UNIVERSITAT POLITÈCNICA DE CATALUNYA - BARCELONATECH

Multi-User Ultra-Massive MIMO for very high frequency bands (mmWave and THz): a resource allocation problem

A Master Thesis Submitted to the Facultat d'Informàtica de Barcelona carried out in

NORTHEASTERN UNIVERSITY

by

Santiago RODRIGO MUÑOZ

in partial fulfilment of the requirements for the

Master in Innovation and Research in Informatics

Advisor:
Dr. Kaushik CHOWDHURY
Genesys Lab (Electrical & Computer Engineering Dept.)
NORTHEASTERN UNIVERSITY (BOSTON, MA, USA)

Reporting advisor (ponent):
Dr. Albert CABELLOS APARICIO
NaNoNetworking Center (Computer Architecture Dept.)
UNIVERSITAT POLITÈCNICA DE CATALUNYA - BARCELONATECH

April 2, 2018

To my family, the best gift I have received, and friends, the family you choose

Acknowledgements

First of all, I would like to thank Prof. Kaushik Chowdhury and Albert Cabellos, the advisors of this project, for the opportunity of carrying out this work and for guiding and helping me during all this time.

To Kaushik and all the people in his group (Carlos, Guillem (*alias* Gülem), Paul, Kunal, Mauro, Stella, Fan, Gokhan, Shamnaz, Yousof...) and all the others in Ell 301, goes my most sincere gratitude: what a good time we all had while I was working there! I hope that someday we will see each other again. Each one of you would deserve a full paragraph here!

To Josep Miquel Jornet, many thanks for devoting me those hours of your time and for all the help, material and *experiential*, that you gave me.

Finally, I would like to express my gratitude to those professors of the UPC who knew how to transmit the enthusiasm over what they do at university, for that is the transmission of an invaluable treasure.

Abstract

Multi-User Ultra-Massive MIMO for very high frequency bands (mmWave and THz): a resource allocation problem

by Santiago RODRIGO MUÑOZ

During the last years, the increase in data traffic and Quality of Service (QoS) requirements for bandwidth consuming applications in future generations (5G and beyond) wireless scenarios has accelerated the exhaustion of the existing technologies. The lack of room for improvement in the microwave band pushes the research to the higher frequency bands, namely mmWave and THz, which are a promising alternative to alleviate the bandwidth scarcity and the need of higher rates. However, both mmWave and THz communications suffer from a major free-space path loss, due to the massive increase in carrier frequency and the higher effect of the molecular absorption. Therefore, a highly directional link is needed in order to span the link more than only a few meters.

Specially in the case of mmWave band, massive MIMO techniques, enabled by the dramatic decrease in wavelength and production cost and usually paired with the use of hybrid beamforming techniques, have been the main answer of the cutting-edge research in this field to the call for high capacity systems that enable multiple ultra fast communications in mobile scenarios as one of the primary goals for 5G. These systems leverage the previously *no man's land* of high frequency communications and the possibility of working with antenna arrays containing dozens or even hundreds of antenna elements, instead of the traditional MIMO devices with only up to 16 antennas. However, to fully unleash the potential of this technique there is a need for the development of multi-user systems that allocate in real-time the huge amount of resources available in these antenna arrays. Among other techniques, the dynamic allocation of different antenna elements to form a separate beam per user constitutes a very promising option, and is seen as a game changer to tackle the existing challenges and push forward the capacity of this technology.

In this work a dynamic subarray allocation algorithm has been fully designed and implemented, in a first approach that uses a simplified scenario and is part of a bigger picture in which a real-time scheduler would be managing the users' demands and the available resources while using this algorithm to allocate antennas every certain period of time. To do so, a systemic cross-layer approach has been used, leveraging knowledge of the system as a whole. This approach makes our proposal capable of adapting to different environments, being a first step in the path to follow in order to design future scheduling techniques that allocate not only upper layer resources but also physical layer ones, having a granularity and control of the whole system never seen before.

