
One of the main new features of 5G wireless networks is the densely spread small cell layer [31]. In small cell configurations as illustrated in Fig. 2 (a), the BS creates a fixed arc-like sectorized beam in the elevation direction. The orientation of the BS beam center in the elevation direction tends to be fixed with an elevation angle α_e [32]. It follows that the probing area in the range direction is restricted and the intensive initial beam searching takes place mainly in the azimuth direction. For notation simplicity, in this paper we only focus on the 2D azimuth plane. Extension to the 3D geometry is conceptually straightforward although may lead to a rather high dimensional search for the initial beam acquisition phase. In the small cell scenario as illustrated in Fig. 2, where the beam shape in the elevation direction is fixed a priori in order to define the cell footprint area, the 2D azimuth geometry is fully justified. We assume that the BS serving simultaneously K UEs. The BS is equipped with a uniform linear array (ULA) of M antennas and M_{RF} RF chains, where $K \leq M_{\text{RF}} \ll M$. Each UE is equipped with a ULA of N antennas and $N_{\text{RF}} \ll N$ RF chains. Since the focus of this paper is the BS architecture, we consider the case of $N_{\text{RF}} = 1$, where the extension to $N_{\text{RF}} > 1$ is straightforward and was considered in our work on BA [16, 18, 19]. The propagation channel between the BS and the k-th UE, $k \in [K]$, consists of $L_k \ll \max\{M, N\}$ *significant* multipath components. As a result, the $N \times M$ baseband equivalent impulse response of the channel at time slot s can be written as

$$\begin{aligned}\mathsf{H}_{s,k}(t,\tau) &= \sum_{l=1}^{L_k} \rho_{s,k,l} e^{j2\pi\nu_{k,l}t} \mathbf{a}_{\text{R}}(\phi_{k,l}) \mathbf{a}_{\text{T}}(\theta_{k,l})^{\mathsf{H}} \delta(\tau - \tau_{k,l}) \\ &= \sum_{l=1}^{L_k} \mathsf{H}_{s,k,l}(t) \delta(\tau - \tau_{k,l}), \qquad (1)\end{aligned}$$



(a) (b)

Fig. 1: Hybrid digital analog (HDA) transmitter architectures: (a) fully-connected (FC), (b) partially-connected with one-stream-per-subarray (OSPS). The "BB" block denotes digital baseband beamforming, K is the number of data streams, M_{RF} is the number of RF chains, and M is the number of antennas.



(a) (b)

Fig. 2: Illustration of a small cell scenario with (a) 3D side view and (b) 2D top view. In this paper, the initial beam alignment refers to the beam training/searching in the azimuth plane as shown in (b).

where $\mathsf{H}_{s,k,l}(t) := \rho_{s,k,l}e^{j2\pi\nu_{k,l}t}\mathbf{a}_{\mathrm{R}}(\phi_{k,l})\mathbf{a}_{\mathrm{T}}(\theta_{k,l})^{\mathsf{H}}$ and $\delta(\cdot)$ denotes the Dirac delta function. Each l-th multipath component is identified by the tuple $(\phi_{k,l}, \theta_{k,l}, \tau_{k,l}, \nu_{k,l})$ of angle of arrival (AoA), angle of departure (AoD), delay, and Doppler shift, respectively. The vectors $\mathbf{a}_{\mathrm{T}}(\theta_{k,l}) \in \mathbb{C}^M$ and $\mathbf{a}_{\mathrm{R}}(\phi_{k,l}) \in \mathbb{C}^N$ are the array response vectors of the BS and the k-th UE at the AoD $\theta_{k,l}$ and the AoA $\phi_{k,l}$, respectively. With the ULA configuration and the assumption that the spacing of the ULA antennas in each array (subarray) equals to a half-wavelength $\lambda/2$, the elements of $\mathbf{a}_{\mathrm{T}}(\theta_{k,l})$ and $\mathbf{a}_{\mathrm{R}}(\phi_{k,l})$ are given by

$$[\mathbf{a}_{\mathrm{T}}(\theta)]_{(i'-1)\cdot\hat{M}+d} = e^{j(d-1)\pi\sin(\theta)} \cdot e^{j\Psi(i',\theta)}, d \in [\hat{M}] \tag{2a}$$

$$[\mathbf{a}_{\mathrm{R}}(\phi)]_n = e^{j(n-1)\pi\sin(\phi)}, n \in [N], \tag{2b}$$

where in (2a) we assume that $(i' \equiv 1, \hat{M} = M)$ for the FC architecture as shown in Fig. 1(a), and $(i' \in [M_{\mathrm{RF}}], \hat{M} = \frac{M}{M_{\mathrm{RF}}})$ for the OSPS architecture as shown in Fig. 1(b). The additional term $\Psi(i', \theta)$ in (2a) takes into account the phase shifts among different subarrays, given by

$$\Psi(i', \theta) = \frac{2\pi}{\lambda}(i'-1) \cdot D_x \cdot \sin(\theta), \tag{3}$$

where i' indicates the index of the subarrays and $D_x \geq 0$ denotes the subarray center-to-center spacing in the scan direction. Hence, in the special case with $D_x = 0$, all the subarrays are co-located[4]; while with $D_x = \frac{M}{M_{\mathrm{RF}}} \cdot \frac{\lambda}{2}$, the antenna element layout in the scan direction for the OSPS architecture is exactly the same as for the FC architecture.

For the sake of modeling simplicity, we assume in (1) that each multipath component has a very narrow footprint over the AoA-AoD-delay domain. The extension to more widely spread multipath clusters is straightforward and will be applied in the numerical simulations. We adopt a block fading model, where the coefficient of the l-th

[4] In this paper, we consider a 2D geometry w.r.t. the azimuth plane as illustrated in Fig. 2 (b). In practice, the co-located layout can be obtained by stacking the arrays on top of each other in the vertical dimension. Strictly speaking this yields a rectangular array configuration, but since each row forms an individually driven array, adaptive beamforming in the elevation direction is not possible, therefore the beamforming geometry is still two-dimensional.

multipath component $\rho_{s,k,l}$ is constant over a short interval (within one slot) and changes from slot to slot according to a wide-sense stationary process statistics characterized by its power spectral density (Doppler spectrum) [33]. When the channel *coherence time* (related to the inverse of the bandwidth of the Doppler spectrum, see [33]) is significantly larger than the slot duration but equal or smaller than the (non-consecutive) slot separation in time, a convenient model is to consider the coefficients as i.i.d. across different slots. Moreover, the Doppler shift $\nu_{k,l}$ as defined in (1) introduces a continuous phase rotation for each channel sample. Each multipath component (channel tap coefficient) is formed by the superposition of a large number of micro-scattering components (e.g., due to rough surfaces) having (approximately) the same AoA-AoD and delay. By the central limit theorem, it is customary to model the superposition of these many small effects as Gaussian [34, 35]. Hence, the multipath component coefficients can be modeled as Rice fading given by

$$\rho_{s,k,l} \sim \sqrt{\gamma_{k,l}} \left(\sqrt{\frac{\eta_{k,l}}{1+\eta_{k,l}}} + \frac{1}{\sqrt{1+\eta_{k,l}}} \breve{\rho}_{s,k,l} \right), \quad (4)$$

where $\gamma_{k,l}$ denotes the overall multipath component strength, $\eta_{k,l} \in [0, \infty)$ indicates the strength ratio between the specular reflection (or LOS) and the scattered components, and $\breve{\rho}_{s,k,l} \sim \mathcal{CN}(0,1)$ is a zero-mean unit-variance complex Gaussian random variable whose value changes in an i.i.d. fashion across different slots. In particular, $\eta_{k,l} \to \infty$ indicates a pure LOS path while $\eta_{k,l} = 0$ indicates a pure scattered path, affected by Rayleigh fading.

The AoA-AoDs $(\phi_{k,l}, \theta_{k,l})$ in (1) can take on arbitrary values in the continuous AoA-AoD domain. Following the widely used approach of [36], known as *beam-domain representation*, we obtain a finite-dimensional representation of the channel response (1). More precisely, we consider the discrete set of AoA-AoDs

$$\Phi := \left\{ \check{\phi} : (1+\sin(\check{\phi}))/2 = \frac{n-1}{N}, \, n \in [N] \right\}, \quad (5a)$$

$$\Theta := \left\{ \check{\theta} : (1+\sin(\check{\theta}))/2 = \frac{m-1}{M}, m \in [M] \right\}. \quad (5b)$$

It follows that the corresponding sets $\mathcal{A}_{\text{R}} := \{\mathbf{a}_{\text{R}}(\check{\phi}) : \check{\phi} \in \Phi\}$ and $\mathcal{A}_{\text{T}} := \{\mathbf{a}_{\text{T}}(\check{\theta}) : \check{\theta} \in \Theta\}$ form discrete dictionaries to represent the channel response. For the ULAs considered in this paper, the dictionaries $\mathcal{A}_{\text{R}}$ and $\mathcal{A}_{\text{T}}$, after suitable normalization, reduce to the columns of unitary *Discrete Fourier Transform* (DFT) matrices $\mathbf{F}_N \in \mathbb{C}^{N\times N}$ and $\mathbf{F}_M \in \mathbb{C}^{M\times M}$, with elements

$$[\mathbf{F}_N]_{n,n'} = \frac{1}{\sqrt{N}} e^{j2\pi(n-1)(\frac{n'-1}{N}-\frac{1}{2})}, n, n' \in [N], \quad (6a)$$

$$[\mathbf{F}_M]_{m,m'} = \frac{1}{\sqrt{M}} e^{j2\pi(m-1)(\frac{m'-1}{M}-\frac{1}{2})}, m, m' \in [M]. \quad (6b)$$

Consequently, based on a subarray basis indexed by i', the beam-domain representation of the channel response (1) is given by [7, 36]

$$\begin{aligned} \check{\mathsf{H}}^{i'}_{s,k}(t,\tau) &= \mathbf{F}^{\mathsf{H}}_N \mathsf{H}_{s,k}(t,\tau) \cdot \left(\mathbf{F}_M \odot \mathbf{1}_{\{(i'-1)\hat{M}+1:i'\hat{M},1:M\}} \right) \\ &= \sum_{l=1}^{L_k} \check{\mathsf{H}}^{i'}_{s,k,l}(t) \delta(\tau - \tau_l), \end{aligned} \quad (7)$$

where $(i' \equiv 1, \hat{M} = M)$ for the FC architecture, and $(i' \in [M_{\text{RF}}], \hat{M} = \frac{M}{M_{\text{RF}}})$ for the OSPS architecture. Here we define $\check{\mathsf{H}}^{i'}_{s,k,l}(t) := \mathbf{F}^{\mathsf{H}}_N \mathsf{H}_{s,k,l}(t) \cdot \left(\mathbf{F}_M \odot \mathbf{1}_{\{(i'-1)\hat{M}+1:i'\hat{M},1:M\}} \right)$ as the beam-domain l-th multipath component between the k-th UE and the BS, where $\mathbf{1}_{\{a_1:a_2,b_1:b_2\}} \in \mathbb{C}^{M\times M}$ is an indicator matrix, with 1 at the components indexed by rows from a_1 to a_2 and by columns from b_1 to b_2, otherwise zero. The indicator matrix takes into account the fact that the number of antenna elements for each subarray in the OSPS architecture is M_{RF} times less than that in the FC architecture.

As shown in our earlier work [16] (and the references therein), for the FC architecture, as the number of antennas M at the BS and N at the UE increases, the DFT basis provides a good sparsification of the propagation channel. As a result, $\check{\mathsf{H}}^{i'}_{s,k}(t,\tau)$ can be approximated as a sparse matrix, with non-zero elements in the locations corresponding to small clusters of discrete AoA-AoD pairs. For the OSPS architecture, note that the indices of non-zero elements in $\check{\mathsf{H}}^{i'}_{s,k}(t,\tau)$ are identical for all $i' \in [M_{\text{RF}}]$. However, the channel sparsity depends on the number of antennas in each subarray. In both cases, we may encounter a grid error in (7) since the AoAs-AoDs do not necessarily fall into the uniform grid $\Phi \times \Theta$. Nevertheless, as shown in [16], the grid error becomes negligible by increasing the number of (subarray) antennas (i.e., the grid resolution). In our simulations, we do not constrain the AoA-AoD pairs of the physical channel to take on values on the discrete grid; therefore, the grid discretization effect is fully taken into account in our numerical results.

B. Signaling Model

Because of space limitation, in this paper we focus on SC signaling. Similar conclusions can be reached for OFDM, although the latter is generally more fragile to frame synchronization errors, large PAPR, and, before BA is achieved, to inter-carrier interference due to the fact that the Doppler spread between the several multipath components may be large [19, 37]. Let $\mathbf{x}_s(t) = [x_{s,1}(t), x_{s,2}(t), ..., x_{s,K}(t)]^{\mathsf{T}}$ denote the continuous-time baseband equivalent signal (either pilot or data signal), transmitted over the s-th slot. With HDA beamforming, the beamformed signal at the output of the transmitter over the s-th slot is generally given by

$$\hat{\mathbf{x}}_s(t) = \sqrt{E_0} \cdot \mathbf{U}^{\text{RF}}_s \cdot \mathbf{W}^{\text{BB}}_s \cdot \mathbf{x}_s(t), \quad (8)$$

where for simplicity of exposition we restrict to the case of uniform power allocation, with $E_0 = \frac{P_{\texttt{tot}} T_c}{K}$ indicating the per-chip energy of each signal stream, where $P_{\texttt{tot}}$ denotes the total radiated power at the BS and $T_c = \frac{1}{B}$ denotes the chip duration with B indicating the signaling bandwidth. In (8), we define $\mathbf{W}_s^{\text{BB}} \in \mathbb{C}^{M_{\text{RF}} \times K}$ and $\mathbf{U}_s^{\text{RF}} \in \mathbb{C}^{M \times M_{\text{RF}}}$ as the baseband (digital) and the RF analog beamforming matrices, respectively. Note that, depending on the transmitter architecture, the analog beamforming matrix $\mathbf{U}_s^{\text{RF}}$ takes on the form

$$[\tilde{\mathbf{u}}_{s,1}, \tilde{\mathbf{u}}_{s,2}, \cdots, \tilde{\mathbf{u}}_{s,M_{\text{RF}}}] \text{ and } \begin{bmatrix} \tilde{\mathbf{u}}_{s,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{u}}_{s,2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{u}}_{s,M_{\text{RF}}} \end{bmatrix} \tag{9}$$

for the FC (left) and the OSPS (right) architectures, respectively, where $\tilde{\mathbf{u}}_{s,i} \in \mathbb{C}^{\hat{M}}$, $i \in [M_{\text{RF}}]$, with $\hat{M} = M$ for the FC architecture and $\hat{M} = \frac{M}{M_{\text{RF}}}$ for the OSPS architecture. Hence, in both cases $\mathbf{U}_s^{\text{RF}}$ has dimension $M \times M_{\text{RF}}$, but FC has a full matrix, while OSPS has a block-diagonal matrix, due to the constrained connectivity. Without loss of generality, the beamforming vectors are normalized as $\sum_{i=1}^{M_{\text{RF}}} \|\mathbf{u}_{s,i}\|^2 = M_{\text{RF}}$.

The beamformed signal (8) goes through the channel as defined in (1). At the UE side, because of the HDA architecture, the UE does not have direct access to each antenna element. Instead, at each slot s, the UE obtains only a projection of the received signal by applying some beamforming vector in the analog domain. We consider a single RF chain at each UE as mentioned before. Thus, the received signal at the k-th UE side is given by

$$\begin{aligned} \hat{y}_{s,k}(t) =& \mathbf{v}_{s,k}^{\mathsf{H}} \mathsf{H}_{s,k}(t,\tau) \circledast \hat{\mathbf{x}}_s(t) + z_{s,k}(t) \\ =& \sqrt{E_0} \mathbf{v}_{s,k}^{\mathsf{H}} \mathsf{H}_{s,k}(t,\tau) \circledast \left(\mathbf{U}_s^{\text{RF}} \cdot \mathbf{W}_s^{\text{BB}} \cdot \mathbf{x}_s(t)\right) \\ & + z_{s,k}(t), \end{aligned} \tag{10}$$

where $\mathbf{v}_{s,k} \in \mathbb{C}^N$ denotes the normalized beamforming vector with $\|\mathbf{v}_{s,k}\| = 1$ at the k-th UE, and $z_{s,k}(t)$ is the continuous-time complex *Additive White Gaussian Noise* (AWGN) at the output of the UE RF chain, with a *Power Spectral Density* (PSD) of N_0 Watt/Hz.

In the following, we will evaluate the performance of different transmitter architectures as shown in Fig. 1. For this purpose, it is useful to first define the channel SNR before beamforming (BBF) $\mathsf{SNR}_{\text{BBF}}$, given by

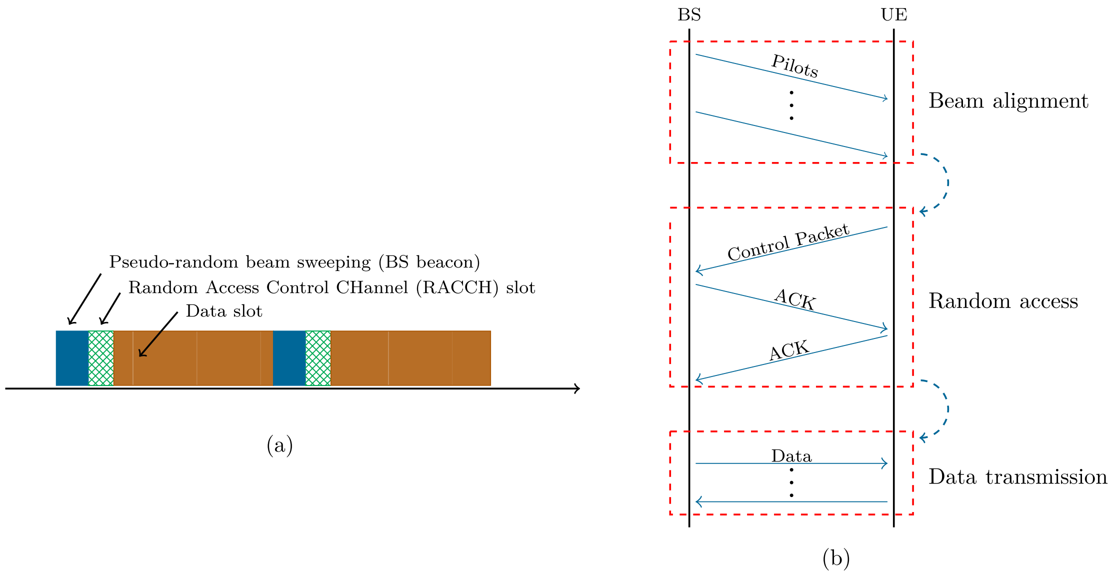
$$\mathsf{SNR}_{\text{BBF},\,k} = \frac{P_{\texttt{tot}} \sum_{l=1}^{L_k} \gamma_{k,l}}{N_0 B}. \tag{11}$$

where k is the index of the UE and $\gamma_{k,l}$ denotes the strength of the l-th multipath component. The SNR in (11) indicates the ratio of the total received signal power (summing over all the multipath components) over the total noise power at the receiver baseband processor input, assuming that the signal is isotropically transmitted by the BS and isotropically received at the k-th UE over the total bandwidth B. As mentioned before, one of the challenges of mmWaves communication is that the SNR before beamforming $\mathsf{SNR}_{\text{BBF}}$ in (11) may be very low.

III. Beam Acquisition and Data Transmission

We evaluate the performance of the FC and OSPS architectures including both the BA phase and the consequent data transmission phase, where the latter uses the beam information obtained by the former. For the BA phase we use the scheme proposed in our previous work [19], that compares favorably with respect to several competing schemes proposed in the literature. For the sake of space limitation, we provide here only a high-level summary of the scheme and invite the reader to consider [19] for the full details. Fig. 3 (a) illustrates the considered frame structure, which consists of three parts: the beacon slot, the random access control channel (RACCH) slot, and the data slot. As shown in Fig. 3 (b), the BS broadcasts its pilot signals periodically over the beacon slots. The measurements are collected at each UE locally and independently of other UEs. Based on measurements accumulated over a sequence of several beacon slots, each UE can estimate a set of strongly coupled AoA-AoD pairs, corresponding to the directions of strong propagation paths between the UE and the BS arrays. These determine the beamforming direction for possible data transmission. During the RACCH slot, the BS stays in listening mode and the UEs send beamformed uplink packets. These packets contain basic information such as the UE ID and the beam indices of the selected BS beam directions. The BS responds with an acknowledgment (ACK) data packet in the data subslot of a next frame, using the indicated beam indices for transmission. From this moment on, the BS and the UE are connected in the sense that, if the procedure is successful, they can communicate by aligning their beams along a small number of multipath components with strong average power transmission.

As explained in details in [19], the BS beacon signals are formed by M_{RF} different PN sequences, each of which undergoes a "multifinger" beam pattern obtained by selecting a subset of the columns (or masked DFT columns as in the case of OSPS). The beamforming patterns send the signal energy uniformly distributed along subsets of the BS AoD grid. The beamforming patterns follow a pre-determined pseudo-random sequence, similar in the spirit to the primary synchronization code of a W-CDMA 3G system for BS identification. During the beacon slot, each UE k receives using its own pseudo-random sequence of multifinger beam patterns, and integrates the received signal energy over the multiple time segments within a beacon slot in order to obtain an estimate of the average received energy. As a result, this fully non-coherent energy measurement yields (approximately) the average energy sum of several multipath components. These multipath components corresponds to the AoA-AoD pairs in the grid



(a)

(b)

Fig. 3: Illustration of (a) the frame structure in the underlying system and (b) the communication process between the BS and a generic UE. The initial beam alignment phase is periodically done over beacon slots, followed by a random access stage to build up the connection between the BS and the active UEs, and consecutively the data communication.

for which the BS transmit directions and the UE receive directions meet. Fig. 4 (a) shows an example of transmit and receive multifinger beam patterns and Fig. 4 (b) shows the corresponding masks of crossing AoA-AoD directions, superimposed with the second moments (channel gain) of the beam-domain channel matrix generated by the QuaDRiGa simulator. The goal of the BA algorithm run at the UE side is to identify the position of the strong components, i.e., the small dark spots in the plot of Fig. 4 (b). It turns out that this problem can be cast as the reconstruction of a sparse non-negative vector from noisy linear measurements, which can be efficiently obtained by solving a non-negative least-squares (NNLS) problem. It can be shown that NNLS naturally induce sparsity in the solution, and it is very efficient to solve by a plethora of well-known algorithms (e.g., projected gradient). The full details of the BA scheme, as well as extensive comparison with other competing state-of-the-art schemes, are provided in [19].

We denote by $\mathbf{\Gamma}_k \in \mathbb{C}^{N\times M}$ as the matrix of second moments of the beam-domain channel coefficients between the BS array and the k-th UE array. An example of $\mathbf{\Gamma}_k$ is illustrated in Fig. 5 (a). Also, Fig. 5 (b) shows the estimate $\mathbf{\Gamma}_k^\star$ of $\mathbf{\Gamma}_k$ provided by the NNLS estimation at UE k. Once the BA algorithm yields $\mathbf{\Gamma}_k^\star$, the k-th UE will send a beamformed control packet to the BS in the RACCH. The UE chooses the beamforming direction corresponding to the strongest AoA direction obtained from $\mathbf{\Gamma}_k^\star$, meanwhile the BS stays in listening mode during the RACCH, using a sectored beamforming configuration. In this way, full beamforming gain at the UE transmit side and a limited sector beamforming gain at the BS receive side can be achieved. Once the RACCH packet is received, the BS can use the transmit beam indicated by UE k to communicate data. In the next section, we focus on the data communication phase assuming that the RACCH has been correctly received, therefore, both the BS and the UE know the indices of the strong components in $\mathbf{\Gamma}_k^\star$. Notice that if the NNLS estimation fails, it is likely that the RAACH will not be received or will be received in error, because the beamforming gain at the UE side will be poor. In this case, the UE will not receive a data packet and after a given time-out will try the BA procedure again. Also in the (very unlikely) case of a collision in the RACCH, the same time-out procedure can be exploited. Therefore, data communication effectively takes place only when a) the strong multipath components in $\mathbf{\Gamma}_k$ are correctly estimated and b) when the RACCH decoding is successful. In [19] we have already argued that the probability that the BA procedure fails is dominated by the error probability in the estimation of the strong components of $\mathbf{\Gamma}_k$. Hence, a sensible system design approach consists of allowing a sufficient number of beacon slots such that the probability of success in identifying the strong components of $\mathbf{\Gamma}_k$ is close to 1, and designing the HDA beamforming scheme in the assumption that the estimation of $\mathbf{\Gamma}_k$ is correct. As a result, we shall compare the FC and OSPS architectures in terms of number of beacon slots needed to achieve a BA success probability near 1, and their achieved spectral efficiency under such condition. In any case, the designed HDA precoders in our simulations are always obtained from the true NNLS estimation $\mathbf{\Gamma}_k^\star$, and not by the

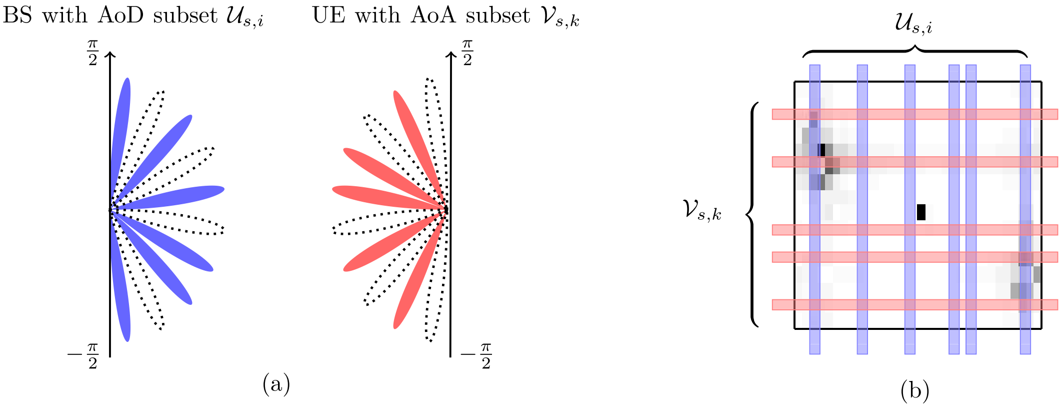


Fig. 4: (a) Illustration of the subset of AoA-AoDs at time slot s probed by the i-th beacon stream transmitted by the BS and received by the k-th UE, for $\hat{M} = N = 10$. The AoD subset is given by $\mathcal{U}_{s,i} = \{1, 3, 4, 6, 8, 10\}$ (numbered counterclockwise) with beamforming vector $\tilde{\mathbf{u}}_{s,i} = \frac{1}{\sqrt{6}}[1, 0, 1, 0, 1, 0, 1, 1, 0, 1]^{\mathsf{T}}$. The AoA subset is given by $\mathcal{V}_{s,k} = \{2, 4, 5, 7, 9\}$ (numbered counterclockwise) with receive beamforming vector $\check{\mathbf{v}}_{s,k} = \frac{1}{\sqrt{5}}[0, 1, 0, 1, 1, 0, 1, 0, 1, 0]^{\mathsf{T}}$. (b) The beam-domain channel gain matrix (with one LOS component and two scattered multipath components indicated by the dark spots, generated by the QuaDRiGa simulator) measured along $\mathcal{V}_{s,k} \times \mathcal{U}_{s,i}$.

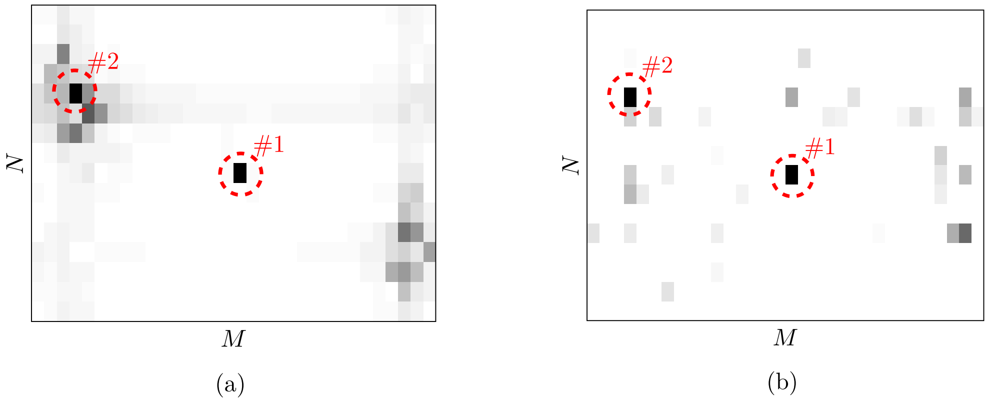


Fig. 5: Illustration of the second moments of the beam-domain channel matrix $\mathbf{\Gamma}_k$: (a) the actual QuaDRiGa generated $\mathbf{\Gamma}_k$, (b) the NNLS estimated $\mathbf{\Gamma}^\star_k$. The dashed circles indicate the top $p = 2$ strongest components in $\mathbf{\Gamma}_k$ and $\mathbf{\Gamma}^\star_k$, respectively. We announce a success in the BA phase if the locations of the strongest component in $\mathbf{\Gamma}_k$ and in $\mathbf{\Gamma}^\star_k$ are consistent.

genie-aided exact knowledge of $\mathbf{\Gamma}_k$.

A. Data Communication Phase

We assume that the BS simultaneously schedules $K = M_{\mathrm{RF}}$ UEs. With the small cell configuration as illustrated in Fig. 2, the distance differences between each UE and the BS are very small, implying that the received power w.r.t. the LOS path for each UE within the BS coverage are similar. Although schedulers such as random or proportionate fair scheduler are commonly used in sub 6 GHz, the directionality of the mmWave channel instead calls for schedulers that select groups of users with good angular separation (directional scheduler) [26]. More precisely, we assume that the selected K UEs have similar received power in terms of the strongest path, and their strongest AoDs in the downlink are at least $\Delta\theta_{\min}$ away from each other.

Let $\mathbf{x}^{\mathrm{d}}(t) = [x^{\mathrm{d}}_1(t), x^{\mathrm{d}}_2(t), ..., x^{\mathrm{d}}_K(t)]^{\mathsf{T}}$ denote the complex baseband data signal,[5] with $x^{\mathrm{d}}_k(t)$, $k \in [K]$, corresponding to the k-th UE, given by

$$x^{\mathrm{d}}_k(t) = \sum_{n=1}^{N_d} d_{k,n} p_r(t - nT_c), \tag{12}$$

where $p_r(t)$ is the unit-energy square-root Nyquist pulse shaping filter, $\{d_{k,n}\}$ denote the unit-energy modulation symbols belonging to a suitable modulation constellation

[5] From now on, we ignore the slot index s for notation simplicity, also because once a successful BA is achieved, the channel statistical property, the precoding vector at the BS, and combining vector at each UE are invariant within many slots. However, note that this invariance holds only until a new updated BA takes place, implying that the underlying channel may encounter large mobility, blockage, etc.

[33], and N_d indicates the number of the transmit symbols. Accordingly, the received data signal at the k-th UE is given by (13), where $\mathbf{w}_{k'}$ denotes the k'-th column of $\mathbf{W}^{\text{BB}}$, $\Delta_{k,n,l} = 2\pi(\check{\nu}_{k,l} + \nu_{k,l} n T_c)$, and $C_{k,k',l,n} := \rho_{k,l} d_{k',n} \sqrt{E_0} (\mathbf{v}_k^{\mathsf{H}} \mathbf{a}_{\text{R}}(\phi_{k,l}) \mathbf{a}_{\text{T}}(\theta_{k,l})^{\mathsf{H}} \mathbf{U}^{\text{RF}} \mathbf{w}_{k'})$. We assume that each UE uses standard timing synchronization with respect to its strongest multipath component indexed by l^1, which is selected by its initial BA. To decode the data signal, each UE performs matched filtering with respect to the symbol pulse $p_r(t)$, sampling at epochs $t = \hat{n} T_c + \tau_{k,l^1}$. It follows that the discrete-time baseband signal received at the k-th UE receiver takes on the form of (14), where $\hat{n}^{\Delta}_{k,k',\hat{n},n,l} := (\hat{n} - n)T_c + \tau_{k,l^1} - \tau_{k',l}$, $\varphi_r[t^{\Delta}] = \varphi_r(t)|_{t=t^{\Delta}} := \int p_r(\tau) p_r^*(\tau - t^{\Delta}) d\tau$, and $z_k^c[\hat{n}]$ denotes the noise at the output of the matched filter with variance $N_0 \cdot \int |p_r(t)|^2 dt = N_0$. As we can see, the first term in (14) corresponds to the desired data symbol $d_{k,n}$ multiplied by a different complex coefficient over each path l.[6] Whereas, the last two terms in (14) correspond to the multiuser interference and noise, respectively. By treating the multiuser interference as noise, the asymptotic ergodic spectral efficiency of the k-th UE is given by (15) and the sum rate reads $R_{\text{sum}} = \sum_{k=1}^{K} R_k$. In all the schemes treated here, coherent communication can be practically achieved by including per-user beamformed pilot symbols at the cost of a very small overhead, as it is quite state of art and usual in virtually any modern wireless communication standard. For simplicity, we shall not take into account this overhead or the degradation of quasi-coherent receivers, which is well known and not a specific feature of the systems under consideration.

1) Hybrid Precoding Formulation

Now the remaining problem is how to define the precoding/combining vectors. We assume that the BS communicates with the k-th UE along its top-p beams. We will show later that the parameter $p \geq 1$ is somehow a tradeoff between the transmitter power spreading, multiuser interference, and the system robustness to potential blockages. To simplify the practical implementation, we define the combining vector at the k-th UE as

$$\mathbf{v}_k = \frac{1}{\sqrt{p}} \mathbf{F}_N \cdot \sum_{p'=1}^{p} \check{\mathbf{v}}_{k,p'}, \tag{16}$$

where $\check{\mathbf{v}}_{k,p'} \in \mathbb{C}^N$ is an all-zero vector with a 1 at the component corresponding to the p'-th strong AoA, i.e., the AoA index of the p'-th strong component in $\boldsymbol{\Gamma}_k^{\star}$. Denoted by $\mathbf{V} \in \mathbb{C}^{NK \times K}$ as the aggregated receive beamforming matrix given by $\mathbf{V} = \text{diag}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K)$. It follows that the receive data signal vector $\bar{\mathbf{y}}(t) = [y_1(t), y_2(t), ..., y_K(t)]^{\mathsf{T}} \in \mathbb{C}^K$ corresponding to the K UEs can be written as

$$\begin{aligned}
\bar{\mathbf{y}}(t) &= \sqrt{E_0} \mathbf{V}^{\mathsf{H}} \cdot \overline{\mathsf{H}}(t,\tau) \circledast \left(\mathbf{U}^{\text{RF}} \cdot \mathbf{W}^{\text{BB}} \cdot \mathbf{x}^{\text{d}}(t)\right) + \bar{\mathbf{z}}(t) \\
&\overset{\text{(a)}}{=} \sqrt{E_0} \left(\mathbf{V}^{\mathsf{H}} \cdot \overline{\mathsf{H}}(t,\tau) \cdot \overline{\mathbf{U}} \cdot \mathbf{A}^{\text{RF}} \cdot \mathbf{W}^{\text{BB}}\right) \circledast \mathbf{x}^{\text{d}}(t) + \bar{\mathbf{z}}(t) \\
&\overset{\text{(b)}}{=} \sqrt{E_0} \left(\widetilde{\mathsf{H}}(t,\tau) \cdot \mathbf{A}^{\text{RF}} \cdot \mathbf{W}^{\text{BB}}\right) \circledast \mathbf{x}^{\text{d}}(t) + \bar{\mathbf{z}}(t),
\end{aligned} \tag{17}$$

where $\bar{\mathbf{z}}(t) \in \mathbb{C}^K$ indicates the noise vector, $\mathbf{U}^{\text{RF}} := \overline{\mathbf{U}} \cdot \mathbf{A}^{\text{RF}}$ is the analog beamforming matrix, $\widetilde{\mathsf{H}}(t,\tau) := \mathbf{V}^{\mathsf{H}} \cdot \overline{\mathsf{H}}(t,\tau) \cdot \overline{\mathbf{U}}$ denotes a constructed effective channel, and $\overline{\mathsf{H}}_s(t,\tau) \in \mathbb{C}^{NK \times M}$ represents the aggregated instantaneous channel of all the K UEs given by

$$\overline{\mathsf{H}}(t,\tau) = \left[\mathsf{H}_1(t,\tau)^{\mathsf{T}}, \mathsf{H}_2(t,\tau)^{\mathsf{T}}, \cdots, \mathsf{H}_K(t,\tau)^{\mathsf{T}}\right]^{\mathsf{T}}, \tag{18}$$

where $\mathsf{H}_k(t,\tau)$, $k \in [K]$, is given in (1). In (17) (a), we define $\overline{\mathbf{U}} \in \mathbb{C}^{M \times pK}$ as the angular support, and $\mathbf{A}^{\text{RF}} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_K] \in \mathbb{C}^{pK \times K}$ as the coefficient tuning for the analog part. More precisely, we assume $\overline{\mathbf{U}} = [\mathbf{U}_1, ..., \mathbf{U}_K]$, where $\mathbf{U}_k \in \mathbb{C}^{M \times p}$, $k \in [K]$, takes on the form

$$\begin{aligned}
\mathbf{U}_k = &\left(\mathbf{F}_M \odot \mathbf{1}_{\{(k'-1)\hat{M}+1:k'\hat{M},1:M\}}\right) \\
&\times \left[\check{\mathbf{u}}_{k,1}, \check{\mathbf{u}}_{k,2}, ..., \check{\mathbf{u}}_{k,p}\right],
\end{aligned} \tag{19}$$

where $(i' \equiv 1, \hat{M} = M)$ for the FC architecture, and $(i' = k, \hat{M} = \frac{M}{M_{\text{RF}}})$ for the OSPS architecture. Also, we define $\check{\mathbf{u}}_{k,p'} \in \mathbb{C}^M$, $p' \in [p]$, as an all-zero vector with a 1 at the component corresponding to the p'-th strongest AoD of $\boldsymbol{\Gamma}_k^{\star}$.

Notice that in order to construct the beamforming vector at each k-th UE and the precoding vectors at the BS, only the AoA-AoD indices of the p strongest components in the estimated channel gain matrix $\boldsymbol{\Gamma}_k^{\star}$ are needed. Then, once these vectors are fixed, the resulting effective channel has much lower dimensions than the original physical $N \times M$ channel (from array to array). Therefore, it can be estimated using orthogonal uplink pilots and channel reciprocity as in regular TDD MU-MIMO (e.g., see [38, 39]). Namely, the constructed effective channel matrix $\widetilde{\mathsf{H}}(t,\tau)$ in (17) (b) has dimension $K \times (pK)$, and can be estimated using pK uplink pilot sub-slots using TDD reciprocity.

2) Beam Steering (BST) Scheme

The BST scheme consists of simply steering the K data streams towards the K UEs along their strongest AoD. Hence, we have $p = 1$ in (16) and in (19), respectively. In such case, the analog tuning matrix and the baseband precoding matrices under the BST precoding scheme turn to be identity, i.e., $\mathbf{A}^{\text{RF}} = \mathbf{W}^{\text{BB}} = \mathbf{I}_K$. Note that in the BST scheme, we do not need any additional uplink channel estimation of $\widetilde{\mathsf{H}}(t,\tau)$. Namely, once the UEs has fed back its strongest AoD control packet, the BS can immediately provide the BST precoder.

[6] Actually, we have shown in our precious work [19] that, the phase perturbations over several strong paths are easy to compensate by standard carrier synchronization techniques given that a successful BA is achieved and the effective channel after BA has a very small time spreading. Due to the space limit, in (14) and also in our simulations, we will keep the phase perturbations such that the numerical results coincide with the conservative worst-case scenario.

$$\begin{aligned}\hat{y}_k(t) &= \sqrt{E_0}\mathbf{v}_k^{\mathsf{H}}\mathsf{H}_k(t,\tau)\circledast\left(\mathbf{U}^{\mathrm{RF}}\cdot\mathbf{W}^{\mathrm{BB}}\cdot\mathbf{x}^{\mathrm{d}}(t)\right)+z_k(t)\\ &=\sum_{n=1}^{N_d}\sum_{k'=1}^{K}\sum_{l=1}^{L_k}d_{k',n}\sqrt{E_0}\mathbf{v}_k^{\mathsf{H}}\mathsf{H}_{k,l}(t)\mathbf{U}^{\mathrm{RF}}\mathbf{w}_{k'}p_r(t-\tau_{k',l}-nT_c)+z_k(t)\\ &=\sum_{n=1}^{N_d}\sum_{k'=1}^{K}\sum_{l=1}^{L_k}C_{k,k',l,n}e^{j\Delta_{k,n,l}}p_r(t-\tau_{k',l}-nT_c)+z_k(t)\end{aligned}\tag{13}$$

$$\begin{aligned}y_k[\hat{n}] &= y_k(t)|_{t=\hat{n}T_c+\tau_{k,l^1}} = \hat{y}_k(t)\circledast p_r^*(-t)\Big|_{t=\hat{n}T_c+\tau_{k,l^1}}\\ &=\sum_{n=1}^{N_d}\sum_{k'=1}^{K}\sum_{l=1}^{L_k}C_{k,k',l,n}e^{j\Delta_{k,n,l}}\varphi_r\left[\hat{n}^{\Delta}_{k,k',\hat{n},n,l}\right]+\sum_{n=1}^{N_d}z_k^c[\hat{n}]\\ &=\sum_{n=1}^{N_d}\left(\sum_{l=1}^{L_k}C_{k,k,l,n}e^{j\Delta_{k,n,l}}\varphi_r\left[\hat{n}^{\Delta}_{k,k,\hat{n},n,l}\right]+\sum_{k'\neq k}\sum_{l=1}^{L_k}C_{k,k',l,n}e^{j\Delta_{k,n,l}}\varphi_r\left[\hat{n}^{\Delta}_{k,k',\hat{n},n,l}\right]+z_k^c[\hat{n}]\right)\end{aligned}\tag{14}$$

$$R_k=\log_2\left(1+\frac{\mathbb{E}\left[\left|\sum_{l=1}^{L_k}C_{k,k,l,n}e^{j\Delta_{k,n,l}}\varphi_r\left[\hat{n}^{\Delta}_{k,k,\hat{n},n,l}\right]\right|^2\right]}{\mathbb{E}\left[\left|\sum_{k'\neq k}\sum_{l=1}^{L_k}C_{k,k',l,n}e^{j\Delta_{k,n,l}}\varphi_r\left[\hat{n}^{\Delta}_{k,k',\hat{n},n,l}\right]\right|^2\right]+N_0}\right)\tag{15}$$

3) Analog Maximum Ratio Transmission (MRT) Scheme

In this scheme, we aims to maximize the desired signal power as well as to increase the scheme blockage robustness. To this end, the baseband precoding matrix remains identity, i.e., $\mathbf{W}^{\mathrm{BB}}=\mathbf{I}_K$, while the k-th analog MRT tuning vector (i.e., the k-th column of $\mathbf{A}^{\mathrm{RF}}$) is given by

$$\mathbf{a}_k=\left(\widetilde{\mathsf{H}}(t,\tau)_{\{k,:\}}\right)^{\mathsf{H}}\odot\mathbf{1}_{\{(k'-1)\hat{p}+1:k'\hat{p}\}}\cdot\Delta^{\mathrm{RF}},\tag{20}$$

where $\widetilde{\mathsf{H}}(t,\tau)_{\{k,:\}}$ indicates the k-th row of $\widetilde{\mathsf{H}}(t,\tau)$, and $\Delta^{\mathrm{RF}}\in\mathbb{R}_+$ denotes the normalizing factor such that $\sum_{i=1}^{M_{\mathrm{RF}}}\|\mathbf{u}_i\|^2=M_{\mathrm{RF}}$. The indicator vector $\mathbf{1}_{\{(k'-1)\hat{p}+1:k'\hat{p}\}}\in\mathbb{C}^{pK}$ has components 1 over the index $\{(k'-1)\hat{p}+1:k'\hat{p}\}$ otherwise 0, where $(k'\equiv 1,\hat{p}=pK)$ for the FC architecture and $(k'=k,\hat{p}=p)$ for the OSPS architecture. Here the indicator vector ensures that, in the OSPS architecture, the analog beamforming matrix $\mathbf{U}^{\mathrm{RF}}=\overline{\mathbf{U}}\cdot\mathbf{A}^{\mathrm{RF}}$ satisfies the block diagonal structure as illustrated in (9).

4) Joint Analog Maximum Ratio and Baseband Zeroforcing (MR-ZF) Scheme

On top of the previous MRT scheme, in this joint MR-ZF scheme, we propose to make use of the baseband precoding to further reduce the multiuser interference. Accordingly, the analog MRT vectors in $\mathbf{A}^{\mathrm{RF}}$ are given by (20), while the baseband ZF matrix $\mathbf{W}^{\mathrm{BB}}$ takes on the form

$$\mathbf{W}^{\mathrm{BB}}=\left(\widetilde{\mathsf{H}}(t,\tau)\mathbf{A}^{\mathrm{RF}}\right)^{\mathsf{H}}\cdot\left(\widetilde{\mathsf{H}}(t,\tau)\mathbf{A}^{\mathrm{RF}}\left(\widetilde{\mathsf{H}}(t,\tau)\mathbf{A}^{\mathrm{RF}}\right)^{\mathsf{H}}\right)^{-1}\cdot\Delta^{\mathrm{ZF}},\tag{21}$$

where $\Delta^{\mathrm{ZF}}\in\mathbb{R}_+$ is the normalizing factor ensuring the total radiated power constraint, i.e., $\sum_{k=1}^{K}\|\mathbf{w}_k\|^2=K$.

IV. Hardware Impairments

In all the above derivations, we have implicitly assumed that all the hardware components work in their ideal range without any distortion or power dissipation. However, in practical hardware systems, such assumption is not trivial to meet. For example, the implementation of HDA transceivers consists of a large number of power dividers and combiners in the analog part, particularly for the FC architecture. The power dissipation caused by these components has a severe impact on the transmit power and the power efficiency. Moreover, due to the superposition of multiple beamformed pilots / data, the input signal at the PAs may encounter a large PAPR. Also, different beamforming vectors will create different power levels for different PAs. As a result, the input power for some individual PAs may exceed their saturation limit (relevant to per-antenna power constraint) and even cause a disruption of the whole transmission. All these hardware impairment have a severe impact on the transmitter performance and should not be neglected. In this section, we will provide the mathematical model to evaluate the hardware efficiency of different transmitter architectures given in Fig. 1.

We assume that each analog path has simultaneous amplitude and phase control as shown in Fig. 1. Refer to (8), let $\tilde{\mathbf{x}} \in \mathbb{C}^M$ denote the pre-amplified beamformed signal[7], given by

$$\tilde{\mathbf{x}} = \sqrt{\alpha_{\text{com}}} \cdot \widetilde{\mathbf{U}}^{\text{RF}} \cdot \sqrt{\alpha_{\text{div}}} \cdot \mathbf{W}^{\text{BB}} \cdot \mathbf{x}, \tag{22}$$

where $\mathbf{x} = [x_1, \cdots, x_K] \in \mathbb{C}^K$ denotes the transmit symbol, with $\mathbb{E}[|x_i|^2] = \epsilon$, $i \in [K]$. The factor α_{div} indicates the power splitting at the divider, with $\alpha_{\text{div}} = \frac{1}{M}$ for the FC architecture as shown in Fig. 1 (a) and $\alpha_{\text{div}} = \frac{M_{\text{RF}}}{M}$ for the OSPS architecture as shown in Fig. 1 (b). Moreover, the factor α_{com} models the power dissipation factor of the combiners, i.e., $\alpha_{\text{com}} = \frac{1}{M_{\text{RF}}}$ for the FC architecture, and $\alpha_{\text{com}} = 1$ for the OSPS architecture. Both α_{div} and α_{com} result from the hardware implementation and are based on the corresponding S-parameters of the dividers and combiners as in [5]. We assume that the baseband beamforming matrix $\mathbf{W}^{\text{BB}}$ is of dimension $K \times K$ with $K = M_{\text{RF}}$, and the analog beamforming matrix $\widetilde{\mathbf{U}}^{\text{RF}} = [\mathbf{u}_1, ..., \mathbf{u}_{M_{\text{RF}}}] \in \mathbb{C}^{M \times M_{\text{RF}}}$ satisfies the specific FC / OSPS architecture as illustrated in (9).