The presented implementation considers an scenario where several users are communicating with a Base Station in the mmWave band, considering only a limited Channel State Information. In order to assess its performance, it has been tested in a great variety of scenarios, rigorously defining the parameters for large testing benchmarks. The results obtained are very promising. Although there is still room for improvement, the behavior of the algorithm in terms of delivered capacity fulfills the theoretical expectations. Despite the fact of requiring extra hardware and computational power, the proposal here presented could be a great alternative for future mobile communications to make a qualitative leap in the resource allocation efficiency and network capacity.

Resumen

Sistemas MIMO masivos para entornos multiusuario en bandas de muy alta frecuencia: un problema de asignación de recursos

por Santiago RODRIGO MUÑOZ

Durante los últimos años, el aumento en el tráfico de datos y los requisitos de calidad de servicio (QoS) de las aplicaciones con alto consumo de ancho de banda y las comunicaciones móviles futuras (5G y posteriores) han acelerado el agotamiento de las tecnologías existentes. La falta de margen de mejora en la banda de microondas empuja la investigación a las bandas de frecuencia más altas, es decir, mmWave y THz, que son una alternativa prometedora para aliviar la escasez de ancho de banda y la necesidad de mayores velocidades. Sin embargo, las comunicaciones tanto en mmWave como en THz sufren grandes pérdidas de señal en el espacio libre, debido al aumento masivo en la frecuencia y al mayor efecto de la absorción molecular. Por lo tanto, se necesita un enlace altamente direccional para extender el enlace más allá de unos pocos metros.

Especialmente en el caso de la banda mmWave, las técnicas MIMO masivas, habilitadas por la disminución dramática en la longitud de onda y el costo de producción, y usualmente combinadas con el uso de técnicas híbridas de formación de haces, han sido la respuesta principal de la investigación de vanguardia en este campo al requerimiento de los sistemas de alta capacidad que permiten múltiples comunicaciones ultrarrápidas en escenarios móviles como uno de los principales objetivos de 5G. Estos sistemas aprovechan lo que hasta hace poco tiempo era *tierra de nadie*, las comunicaciones de muy alta frecuencia, y la posibilidad de trabajar con sistemas de antenas que contienen docenas o incluso cientos de elementos, en lugar de los dispositivos MIMO tradicionales con 16 antenas como máximo. Sin embargo, para liberar completamente el potencial de esta técnica, existe la necesidad del desarrollo de sistemas multiusuario que asignen en tiempo real la gran cantidad de recursos disponibles en estos sistemas de antenas. Entre otras técnicas, la asignación dinámica de diferentes elementos para formar un haz separado para cada usuario constituye una opción muy prometedora, y se considera como un elemento de cambio para abordar los desafíos existentes e impulsar la capacidad de esta tecnología.

En este trabajo, un algoritmo dinámico de asignación de antenas ha sido completamente diseñado e implementado, en un primer enfoque que utiliza un escenario simplificado y como parte de un sistema más grande en el que un planificador gestionaría en tiempo real las demandas de los usuarios y los recursos disponibles para asignar antenas, renovándola cada cierto período de tiempo. Para hacerlo, se ha utilizado el enfoque sistémico de coordinación entre las capas existentes, aprovechando el conocimiento omnisciente del sistema. Este enfoque hace que nuestra propuesta sea capaz de adaptarse a diferentes entornos, siendo un primer paso en el camino a seguir para diseñar futuras técnicas de planificación que asignen no solo recursos de capa superior sino también de capa física, teniendo una granularidad y un control del sistema nunca antes vistos.

La implementación presentada considera un escenario en el que varios usuarios se comunican con una estación base en la banda mmWave, considerando solo una información de estado de canal limitada. Para evaluar su rendimiento, ha sido probado en una gran variedad de escenarios, definiendo rigurosamente los parámetros para grandes pruebas de referencia. Los resultados obtenidos son muy prometedores. Aunque todavía hay margen de mejora, el comportamiento del algoritmo en términos de capacidad entregada cumple las expectativas teóricas. A pesar de requerir hardware adicional y potencia computacional, la propuesta aquí presentada podría ser una gran alternativa para futuras comunicaciones móviles para dar un salto cualitativo en la eficiencia de la asignación de recursos y la capacidad de la red.