We consider the rather simple BST precoding with $\mathbf{W}^{\text{BB}} = \mathbf{I}_K$. To first meet the total power constraint, for any $i \in [M_{\text{RF}}]$, we have $\|\mathbf{u}_i\|^2 = M$ for the FC architecture and $\|\mathbf{u}_i\|^2 = \frac{M}{M_{\text{RF}}}$ for the OSPS architecture, respectively. It follows that the effective pre-amplified radiated power of the beamformed signal $\tilde{\mathbf{x}}$ in (22) can be written as

$$\begin{aligned}\tilde{P} &= \mathbb{E}[\tilde{\mathbf{x}}^{\mathsf{H}}\tilde{\mathbf{x}}] = \alpha_{\text{com}}\alpha_{\text{div}} \cdot \mathbb{E}[\mathbf{x}^{\mathsf{H}}(\widetilde{\mathbf{U}}^{\text{RF}})^{\mathsf{H}}\widetilde{\mathbf{U}}^{\text{RF}}\mathbf{x}] \\ &= \alpha_{\text{com}}\alpha_{\text{div}} \cdot \operatorname{tr}\left(\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{H}}] \cdot (\widetilde{\mathbf{U}}^{\text{RF}})^{\mathsf{H}}\widetilde{\mathbf{U}}^{\text{RF}}\right). \end{aligned} \tag{23}$$

Accordingly, the pre-amplified radiated power for the FC and the OSPS architectures reads $\tilde{P}_{\text{FC}} = \epsilon M_{\text{RF}} \frac{1}{M_{\text{RF}}}$ and $\tilde{P}_{\text{OSPS}} = \epsilon M_{\text{RF}}$, respectively. As we can see, in order to achieve the same output power, the FC transmitter should compensate for an additional combiner power dissipation. More precisely, the transmitter should either boost the input signal as $M_{\text{RF}}\mathbf{x}$ or choose PAs with larger gain for the amplification stage. We consider the former approach and mathematically include the potential boosting factor M_{RF} as well as the factors $(\alpha_{\text{com}}, \alpha_{\text{div}})$ into the beamforming matrix $\widetilde{\mathbf{U}}^{\text{RF}}$. Denoted by $\mathbf{U}^{\text{RF}}$ as the integrated analog beamforming matrix, such that the pre-amplified beamformed signal in (22) can be written as $\tilde{\mathbf{x}} = \mathbf{U}^{\text{RF}} \cdot \mathbf{W}^{\text{BB}} \cdot \mathbf{x}$, which is consistent with our assumptions and formulations in Section II.

The beamformed signal (22) then goes through the amplification stage, where at each antenna branch a PA amplifies the signal before transmission. We assume that the PAs in different antenna branches have the same input-output relation. For any given antenna in the transmitter array, let P_{rad} denote the radiated power of the antenna, and P_{cons} denote the consumed power of the corresponding PA, which includes both the radiated power and the dissipated power. Following the approach in [28], the power consumed by the PA takes on the form

$$P_{\text{cons}} = \frac{\sqrt{P_{\max}}}{\eta_{\max}}\sqrt{P_{\text{rad}}}, \tag{24}$$

where $P_{\max}$ is the maximum output power of the PA with $P_{\text{rad}} \leq P_{\max}$ and $\eta_{\max}$ is the maximum efficiency of the PA. Note that this relation holds for the most common PA implementations and is therefore a good choice for the following calculation. Considering that the PAs are often the predominant power consumption part, we define η_{eff} given by

$$\eta_{\text{eff}} = \frac{P_{\text{rad}}}{P_{\text{cons}}} \tag{25}$$

as the metric to effectively compare the power efficiency of the two transmitter architectures shown in Fig. 1. Note that due to the superposition of multiple beamforming vectors (particularly for the FC architecture) and the potentially high PAPR of the time-domain transmit waveform $\tilde{\mathbf{x}}$ in (22) (particularly with OFDM signaling), the input power for some individual PA may exceed its saturation limit. This would result in non-linear distortion and even the disruption of the whole transmission. To compare the two transmitter architectures and ensure that all the underlying M PAs simultaneously work in their linear range, we generally have two options:

Option I: Both the FC architecture and the OSPS architecture utilize the same PA but apply a different input back-off $\alpha_{\text{off}} \in (0, 1]$, such that the peak power of the radiated signal is smaller than $P_{\max}$. As a reference, we denote by $(P_{\text{rad},0}, \eta_{\max,0})$ as the parameters of a reference PA under the reference precoding/beamforming strategy with a power backoff factor $\alpha_{\text{off},0}$ (as illustrated later in Section V). For different scenarios (with certain α_{off}) the average radiated power and the consumed power take the form $P_{\text{rad}} = \frac{\alpha_{\text{off}}}{\alpha_{\text{off},0}} P_{\text{rad},0}$, $P_{\text{cons}} = \frac{\sqrt{P_{\max,0}}}{\eta_{\max,0}}\sqrt{P_{\text{rad}}}$. The transmitter power efficiency is given by

$$\eta_{\text{eff}} = \frac{P_{\text{rad}}}{P_{\text{cons}}} = \frac{\sqrt{P_{\text{rad}}} \cdot \eta_{\max,0}}{\sqrt{P_{\max,0}}}. \tag{26}$$

Option II: We choose to deploy different PAs for the FC architecture and the OSPS architecture. More precisely, we assume that the underlying PA has a maximum output power of $P_{\max} = \frac{\alpha_{\text{off},0}}{\alpha_{\text{off}}} P_{\max,0}$, where α_{off} has the same value as in *Option I*. Consequently, the average radiated power and the consumed power of the underlying PA can be written as $P_{\text{rad}} = P_{\text{rad},0}$, $P_{\text{cons}} = \frac{\sqrt{P_{\max,0} \cdot \alpha_{\text{off},0} / \alpha_{\text{off}}}}{\eta_{\max}}\sqrt{P_{\text{rad}}}$. The transmitter power efficiency is given by

$$\eta_{\text{eff}} = \frac{P_{\text{rad}}}{P_{\text{cons}}} = \frac{\sqrt{P_{\text{rad}}} \cdot \eta_{\max}}{\sqrt{P_{\max,0} \cdot \alpha_{\text{off},0}}} \cdot \sqrt{\alpha_{\text{off}}}. \tag{27}$$

Note that the characteristics ($P_{\max}$ and $\eta_{\max}$) of different PAs highly depend on the operation frequency,

[7] For notation simplicity, here we ignored the slot index s and the time index t.

implementation, and technology. Aiming at illustrating how to apply the proposed analysis framework in practical system design, we will exemplify a set of PA parameters in Section V to evaluate the efficiency η_{eff} of the two architectures given in Fig. 1. For the comparison of BA and data communication algorithms, we are interested in the performance of the corresponding algorithms using the different transmitter architectures but with the same channel condition as in (11). Therefore, we assume the same total radiated power $P_{\texttt{tot}}$ constraint for both architectures in Fig. 1. In practical systems, this assumption can be satisfied by applying a certain power backoff as in *Option I* or chosing different PAs as in *Option II*. This in addition fulfills the per-antenna power constraint, such that all the underlying PAs work in their linear range with an identical scalar gain. However, we will show in Section V that, under the same radiated power constraint, different architectures may have a different power efficiency.

V. Numerical Evaluation

We now present the numerical results to evaluate the proposed precoding schemes and to illustrate the performance of different transmitter architectures as shown in Fig. 1. The BA scheme was already extensively studied in [16, 18, 19] in terms of complexity, system-level scalability, and robustness to fast channel time-variations / large Doppler spread. Hence, here we focus only on the difference in time-to-successful BA required by the two BS architectures under comparison. We consider a system with a BS using $M = 128$ antennas and $M_{\text{RF}} = 4$ RF chains. The BS simultaneously schedules $K = M_{\text{RF}} = 4$ UEs, each of which uses $N = 16$ antennas and $N_{\text{RF}} = 1$ RF chain. We assume a short preamble structure used in IEEE 802.11ad [1, 40], where the beacon slot is of duration $t_0 S = 1.891\,\mu$s. The system is assumed to work at $f_0 = 40\,$GHz with a bandwidth of $B = 0.8\,$GHz, namely, each beacon slot amounts to more than 1500 chips.

In the following simulations, otherwise stated, we will assume a fixed total radiated power constraint $P_{\texttt{tot}}$, where all the underlying PAs working in their linear range (w.r.t., per-antenna power constraint $P_{\max}$) with an identical scalar gain. The MU-MIMO channel is generated in two ways:

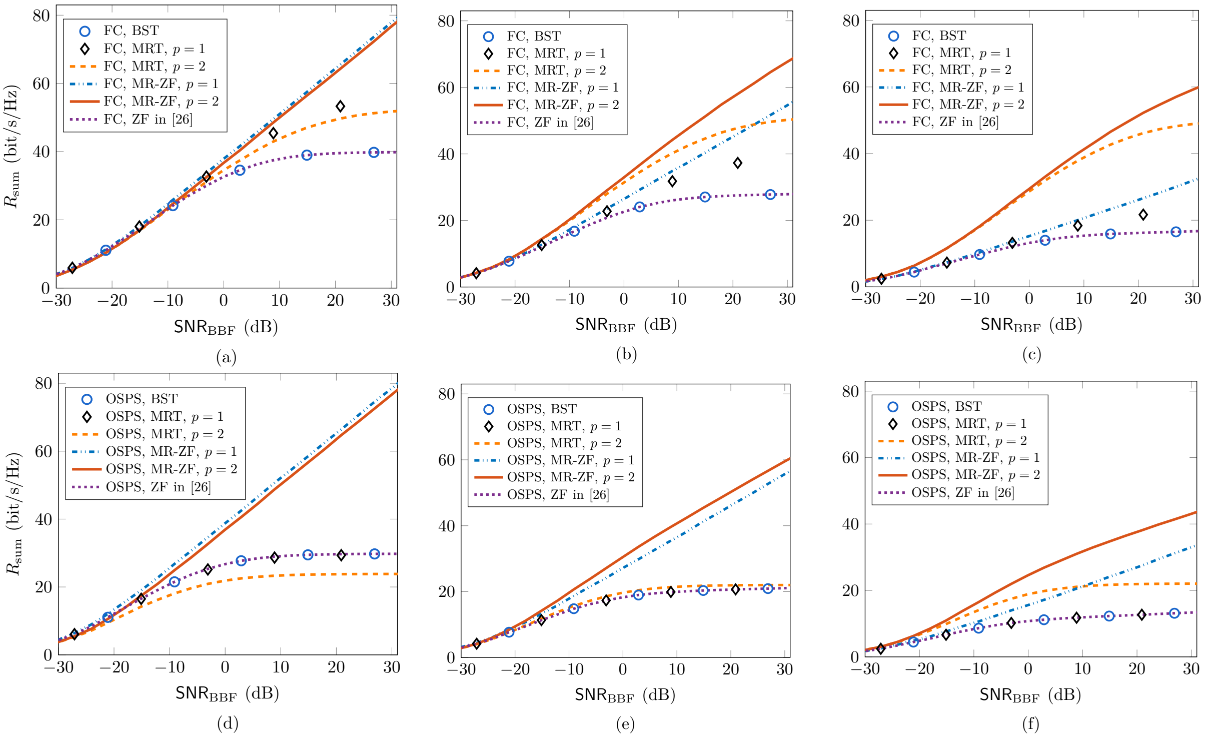
1) In Section V-A and Section V-B, we use the channel model in (1) to generate the channel matrix between each UE k and the BS. Based on the practical mmWave MIMO channel measurements in [29], we assume $L_k = 3$, $k \in [K]$, multipath components for each UE, given by $(\gamma_{k,1} = 1, \eta_{k,1} = 100)$, $(\gamma_{k,2} = 0.6, \eta_{k,2} = 10)$ and $(\gamma_{k,3} = 0.4, \eta_{k,3} = 0)$ with respect to (4). Thus, the first link can be roughly regarded as the LOS path, while the remaining links represent the NLOS paths. We also assume that the LOS paths for the simultaneously scheduled UEs are well separated in the beam domain, while all the NLOS paths are generated in a random way.

2) In Section V-C, we use the quasi-deterministic radio channel generator (QuaDRiGa) to generate the propagation channel matrix. The channel model is based on the 3GPP 38.901 standard and takes into account the spatial consistency [30, 41]. In this case, the height of the BS antenna array is set to 10 m. The beam center of the BS orientates to the ground with an elevation angle[8] of $\alpha_e = -20°$ as shown in Fig. 2 (a). The simultaneous scheduled UEs are set to 1.5 m in height and $18 \sim 25\,$m horizontally away from the BS with a downlink AoD difference of $\Delta\theta_{\text{min}} \approx 8°$ [42]. Each UE k moves towards the BS at a speed of $\Delta v_k = 1\,\text{m/s}$. We will show that the numerical results based on our proposed channel model (1) are consistent with the results based on the QuaDRiGa generator, implying that the proposed work not only theoretically but also practically provides valuable references for mmWave system design.

A. Evaluation of the Proposed Precoding Schemes

The efficiency of the proposed precoding schemes are illustrated in Fig. 6. As a comparison, we also simulate the ZF precoder proposed in [26], where the effective channel is approximated by the initial BA vectors, and only a single path is selected between each UE and the BS. As we can see from Fig. 6 (a), for the FC architecture with no blockage, all the schemes coincide with each other in the range of $\mathsf{SNR}_{\text{BBF}} \leq 0\,$dB. Whereas when $\mathsf{SNR}_{\text{BBF}} > 0\,$dB, the performance ranking of the underlying precoding schemes is as follow (MR-ZF, $p = 2$) ≈ (MR-ZF, $p = 1$) > (MRT, $p = 1$) > (MRT, $p = 2$) > (BST) ≈ (ZF in [26]). Here the MRT scheme with $p = 2$ performs worse than with $p = 1$ due to the fact of power spreading and the fact that with multiple receiving directions, the UE tends to have more interference. However, this effect is not observable in the MR-ZF scheme because of the further power coefficient tuning and interference cancellation, which results from the baseband zeroforcing. Next, by increasing the blockage probability of the strongest path while remaining unblocked for all the less strong paths between each UE and the BS, as shown in Fig. 6 (b) and Fig. 6 (c), the curves with $p = 2$ drops much less than the others (equivalent to $p = 1$), and the scheme of MR-ZF with $p = 2$ achieves the best performance. For the OSPS architecture, when there is no blockage as shown in Fig. 6 (d), in the low SNR range ($\mathsf{SNR}_{\text{BBF}} \leq -10\,$dB), all the curves (roughly) coincide with each other. Whereas, by increasing $\mathsf{SNR}_{\text{BBF}} > -10\,$dB, the precoding schemes rank (MR-ZF, $p = 2$) ≈ (MR-ZF, $p = 1$) > (MRT, $p = 1$) ≈ (BST) ≈ (ZF in [26]) > (MRT, $p = 2$). Similar with the FC case, the MR-ZF scheme for the OSPS architecture achieves the best performance when increasing the blockage probability as shown in Fig. 6 (e) and Fig. 6 (f). As a brief summary w.r.t. the given scenario, for both architectures, when the channel SNR is weak and there is no blockage, we claim that the BST scheme is

[8] In QuaDRiGa, the elevation angle 90° points to the zenith and 0° points to the horizon.



(a) (b) (c)

(d) (e) (f)

Fig. 6: The sum spectral efficiency vs. increasing $\mathsf{SNR}_{\text{BBF}}$. The blockage probability of the strongest path is given by (a) 0.0, (b) 0.3, (c) 0.6 for the FC architecture, and (d) 0.0, (e) 0.3, (f) 0.6 for the OSPS architecture.

preferred since it is rather simple but adequate to achieve good performance. However, when the channel SNR is not too weak or there are potential blockages, the MR-ZF scheme with $p > 1$ outperforms the other schemes. As a side note, in practical implementation, the choice of p should not be too large since it plays a trade-off between blockage robustness, power spreading and the overhead for additional channel estimation.

B. Fully-Connected (FC) or One-Stream-Per-Subarray (OSPS)?

Note that the performance of different architectures highly depends on the channel condition and the underlying precoders. On top of the given scenario in this paper, we jointly evaluate the architecture performance in three aspects:

Training efficiency for the initial BA phase. Let P_D denote the detection probability, i.e., the probability of finding the strongest AoA-AoD pair between the BS and a generic UE. The BA results are illustrated in Fig. 7 (a). As a comparison, we also simulate a recent time-domain BA algorithm proposed in [43], which focuses on estimating the instantaneous channel coefficients with an orthogonal matching pursuit (OMP) technique. As we can see, the proposed BA scheme requires much less training overhead than that in [43]. In addition, due to the fact that the OSPS architecture has lower angular resolution and encounters larger sidelobe power leakage than the FC case, the former requires moderately ~ 10 more beacon slots than the latter for $P_D \geq 0.95$.

Spectral efficiency for the data communication phase. To compare the spectral efficiency of the two transmitter architectures as shown in Fig. 1, we consider a no-blockage scenario and focus on two precoding schemes, i.e., the simple BST scheme and the high-performance MR-ZF scheme with $p = 2$. As we can see in Fig. 7 (b), in the range of $\mathsf{SNR}_{\text{BBF}} \leq -10\,\text{dB}$, which is more relevant in mmWave channels, all the 4 curves coincide with each other. Namely, for either the MR-ZF scheme or the BST scheme, the two architectures achieve a rather similar spectral efficiency. In contrast, when $\mathsf{SNR}_{\text{BBF}} > -10\,\text{dB}$, the MR-ZF scheme performs better. The two architectures with the MR-ZF precoding again achieve a rather similar performance.

Hardware power efficiency. To evaluate the architecture power efficiency, otherwise stated, we will consider the simple BST precoder. Also, since the modulation highly affects the power efficiency, we will take into account both the SC and the OFDM signaling in this section. We first assume a reference scenario as the baseline, i.e, the OSPS architecture using the BST precoder and a SC modulation. We use reference PAs with $P_{\text{max},0} = 6$ dBm and $\eta_{\text{max},0} = 0.3$. The backoff factor with respect to different waveforms and transmitter architectures can be written as $\alpha_{\text{off}} = 1/(P_{\text{PAPR}})$, where P_{PAPR} represents the PAPR of the input signal at a PA. The investigation

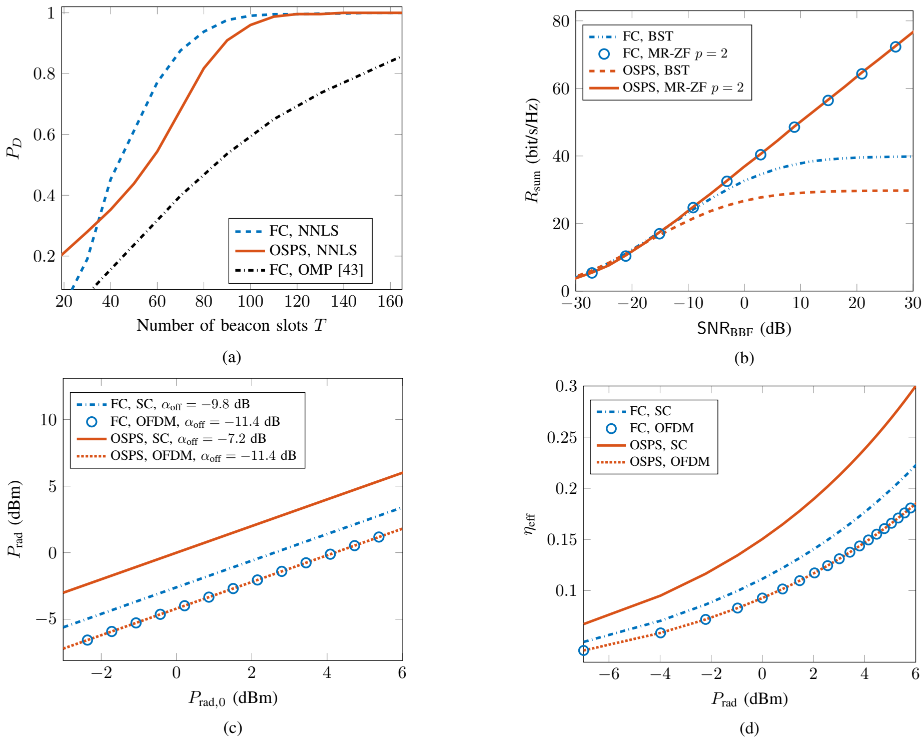


Fig. 7: The performance comparison of different transmitter architectures. (a) The initial BA detection probability vs. the training overhead with $\text{SNR}_{\text{BBF}} = -19$ dB. (b) The sum spectral efficiency vs. increasing SNR_{BBF}, without blockage. (c) The actual radiated power under *Option I* vs. the radiated power of the reference scenario. (d) The power efficiency under *Option II* vs. the actual radiated power.

for 3GPP LTE in [37] showed that with a probability of 0.999, the PAPR of the LTE SC waveform (known as SC-FDMA) is smaller than ~ 7.2 dB and the PAPR of the LTE OFDM waveform (with 512 subcarriers employing QPSK) is smaller than ~ 11.4 dB. We set P_{PAPR} to these values for the OSPS architecture. For the FC architecture, however, the input signals of the PAs are the sum of the signals from different RF chains. Since each OFDM signal can be modeled as a Gaussian random process [37] and the signals from different RF chains are independent, the PAPR of the sum is the same as of one RF chain. For the case of SC signaling, there is no clear work in the literature that shows how the sum of SC signals behaves. We simulated the sum of $M_{\text{RF}} = 4$ SC signals using the same parameters as in [37]. The result shows that with probability of 0.999 the PAPR of the sum is smaller than ~ 9.8 dB. We apply these values and without loss of generality, we choose $\alpha_{\text{off},0} = -7.2$ dB as the reference scenario. As shown in (26), by deploying the same PAs (*Option I*), the two architectures achieve the same efficiency for a given P_{rad}. However, as illustrated in Fig. 7 (c), the OSPS architecture with SC signaling (OSPS, SC) achieves the highest P_{rad}, followed by (FC, SC), (OSPS, OFDM), and (FC, OFDM). In contrast, by deploying different PAs (*Option II*)[9], Fig. 7 (d) shows that (OSPS, SC) achieves the highest power efficiency, followed by (FC, SC), (OSPS, OFDM) and (FC, OFDM).

To sum up, given the parameters in this paper, the two architectures achieve a similar sum spectral efficiency with certain precoders, but the OSPS architecture outperforms the FC case in terms of hardware complexity and power efficiency, only at the cost of a slightly longer latency for the initial BA.

C. *Simulations Based on QuaDRiGa*

In this section, we resort to the 3D geometry based channel generator QuaDRiGa [30] to show that our numerical results are quite consistent with practical mmWave communication channels.[10] More precisely, we apply our BA and precoding schemes over $\sim 3 \times 10^5$ channel snapshots generated by QuaDRiGa. These channel snapshots correspond to a short segment of time evolution, where the BS is stationary and the speed of each UE along its moving direction is 1 m/s. The simulation results with respect to different transmitter architectures are shown

[9] Since the η_{max} of different PAs highly depends on the technology, for simplicity, we assume that different PAs working in their linear range have roughly the same maximum efficiency $\eta_{\text{max},0}$.

[10] Due to the QuaDRiGa generator limits, only the no-blockage scenario is considered in this section.