Resum

Sistemes MIMO massius per a entorns multiusuari en bandes de molt alta freqüència: un problema d'assignació de recursos

per Santiago RODRIGO MUÑOZ

Durant els últims anys, l'augment del trànsit de dades i els requisits de qualitat de servei (QoS) per a aplicacions amb alt consum d'ample de banda i les comunicacions mòbils futures (5G i més enllà) han accelerat l'esgotament de les tecnologies existents. La manca de marge de millora en la banda de microones empeny a la recerca a bandes de freqüència més elevades, a saber, mmWave i THz, que són una alternativa prometedora per pal·liar l'escassetat d'ample de banda i la necessitat de velocitats més elevades. Tanmateix, tant les comunicacions mmWave com THz pateixen una gran pèrdua de senyal en l'espai lliure, a causa de l'augment massiu de la freqüència del portador i l'efecte més alt de l'absorció molecular. Per tant, cal un enllaç altament direccional per tal de tenir un enllaç més llarg que només uns pocs metres.

Especialment en el cas de la banda mmWave, les tècniques MIMO massives, habilitades per la dramàtica disminució de la longitud d'ona i el cost de producció, i generalment emparejades amb l'ús de tècniques híbrides de formigó, han estat la principal resposta de la recerca d'avantguarda en aquest camp a la convocatòria per a sistemes d'alta capacitat que permetin múltiples comunicacions ultra ràpides en escenaris mòbils com un dels objectius principals de 5G. Aquests sistemes aprofiten les comunicacions d'alta freqüència i la possibilitat de treballar amb matrius d'antenes que contenen desenes o fins i tot centenars d'antenes, en comptes dels dispositius MIMO tradicionals amb 16 antenes com a màxim. Tanmateix, per desencadenar completament el potencial d'aquesta tècnica, cal desenvolupar sistemes multiusuari que assignin en temps real l'enorme quantitat de recursos disponibles en aquestes matrius d'antenes. Entre altres tècniques, l'assignació dinàmica de diferents antenes per formar un feix separat per usuari constitueix una opció molt prometedora i es considera un element diferenciador per afrontar els reptes existents i impulsar la capacitat d'aquesta tecnologia.

En aquest treball, s'ha dissenyat i implementat un algorisme de distribució dinàmic de antenes, en un primer enfocament amb un escenari simplificat i com a part d'una imatge més gran en què un planificador en temps real gestionaria les demandes dels usuaris i els recursos disponibles utilitzant aquest algorisme per assignar antenes cada cert període de temps. Per fer-ho, s'ha utilitzat l'enfocament sistèmic de coordinació entre les capes existents, aprofitant el coneixement del sistema en general. Aquest enfocament fa que la nostra proposta sigui capaç d'adaptar-se a diferents entorns, sent un primer pas en el camí a seguir per dissenyar futures tècniques de programació que assignen no només recursos de capa superior, sinó també de la cape física, amb una granularitat i un control de tot el sistema mai vist abans.

La implementació presentada considera un escenari en què diversos usuaris es comuniquen amb una estació base a la banda mmWave, considerant només una informació limitada de l'estat del canal. Per tal d'avaluar el seu rendiment, s'ha provat en una gran varietat d'escenaris, definint rigorosament els paràmetres per a grans assaigs de referència. Els resultats obtinguts són molt prometedors. Tot i que encara hi ha marge de millora, el comportament de l'algorisme en termes de capacitat entregada compleix les expectatives teòriques. Malgrat el fet de requerir un maquinari addicional i gran potència computacional, la proposta aquí presentada podria ser una gran alternativa per a les futures comunicacions mòbils per fer un salt qualitatiu en l'eficiència de l'assignació de recursos i la capacitat de la xarxa.

Contents

List of Figures

List of Abbreviations

QoS	**Q**uality **o**f **S**ervice
IoT	**I**nternet **o**f **T**hings
EB	**E**xa**B**ytes
UHF	**U**ltra **H**igh **F**requency
SHF	**S**uper **H**igh **F**requency
FCC	**F**ederal **C**ommunications **C**ommission
PAN	**P**ersonal **A**rea **N**etwork
WLAN	**W**ireless **L**ocal **A**rea **N**etwork
ME	**M**obile **E**quipment
AP	**A**ccess **P**oint
BS	**B**ase **S**tation
GA	**G**enetic **A**lgorithm
PSO	**P**article **S**warm **O**ptimization
GPS	**G**eneralized **P**attern **S**earch

Chapter 1

Introduction

1.1 The end of the microwave era

During the last years, the increase in data traffic and Quality of Service (QoS) requirements for bandwidth consuming applications in wireless scenarios has accelerated the exhaustion of the existing technologies. Good examples of these challenging scenarios are ultra-high quality video streaming or data centers back-haul links, as well as the increasing number of mobile devices with wireless capabilities deployed as part of the Internet of Things (IoT).

According to Cisco traffic forecasts, traffic coming from wireless and mobile devices will account for more than 63 % of the total IP traffic by 2021, while in 2016 it was only the 49 %. The total IP traffic will also experiment an increase by more than two times of the current figures: from 96 exabytes (EB) per month to more than 275 EB [1]. Although the different wireless communication protocols have continuously increased their capacity and have effectively "cut the cord" in our daily life, these highly demanding applications are in need of a step forward in the offered capacity.

Nevertheless, no room for further improvement is available in the microwave band, i.e. the frequency band under the 6 GHz that is used today for mainstream wireless protocols. We have almost approached the capacity of current wireless systems [2]. Higher frequency bands such as mmWave (30 - 300 GHz) and Terahertz band (0.3 - 10 THz) are a promising alternative to alleviate the bandwidth scarcity and the need of higher rates: 2 GHz wide channels are commonly used in systems working in the 60 GHz mmWave band [3]; in the existing IEEE 802.11ad standard rates up to 7 Gbps are supported [4], while rates up to 10 Gbps are expected under certain conditions in a close future [5, 6]. On the other hand, the THz-band is expected to provide speeds of tens or even hundreds of Tbps and bandwidths of tens of GHz [7–9].

Each of these frequency bands have quite different characteristics and development status. Due to their expected performance, these technologies are a promising alternative both at macro-scale (ultra-high-speed small cell systems, data centers' back-haul links, secure communications...) and at nano-scale (nano-sensors, intra-body communications, Wireless Network on Chip...), but they will provide solutions using specific approaches and each one is on a different development stage.

1.1.1 The short-term solution: mmWave

The frequency band ranging between 30 and 300 GHz is usually referred to as millimeter Wave (mmWave), because the wavelength in these frequencies is within the mm

scale (1 cm - 1 mm). This band is adjacent to the highly scarce microwave band (approximately the union of Ultra High Frequency –UHF, 300 MHz to 3 GHz– and Super High Frequency –SHF, 3 GHz to 30 GHz–), greatly packed with the radio technologies that have been around for several years, such as TV, cellular telephony, navigation services, the most common wireless networks such as WiFi, Bluetooth and others... In the context of wireless communications, the term mmWave corresponds to the spectrum bands centered in the 38 GHz, 60 GHz, 94 GHz and the E-Band (70-90 GHz).

Millimeter wave communications were already part of the first experiments with electromagnetic waves in the late 19th century [10]. They were also widely used for the first commercial standardized consumer radios, which worked in the 60 GHz unlicensed band. The main applications during the 1960s and 1970s were in Radio Astronomy and military. The first consumer oriented use of millimeter wave was its application on collision avoidance radars in cars, working at 77 GHz. The Federal Communications Commission (FCC) declared in 1995 the frequency band ranging between 59 and 64 GHz as open for unlicensed wireless communications. Almost ten years later, the 71-76 GHz and the 81-86 GHz bands were opened for licensed point-to-point communications. This available bandwidth has enabled the industry and the academia to develop technologies that take advantage of this frequency.

Among all the frequency bands included in mmWave, the E-bands are more suited for efficient highly-performing point-to-point communications, containing more spectral bandwidth than all the one contained in the entire microwave band. This huge amount of available bandwidth allows rates of several Gbps without the need for very complicated modulation schemes.