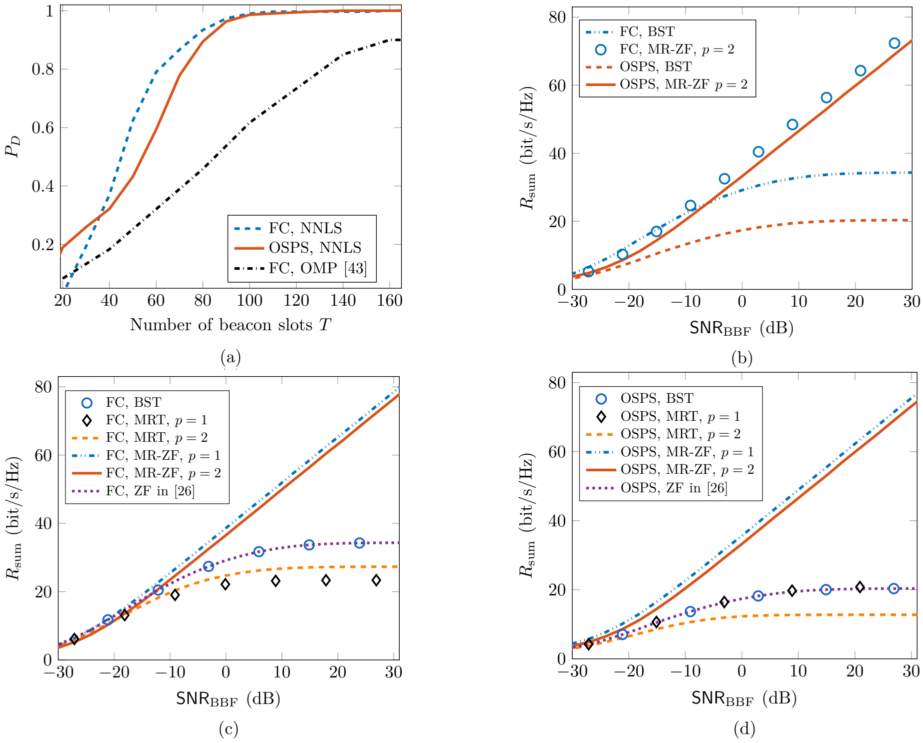


Fig. 8: The simulations based on QuaDRiGa: (a) The initial BA detection probability vs. the training overhead, with $\text{SNR}_{\text{BBF}} = -19$ dB. (b) The sum spectral efficiency of different transmitter architectures vs. increasing SNR_{BBF}. (c) The sum spectral efficiency of the FC architecture vs. increasing SNR_{BBF}. (d) The sum spectral efficiency of the OSPS architecture vs. increasing SNR_{BBF}.

in Fig. 8. As we can see from Fig. 8 (a), for the initial BA with $P_D \geq 0.95$, the FC architecture requires ~ 10 less beacon slots than the OSPS case. Whereas, for the data communication phase as shown in Fig. 8 (b), by using either the BST or the MR-ZF precoder in the low SNR range ($\text{SNR}_{\text{BBF}} \leq -15\,\text{dB}$), and using the MR-ZF precoder in the high SNR range ($\text{SNR}_{\text{BBF}} > -15\,\text{dB}$), the two architectures achieve a quite similar performance. In addition, for both architectures as shown in Fig. 8 (c) and Fig. 8 (d), respectively, all the curves coincides with each other in the low SNR range, whereas the MR-ZF precoder outperforms the rest in the high SNR range. As we can see, all the results based on the QuaDRiGa generator are quite consistent with the results based on our proposed channel model. This consistency implies that our models, schemes, results and statements are not only theoretically reliable but also practically applicable.

VI. Conclusion

In this paper, we proposed an analysis framework to evaluate the performance of typical hybrid transmitters at mmWave frequencies. In particular, we focused on the comparison of a fully-connected (FC) architecture and a partially-connected architecture with one-stream-per-subarray (OSPS) for a MU-MIMO base station using HDA beamforming. We jointly evaluated the performance of the two architectures in terms of the initial beam alignment (BA), the data communication, and the transmitter power efficiency. We used our recently proposed BA scheme and further proposed three simple precoding schemes on top of the effective channel after the BA. The precoding schemes are based on beam steering (BST), analog maximum ratio transmitting (MRT), and joint analog maximum ratio and baseband zero-forcing (MR-ZF), respectively. Particularly, both the BA scheme and the MR-ZF precoding scheme outperform the state-of-the-art counterparts in the literature. Given the parameters in this paper, our simulation results show that the two architectures achieve a similar sum spectral efficiency, but the OSPS architecture outperforms the FC case in terms of hardware complexity and power efficiency, only at the cost of a slightly longer latency for the initial BA. Therefore, the OSPS architecture emerges as a good choice for a simple and efficient design of MU-MIMO base stations operating at mmWave.

References

[1] K. Venugopal, A. Alkhateeb, N. G. Prelcic, and R. W. Heath, “Channel estimation for hybrid architecture-based wideband millimeter wave systems,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1996–2009, 2017.

[2] F. Sohrabi and W. Yu, “Hybrid digital and analog beamforming

design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, April 2016.

[3] A. Li and C. Masouros, "Hybrid analog-digital millimeter-wave MU-MIMO transmission with virtual path selection," *IEEE Communications Letters*, vol. 21, no. 2, pp. 438–441, 2017.

[4] S. S. Ioushua and Y. C. Eldar, "Hybrid analog-digital beamforming for massive MIMO systems," *arXiv preprint arXiv:1712.03485*, 2017.

[5] J. Du, W. Xu, H. Shen, X. Dong, and C. Zhao, "Hybrid precoding architecture for massive multiuser MIMO with dissipation: sub-connected or fully connected structures?" *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5465–5479, 2018.

[6] P. L. Cao, T. J. Oechtering, and M. Skoglund, "Precoding design for massive MIMO systems with sub-connected architecture and per-antenna power constraints," in *WSA 2018; 22nd International ITG Workshop on Smart Antennas*, March 2018, pp. 1–6.

[7] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[8] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-complexity technologies to enable millimeter-wave MIMO with large antenna array for 5G wireless communications," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 211–217, APRIL 2018.

[9] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 134–141, 2017.

[10] M. Majidzadeh, A. Moilanen, N. Tervo, H. Pennanen, A. Tölli, and M. Latva-aho, "Hybrid beamforming for single-user MIMO with partially connected RF architecture," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–6.

[11] M. R. Castellanos, V. Raghavan, J. H. Ryu, O. H. Koymen, J. Li, D. J. Love, and B. Peleato, "Hybrid multi-user precoding with amplitude and phase control," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[12] V. Raghavan, A. Partyka, A. Sampath, S. Subramanian, O. H. Koymen, K. Ravid, J. Cezanne, K. Mukkavilli, and J. Li, "Millimeter-wave MIMO prototype: Measurements and experimental results," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 202–209, 2018.

[13] Z. Gao, L. Dai, and Z. Wang, "Channel estimation for mmwave massive MIMO based access and backhaul in ultra-dense network," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, Conference Proceedings, pp. 1–6.

[14] J. Rodríguez-Fernández, N. González-Prelcic, K. Venugopal, and R. W. Heath Jr, "Frequency-domain compressive channel estimation for frequency-selective hybrid mmWave MIMO systems," *arXiv preprint arXiv:1704.08572*, 2017.

[15] S. Haghighatshoar and G. Caire, "The beam alignment problem in mmWave wireless networks," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 741–745.

[16] X. Song, S. Haghighatshoar, and G. Caire, "A scalable and statistically robust beam alignment technique for mm-Wave systems," *IEEE Trans. on Wireless Comm.*, vol. PP, pp. 1–1, 2018.

[17] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5014–5029, 2017.

[18] X. Song, S. Haghighatshoar, and G. Caire, "A robust time-domain beam alignment scheme for multi-user wideband mmWave systems," in *WSA 2018; 22th International ITG Workshop on Smart Antennas (to be published)*, March 2018, pp. 1–7.

[19] ——, "Efficient beam alignment for mmWave single-carrier systems with hybrid MIMO transceivers," *IEEE Transactions on Wireless Communications*, 2019.

[20] R. J. Weiler, M. Peter, W. Keusgen, and M. Wisotzki, "Measuring the busy urban 60 GHz outdoor access radio channel," in *2014 IEEE International Conference on Ultra-WideBand (ICUWB)*, Sept 2014, pp. 166–170.

[21] P. A. Eliasi, S. Rangan, and T. S. Rappaport, "Low-rank spatial channel estimation for millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2748–2759, 2017.

[22] O. El Ayach, R. W. Heath, S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," in *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, Conference Proceedings, pp. 3476–3480.

[23] D. Zhang, Y. Wang, X. Li, and W. Xiang, "Hybridly connected structure for hybrid beamforming in mmWave massive MIMO systems," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 662–674, 2018.

[24] H.-L. Chiang, W. Rave, T. Kadur, and G. Fettweis, "Hybrid beamforming based on implicit channel state information for millimeter wave links," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 2, pp. 326–339, 2018.

[25] V. Raghavan, S. Subramanian, J. Cezanne, A. Sampath, O. Koymen, and J. Li, "Directional hybrid precoding in millimeter-wave MIMO systems," in *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE, Conference Proceedings, pp. 1–7.

[26] V. Raghavan, S. Subramanian, J. Cezanne, A. Sampath, O. H. Koymen, and J. Li, "Single-user versus multi-User precoding for millimeter wave MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1387–1401, June 2017.

[27] X. Song, T. Kühne, and G. Caire, "Fully-connected vs. sub-connected hybrid precoding architectures for mmWave MU-MIMO," in *2019 IEEE International Conference on Communications (ICC) (accepted)*.

[28] N. N. Moghadam, G. Fodor, M. Bengtsson, and D. J. Love, "On the energy efficiency of MIMO hybrid beamforming for millimeter wave systems with nonlinear power amplifiers," *arXiv preprint arXiv:1806.01602*, 2018.

[29] T. Hälsig, D. Cvetkovski, E. Grass, and B. Lankl, "Statistical properties and variations of LOS MIMO channels at millimeter wave frequencies," *arXiv preprint arXiv:1803.07768*, 2018.

[30] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 6, pp. 3242–3256, 2014.

[31] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[32] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.

[33] J. G. Proakis and M. Salehi, *Digital communications*. McGraw-Hill, 2008.

[34] P. Bello, "Characterization of randomly time-variant linear channels," *IEEE Transactions on Communications Systems*, vol. 11, no. 4, pp. 360–393, 1963.

[35] A. Goldsmith, *Wireless communications*. Cambridge University Press, 2005.

[36] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2563–2579, 2002.

[37] H. G. Myung, J. Lim, and D. J. Goodman, "Peak-to-average power ratio of single carrier FDMA signals with pulse shaping," in *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on*. IEEE, Conference Proceedings, pp. 1–5.

[38] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 2012, pp. 53–64.

[39] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of massive MIMO*. Cambridge University Press, 2016.

[40] E. Perahia, C. Cordeiro, M. Park, and L. L. Yang, "IEEE 802.11 ad: Defining the next generation multi-Gbps Wi-Fi," in *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE*. IEEE, Conference Proceedings, pp. 1–5.

[41] T. S. Rappaport, G. R. MacCartney, S. Sun, H. Yan, and S. Deng, "Small-Scale, Local Area, and Transitional Millimeter Wave Propagation for 5G Communications," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6474–6490, Dec 2017.

[42] S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, and F. Burkhardt, "Quasi deterministic radio channel generator user manual and

documentation," *Fraunhofer Heinrich Hertz Institute Wireless Communications and Networks*, 2016.

[43] K. Venugopal, A. Alkhateeb, R. W. Heath, and N. G. Prelcic, "Time-domain channel estimation for wideband millimeter wave systems with hybrid architecture," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, Conference Proceedings, pp. 6493–6497.

Xiaoshen Song (S'17) received the B.Sc. degree in Communication Engineering from Northwestern Polytechnical University, Xi'an, China, in 2013, and the M.Sc. degree in Communication and Information Systems from the Institute of Electronics, University of Chinese Academy of Sciences, Beijing, China, in 2016. Her master's thesis focuses on video synthetic aperture radar (VideoSAR) system design and imaging algorithms. She is currently pursuing the Ph.D. degree with the Communications and Information Theory (CommIT) group at Technische Universität Berlin, Berlin, Germany. Her research interests include wireless communication, mmWave MIMO, and compressed sensing.

Thomas Kühne received his university degree (5-year Dipl.-Ing. equivalent to a M.Sc.) in Electrical Engineering from the University of Technology Dresden. During his master studies he focused on communication systems and circuit design. He gained professional research experience while working 3 years for the Fraunhofer Heinrich-Hertz-Institute in Berlin Germany. At the Heinrich-Hertz-Institute he developed prototypes for mm-wave communication and measurement devices for mm-wave channels. Since 2015 he works for the Communications and Information Theory group of Prof. Caire at the Technische Universität Berlin. His research interests include hardware software co-design, wireless communication systems, and signal processing.

Giuseppe Caire (S'92 – M'94 – SM'03 – F'05) was born in Torino in 1965. He received the B.Sc. in Electrical Engineering from Politecnico di Torino in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992, and the Ph.D. from Politecnico di Torino in 1994. He has been a post-doctoral research fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands) in 1994-1995, Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, and he is currently an Alexander von Humboldt Professor with the Faculty of Electrical Engineering and Computer Science at the Technical University of Berlin, Germany.

He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society & Information Theory Society Joint Paper Award in 2004 and in 2011, the Leonard G. Abraham Prize for best IEEE JSAC paper in 2019, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, the Vodafone Innovation Prize in 2015, and an ERC Advanced Grant in 2018. Giuseppe Caire is a Fellow of IEEE since 2005. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications theory, information theory, channel and source coding with particular focus on wireless communications.

6

Beam Scheduling for mmWave Relay Networks

6.1 Introduction

In mmWave communication, one effective way to mitigate the severe path loss, the sensitivity to blockages and meanwhile to increase the communication range is beamforming in combination with relaying. Having studied the beamforming issues in the previous chapters, this chapter focus on the beam scheduling problem for mmWave half-duplex (HD) relay networks. Two practical beam scheduling schemes, i.e., the deterministic edge coloring (EC) scheduler and the adaptive backpressure (BP) scheduler, will be presented to stabilize the network within its capacity range, meanwhile to guarantee small queuing backlog and end-to-end delay.

6.2 Clarification of each authors' contributions

This chapter is a journal manuscript, which is a joint work with Yahya H. Ezzeldin, Giuseppe Caire, and Christina Fragouli. I wrote this journal manuscript as the first author. This manuscript will be submitted to the journal IEEE TWC in a short time. Currently the manuscripy is still under modifications by the co-authors. The citation information is in below:

Xiaoshen Song, Yahya H. Ezzeldin, Giuseppe Caire, Christina Fragouli,"Efficient Beam Scheduling for Half-Duplex mmWave Relay Networks," IEEE Transactions on Wireless Communications, 2020. (to be submitted).

All the authors contributed to this paper. I authored the beam scheduling sections. I proposed the underlying beam scheduling methods and implemented the simulations for different beam schedulers. I also wrote the complete first draft (including all sections) of this paper.

Yahya H. Ezzeldin authored the network capacity section. He implemented the simulations for the network capacity.

Giuseppe Caire, who is my PhD supervisor, provided valuable discussions in each meeting of this work. He will also do a final modification together with Christina Fragouli for the overall draft.

6.3 Original journal article

The following article is a reprint of the original journal manuscript. It is the latest version of our work. The copyright information is given in page xii of this thesis as well as in the first page of the reprinted paper

Efficient Beam Scheduling for Half-Duplex mmWave Relay Networks

Xiaoshen Song[†], *Student Member, IEEE,* Yahya H. Ezzeldin[*], *Student Member, IEEE,* Giuseppe Caire[†], *Fellow, IEEE*, Christina Fragouli[*], *Fellow, IEEE*

***Abstract*—Millimeter wave (mmWave) communication is expected to play a central role in next generation mobile systems (5G) and beyond, by providing multi-Gbps data rates. However, the severe pathloss and sensitivity to blockages at mmWave frequencies significantly challenge practical implementations. One effective way to mitigate these effects and to increase the communication range is beamforming in combination with relaying. In this paper, we study the beam scheduling problem for mmWave half-duplex (HD) relay networks, where the relay topology can be arbitrary. Based on theoretically optimal schedule results, we first implement a network simplification procedure to reduce the network topology complexity, and then propose two practically relevant beam scheduling schemes: the deterministic edge coloring (EC) scheduler and the adaptive backpressure (BP) scheduler. The former consists of a very simple one-time computation of the sequence of scheduling states, which is then repeated periodically. The one-time computation depends on the underlying network topology, and therefore it must be repeated when such topology changes. As such, this approach is more suited to quasi-static scenarios. The latter is an "online" approach which updates scheduling weights and solves at each time slots a weighted sum rate maximization. Hence, its computational complexity may be significantly higher than that of EC, but it is better suited to dynamic time-varying scenarios. With the aid of computer simulations, we show that both the proposed schedulers guarantee network stability within the network capacity. Particularly, in comparison with a baseline scheme, the proposed schedulers achieve much smaller queuing backlogs, much smaller backlog fluctuations, and much lower packet end-to-end delays.**

***Index Terms*—mmWave, relay network, scheduling, network stability, end-to-end delay, network capacity**

I. Introduction

Migration towards millimeter wave (mmWave) bands (30-300 GHz) is considered a key enabler for next generation (5G) mobile networks and beyond [1–4]. Thanks to the large available bandwidth, a mmWave transceiver can potentially achieve individual link rates in tens of Gbps. However, compared with the traditional sub-6 GHz frequencies, mmWave communication has three main characteristics [4–6]: 1) High free-space isotropic propagation loss; 2) Highly directional propagation along the line of sight (LoS) and a small number of specular paths; 3) Vulnerability to obstacles. One effective way to mitigate these effects is beamforming in combination with relaying [2], where the former is achieved by utilizing large antenna arrays at both the transmitter (Tx) and receiver (Rx) sides and pointing their beams towards each other, and the latter refers to using intermediate nodes to relay the source signal to the destination [3].

The beamforming problem in a small cell mmWave scenario with one base station (BS) and multiple user equipments (UEs) has been studied in our previous work [7–9], in which we proposed an efficient initial beam alignment scheme to find the strongest beam pair connecting each UE and the BS, such that the consequent data communication phase can achieve large directivity and beamforming gain. Under this directional communication, the multi-user interference among different links is negligible [3, 10–12], and thus concurrent transmissions (i.e., spatial reuse) can be fully utilized to improve the transmission efficiency and to increase the network capacity.

With the increasing interest in developing small cells for mmWave communication, how to use relays to increase the coverage and to support high-rate mmWave wireless connections for dense small cell deployments remains a major challenge [3]. The relay network problem at sub-6 GHz frequencies has been well studied in the past decades [13–15]. A single source single destination relay network is a classical information theoretic model [16], and represents one step towards the understanding and designing of general multiple multicast networks. In its own right, there are important situations where one node wishes to communicate a common message to a set of other nodes (single multicast relay network), e.g., vehicle to vehicle (V2V) communication for platooning, where the head of the platoon sends commands to the other vehicles, or vehicle to everything (V2X) fast emergency control, where a road-side base station wishes to send emergency control messages to all the vehicles in a certain area [17, 18]. By taking into account the unique characteristics at mmWave frequencies, the relay nodes in a mmWave network divide the long link into some short but very high-rate links to overcome the mmWave sensitivity to blockages. In such a case, a link is active only if both nodes focus their beams to face each other, which is

[†]X. Song and G. Caire are with the Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany. [*]Y. H. Ezzeldin and C. Fragouli are with the University of California, Los Angeles, CA 90095, USA.

determined by the underlying beam scheduling scheme. The source and destination cannot communicate with each other directly because the distance between them is too large to achieve the required data rate and / or some obstacles are in between preventing direct communication. Consider a general half-duplex (HD) relay network, where all the nodes are assumed to work in HD mode thus cannot simultaneously transmit and receive [1]. Although the optimal beam directions for each node pair can be obtained through an initial BA phase, how to efficiently schedule the beams, in terms of avoiding too large queuing backlogs at the intermediate nodes as well as assuring a small end-to-end delay, becomes an important concern in practical network operations [12].

In this paper, we study the beam scheduling problem for HD mmWave relay networks with arbitrary topology. Our study will focus on developing practically relevant scheduling algorithms guided by theoretical results on the network *approximate capacity* $C_{\text{cs,iid}}$ and the optimal scheduling in mmWave network models [19].

A. Related work

While relays on sub-6 GHz bands suffer from severe interference due to their ominidirectional transmissions, the directivity of mmWave antennas significantly mitigates interference [10, 11], especially in backhaul systems [3, 12]. A large body of efforts have been made to study the mmWave relay network regime with an emphasis on one or several aspects, i.e., relay selection, congestion control, routing, scheduling and so on. However, we observe that the existing works lack the fundamental understanding of the information theoretic limit of the underlying relay network model.

The work in [1, 20] studied the relay selection problem, in which once a direct LoS link is blocked, a relay selection scheme would be activated to search a best relay path in terms of the achievable data rate. The work in [1, 20] can effectively handle an accidental blockage, however, one should note that mmWave relay network settings have potentially much more advantages than only passively dealing with blockages. The work in [3] and [21] focused on designing a multi-hop mmWave network for backhauling, range extension and improved robustness from path diversity. A main limitation in [3, 21], however, is that it considers only single path streaming [12], i.e., the selection of a single relay-path with the highest throughput for each UE. Although a claim is made to maximize the network capacity, we observe that a more fundamental capacity exploitation between the source-destination pair with possibly multiple relay paths is not taken into consideration. Actually, the throughput improvement with multiple relay paths (flows) for a single source-destination pair has been proved in [22]. The underlying idea is to inject as much traffic demands as possible so as to activate more concurrent transmission flows. Unfortunately the authors in [22] have ignored a crucial congestion control procedure, which may result in large queuing aggregation and network instability. A recent work in [12] used a network utility maximization (NUM) framework to study the operation regime of mmWave relay networks, subject to an upper delay bound and network stability. As mentioned in this paper, one important suggestion can be to randomly re-select some paths from the set of all available paths and then shift among the links with higher payoff (e.g., the minimum power consumption or the highest throughput). However, without a prior topology simplification to remove unnecessary links, the underlying method in [12] is very likely to split data into too many paths, resulting in increased signaling overhead and traffic congestion.

In general, the Shannon capacity of an arbitrary HD mmWave relay network is unknown and is notoriously hard to study, since for a network with N nodes, each of which can either transmit or receive, there exist as many as 2^N possible states. The classical network optimization scheme uses a NUM framework [12, 21, 23–25], which includes a joint congestion control and routing / scheduling, so as to accept data into the network to maximize certain utilities and to make scheduling decisions at each node, such that all accepted data are delivered to intended destinations without overflowing any queue in intermediate nodes. However, since the network capacity is unknown, all the existing algorithms suffer from the complexity of a multi-parameter tuning procedure to tackle the fundamental utility-delay tradeoff [23–25]. A recent progress in information theory [19] proposed a Gaussian HD 1-2-1 model, which corresponds to an idealized and simplified information theoretic relay network. In this model, all the nodes work in HD mode. A potential link is active only if the transmitter beam and the receiver beam are pointing at each other. In this way, the fundamental characteristic of directional transmission and necessity of two-sided (Tx and Rx) beamforming to “close the link” (i.e., achieving a sufficient received signal power after beamforming), are captured by the 1-2-1 model. The authors in [19] designed an algorithm that computes the optimal schedule to achieve the *approximate capacity* in polynomial time. The *approximate capacity* is information theoretically optimal for the Gaussian HD 1-2-1 model within a gap that depends only on the network size N but not on the topological and operating signal to noise ratio (SNR). Moreover, this *approximate capacity* can be achieved by activating only a subset of all the available links.