However, the propagation characteristics of millimeter waves are highly affected by two main components: the path loss and the absorption loss. The path loss is inversely proportional to the wavelength of the signal traveling in a certain medium. Therefore, millimeter waves experience a much higher degradation of the signal when compared to traditional microwaves, significantly shortening the length of the practical communication link. On the other hand, the absorption loss, which is almost inexistent in lower frequencies, is mainly caused by the presence of water vapor (air humidity, fog, clouds...) and specially rain [11].

The research in millimeter wave on consumer-oriented applications has been developed already for more than twenty years. The first fruits of this work are now available for practical use. Among these technologies already accessible the most prominent are Wireless HD, a personal area network (PAN) technology used mainly to transport uncompressed high definition video, and the Wireless Local Area Network (WLAN) standard 802.11ad. Other communication standards intended to make a more efficient use of the mmWave band are being developed, such as the 802.11ay. These and other applications, such as metro network services, micro-cells in future (5G and beyond) cellular networks and data center backhauling take advantage of the characteristics found in mmWave systems to provide unforeseen performance.

1.1.2 Looking to the future: the Terahertz band

The THz band is the spectral band that spans the frequencies between 0.3 and 10 THz, although some studies consider that it should include also the frequencies between 0.1 and 0.3 THz. The millimeter wave is therefore immediately inferior to this band, and the far infrared is contiguous to the upper limit. Although both extremes have been

subject of extensive reasearch, the THz band remained as a no-man's-land until only some years ago.

Despite the fact that millimeter waves are still an ongoing research topic, and there is still a lot of room for improvement in its applications to wireless communications, the dramatic increase of the wireless data traffic leads us to think that sometime soon the mmWave will end up being insufficient. If we are looking not for Gbps speeds, but for Tbps speeds, we have to move up in the frequency spectrum to the region over 0.1 THz, but not too far, as when we reach 10 THz several practical issues heavily limit the speed. Moreover, the THz band can provide wireless systems with massive amounts of bandwidth (from tens of GHz to several THz, depending on the transmission distance).

The research on the application of THz band to wireless communications faces several challenges: the technology required exceeds in many aspects the possibilities of the techniques used for lower frequency communications, although in the last few years the needed innovations are becoming a reality. As we mentioned in the case of millimeter waves, the higher the frequency, the worse the path loss. Moreover, this extremely large path loss is combined with the absorption loss, which in these frequency range becomes a major constraint. To this we have to add their selectivity in frequency, which makes even more challenging the design of transmission schemes, specific modulations and communication protocols [9].

Due to their early development stage, the THz band has not yet been regulated: the FCC has not allocated any frequency over 275 GHz [12]. However, some groups are already actively defining the future possibilities of this frequency band, such as the IEEE 802.15 WPAN Study Group 100 Gbps Wireless (SG100G), and will be the ones ultimately designing the first standard for THz band communications.

The applications of the Terahertz waves comprehend both macro and micro-scale communication scenarios, although the higher losses prevent such communications to be used for long range links. In any case, the work on this frequency band is on its infancy, and in order to have a practical wireless link some *a priori* elements must be developed further, such as transceivers capable of resonating at such high frequency, ultra-broadband antennas and accurate channel models. Much work has been done during the last few years on the topic (e.g. [13–19]), leading to a promising growth and making this not long ago unknown spectrum band much closer to become reality.

1.2 Facing the losses

Both mmWave and THz communications suffer from a major free-space path loss, due to the massive increase in carrier frequency and the higher effect of the molecular absorption. To give some rough figures of the impact of such high losses on the communication, see Fig. 1.1. With such numbers, if we do not apply any technique on top of a plain communication link using a quasi-omnidirectional antenna, the link cannot span more than a few meters in the best case.