By noticing the great similarities between the Gaussian HD 1-2-1 model [19] and the HD mmWave relay network (i.e., very high pathloss and strong directivity), in this paper, we introduce this information theoretical result from [19] into the operation regime of HD mmWave relay networks. This helps us to understand the maximum

[1] We assume each node is equipped with an electronically steerable antenna array to beamform in the transmitting or receiving directions, so each node works in the half-duplex (HD) mode

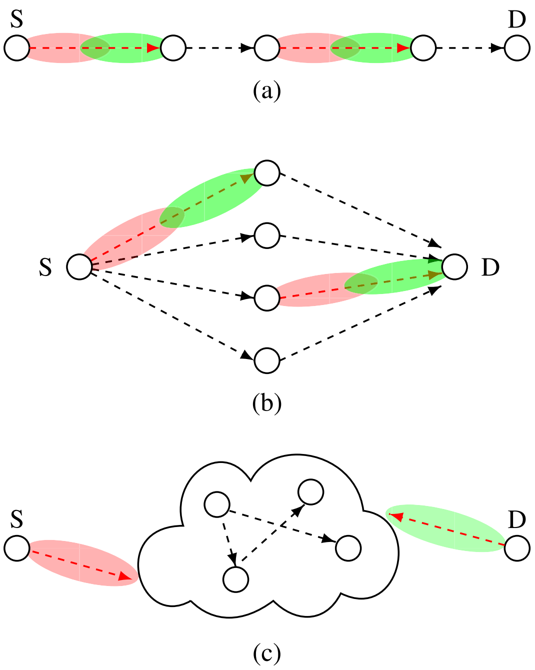


Fig. 1: An illustration of the half-duplex (HD) mmWave relay network topologies: (a) Line network; (b) Diamond network; (c) A general network. In all the networks, a potential link is activated only if the corresponding two beams are pointing at each other (e.g., the red dashed link covered by a beam pair).

potential (in terms of information theoretic capacity) of the underlying networks, and gives us insights on how to overcome the single relay-path limitation as discussed above [1, 3, 20, 21]. To the best of our knowledge, the present work is the first to exploit the information theoretic network capacity and the related network simplification into the design of practically relevant network scheduling strategies.

B. Contributions

We propose two beam scheduling schemes to approach the optimal network *approximate capacity* as well as to ensure network stability and small end-to-end delays. The main contributions of this paper are summarized as follows.

1) More general relay network topologies. In a previous work [26], the beam scheduling problem was studied for special topology HD mmWave relay networks, particularly, the line network and the diamond network shown in Fig. 1 (a) and (b), respectively. As an extension, in this paper we consider a HD mmWave relay network with an arbitrary topology as shown in Fig. 1 (c). For the convenience of explanation, we will focus on a single source-destination unicast transmission. However, the proposed work can be readily extended to multicast.

2) Exploitation of network simplification in the algorithm design. Although an arbitrary topology is considered, only a subset of the links are essential for achieving the information theoretically optimal network *approximate capacity* [19]. As a pre-processing procedure, we employ the algorithm proposed in [19] to extract the essential links and mute the non-essential ones. This pre-processing can significantly simplify the network topology and more importantly, reduce the consequent scheduling complexity. Furthermore, in traditional cross-layer policy, since the network capacity (stability) region is in general unknown, the congestion control phase usually suffers from a complex multi-parameter tuning procedure [23–25]. In contrast, in this paper we can compute the optimal network *approximate capacity* as a priori [19], and then make use of it to simplify the congestion control step.

3) New simple deterministic edge-coloring (EC) beam scheduler. By noticing the similarities between network states in HD and edge coloring in a graph, we propose a heuristic edge-coloring (EC) scheduler. The EC scheduler consists of two procedures: the path partition (PP) procedure and the alternately coloring (AC) procedure. The underlying idea is to maximally reduce the packet end-to-end delay as well as to ensure the network stability. We will show that, for most of the cases, the EC scheduler can achieve the optimal network *approximate capacity*. More importantly, the EC scheduler is rather simple, since it employs one-time computation and then periodical state repetition.

4) Efficient adaptive backpressure (BP) beam scheduler. In dynamic scenarios, the links may encounter accidental blocking. Since the aforementioned EC scheduler is more suitable for static scenarios, as an alternative approach, we propose an "online" adaptive backpressure (BP) scheduler, which would update the network state information and the scheduling decision in every time slot. We apply the concept of backpressure algorithm [23] with proper adaptation of the constraint set. We show via simulation that, benefiting from the prior topology simplification, the network queuing backlog and packet end-to-end delay performance of the proposed BP scheduler significantly outperforms the baseline scheme.

The rest of the paper is organized as follows. Section II describes the mmWave network model considered in this paper as well as the preliminary capacity results in the literature. Section III describes the two proposed beam scheduling algorithms. The algorithms are compared and evaluated and compared with a baseline approach in Section IV. The paper contributions and results are summarized in Section V.

Notation: We denote vectors, matrices, and scalars by $\mathbf{a}$, $\mathbf{A}$, and a (A), respectively. The (j,i)-th element of the matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{[j,i]}$. For an integer $K \in \mathbb{Z}$, $[K]$ denotes the index set $\{1, ..., K\}$. The notation $[a, b]$ is used to denote the closed set from $a, b \in \mathbb{R}$. We represent sets by calligraphic $\mathcal{A}$ and their cardinality with $|\mathcal{A}|$. We use $\equiv$ for identically equal, $\odot$ for Hadamard product, $| \cdot |$ for absolution and $\| \cdot \|_a$ for l_a-norm.

II. System Model

A. Channel model

We consider a general topology for a HD mmWave N-node network denoted by $\mathcal{N}_0$. The network, as shown in Fig. 1 (c), consists of $N-2$ relays assisting the communication between a source node (node 1) and a destination node (node N) [2]. We assume that the network operates in slotted time, denoted by $t \geq 0$. At any time slot, each node can point its transmitting/receiving beam towards at most one other node along the links corresponding to the edges of the network graph. In addition, all the relay nodes operate in HD mode, namely, at any time slot, each relay can be either transmitting to or receiving from at most one node. Note that the network graph describes the ensemble of potential links, i.e., the links that can transmit information provided that the beam pointing condition is satisfied. This captures the notions of blocking and distance, i.e., two nodes in the graph are connected by an edge if they are sufficiently close and there is no blocking object between them. The potential links are actually "active" when the beams of the Tx and Rx nodes connected by the link are "aligned". This captures the fact that even in LoS/proximity condition, isotropic transmission is not sufficient to achieve the desired SNR over the link, and that beam alignment is necessary. At any point in time, the network state is determined by where the node beams are pointing and whether the node is transmitting or receiving. We denote the network state by s. We can mathematically model the aforementioned network operational features by introducing two discrete set variables $\mathcal{S}_{i,t}$ and $\mathcal{S}_{i,r}$, for each node $i \in [N]$ in state s. The set variable $\mathcal{S}_{i,t}$ (respectively, $\mathcal{S}_{i,r}$) indicates the node towards which node i is pointing its Tx (respectively, Rx) beam in state s. With this, we have

$$\mathcal{S}_{i,t} \subseteq [N] \setminus \{1, i\}, |\mathcal{S}_{i,t}| \leq 1, \tag{1a}$$

$$\mathcal{S}_{i,r} \subseteq [N] \setminus \{i, N\}, |\mathcal{S}_{i,r}| \leq 1, \tag{1b}$$

$$|\mathcal{S}_{i,t}| + |\mathcal{S}_{i,r}| \leq 1, \tag{1c}$$

where $\mathcal{S}_{1,r} = \mathcal{S}_{N,t} = \emptyset$ since the source node always transmits and the destination node always receives, and where (1c) follows the HD operation, i.e., for any relay node i, if $\mathcal{S}_{i,t} \neq \emptyset$, then $\mathcal{S}_{i,r} = \emptyset$, and vice versa.

We denote by $\mathbf{H}_0 \in \mathbb{C}^{N \times N}$, the matrix of complex channel coefficients between nodes in the network, with element $\mathbf{H}_{0,[j,i]} = h_{j,i}$, $i, j \in [N]$, representing the complex channel coefficient from node i to node j. Also, since the source node can only transmit and the destination can only receive, we have $h_{1,i} = h_{j,N} \equiv 0$ for all $i, j \in [N]$. Aside from these restrictions, the node connection and channel coefficients can be arbitrary. Denote the point-to-point link capacity from node i to node j by $\mathbf{L}_0 \in \mathbb{C}^{N \times N}$, with elements $\mathbf{L}_{0,[j,i]} = l_{j,i}$, $i, j \in [N]$. Suppose that the channel inputs satisfy a unit average power constraint, hence the link capacity $l_{j,i}$ can be written as

$$l_{j,i} = \log(1 + G \cdot |h_{j,i}|^2), \quad \forall i, j \in [N], \tag{2}$$

where we assume the additive Gaussian noise at each node is independent and identically distributed (i.i.d.) as $\mathcal{CN}(0,1)$. The factor G indicates the combined BF gain of the Tx and Rx beams in alignment condition. Following [27], we refer to the HD mmWave network described above as a Gaussian HD 1-2-1 network.

Note that the Gaussian capacity in (2) is fully justified in light of our previous results in [8], where we have shown that effectively, after beam alignment, the channel for each link is reduced to a pure delay and Doppler shift (all multipath is killed by directional beamforming), hence, timing and frequency synchronization after beamforming can be easily implemented. Therefore, the unfaded Gaussian capacity for the links after beamforming (2) is a good first-order model.

B. Network capacity results

The Shannon capacity C of the considered Gaussian HD 1-2-1 network is in general unknown. However, the work in [27] has proved that C can be approximated by $C_{\text{cs,iid}}$ as follows,

$$C_{\text{cs,iid}} \leq C \leq C_{\text{cs,iid}} + \mathsf{GAP}, \tag{3a}$$

$$C_{\text{cs,iid}} = \max_{\substack{\lambda_s : \lambda_s \geq 0 \\ \sum_s \lambda_s = 1}} \min_{\substack{\bar{\mathcal{A}} \subseteq [N-1], \\ \mathcal{A} = \bar{\mathcal{A}} \cup \{1\}}} \sum_{\substack{(j,i): i \in \mathcal{A}, \\ j \in \mathcal{A}^c}} \left(\sum_{\substack{s: \\ j \in \mathcal{S}_{i,t}, \\ i \in \mathcal{S}_{j,r}}} \lambda_s \right) l_{j,i}, \tag{3b}$$

$$\mathsf{GAP} = O\big(N \log N\big), \tag{3c}$$

where (i) $\mathcal{A}$ enumerates all possible cuts in the graph representing the network topology, the source node 1 always belongs to $\mathcal{A}$ and $\mathcal{A}^c = [N] \setminus \mathcal{A}$; (ii) s represents all possible network states of the HD 1-2-1 network, with each network state s corresponding to specific values for the set variables $\mathcal{S}_{i,t}$ and $\mathcal{S}_{i,r}$ as defined in (1); (iii) $\{\lambda_s\}$, i.e., the optimization variables, are the fraction of time for which state s is active. We refer to a schedule as the collection of $\{\lambda_s\}$ for all feasible states, such that they sum up at most to 1. The expression in (3b) can be explained as maximizing a graph-theoretical min-cut over all possible feasible schedules of the HD 1-2-1 network. For any Gaussian HD 1-2-1 networks, $C_{\text{cs,iid}}$ is the *approximate capacity* of the network, where there exist a gap in comparison with the Shannon capacity C and the gap depends only on the network size N as shown in (3c).

In [19], it was shown that $C_{\text{cs,iid}}$ in (3b) can be efficiently computed by solving an equivalent linear program (LP), where the state activation times $\{\lambda_s\}$ are replaced by link

[2] For clarity, we focus on a single source-destination pair. However, the proposed work can be readily extended to multicasting as long as the (approximate) network capacity and the corresponding optimal scheduling are known.

activation times $\{\lambda_{j,i}\}$. Let $\bar{\mathbf{\Lambda}} \in \mathbb{C}^{N\times N}$ be the average link activation time fraction matrix with elements $\bar{\mathbf{\Lambda}}_{[j,i]} = \lambda_{j,i}$. Then, it follows that

$$
\begin{aligned}
C_{\text{cs,iid}} &= \max \sum_{j=1}^{N} F_{j,1}, && \text{(4a)}\\
s.t.\quad & 0 \le F_{j,i} \le \lambda_{j,i} l_{j,i}, \quad i,j \in [N], && \text{(4b)}\\
& \sum_{j\in[N]} F_{j,i} = \sum_{k\in[N]} F_{i,k},\ i \in [N-1]\setminus\{1\}, && \text{(4c)}\\
& \lambda_{j,i} \ge 0, \quad i,j \in [N], && \text{(4d)}\\
& \hat{\lambda}_{i,j} = \lambda_{j,i} + \lambda_{i,j},\ i \in [N-1], j \in [N]\setminus[i], && \text{(4e)}\\
& \hat{\lambda}_{i,j} \ge 0, \quad i \in [N-1], j \in [N]\setminus[i], && \text{(4f)}\\
& \sum_{\substack{(i,j):i=k \text{ or } j=k\\ i<j}} \hat{\lambda}_{i,j} \le 1, \quad k \in [N], && \text{(4g)}\\
& \sum_{\substack{i\in\mathcal{S},j\in\mathcal{S}\\ i<j}} \hat{\lambda}_{i,j} \le \frac{|\mathcal{S}|-1}{2},\ \mathcal{S} \subseteq [N], |\mathcal{S}| \text{ is odd}, && \text{(4h)}
\end{aligned}
$$

where, $F_{j,i}$ represents the data flow through the link of capacity $l_{j,i}$ and $\lambda_{j,i}$ represents the fraction of time in which the link from node i to node j is active. Note that although the relay links satisfy reciprocity with $l_{j,i} = l_{i,j}$, $i,j \in [N-1]\setminus\{1\}$, the corresponding link activation time $\lambda_{i,j}$ and $\lambda_{j,i}$ are not necessarily equal.

Remark *1:* Although the LP in (4) has an exponential number of constraints, it has been shown in [19] that using the ellipsoid method, the optimal solution for (4) can be found in polynomial-time in N. The approach relies on constructing a polynomial-time separation oracle for the ellipsoid method for the HD 1-2-1 network using the concept of Gomory-Hu trees [28]. We refer to [19] for more comprehensive details. Throughout this work, we will use the *approximate capacity* $C_{\text{cs,iid}}$ (4) as a prior to bound the network capacity. ◊

C. Network stability and end-to-end delay

All the exogenous arrivals first enter the transport layer at the source node, and this data is held in storage reservoirs to await acceptance to the network layer. The resulting source admission rate is determined by a congestion control mechanism. Assume that the transport layer reservoir at the source node 1 is infinitely backlogged. We denote by $x_1(t)$, the source admission rate at slot t. We say that a network is stable for an average admission rate $\bar{x}_1 = \lim\limits_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_1(t)$ if there exists a scheduling strategy such that the average backlog of all queues is finite. A well known result [23] is that the network could be stable for any $\bar{x}_1 < C$, where C is the Shannon capacity of the network. Consider a first-in-first-out (FIFO) system, we assume that only the packets currently in the node i at the beginning of slot t can be transmitted during that slot. Let $D_i(t)$ and $A_i(t)$ be the transmitting and arriving processes at node i, respectively. The arriving process $A_i(t)$ is composed of random exogenous arrivals as well as endogenous arrivals resulting from routing and transmission decisions from other nodes of the network. We assume that the $A_i(t)$ arrivals occur at the end of each slot t, so that they cannot be transmitted during that slot. Accordingly, the slot-to-slot dynamics of the queuing backlog $U_i(t)$ satisfies the following

$$
U_i(t+1) = \max\big\{U_i(t) - D_i(t), 0\big\} + A_i(t). \tag{5}
$$

To evaluate the network stability under a certain scheduling scheme, we define the network average sum backlog given by

$$
\begin{aligned}
\bar{U} &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N-1} U_i(t)\\
&= \sum_{i=1}^{N-1} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} U_i(t) \right\} = \sum_{i=1}^{N-1} \bar{U}_i,
\end{aligned} \tag{6}
$$

where $\bar{U}_i = \lim\limits_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} U_i(t)$ denotes the time average backlog in the queue of node i. Here we have implicitly ignored the destination node since the backlog at the destination node $U_N(t)$ is always zero.

Note that if the average admission rate at the source node $\bar{x}_1$ exceeds the network capacity, the network would surely become unstable regardless of the underlying scheduling schemes. However, within the network capacity region, a superior beam scheduling scheme should achieve a smaller average backlog as defined in (6). Actually by Littles's theorem [29], a small average backlog indicates also a small end-to–end delay. Here the end-to-end delay refers to the time taken for a packet to be transmitted across the network from the source node 1 to the destination node N. The end-to-end delay comes from several sources including transmission delay, propagation delay, processing delay and queuing delay. We assume that the slot duration is long enough such that the aforementioned transmission, propagation and processing time are included within each slot. Moreover, the slot duration remains constant regardless of the coding and scheduling policies. Accordingly, the most time consuming part is the queuing delay [30]. By Little's theorem [29], the average queuing delay time $\bar{\omega}$ that a packet spends in the network satisfies $\bar{\omega} = \bar{U}/\bar{x}_1$. In the later simulation section, we will evaluate the network stability as well as the packet end-to-end delay performance with respect to (w.r.t.) different scheduling schemes.

D. The NUM framework for joint congestion control and scheduling

When the exogenous arrival rates are outside the network capacity region, the network cannot be stabilized without a congestion control mechanism to limit the amount of data that is admitted. The classical NUM (network utility maximization) framework controls the admission

Algorithm 1: The network utility maximization (NUM) framework for joint congestion control and scheduling

Initialization:

Choose $V > 0$ and $x_{\max} > 0$ as constant parameters. Initialize the queue backlog at the beginning of time slot $t = 1$ as $U_i(1) = 0$, $\forall i \in [N]$.

Iteration:

In each time slot $t \geq 1$, repeat the following three steps.

1. **Scheduling**: At the beginning of each slot t, define the differential backlog weight matrix $\mathbf{W}(t)$, with elements $\mathbf{W}(t)_{[j,i]} = \max\{U_i(t) - U_j(t), 0\}$ for all $j, i \in [N]$. Then choose the scheduling decision matrix $\mathbf{\Lambda}(t) \in \mathbb{C}^{N\times N}$ and the link rate allocation matrix $\mathbf{R}(t) \in \mathbb{C}^{N\times N}$ as the solution to the following optimization problem

$$\mathbf{\Lambda}(t), \mathbf{R}(t) = \arg\max \sum_{j=1}^{N}\sum_{i=1}^{N} (\mathbf{W}(t) \odot \mathbf{\Lambda}(t) \odot \mathbf{R}(t))_{[j,i]} \tag{7a}$$

$$s.t. \quad \mathbf{R}(t)_{[j,i]} \leq l_{j,i}, \quad \forall i, j \in [N] \tag{7b}$$

$$\mathbf{\Lambda}(t) \in \mathcal{I}. \tag{7c}$$

where $l_{j,i}$ is the link capacity defined in (2), $\mathcal{I}$ consists of all feasible link activation sets, i.e., all sets of links that can be simultaneously activated.

2. **Congestion control**: For the source node $i = 1$, calculate the admission rate $x_1(t)$ as the solution to the following optimization problem

$$x_1(t) = \arg\max V\cdot g_1(x_1(t)) - x_1(t) \cdot U_1(t) \tag{8a}$$

$$s.t. \quad x_1(t) \in [0, x_{\max}], \tag{8b}$$

where the utility function $g_1(\cdot)$ is assumed to be non-decreasing and concave, $x_{\max}$ is a large constant number.

3. **Queuing update**: For each node $i \in [N-1]$, update the queue backlogs for the next time slot as

$$U_i(t+1) = U_i(t) - \sum_{j\in\mathcal{O}(i)} \mathbf{R}(t)_{[j,i]} + \sum_{j\in\mathcal{I}(i)} \mathbf{R}(t)_{[i,j]} + x_1(t) \cdot \mathbf{1}_{\{i=1\}}, \tag{9}$$

where $\mathcal{O}(i)$ and $\mathcal{I}(i)$ represent the sets of outgoing links and incoming links of node i, respectively. $\mathbf{1}_{\{\cdot\}}$ is an indicator function that takes the value 1 if the underlying condition is true, otherwise 0.

congestion via an optimization of the utility function $g_0(x_1(t))$which represents the "satisfaction" received by sending the commodity data from source node 1 to the destination node N at an admission rate of $x_1(t)$. The network is then stabilized by applying the backpressure algorithm at each time slot t [12, 23, 25, 31]. Define $\mathbf{R}(t) \in \mathbb{C}^{N\times N}$ and $\mathbf{\Lambda}(t) \in \mathbb{C}^{N\times N}$ as the link rate allocation and the scheduling decision matrices at time slot t, respectively. The scheduling decision matrix has elements $\mathbf{\Lambda}(t)_{[j,i]} = 1$ if link (i,j) is activated, otherwise 0. We summarize the conceptual NUM framework in Algorithm 1. As discussed before, in most of the literature it is not clear how to tune the algorithm parameters V and $x_{\max}$, which often needs an empirical trial-and-error procedure. In contrast, by knowing $C_{\text{cs,iid}}$ in our proposed scheme, we can easily get rid of the parameters V and $x_{\max}$, setting the source admission rate as a simple constant. Moreover, we exploit the insight on the underlying optimization problem to do network simplification so as to significantly reduce the network topology / scheduling complexity. In what follows we will present our scheduling schemes in more detail.

III. Proposed Beam Scheduling methods

In this section we first introduce a pre-processing procedure to simplify the network topology, on top of which two beam scheduling schemes are provided: the deterministic edge coloring (EC) scheduler and the adaptive backpressure (BP) scheduler.

A. The prior network topology simplification

The topology of the original network $\mathcal{N}_0$ hinges on the link capacity matrix $\mathbf{L}_0$. Namely, a link connection (i,j) exists only if $\mathbf{L}_{0,[j,i]} = l_{j,i} > 0$. However, based on the link activation time $\bar{\mathbf{\Lambda}}_{[j,i]} = \lambda_{j,i}$ as calculated in (4), in order to approach the network *approximate capacity* $C_{\text{cs,iid}}$, some links are not necessary to be activated at all. Aiming at reducing the scheduling complexity, we define a new associate N-node network $\mathcal{N}$ with link capacity $\mathbf{L}$ given by

$$\mathbf{L}_{[j,i]} = \begin{cases} l_{j,i}, & \lambda_{j,i} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The new network $\mathcal{N}$ is a simplified version of the original network $\mathcal{N}_0$ and contains only the links that are necessary to use w.r.t. the network *approximate capacity* $C_{\text{cs,iid}}$. We consider a running example as shown in Fig. 2 (a). Without loss of generality, we assume that the link capacities $l_{j,i}$ are in the unit of packet per slot (packet/slot). The link

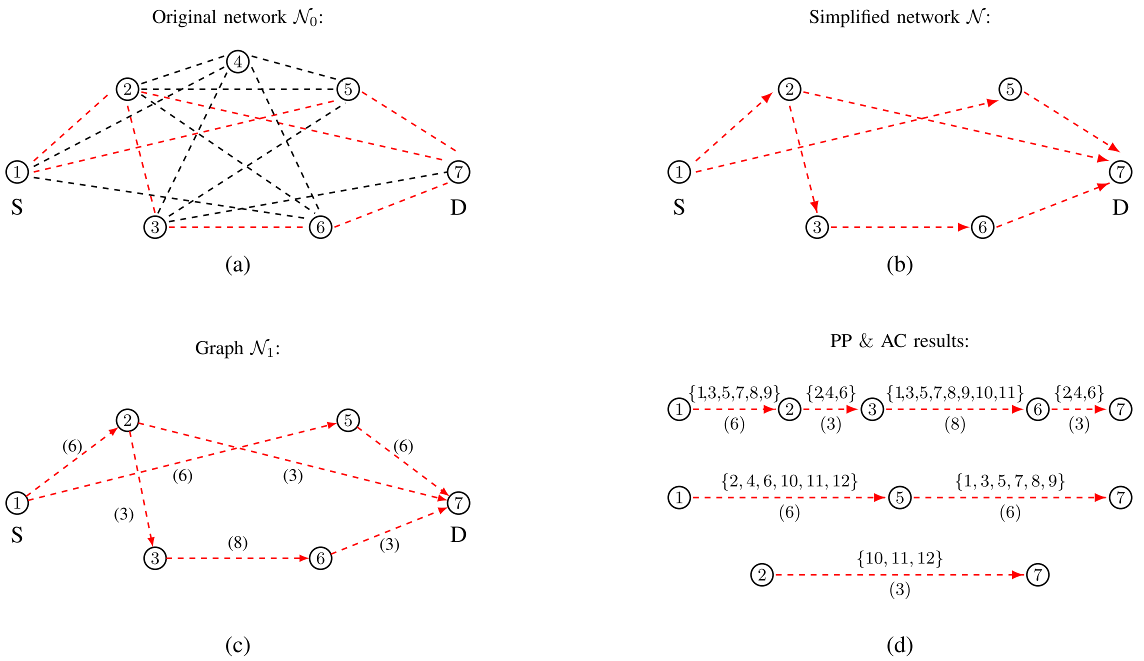


Fig. 2: An illustration of network topology simplification and the edge coloring procedures: (a) The topology of the original network $\mathcal{N}_0$; (b) The topology of the simplified network $\mathcal{N}$; (c) The associate graph $\mathcal{N}_1$, where digits in the parentheses $(\cdot)$ indicates the number of parallel edges $n_{j,i}$ assigned to link (i, j); (d) The path partitioning (PP) and alternately coloring (AC) results, where the digits inside the braces $\{\cdot\}$ indicate the color set $\mathcal{E}_{j,i}$ assigned to link (i, j).

capacity matrix $\mathbf{L}_0$ of the original network $\mathcal{N}_0$ is given by

$$\mathbf{L}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 14 & 0 & 16 & 14 & 6 & 10 & 0 \\ 0 & 16 & 0 & 10 & 10 & 6 & 0 \\ 12 & 14 & 10 & 0 & 12 & 2 & 0 \\ 16 & 6 & 10 & 12 & 0 & 0 & 0 \\ 12 & 10 & 6 & 2 & 0 & 0 & 0 \\ 0 & 12 & 6 & 0 & 16 & 16 & 0 \end{bmatrix}. \tag{11}$$

Based on the approach in (4)(10), the link activation time fraction $\bar{\mathbf{\Lambda}}$ and the link capacity $\mathbf{L}$ of the simplified network $\mathcal{N}$ w.r.t. the overall network approximate capacity $C_{\text{cs,iid}}$ satisfy

$$\bar{\mathbf{\Lambda}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{2} & \frac{1}{4} & 0 \end{bmatrix} \tag{12}$$

and

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 14 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 16 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 & 16 & 16 & 0 \end{bmatrix}. \tag{13}$$

As we can see from Fig. 2 (b), most of the links in the simplified network $\mathcal{N}$ are muted ($\mathbf{L}_{[j,i]} = 0$) since they would never be activated ($\lambda_{j,i} = 0$). In what follows, we will completely ignore the muted links and concentrate only on the simplified network $\mathcal{N}$, because $\mathcal{N}$ is sufficient to approach the *approximate capacity* $C_{\text{cs,iid}}$ of the original network given in (4). More importantly, focusing on the simplified network $\mathcal{N}$ can significantly reduce the scheduling complexity.

Remark 2 *(Congestion control with known approximate capacity):* Having obtained the *approximate capacity* $C_{\text{cs,iid}}$ in (4), the congestion control in our case reduces to a constant threshold, given by

$$x_1(t) = \hat{x}_1 \leq C_{\text{cs,iid}}, \tag{14}$$

where $\hat{x}_1$ represents a non-negative constant. Note that in the case where the source has an infinite reservoir of information bits (e.g., video server, web server, database, where the bits are pre-stored at the source, and not generated at random according to a random arrival process), it makes sense to set $x_1(t)$ as some constant number of admired bits per unit time, since this makes the arrival process deterministic and yields less jitters (fluctuations) that tend to increase the average queuing backlog. $\Diamond$

Given this cross-layer congestion control described above, in next subsections, we present two beam scheduling schemes to stabilize the network as well as to achieve small end-to-end delays.

B. The deterministic edge coloring (EC) beam scheduler

The EC scheduler leverages the similarities between network states in HD and edge coloring in a graph [32, 33]. In particular, an edge coloring assigns colors to edges in a graph such that no two adjacent edges are colored with the same color. Similarly in HD, a network state cannot be a receiver and a transmitter simultaneously. Consider the same running example as illustrated in Fig. 2 with link capacity matrix $\mathbf{L}$ (after a prior network simplification) and its associated link activation times matrix $\bar{\mathbf{\Lambda}}$, as defined in (12). Let M be a common multiple of the denominators in $\bar{\mathbf{\Lambda}}$. We construct an associate multigraph $\mathcal{N}_1$ w.r.t. the network $\mathcal{N}$, as illustrated in Fig. 2 (c), where the set of nodes is the same as in $\mathcal{N}$ and each link (i, j) with capacity $\mathbf{L}_{[j,i]} > 0$ is replaced by $n_{j,i}$ parallel edges, given by

$$n_{j,i} = \begin{cases} M \cdot \bar{\mathbf{\Lambda}}_{[j,i]}, & \mathbf{L}_{[j,i]} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

It is not difficult to see that $n_{j,i} \in \mathbb{Z}, \forall i, j \in [N]$. It follows that the maximum node degree Δ of the graph $\mathcal{N}_1$ can be written as

$$\Delta = \max_{i,j,k \in [N]} \{n_{j,k} + n_{k,i}\}. \tag{16}$$

In the running example we have $M = 12$, $\Delta = 12$. The values of $n_{j,i}$ are labeled aside each edge in Fig. 2(c).

Our proposed EC scheduler is applied on $\mathcal{N}$ and $\mathcal{N}_1$ consecutively, and consists of two procedures, namely, the Path Partitioning (PP) and the Alternating Coloring (AC) procedures described below.

1) Path Partitioning (PP): The PP procedure is based on network $\mathcal{N}$ and gives a partition of the network links into independent paths, such that each link in $\mathcal{N}$ appears only in one path. The main motivation of the PP procedure is to provide a logical order for the consequent coloring procedure. Also, each path resulted from the PP procedure corresponds to a simple line network with a single flow direction, that logically align with the overall data flow from the source node 1 to the destination node N. This enables us to implement a consequent alternating coloring (AC) procedure (as provided in the next paragraph) to reduce the packet delay [34]. The PP procedure can be applied as follows: Choose a node in $\mathcal{N}$ with no incoming edges. Traverse an edge from that node to another, and erase that edge. Continue traversing and erasing edges until a node with no outgoing edge is reached. This gives a path in the partition. Then choose a new start node and repeat the process. Do this until no possible start node remains. We summarize the PP procedure in Algorithm 2 with step 1). Define $\mathcal{P}$ as the set disjoint paths from the PP procedure and let P be the number such paths. For the running example we have $P = 3$ and the paths in $\mathcal{P}$ are illustrated in Fig. 2(d).

2) Alternating Coloring (AC): For each path in $\mathcal{P}$, replace each link in the path with its corresponding parallel edges in $\mathcal{N}_1$ with the number defined as in (15). We color the edges in an alternative manner, such that any data packets entering into the network $\mathcal{N}$ will be transmitted towards the destination node as soon as possible. More precisely, for each path, extract one non-colored edge for each link if there exists. Start from the first non-colored edge. Consecutive edges in the path are alternately colored with the smallest legal color. Continue this extracting and alternately coloring process until no no-colored edge remains. We summarize the AC procedure in Algorithm 2 with step 2). The color assignments for the running example are shown in Fig. 2(d). Note that, as illustrated in Fig. 2(d), the coloring is done greedily. The first path p_1 is colored before moving on to color the second path p_2 and so on. When coloring the edge in p_2 for the first hop, the smallest legal color is #2 since color #1 is already used for an edge connected to node 1 in p_1.

Define K as the total number of unique colors used in Algorithm 2, and $\mathcal{E}_{j,i}$ as the set of colors assigned to link (i, j). Once the two procedures in Algorithm 2 have finished, the consequent beam scheduling would reduce to a simple deterministic repetition among the K states, where each state indexed with $k \in [K]$ corresponds to activating the links associated to the k-th color. Particularly, for each time slot $t \geq 1$, the scheduling decision is given by

$$\mathbf{\Lambda}(t)_{[j,i]} = \mathbf{1}_{\{\hat{t} \in \mathcal{E}_{j,i}\}}, \quad i, j \in [N], \;\; \hat{t} = ((t-1) \bmod K) + 1, \tag{17}$$

where $\mathbf{\Lambda}(t)_{[j,i]} = 1$ indicates that link (i, j) is activated at slot t, and it is idle otherwise. The actual transmission rate for link (i, j) at slot t is given by

$$\mathbf{R}(t)_{[j,i]} = \min\{\mathbf{L}(t)_{[j,i]} \cdot \mathbf{\Lambda}(t)_{[j,i]}, U_i(t)\}. \tag{18}$$

Accordingly, the slot-to-slot queuing evolution follows (9).

Lemma *1: The proposed EC scheduler can achieve at least $\frac{1}{2}$ of the network approximate capacity, i.e.,*

$$\frac{1}{2} C_{\text{cs,iid}} < C_{\max} \leq C_{\text{cs,iid}}, \tag{19}$$

where $C_{\max}$ denotes the maximum achievable data rate under the EC scheduler.

Proof. With a total number of K colors used in the EC scheduler, we have $C_{\max} = \frac{\Delta}{K} C_{\text{cs,iid}}$, where $K \geq \Delta$ since no two incident edges have the same color. Accordingly, we have proved the upper bound in (19) with $C_{\max} \leq C_{\text{cs,iid}}$. On the other hand, the number of colors K w.r.t. network graph $\mathcal{N}_1$ satisfies $K \leq 2\Delta - 1$. The proof is straightforward, since for any given edge, there are at most $\Delta - 1$ colored edges incident to each of its endpoints; thus, even if all $2\Delta - 2$ edges have different colors, there is still a single usable color. Accordingly, the proposed EC scheduler is guaranteed to achieve at least $\frac{1}{2}$ of the *approximate capacity* $C_{\text{cs,iid}}$, with $C_{\max} > \frac{1}{2} C_{\text{cs,iid}}$. This is a very nice performance guarantee, given the low complexity and simplicity of scheduling and the fact that it yields very low latency (as shown later). Actually, a

Algorithm 2: The two procedures for the edge coloring (EC) beam scheduler

```
1) Procedure path partition (PP)
Initialization: Make $\mathcal{P}$ an empty list; Make $\mathcal{V}$ a set that contains all the nodes in the network $\mathcal{N}$;
while $\mathcal{V}$ is nonempty do
    let $v$ be the first node in $\mathcal{V}$ that has no incoming links;
    delete $v$ from $\mathcal{V}$;
    if node $v$ has nonzero outgoing links then
        make a new path $\rho$ empty;
        $\hat{v} := v$;
        while node $\hat{v}$ has nonzero outgoing links do
            let $(w, \hat{v})$ be an outgoing link of $\hat{v}$;
            delete $(w, \hat{v})$ from $\mathcal{N}$;
            put $(w, \hat{v})$ in $\rho$;
            $\hat{v} := w$;
        end
        put path $\rho$ in $\mathcal{P}$;
    end
    if node $v$ has nonzero degree then
        put $v$ in $\mathcal{V}$
    end
end
2) Procedure alternately coloring (AC)
Initialization: Define $P$ as the number of paths in $\mathcal{P}$; Define $\bar{v}_p$, $p \in [P]$, as the number of nodes in the $p$-th path; Define
  $\mathcal{E}_{j,i}$ as the set of colors assigned to link $(i, j)$, $\mathcal{E}_{j,i}$ are initialized as empty; Replace each link in the $p$-th path with parallel
  edges as defined in (15);
for each path $p$ in $\mathcal{P}$ do
    while there still exists non-colored edge in the $p$-th path do
        for $k$ in $[\bar{v}_p - 1]$ do
            assign the smallest legal color $e \in \mathbb{Z}_+$ to one of the non-colored edges in the $k$-th hop of path $p$;
            put $e$ in the corresponding set $\mathcal{E}_{j,i}$;
        end
    end
end
```

classical upper bound [35] on coloring the multigraph $\mathcal{N}_1$ states that an optimal coloring scheme (one that uses the minimum number of colors possible) uses at most $\Delta + \mu$ colors, where μ is the multiplicity of graph $\mathcal{N}_1$, i.e, the maximum number of edges in any bundle of parallel edges. Although not theoretically proven, we have observed in our simulations that the number of colors K used by our EC scheduler satisfies that $K \leq \Delta + \mu$. Hence in most of the cases, the EC scheduler can guarantee much more than $\frac{1}{2}C_{\text{cs,iid}}$. □

C. The adaptive backpressure (BP) beam scheduler

The EC scheduler described in the previous subsection is rather simple, since once the K-color states are obtained, the network scheduling becomes deterministic, namely, the scheduler just needs to periodically repeat the K states defined by (17). However, since the EC scheduler is one-time predetermined by the network link capacities $l_{j,i}$, the scheduler is mostly favorable for quasi-static scenarios, and needs to be recomputed whenever some significant changes in the network topology or potential link capacities occur. As an alternative approach for time-varying scenarios, we will consider "online" dynamic scheduling policies that are guaranteed to achieve stability for all $x_1(t) \leq C_{\text{cs,iid}}$. In particular, we consider the well-known BP algorithm [23] which is well understood to stabilize the network whenever the source admission rate lies within the capacity region of the network.

Define the differential backlog weight matrix $\mathbf{W}(t) \in \mathbb{C}^{N\times N}$ with elements given by

$$\mathbf{W}(t)_{[j,i]} = \begin{cases} \max\{U_i(t) - U_j(t), 0\}, & \lambda_{j,i} > 0 \\ 0, \quad \text{otherwise}, \end{cases} \tag{20}$$

where as mentioned before, we have intentionally ignored all the links that would never be activated ($\lambda_{j,i} = 0$) in terms of achieving the network *approximate capacity* $C_{\text{cs,iid}}$. As a consequence, the scheduler only needs to deal with a much smaller set of links, which can significantly reduce the scheduling complexity. Similarly, define the candidate transmit rate matrix $\hat{\mathbf{R}}(t) \in \mathbb{C}^{N\times N}$ with elements given by

$$\hat{\mathbf{R}}(t)_{[j,i]} = \begin{cases} \min\{U_i(t), \mathbf{L}_{[j,i]}\}, & \lambda_{j,i} > 0 \\ 0, \quad \text{otherwise}. \end{cases} \tag{21}$$

Then choose the scheduling matrix $\mathbf{\Lambda}(t)$ at slot t as the solution to the following binary integer program (BIP) optimization problem

$$\mathbf{\Lambda}(t) = \arg\max \sum_{j=1}^{N}\sum_{i=1}^{N} \Big(\mathbf{W}(t) \odot \hat{\mathbf{R}}(t) \odot \mathbf{\Lambda}(t)\Big)_{[j,i]} \tag{22a}$$

$$\begin{aligned} s.t. \quad & \mathbf{\Lambda}(t)_{[j,i]} \in \{0,1\}, && \text{(22b)} \\ & \|\mathbf{\Lambda}(t)_{[:,i]}\|_1 + \|\mathbf{\Lambda}(t)_{[i,:]}\|_1 \leq 1, i \in [N], && \text{(22c)} \end{aligned}$$

where (22b) denotes the binary activation constraint, and (22c) indicates the HD 1-2-1 network operating constraint, i.e., a node can at most receive from (or transmit to) one node and cannot do both simultaneously. Accordingly, the actual link transmission rate due to (22) is given by

$$\mathbf{R}(t)_{[j,i]} = \hat{\mathbf{R}}(t)_{[j,i]} \cdot \mathbf{\Lambda}(t)_{[j,i]}, \tag{23}$$

and the slot-to-slot queuing evolution follows the procedure of (9).

It is well-known that BP is able to stabilize the network the source admission rate lies within the capacity region of the network [23]. In the following section, we compare the performance of our two proposed algorithms with a "standard" baseline scheme, and show how applying the EC and BP schedulers on top of the simplified network $\mathcal{N}$ can significantly reduce the scheduling complexity, thus, bearing smaller queuing backlogs. Also, the packets will experience much smaller packet end-to-end delays.

Remark *3:* Note that although (22) is an integer linear program, the convex hull of its feasible points can be represented by a set of linear inequalities using Edmonds [36] *matching polytope*. The matching polytope, although having an exponential number of constraints, can be efficiently solved in polynomial-time using the ellipsoid method [19]. ♢

IV. Numerical Results

In this section, we investigate the numerical performance of the proposed EC and BP schedulers, and compare them with a "standard" baseline scheme. We start off by presenting our simulation scenarios and then discuss our baseline comparison before delving into the simulation results.

Simulated Examples. We consider two running examples (two random network topologies) and denote the two examples by Exp1 and Exp2, respectively. The first running example Exp1 is the same network $\mathcal{N}_0$ used in the previous subsections with total number of nodes $N = 7$ and the link capacity matrix $\mathbf{L}_0$ given by (11). By solving (4) and (10) for Exp1, the network *approximate capacity* is $C_{\text{cs,iid}} = 15$ packets/slot. The link activation time fraction matrix $\bar{\mathbf{\Lambda}}$ and the link capacity matrix $\mathbf{L}$ for the simplified network $\mathcal{N}$ are shown in (12). The second running example Exp2 again has $N = 7$ nodes. The link capacity matrix for Exp2 is given by

$$\mathbf{L}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 6 & 9 & 8 & 9 & 0 \\ 7 & 6 & 0 & 7 & 6 & 6 & 0 \\ 9 & 9 & 7 & 0 & 6 & 7 & 0 \\ 9 & 8 & 6 & 6 & 0 & 6 & 0 \\ 8 & 9 & 6 & 7 & 6 & 0 & 0 \\ 0 & 6 & 7 & 6 & 6 & 7 & 0 \end{bmatrix}. \tag{24}$$

Following the approach in (4)(10), the link activation time fraction $\bar{\mathbf{\Lambda}}$ and the link capacity $\mathbf{L}$ of the simplified network $\mathcal{N}$ satisfy

$$\bar{\mathbf{\Lambda}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{7}{18} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \end{bmatrix} \tag{25}$$

and

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 & 7 & 0 \end{bmatrix}, \tag{26}$$

respectively. The network *approximate capacity* reads $C_{\text{cs,iid}} = 7$ packets/slot. Note that unless otherwise stated, we will assume that simulated network is static.

Baseline scheme. As a comparison, we also consider a baseline backpressure scheme from [23] and denote it by BPo. The baseline scheme BPo has been commonly used in the literature [12, 21, 31], which uses the same NUM framework as in Algorithm 1 and the underlying scheduling is also based on the concept of backpressure. In contrast with our proposed beam schedulers, the BPo does not exploit the knowledge of the network *approximate capacity* and the resulting network simplification. As a result, the congestion control in BPo requires a complex multi-parameter $(V, x_{\max})$ tuning procedure so as to tackle the fundamental utility-delay tradeoff as illustrated in (8). Moreover, the scheduling procedure in BPo is directly implemented on the original network topology $\mathcal{N}_0$, which consequently encounters a larger scheduling complexity.

In what follows, we provide numerical results that: 1) Evaluate the performance of our proposed schemes; 2) Compare the performance of our proposed schemes and the aforementioned baseline scheme. We also separately provide simulation results for time-varying scenarios, where the network encounters accidental blockages.

A. Performance evaluation of the proposed schemes

We consider three performance metrics, i.e., the network stability, the packet end-to-end delay and the queuing evolution in time-varying scenarios with accidental blockages.

In terms of network stability, Fig. 3 (a) illustrates the numerical performance for the example network Exp1. Here the network *approximate capacity* is $C_{\text{cs,iid}} = 15$ packets/slot , shown by the vertical dotted line. As we can see from Fig. 3 (a), within the network capacity region $x_1(t) \leq C_{\text{cs,iid}}$, both the EC scheduler and the BP scheduler can guarantee a finite average backlog $\bar{U}$, i.e., can effectively stabilize the network. In particular,

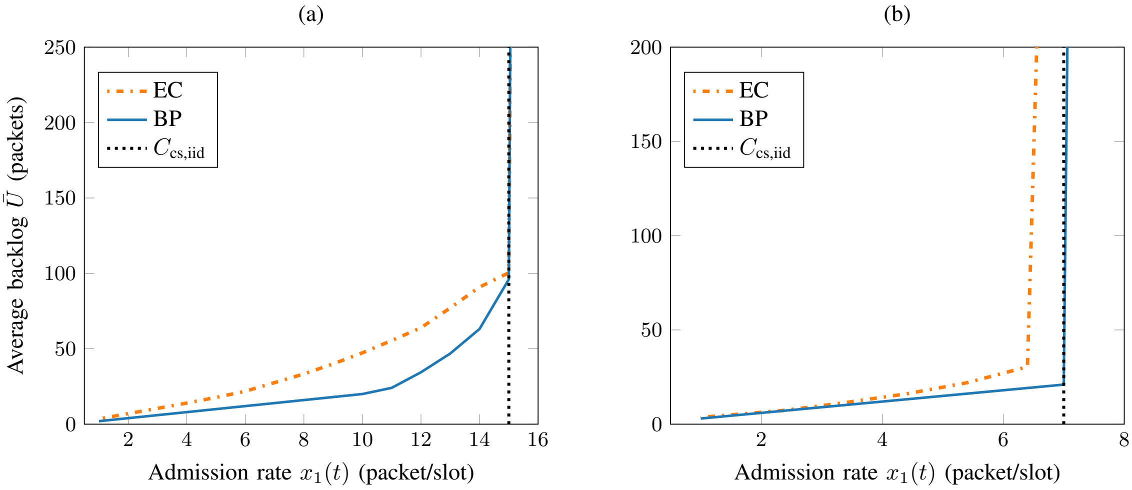


Fig. 3: The average backlog $\bar{U}$ with respect to different source admission rate $x_1(t)$. (a) Evaluation on the running example Exp1, where the network *approximate capacity* reads $C_{\text{cs,iid}} = 15$ packets/slot, the maximum node degree of the corresponding associate multigraph is $\Delta = 12$, and the total number of unique colors used in the EC scheduler reads $K = 12 = \Delta$. (b) Evaluation on the running example Exp2, with $C_{\text{cs,iid}} = 7$ packets/slot, $\Delta = 18$ and $K = 19 > \Delta$.

since the total number of unique colors used in the EC scheduler equals the maximum node degree of the corresponding associate multigraph $\mathcal{N}_1$ with $K = \Delta = 12$, the maximum achievable data rate under the EC scheduler reaches exactly the maximum network capacity point with $C_{\max} = \frac{\Delta}{K} C_{\text{cs,iid}} = C_{\text{cs,iid}}$. The stability evaluation w.r.t. the example network Exp2 is shown in Fig. 3 (b), where the network *approximate capacity* is $C_{\text{cs,iid}} = 7$ packets/slot. In this particular example, the total number of unique colors used in the EC scheduler is slightly greater than the maximum node degree of the corresponding associate multigraph $\mathcal{N}_1$ with $K = 19$ colors and $\Delta = 18$ maximum degree. As a result, the maximum achievable data rate for Exp2 using the EC scheduler is $C_{\max} = \frac{\Delta}{K} C_{\text{cs,iid}} < C_{\text{cs,iid}}$. In addition to this, the numerical performance in Exp2 is similar to that in Exp1. As shown in Fig. 3 (b), within the corresponding capacity ranges, i.e., $x_1(t) \leq C_{\max}$ for the EC scheduler and $x_1(t) \leq C_{\text{cs,iid}}$ for the BP scheduler, both the two schedulers can efficiently stabilize the network with finite average backlog $\bar{U}$. With the same source admission $x_1(t) \leq C_{\max}$, the average backlog $\bar{U}$ with the BP scheduler is slightly smaller than that with the EC scheduler.

Note that, although the BP scheduler shows slight apparent benefits over the EC scheduler as seen in Fig. 3 (a)-(b), it should be contrasted with its operational complexity. The BP scheduler must solve a weighted sum rate maximization (22) at each time slot, while the EC scheduler uses only one-time computation and then periodical state repetition.