Therefore, directional communications are required in order to reduce the impact of the losses in the link and reach the expected performance. This can be achieved with the implementation of large antenna arrays at both the transmitter and the receiver. Putting several antennas together in a certain disposition so that all of them transmit the same information creates constructive interference and transforms the wide beam

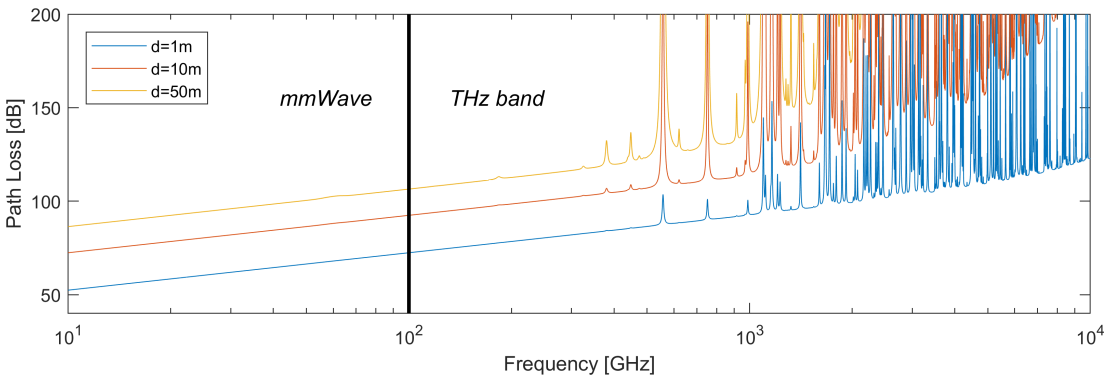

FIGURE 1.1: Path loss values for different link distances and frequencies in mmWave and THz bands

of a single antenna into a thinner beam that can be directed to a given angle by means of a certain phase pattern among the different antennas. The theory behind this technique is out of the scope of this thesis, but it is sufficient to know that antenna arrays are useful to create a directional link, and provide the communication path with an additional gain usually referred to as beamforming gain.

Using large antenna arrays is something applicable to any frequency. However, the antennas placed in this fashion should be separated typically by a distance equivalent to half of the wavelength. Moreover, each antenna element's size is also proportional to the wavelength. Therefore, antenna arrays working at low frequencies will occupy necessarily a large area, which renders this technique less useful. For instance, if working at 2.4 GHz (used by mainstream Wi-Fi technologies), an array made of 25 antennas arranged in a square would take more than 0.25 m^2, an impractical size for both Access Points (APs) and Mobile Equipment (ME). However, when working at higher frequencies, it is possible to pack dozens or even hundreds of antenna elements in reasonably sized arrays, thanks to the dramatic shortening of the wavelength in these bands (10 - 1 mm in the case of mmWave and 1 mm - 0.03 mm for Terahertz waves), as it has been already experimentally shown, e.g. [20].

When using antenna arrays at both transmitter and receiver, the communication link would be in fact a Multiple Input Multiple Output (MIMO) system. MIMO systems working with arrays containing such a large number of antennas are usually referred to as massive MIMO systems. Looking at the communication link from this perspective leads us to realize that, apart from the already mentioned *beamforming gain*, we can leverage as well the *spatial multiplexing gain*: sending not only one stream of information identical to all antennas, but different streams that can be mixed up in the transmitter and separated thereafter in the receiver.

1.2.1 Hybrid beamforming

When implementing antenna arrays in hardware, several structures can be used. Three different cases could be differentiated:

- **Fully digital structure:** each antenna element is connected to a separate RF chain, providing it with the capability of transmitting independent information. This design usually allows to take special advantage of the spatial multiplexing gain, but is more power consuming and has an elevated production cost.

- **Fully analog structure:** all the antenna elements in the array are connected to a single RF chain, i.e. the system can only transmit one stream at a time. Therefore, the system is much simpler and efficient, and is focused on the beamforming gain.
- **Hybrid structures:** although there are several different variations that may be grouped into this class, the key characteristic of these structures is the number of RF chains, N_{RF} is such that $1 < N_{RF} < N_{ant}$, where N_{ant} is the number of antenna elements in the array. In this way, we reach a compromise between the power consumption and the array capabilities.