In terms of packet end-to-end delay, Fig. 4 (a)-(b) illustrates the numerical performance w.r.t. Exp1. Here the end-to-end delay indicates how long the packets are delayed in the queues during the transmission from the source node to the destination node. The cumulative density function (CDF) of the packet delay in Fig. 4 indicates the probability that the packet end-to-end delay is smaller than the specified delay. The packet delay distribution for example Exp1 under the EC scheduler is shown in Fig. 4 (a), where the source admission rate is set as $x_1(t) = 12$ packets/slot. As we can see, with probability 1 the end-to-end delay of each individual data packet is smaller than 13 slots. By increasing the source admission rate from $x_1(t) = 12$ packets/slot to $x_1(t) = 15$ packets/slot, the maximum end-to-end delay increases to 15 slots. The BP scheduler achieves similar performance as shown in Fig. 4 (b). As we can see, with probability 1 the packet end-to-end delay is smaller than 3 slots for source admission $x_1(t) = 12$ packets/slot. This maximum delay shifts to 13 slots for source admission $x_1(t) = 15$ packets/slot.

The delay performance for Exp2 is similar as shown in Fig. 4 (c)-(d). Namely, with probability 1 the packet end-to-end delay is smaller than 10 slots under EC scheduler with $x_1(t) = 4$ packets/slot, 14 slots under EC scheduler with $x_1(t) = 6$ packets/slot, 4 slots under BP scheduler with $x_1(t) = 4$ packets/slot, and finally 5 slots under BP scheduler with $x_1(t) = 7$ packets/slot. A short conclusion is that for either of the proposed two schedulers, by increasing the source admission rate $x_1(t)$, packet delay CDF curve translates more to the right side, namely, the packets experience longer delays.

The queuing evolution regarding the instantaneous (i.e., not time-averaged) sum backlog $\sum_{i=1}^{N-1} U_i(t)$ w.r.t. Exp1 is show in Fig. 5 (a) and (b) for the static and the time-varying scenarios, respectively. Here we assume that in the static scenario, the link capacities are constant with no accidental blockages, however, in the time-varying

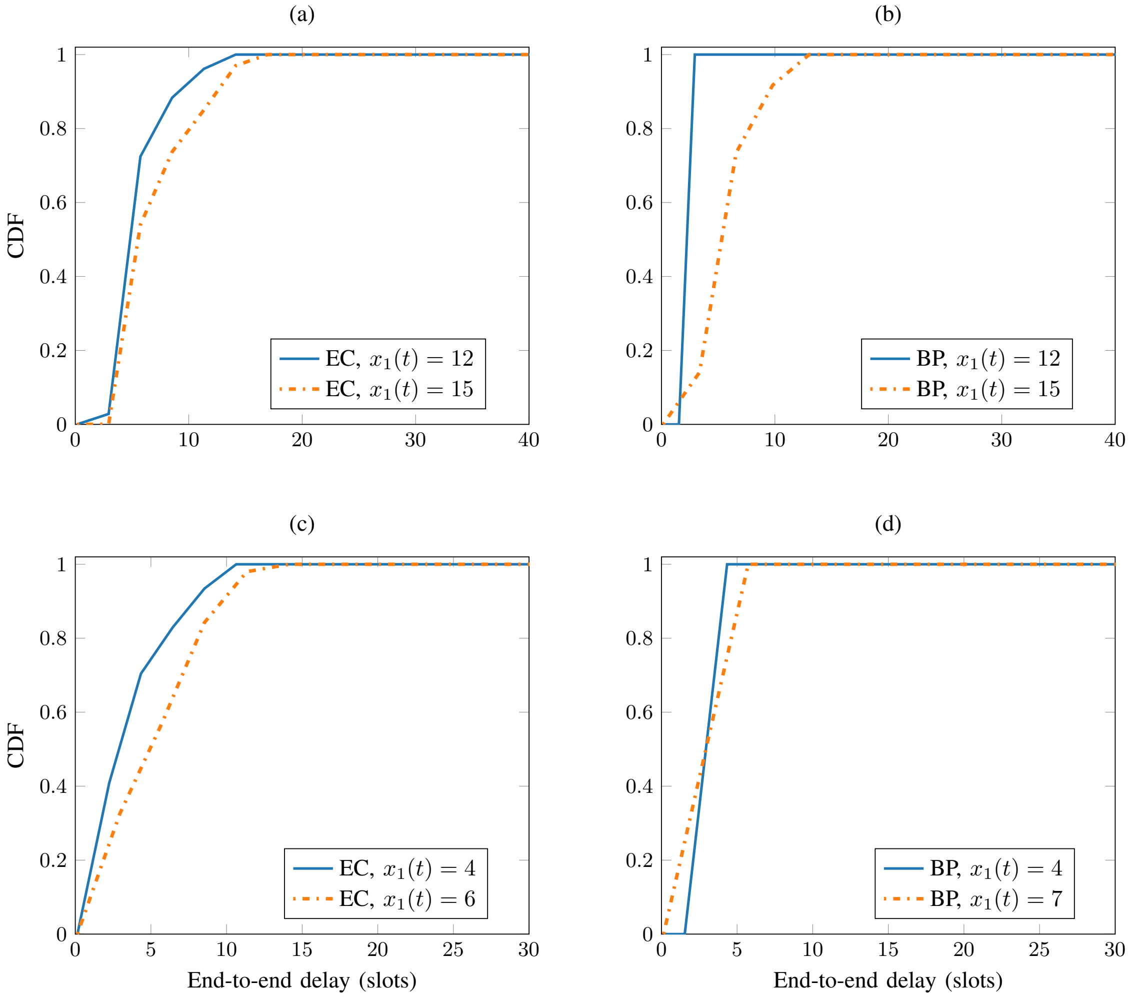


Fig. 4: The packet end-to-end delay distribution under the proposed EC and BP schedulers. (a) The delay distribution in Exp1 under the EC scheduler, with admission rate $x_1(t) = 12$ and $x_1(t) = 15$, respectively. (b) The delay distribution in Exp1 under the BP scheduler, with admission rate $x_1(t) = 12$ and $x_1(t) = 15$, respectively. (c) The delay distribution in Exp2 under the EC scheduler, with admission rate $x_1(t) = 4$ and $x_1(t) = 6$, respectively. (d) The delay distribution in Exp2 under the BP scheduler, with admission rate $x_1(t) = 4$ and $x_1(t) = 7$, respectively. For all the cases, the maximum delay will increase by increasing the source admission rate $x_1(t)$.

scenario, link $(7, 6)$ will be blocked every $T_0 = 200$ slots and each time the blocking will last for 80 slots. We assume that the network state, the computation of the network *approximate capacity* and the overall scheduling decisions will be updated every $T_{\text{EC}} = 50$ slots for the EC scheduler and every $T_{\text{BP}} = 1$ slot for the BP scheduler, respectively. As we can see from Fig. 5 (a), in the static scenario, the sum backlog and its fluctuations under the EC scheduler are slightly larger than that under the BP scheduler. In the time-varying scenario, however, the sum backlog and its fluctuation under the EC scheduler are much larger than that under the BP scheduler as illustrated in Fig. 5 (b). We can observe a similar performance in the Exp2 network as illustrated in Fig. 5 (c) and (d) for the static and the time-varying scenarios, respectively. Here we assume that in the static scenario, the link capacities are constant with no accidental blockages, while in the time-varying scenario, link $(1, 0)$ will be blocked every $T_0 = 200$ slots and each time the blocking will last for 40 slots. The scheduler updates are the same as in Exp1. As we can see, the performance difference between the proposed two schedulers in the static scenario is very moderate. However, in the time-varying scenario, the BP scheduler again outperforms the EC scheduler in terms the amount of queuing backlog and its fluctuations. Therefore, we claim that the EC scheduler is more suitable for static scenarios with mulch less computation and slightly larger sum backlog than that under the BP scheduler. In contrast, the BP scheduler will be updated in every time slot and react very fast to blockages, thus is more favorable for time-varying scenarios.

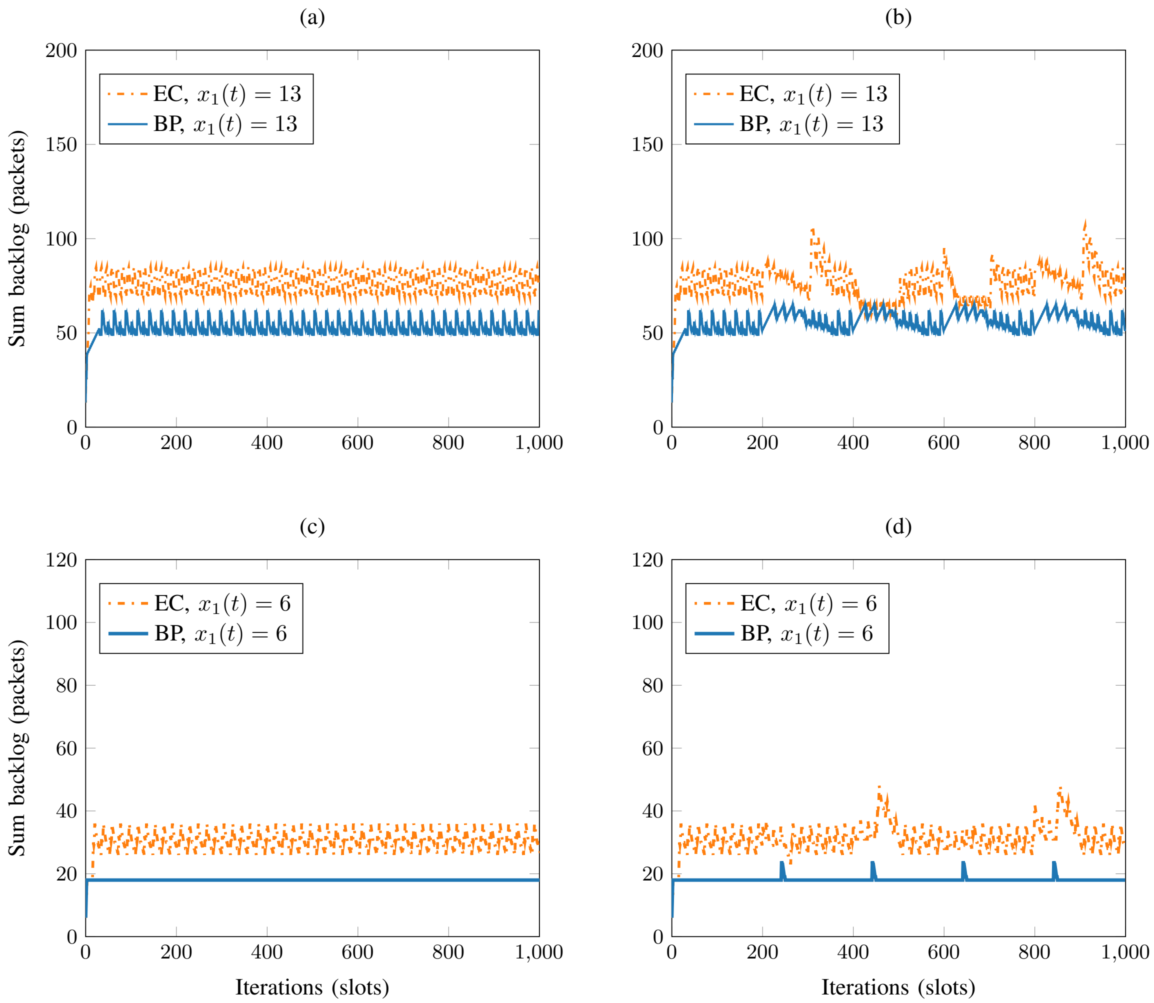
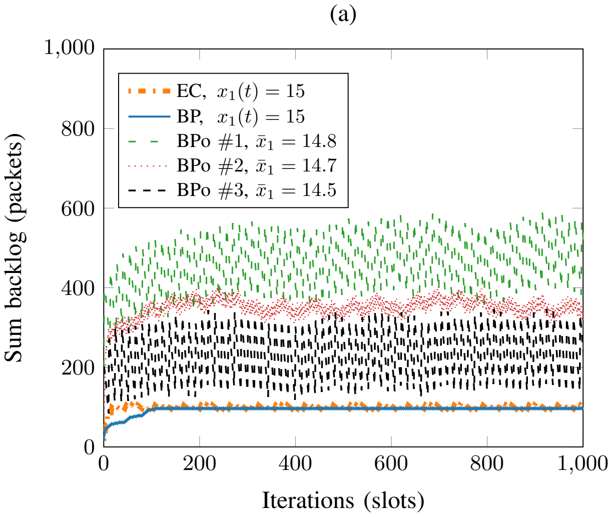


Fig. 5: The instantaneous sum backlog $\sum_{i=1}^{N-1} U_i(t)$ w.r.t. increasing iterations (slots): (a) Exp1 in static scenario. (b) Exp1 in time-varying scenario with accidental blockages. (c) Exp2 in static scenario. (d) Exp2 in time-varying scenario with accidental blockages.

B. The performance comparison with the baseline scheme

Fig. 6 (a) compares the instantaneous sum backlog $\sum_{i=1}^{N-1} U_i(t)$ between the proposed schemes (EC, BP) and the baseline scheme (BPo) w.r.t. the running example Exp1. For the baseline scheme BPo, we choose the sum-rate utility as follows: since there is only one commodity, then in the NUM framework in Algorithm 1, we have $g_1(x_1(t)) = x_1(t)$. Aiming at on one hand to approach the network capacity (w.r.t. large value of $x_{\max}$), and on the other hand to handle the utility-delay tradeoff (w.r.t. $(O(V), O(1/V))$), we choose three sets of parameter for the baseline BPo scheme with $(V, x_{\max}) = (200, 200)$, $(V, x_{\max}) = (200, 50)$ and $(V, x_{\max}) = (50, 200)$, respectively. For our proposed schemes EC and BP, since we have managed to compute the network *approximate capacity* $C_{\text{cs,iid}}$ as shown in (4), the congestion control reduces to a simple constant threshold given by (14). Hence we do not need to suffer from a complex multi-parameter $(V, x_{\max})$ tuning procedure. We pick the point with the maximum achievable data rate (source admission rate) for the proposed schemes, i.e., $x_1(t) = 15$ packets/slot for both of the EC and BP schedulers. As we can see from Fig. 6 (a), all the underlying schemes can stabilize the network since they all converge to finite backlogs. The baseline scheme BPo can approximately approach $C_{\text{cs,iid}}$ in a long-term average sense indicated by $\bar{x}_1$. It's worth noting that the fluctuation ranges of instantaneous sum backlog converge to $[96, 105]$ packets under the EC scheduler and $[93, 98]$ packets under the BP scheduler, respectively. However, this fluctuations increase to the ranges of $[326, 529]$, $[306, 404]$, and $[118, 346]$ packets under the baseline BPo scheme with $(V, x_{\max}) = (200, 200)$, $(V, x_{\max}) = (200, 50)$, and $(V, x_{\max}) = (50, 200)$, respectively. Hence, the proposed schemes achieve much smaller backlog and much smaller backlog fluctuations compared with the baseline scheme.

As for the packet end-to-end delay, Fig. 6 (b) illustrates



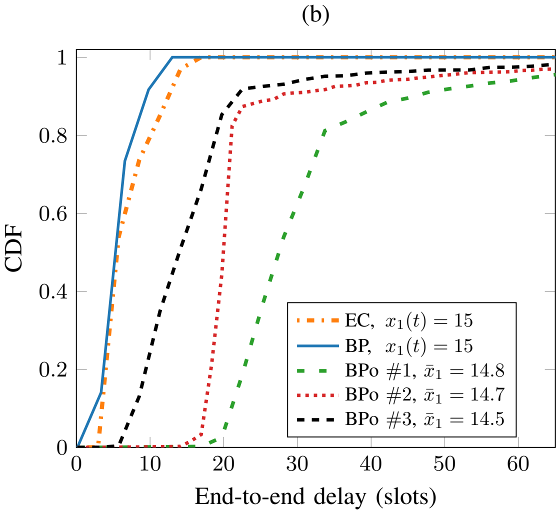


Fig. 6: The performance comparison between the proposed schemes (EC, BP) and the baseline scheme (BPo) w.r.t. the first running example Expl. (a) The instantaneous sum backlog $\sum_{i=1}^{N-1} U_i(t)$ w.r.t. increasing iterations (slots). (b) The packet end-to-end delay distribution. The multi-parameter sets in the BPo scheme are $(V, x_{\max}) = (200, 200)$, $(V, x_{\max}) = (200, 50)$ and $(V, x_{\max}) = (50, 200)$ for #1, #2 and #3, respectively.

the delay distributions w.r.t. different schemes in the running example Exp1. As we can see, when the source admission rate is set as $x_1(t) = 15$ packets/slot, with probability 1 the end-to-end delay of each individual data packet is smaller than 15 slots under the EC scheduler and smaller than 13 slots under the BP scheduler, respectively. However, all the curves w.r.t. the BPo scheme significantly shift to the right side. Namely, the maximum packet end-to-end delays under the BPo scheme with different parameter sets are much larger than that under the proposed schemes (> 68 slots).

As illustrated in Fig. 7, the numerical results in the running example Exp2 achieve similar performance as that in Exp1. Again we choose three sets of parameter for the baseline scheme BPo with $(V, x_{\max}) = (40, 50)$, $(V, x_{\max}) = (40, 20)$ and $(V, x_{\max}) = (10, 50)$, respectively. For the proposed schemes, we pick the point with the maximum achievable data rate (source admission rate) , i.e., $x_1(t) = 6$ packets/slot for the EC scheduler and $x_1(t) = 7$ packets/slot for the BP scheduler, respectively. As we can see from Fig. 7 (a), the fluctuation ranges of instantaneous sum backlog converge to $[26, 36]$ packets under the EC scheduler, $[21, 21]$ packets under the BP scheduler, $[224, 326]$ packets under the BPo with $(V, x_{\max}) = (40, 50)$, $[226, 279]$ packets under the BPo with $(V, x_{\max}) = (40, 20)$, and $[50, 149]$ packets under the BPo with $(V, x_{\max}) = (10, 50)$. Hence, the proposed schemes achieve much smaller backlog and much smaller backlog fluctuations compared with the baseline scheme. The packet end-to-end delay distribution is illustrated in Fig. 7 (b). As we can see, the maximum end-to-end delay under the proposed schemes are 14 slots (EC, $x_1(t) = 6$) and 5 slots (BP, $x_1(t) = 7$), respectively. However, the packets under the baseline BPo scheme with different parameter sets experience much longer delays ($\gg 35$ slots).

V. Conclusion

In this paper, we studied the beam scheduling problem for HD mmWave relay networks with arbitrary topology. Our study focused on developing practically relevant scheduling algorithms guided by theoretical results on the *approximate capacity* $C_{\text{cs,iid}}$ and optimal scheduling in mmWave network models [19]. Based on the theoretically optimal schedule results, we first implemented a network simplification procedure to reduce the network topology complexity. Accordingly, using this simplified topology, we proposed two practical and very simple beam scheduling schemes; the deterministic edge coloring (EC) scheduler and the adaptive backpressure (BP) scheduler. The former is a very simple one-time computation followed by a periodic repetitive schedule, hence is more suitable for quasi-static scenarios. The later is an "online" approach which will update in every time slot, thus is more favorable for time-varying scenarios. We have shown through simulation that both the proposed schedulers can guarantee the network stability within a certain operating range of the input rate. In particular, the EC scheduler guarantees stability for input rates less than $\frac{\Delta}{K} C_{\text{cs,iid}}$, where Δ and K denote the maximum degree and the number of colors used in EC for an associate multigraph, respectively; The BP scheduler guarantees stability for rates less than $C_{\text{cs,iid}}$. Moreover, in comparison with a standard baseline scheme, which consists of applying classical BP-based NUM over the whole network (without network simplification), the proposed schedulers do not require the

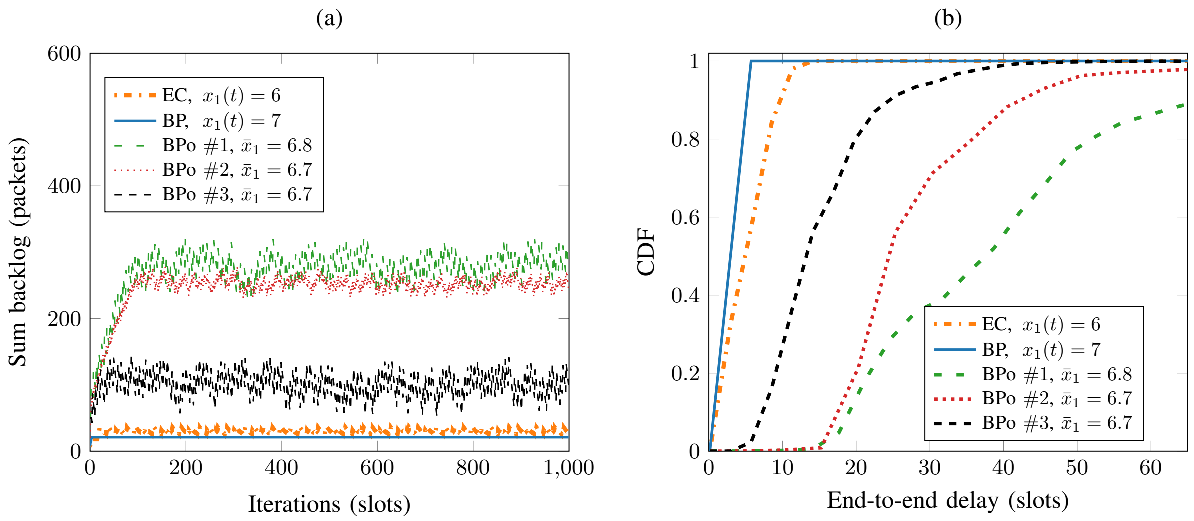


Fig. 7: The performance comparison between the proposed schemes (EC, BP) and the baseline scheme (BPo) w.r.t. the second running example Exp2. (a) The instantaneous sum backlog $\sum_{i=1}^{N-1} U_i(t)$ w.r.t. increasing iterations (slots). (b) The packet end-to-end delay distribution. The multi-parameter sets in the BPo scheme are $(V, x_{\max}) = (40, 50)$, $(V, x_{\max}) = (40, 20)$ and $(V, x_{\max}) = (10, 50)$ for #1, #2 and #3, respectively.

empirical tuning of the BP control parameters and achieve a much smaller queuing backlogs and packet end-to-end delays.

References

[1] Y. Niu, W. Ding, H. Wu, Y. Li, X. Chen, B. Ai, and Z. Zhong, "Relay-Assisted and QoS Aware Scheduling to Overcome Blockage in mmWave Backhaul Networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1733–1744, 2019.

[2] A. Dimas, D. S. Kalogerias, and A. P. Petropulu, "Cooperative beamforming with predictive relay selection for urban mmWave communications," *IEEE Access*, vol. 7, pp. 157 057–157 071, 2019.

[3] Y. Yan, Q. Hu, and D. M. Blough, "Path Selection with Amplify and Forward Relays in mmWave Backhaul Networks," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2018, pp. 1–6.

[4] M. Shafi, J. Zhang, H. Tataria, A. F. Molisch, S. Sun, T. S. Rappaport, F. Tufvesson, S. Wu, and K. Kitao, "Microwave vs. Millimeter-Wave Propagation Channels: Key Differences and Impact on 5G Cellular Systems," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 14–20, 2018.

[5] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 134–141, 2017.

[6] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[7] X. Song, S. Haghighatshoar, and G. Caire, "A scalable and statistically robust beam alignment technique for mm-Wave systems," *IEEE Trans. on Wireless Comm.*, vol. PP, pp. 1–1, 2018.

[8] X. Song, S. Haghighatshoar, and G. Caire, "Efficient Beam Alignment for Millimeter Wave Single-Carrier Systems With Hybrid MIMO Transceivers," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1518–1533, 2019.

[9] X. Song, T. Kühne, and G. Caire, "Fully-/Partially-Connected Hybrid Beamforming Architectures for mmWave MU-MIMO," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1754–1769, 2020.

[10] Y. Xu, H. Shokri-Ghadikolaei, and C. Fischione, "Distributed Association and Relaying With Fairness in Millimeter Wave Networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 7955–7970, 2016.

[11] Y. Xu, G. Athanasiou, C. Fischione, and L. Tassiulas, "Distributed Association Control and Relaying in Millimeter Wave Wireless Networks," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.

[12] T. K. Vu, M. Bennis, M. Debbah, and M. Latva-Aho, "Joint Path Selection and Rate Allocation Framework for 5G Self-Backhauled mm-wave Networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2431–2445, 2019.

[13] J. Chang and Y. Chen, "A cluster-based relay station deployment scheme for multi-hop relay networks," *Journal of Communications and Networks*, vol. 17, no. 1, pp. 84–92, 2015.

[14] Y. Wei, Y. Hou, L. Li, and M. Song, "Energy efficient topology control for multi-hop relay cellular networks based on flow management," *Journal of Communications and Networks*, vol. 19, no. 6, pp. 618–626, 2017.

[15] X. Song, R. Zhang, J. Pan, and J. Liu, "A statistical geometric approach for capacity analysis in two-hop relay communications," in *2013 IEEE Global Communications Conference (GLOBECOM)*, Conference Proceedings, pp. 4823–4829.

[16] T. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, 1979.

[17] W. Yi, Y. Liu, Y. Deng, A. Nallanathan, and R. W. Heath, "Modeling and analysis of mmwave v2x networks with vehicular platoon systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 12, pp. 2851–2866, 2019.

[18] S. Lien, Y. Kuo, D. Deng, H. Tsai, A. Vinel, and A. Benslimane, "Latency-optimal mmwave radio access for v2x supporting next generation driving use cases," *IEEE Access*, vol. 7, pp. 6782–6795, 2019.

[19] Y. H. Ezzeldin, M. Cardone, C. Fragouli, and G. Caire, "Polynomial-time Capacity Calculation and Scheduling for Half-Duplex 1-2-1 Networks," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 460–464.

[20] R. Abdel-Raouf, H. Esmaiel, and O. A. Omer, "Fuzzy logic based relay selection for mmWave communications," in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, March 2019, pp. 263–267.

[21] J. García-Rois, F. Gómez-Cuba, M. R. Akdeniz, F. J. González-Castao, J. C. Burguillo, S. Rangan, and B. Lorenzo, "On the analysis of scheduling in dynamic duplex multihop mmwave

cellular systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6028–6042, 2015.

[22] B. P. S. Sahoo, C. Yao, and H. Wei, "Millimeter-Wave Multi-Hop Wireless Backhauling for 5G Cellular Networks," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, 2017, pp. 1–5.

[23] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[24] S. Wang and N. Shroff, "Towards fast-convergence, low-delay and low-complexity network optimization," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, p. 34, 2017.

[25] H. Yu and M. J. Neely, "A new backpressure algorithm for joint rate control and routing with vanishing utility optimality gaps and finite queue lengths," *IEEE/ACM Transactions on Networking*, vol. 26, no. 4, pp. 1605–1618, 2018.

[26] X. Song and G. Caire, "Queue-Aware Beam Scheduling for Half-Duplex mmWave Relay Networks," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1611–1616.

[27] Y. H. Ezzeldin, M. Cardone, C. Fragouli, and G. Caire, "Gaussian 1-2-1 networks: Capacity results for mmWave communications," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 2569–2573.

[28] R. E. Gomory and T. C. Hu, "Multi-Terminal Network Flows," *Journal of the Society for Industrial and Applied Mathematics*, vol. 9, no. 4, pp. 551–570, 1961.

[29] J. D. Little, "A proof for the queuing formula: $L = W$," *Operations research*, vol. 9, no. 3, pp. 383–387, 1961.

[30] D. Park, "A throughput-optimal scheduling policy for wireless relay networks," in *2010 IEEE Wireless Communication and Networking Conference*, April 2010, pp. 1–5.

[31] J. Wang, L. He, and J. Song, "Stochastic Optimization Based Dynamic User Scheduling and Hybrid Precoding for Broadband MmWave MIMO," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[32] H. N. Gabow, "Using Euler Partitions to Edge Color Bipartite Multigraphs," *International Journal of Computer & Information Sciences*, vol. 5, no. 4, pp. 345–355, 1976.

[33] C. Sinnamon, "Fast and Simple Edge-Coloring Algorithms," *preprint arXiv:1907.03201*, 2019.

[34] X. Song and G. Caire, "Queue-Aware Beam Scheduling for Half-Duplex mmWave Relay Networks," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1611–1616.

[35] V. G. Vizing, "On an estimate of the chromatic class of a p-graph," *Discret Analiz*, vol. 3, pp. 25–30, 1964.

[36] J. Edmonds, "Maximum matching and a polyhedron with 0, 1-vertices," *Journal of research of the National Bureau of Standards B*, vol. 69, no. 125-130, pp. 55–56, 1965.

7 Conclusions

7.1 Summary of this thesis

In the past decades, tremendous fundings and research efforts have been dedicated to the investigation of millimeter wave (mmWave) wireless communication, since the use of mmWaves will solve the spectrum shortage in current sub-6 GHz cellular communication systems and offer unprecedented multi-Gbps date rates for each mobile devices in the next generation (5G) mobile communication systems. This thesis has proposed several enabling schemes to address the challenges in mmWave communication including the initial access, the data communication and the relay networking.

For the initial access, we proposed two efficient beam alignment (BA) schemes for mmWave OFDM (orthogonal frequency division multiplexing) system and mmWave SC (single-carrier) system, respectively. The proposed schemes are based on quadratic channel measurements and the non-negative Least Squares (NNLS) technique in compressed sensing (CS). These schemes can operate in much more realistic conditions than existing schemes in the literature, are strongly scalable for multi-user scenarios and are very robust to fast channel variations cased by Doppler spread.

For the data communication after BA is achieved, we defined two “extreme” hybrid digital analog (HDA) antenna architectures, i.e., the fully-connected (FC) architecture and the one-stream-per-subarray (OSPS) architecture. We provided a joint performance evaluation of the initial access and data communication phases with more realistic channel and hardware conditions. In each phase, we proposed our own BA and precoding schemes that outperform the counterparts in the literature. We have observed that the proposed two architectures achieve similar sum spectral efficiency, but the OSPS architecture outperforms the FC case in terms of hardware complexity and power efficiency, only at the cost of a slightly longer time of initial beam acquisition.

On top of the above beamforming work, we further extended our work into mmWave relay networking. For a general half-duplex (HD) mmWave relay network with arbitrary relay connections, we proposed two beam scheduling schemes to approach the approximate information theoretical Shannon capacity, namely, the deterministic edge coloring (EC) scheduler and the adaptive backpressure (BP) scheduler. The EC scheduler is more suitable for static scenarios since it is one-time computation and then periodically state repetition. In contrast, the BP scheduler is more favorable for time-varying scenarios because it updates in every time slots. Both the proposed schedulers can effectively stabilize the network, meanwhile achieve much smaller queuing backlogs, much smaller backlog fluctuations, and much lower packet end-to-end delays in comparison with the reference baseline scheme.

7.2 Future directions

One interesting direction to go on top of this thesis is mmWave vehicle-to-vehicle (V2V) and vehicle-to-everything (V2X) communication. mmWave V2V and V2X communication can provide NLOS information about the surrounding environment, thus improve the safety and traffic efficiency of cooperative automated driving. In the V2V and V2X scenarios, the nodes will create a relay network and route data packets through multi-hop transmission. An essential component in these networks consists of selecting, at each time slot, which beams are active, and in which direction they should be pointed. Therefore, the problem of beamforming (proposed in Chapter 4-6) is intrinsically connected with the problem of beam scheduling (proposed in Chapter 7), since transmission is not isotropic as in conventional wireless networks, rather highly directional.

Also, the development of mmWave massive MIMO communication technology is now in the hands of the product departments of companies such as Huawei, qualcomm, Ericsson, Nokia, etc.. A large number of communication, signal processing, and optimization algorithms have been development over the years and it remains to be seen which ones will work well in practice. If 5G becomes a commercial success, massive digitally controllable antenna arrays will be deployed “everywhere” for countless applications at mmWave frequencies and even much higher THz frequencies (6G). Thus, we can expect a future where extremely large aperture array with thousands of antenna elements are used to serve a set of users. There are, however, practical limits to how many antennas can be deployed at conventional towers and rooftop locations.

In addition, there is a recent surge of papers applying machine learning (ML) to various problems in communications. ML is especially powerful when a system has characteristics that are hard to model or analyze by conventional approaches. Thus it would be an exciting possibility to use ML in the future mmWave wireless systems whenever a good model is lacking, or a model is available but it is intractable for analysis.

However, before ML can be successfully used in communication systems, many obstacles, like the acquisition of training data, the hard real-time constraints and so on still need to be overcome.

A

Acronyms and Abbreviations

AoA angle of arrival 7

AoD angle of departure 7

AR augmented reality 2

AWGN additive white Gaussian noise 16

BA beam alignment 7

BBF before beamforming 17

BP backpressure 9

BS base station 8

CSI full channel state information 7

D-RoF digital radio-over-fiber 12

DFT discrete Fourier transform 15

EB exabytes 1

EC edge coloring 9

eMBB enhanced mobile broadband 2

FC fully-connected 7

Gbps gigabit per second 2

HD half-duplex 9

HD high definition 1

HDA hybrid digital analog 9

IC integrated circuit 6

IMT international mobile telecommunications 1

IP Internet protocol 11

IQ in-phase and quadrature 6

LAN local area network 6

MAC media access control 6

MIMO multiple input multiple output 2

mMTC massive machine type communications 2

mmWave millimeter wave 2

OFDM orthogonal frequency division multiplexing 6

OSPS one-stream-per-subarray 9

PA power amplifier 6

PSD power spectral density 16

RF radio frequency 5

SC single carrier 6

SNR signal to noise ratio 6

UE user equipment 9

ULA uniform linear array 13

uRLLC ultra-reliable and low-latency communications 2

V2V vehicle to vehicle 96

V2X vehicle to everything 96

VR virtual reality 2

WPAN wireless personal area network 6

Bibliography

[1] Ericsson Inc. *Ericsson Mobility Report*. Mar. 2018. URL: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-november-2018.pdf.

[2] Erik Dahlman, Stefan Parkvall, and Johan Skold. *5G NR: The next generation wireless access technology*. Academic Press, 2018. ISBN: 012814324X.

[3] Xingqin Lin et al. "5G New Radio: Unveiling the essentials of the next generation wireless access technology". In: *IEEE Communications Standards Magazine* 3.3 (2019), pp. 30–37. ISSN: 2471-2825.

[4] Khagendra Belbase. "Analysis of Millimeter Wave Wireless Relay Networks". PhD thesis. University of Alberta, 2019.

[5] Theodore S Rappaport et al. "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models". In: *IEEE Transactions on Antennas and Propagation* 65.12 (2017), pp. 6213–6230. ISSN: 0018-926X.

[6] M. Shafi et al. "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice". In: *IEEE Journal on Selected Areas in Communications* 35.6 (2017), pp. 1201–1221. ISSN: 0733-8716. DOI: 10.1109/JSAC.2017.2692307.

[7] M Series. "IMT Vision–Framework and overall objectives of the future development of IMT for 2020 and beyond". In: *Recommendation ITU* (2015), pp. 2083–.

[8] Huawei Technologies Co. "White Paper: 5G Network Architecture - A High-Level Perspective". In: (2016).

[9] Naga Bhushan et al. "Network densification: the dominant theme for wireless evolution into 5G". In: *IEEE Communications Magazine* 52.2 (2014), pp. 82–89. ISSN: 0163-6804.

[10] A. Gupta and R. K. Jha. "A Survey of 5G Network: Architecture and Emerging Technologies". In: *IEEE Access* 3 (2015), pp. 1206–1232. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2015.2461602.

[11] M. Agiwal, A. Roy, and N. Saxena. "Next Generation 5G Wireless Networks: A Comprehensive Survey". In: *IEEE Communications Surveys Tutorials* 18.3 (thirdquarter 2016), pp. 1617–1655. ISSN: 1553-877X. DOI: 10.1109/COMST.2016.2532458.

[12] Erik G Larsson et al. "Massive MIMO for next generation wireless systems". In: *IEEE communications magazine* 52.2 (2014), pp. 186–195. ISSN: 0163-6804.

[13] Emil Björnson, Jakob Hoydis, and Luca Sanguinetti. *Massive MIMO networks: Spectral, energy, and hardware efficiency.* Vol. 11. 3-4. 2017, pp. 154–655.

[14] M. R. Akdeniz et al. "Millimeter wave channel modeling and cellular capacity evaluation". In: *IEEE Journal on Selected Areas in Communications* 32.6 (June 2014), pp. 1164–1179. ISSN: 0733-8716. DOI: 10.1109/JSAC.2014.2328154.

[15] Robert W Heath et al. "An overview of signal processing techniques for millimeter wave MIMO systems". In: *IEEE journal of selected topics in signal processing* 10.3 (2016), pp. 436–453. ISSN: 1932-4553.

[16] M. Shafi et al. "Microwave vs. Millimeter-Wave Propagation Channels: Key Differences and Impact on 5G Cellular Systems". In: *IEEE Communications Magazine* 56.12 (2018), pp. 14–20.

[17] Yong Niu et al. "Relay-Assisted and QoS Aware Scheduling to Overcome Blockage in mmWave Backhaul Networks". In: *IEEE Transactions on Vehicular Technology* 68.2 (2019), pp. 1733–1744. ISSN: 0018-9545.

[18] A. Dimas, D. S. Kalogerias, and A. P. Petropulu. "Cooperative beamforming with predictive relay selection for urban mmWave communications". In: *IEEE Access* 7 (2019), pp. 157057–157071. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2950274.

[19] Y. Yan, Q. Hu, and D. M. Blough. "Path Selection with Amplify and Forward Relays in mmWave Backhaul Networks". In: *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. Sept. 2018, pp. 1–6. DOI: 10.1109/PIMRC.2018.8580768.

[20] Yilin Li et al. "Radio resource management considerations for 5G millimeter wave backhaul and access networks". In: *IEEE Communications Magazine* 55.6 (2017), pp. 86–92. ISSN: 0163-6804.

[21] Yilin Li. "Efficient Data Delivery in 5G Mobile Communication Networks". PhD thesis. Technischen Universität Berlin, 2019.

[22] TE Bogale, X Wang, and LB Le. "mmWave communication enabling techniques for 5G wireless systems: A link level perspective". In: *mmWave Massive MIMO*. Elsevier, 2017, pp. 195–225.

[23] Farooq Khan and Zhouyue Pi. “mmWave mobile broadband (MMB): Unleashing the 3–300GHz spectrum”. In: *34th IEEE Sarnoff Symposium*. IEEE, pp. 1–6. ISBN: 1612846807.

[24] M. Jacob et al. “Diffraction in mm and Sub-mm Wave Indoor Propagation Channels”. In: *IEEE Transactions on Microwave Theory and Techniques* 60.3 (2012), pp. 833–844.

[25] Z. Shi et al. “Three-dimensional spatial multiplexing for directional millimeter-wave communications in multi-cubicle office environments”. In: *2013 IEEE Global Communications Conference (GLOBECOM)*. 2013, pp. 4384–4389.

[26] T. S. Rappaport, J. N. Murdock, and F. Gutierrez. “State of the Art in 60-GHz Integrated Circuits and Systems for Wireless Communications”. In: *Proceedings of the IEEE* 99.8 (2011), pp. 1390–1436.

[27] Z. Qingling and J. Li. “Rain Attenuation in Millimeter Wave Ranges”. In: *2006 7th International Symposium on Antennas, Propagation EM Theory*. 2006, pp. 1–4.

[28] S. Joshi and S. Sancheti. “Foliage loss measurements of tropical trees at 35 GHz”. In: *2008 International Conference on Recent Advances in Microwave Theory and Applications*. 2008, pp. 531–532.

[29] Ahmed M Al-samman, Marwan Hadri Azmi, and Tharek Abd Rahman. “A survey of millimeter wave (mm-Wave) communications for 5G: Channel measurement below and above 6 GHz”. In: *International Conference of Reliable Information and Communication Technology*. Springer, pp. 451–463.

[30] T. Nitsche, C. Cordeiro, A. B. Flores, E. W. Knightly, E. Perahia, and J. C. Widmer. “IEEE 802.11 ad: directional 60 GHz communication for multi-Gigabit-per-second Wi-Fi”. In: *IEEE Communications Magazine* 52.12 (2014), pp. 132–141. ISSN: 0163-6804.

[31] Ahmed Alkhateeb et al. “Channel estimation and hybrid precoding for millimeter wave cellular systems”. In: *Selected Topics in Signal Processing, IEEE Journal of* 8.5 (2014), pp. 831–846.

[32] Matthew Kokshoorn et al. “Millimeter wave MiMo channel estimation using overlapped beam patterns and rate adaptation”. In: *IEEE Transactions on Signal Processing* 65.3 (2016), pp. 601–616. ISSN: 1053-587X.

[33] S. Noh, M. D. Zoltowski, and D. J. Love. “Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment”. In: *IEEE Transactions on Wireless Communications* 16.9 (Sept. 2017), pp. 5689–5701. ISSN: 1536-1276. DOI: 10.1109/TWC.2017.2713357.

[34] M. Hussain and N. Michelusi. “Throughput optimal beam alignment in millimeter wave networks”. In: *2017 Information Theory and Applications Workshop (ITA)*. Feb. 2017, pp. 1–6. DOI: 10.1109/ITA.2017.8023460.

[35] J. Palacios and D. De Donno and J. Widmer. “Tracking mm-Wave channel dynamics: Fast beam training strategies under mobility”. In: *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. May 2017, pp. 1–9. DOI: 10.1109/INFOCOM.2017.8056991.

[36] A. Alkhateeb, G. Leus, and R. W. Heath. “Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?” In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2015, pp. 2909–2913. DOI: 10.1109/ICASSP.2015.7178503.

[37] J. Rodríguez-Fernández, N. González-Prelcic, K. Venugopal, and R. W. Heath Jr. “Frequency-domain compressive channel estimation for frequency-selective hybrid mmWave MIMO systems”. In: *arXiv preprint arXiv:1704.08572* (2017).

[38] Kiran Venugopal et al. “Channel estimation for hybrid srchitecture-based wideband millimeter wave systems”. In: *IEEE Journal on Selected Areas in Communications* 35.9 (2017), pp. 1996–2009. ISSN: 0733-8716.

[39] Kiran Venugopal et al. “Time-domain channel estimation for wideband millimeter wave systems with hybrid architecture”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6493–6497. ISBN: 1509041176.

[40] O. El Ayach, R. W. Heath, S. Rajagopal, and Z. Pi. “Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays”. In: *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, pp. 3476–3480. ISBN: 1479913537.

[41] Didi Zhang et al. “Hybridly connected structure for hybrid beamforming in mmWave massive MIMO systems”. In: *IEEE Transactions on Communications* 66.2 (2018), pp. 662–674. ISSN: 0090-6778.

[42] P. L. Cao, T. J. Oechtering, and M. Skoglund. “Precoding design for massive MIMO systems with sub-connected architecture and per-antenna power constraints”. In: *WSA 2018; 22nd International ITG Workshop on Smart Antennas*. Mar. 2018, pp. 1–6.

[43] M. Majidzadeh et al. “Hybrid beamforming for single-user MIMO with partially connected RF architecture”. In: *2017 European Conference on Networks and Communications (EuCNC)*. June 2017, pp. 1–6. DOI: 10.1109/EuCNC.2017.7980696.

[44] Shahar Stein Ioushua and Yonina C Eldar. “Hybrid analog-digital beamforming for massive MIMO systems”. In: *arXiv preprint arXiv:1712.03485* (2017).

[45] F. Sohrabi and W. Yu. “Hybrid digital and analog beamforming design for large-scale antenna arrays”. In: *IEEE Journal of Selected Topics in Signal Processing* 10.3 (Apr. 2016), pp. 501–513. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2016.2520912.

[46] Ang Li and Christos Masouros. “Hybrid analog-digital millimeter-wave MU-MIMO transmission with virtual path selection”. In: *IEEE Communications Letters* 21.2 (2017), pp. 438–441. ISSN: 1089-7798.

[47] Jingbo Du et al. “Hybrid precoding architecture for massive multiuser MIMO with dissipation: Sub-connected or fully-connected structures?” In: *arXiv preprint arXiv:1806.02857* (2018).

[48] T. K. Vu et al. “Joint Path Selection and Rate Allocation Framework for 5G Self-Backhauled mm-wave Networks”. In: *IEEE Transactions on Wireless Communications* 18.4 (2019), pp. 2431–2445.

[49] R. Abdel-Raouf, H. Esmaiel, and O. A. Omer. “Fuzzy logic based relay selection for mmWave communications”. In: *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*. Mar. 2019, pp. 263–267. DOI: `10.1109/IEMECONX.2019.8877074`.

[50] J. García-Rois et al. “On the Analysis of Scheduling in Dynamic Duplex Multihop mmWave Cellular Systems”. In: *IEEE Transactions on Wireless Communications* 14.11 (2015), pp. 6028–6042.

[51] B. P. S. Sahoo, C. Yao, and H. Wei. “Millimeter-Wave Multi-Hop Wireless Backhauling for 5G Cellular Networks”. In: *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*. 2017, pp. 1–5.

[52] X. Gao, L. Dai, and A. M. Sayeed. “Low RF-complexity technologies to enable millimeter-wave MIMO with large antenna array for 5G wireless communications”. In: *IEEE Communications Magazine* 56.4 (Apr. 2018), pp. 211–217. ISSN: 0163-6804. DOI: `10.1109/MCOM.2018.1600727`.

[53] J. Palacios, N. González-Prelcic, and J. Widmer. “Managing Hardware Impairments in Hybrid Millimeter Wave Mimo Systems: A Dictionary Learning-Based Approach”. In: *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, pp. 168–172. ISBN: 1728143004.

[54] Nima N Moghadam et al. “On the energy efficiency of MIMO hybrid beamforming for millimeter wave systems with nonlinear power amplifiers”. In: *arXiv preprint arXiv:1806.01602* (2018).

[55] Xiaoshen Song, Saeid Haghighatshoar, and Giuseppe Caire. “A scalable and statistically robust beam alignment technique for mm-Wave systems”. In: *IEEE Trans. on Wireless Comm.* PP (2018), pp. 1–1.

[56] X. Song, S. Haghighatshoar, and G. Caire. “Efficient Beam Alignment for Millimeter Wave Single-Carrier Systems With Hybrid MIMO Transceivers”. In: *IEEE Transactions on Wireless Communications* 18.3 (2019), pp. 1518–1533.

[57] X. Song, T. Kühne, and G. Caire. “Fully-/Partially-Connected Hybrid Beamforming Architectures for mmWave MU-MIMO”. In: *IEEE Transactions on Wireless Communications* 19.3 (2020), pp. 1754–1769.

[58] X. Song et al. “Joint Topology Simplification and Beam Scheduling for Half-Duplex mmWave Relay Networks”. In: *IEEE Transactions on Wireless Communications.* (2020 (to be submitted)).

[59] P. Schniter and A. Sayeed. “Channel estimation and precoder design for millimeter-wave communications: The sparse way”. In: *2014 48th Asilomar Conference on Signals, Systems and Computers.* Nov. 2014, pp. 273–277. DOI: `10.1109/ACSSC.2014.7094443`.

[60] John G.. Proakis and Masoud Salehi. *Digital communications.* McGraw-Hill, 2008.

[61] Philip Bello. “Characterization of randomly time-variant linear channels”. In: *IEEE Transactions on Communications Systems* 11.4 (1963), pp. 360–393.

[62] Andrea Goldsmith. *Wireless communications.* Cambridge University Press, 2005.

[63] Akbar M Sayeed. “Deconstructing multiantenna fading channels”. In: *IEEE Transactions on Signal Processing* 50.10 (2002), pp. 2563–2579. ISSN: 1053-587X.