Traditional MIMO systems working in frequency bands below 6 GHz usually present simple structures with 2, 4 or at most 8 antennas. Providing each of this antenna element with an RF chain does not increase greatly the production cost, and working at these lower frequencies it is more likely to have rich scattering environments which will increase the performance of the spatial multiplexing techniques. However, massive MIMO systems in high frequency bands, presenting tens or even hundreds of antenna elements, become incompatible with fully digital structures, for two reasons mainly [21]. On the one hand, attaching RF chains to every single antenna in these large arrays is not cost-effective, both in terms of production and power consumption. On the other hand, the sparsity of the mmWave channel and the high losses demand for leveraging beamforming gain as well. Although fully analog structures could be used, limited scattering present specially in mmWave makes hybrid structures the most fitting approach in this case.

MIMO systems working with hybrid structures are said to make use of *hybrid beamforming*, i.e. they rely in both digital baseband (to mix the different streams among the available RF chains) and analog processing (usually using only constant-modulus phase shifters attached to each antenna element).

Although hybrid architectures were first proposed for generic MIMO systems more than ten years ago [22, 23], the interest on this technique has raised again recently for its properties and advantages when applied to massive MIMO systems. However, the reduction in hardware complexity achieved by these hybrid MIMO systems leads to an increase in the signal processing complexity required to achieve similar performance to that obtained by standard fully digital architectures. This complexity greatly increases when multi-user scenarios are considered, where the Inter-User Interference (IUI) cancellation plays an important role.

For that reason, over the past few years, most of the work on the matter has been focused on a signal processing approach [24–31], that allows to leverage the characteristics of the channel to maximize the capacity of the link. This approach leads to a clean and quasi-optimal design, but at the expense of flexibility: the analysis is typically focused on a very specific context over the channel and the array structure, thus resulting in a closed-form solution which validity is limited to some set of assumptions.

Another important concern present when designing massive MIMO systems is the reduction of the power consumption. The usage of hybrid structures is already part of the solution to this issue. However, different hybrid approaches exist, with clear differences in performance depending on the characteristics of the final environment.

- Fully-connected structures have connections between every RF chain and all the antenna elements in the array, obtaining performance results close to those of fully digital architectures.
- A reduced-complexity version connects to each RF only a subset of the antennas, with the consequent reduction in power consumption and efficiency [30]
- Recently, a halfway solution was proposed ([32, 33]), where the RF chains can be dynamically connected to different antennas thanks to a switching network, which results in a better balance between both power consumption and flexibility.

1.3 Massive MIMO in Multi-User scenarios

Given the future significance of mmWave massive MIMO technology in cellular networks, it is of the greatest importance to focus on the design of multi-user scheduling schemes [30]. A generic MIMO system can easily be adapted to support several users, by transmitting multiple streams using either spatial multiplexing or separate beams per user. This second option is specially interesting when using a massive array, as the number of antenna elements is sufficiently large so as to be able to group them into several subarrays that eventually will serve different users each. Hybrid structures make it even easier to do this grouping, having already a *physical* division of antenna elements through the connection to the RF chains.

Using the dynamic-adaptive subarray formation mentioned before, and found in, e.g. [32, 33], we could eventually choose any subset of antennas and group them as an independent subarray which will serve a single user in the system. In this approach the antenna elements become a new resource to be allocated along with time/frequency slots, codes, computational resources... Therefore, we face an scheduling or resource allocation problem, something easy to formulate, rather than a complicated system with the added complexity of adaptive anntenna elements selection.

Nevertheless, the theoretical approaches found in most of the existing works on mmWave massive MIMO systems are not focused in the scheduling problem, but rather in the signal processing approach. The first aims at solving the usual case in cellular networks where many users are trying to access the same resources at the same time so that the BS must decide which of them to allocate resources and the amount granted to each of them taking into account priorities in the form of QoS requirements, while the second struggles with the development of new techniques that may guarantee a better Signal to Noise Ratio (SNR) in the final user, without taking too much into account the limited resources available.

Therefore, a systemic cross-layer approach, which leverages knowledge of the system as a whole and is able to adapt to different environments, might be the path to follow in order to design such a scheduling technique that allocates not only upper layer resources but also physical layer ones, having a granularity and control of the system never seen before.

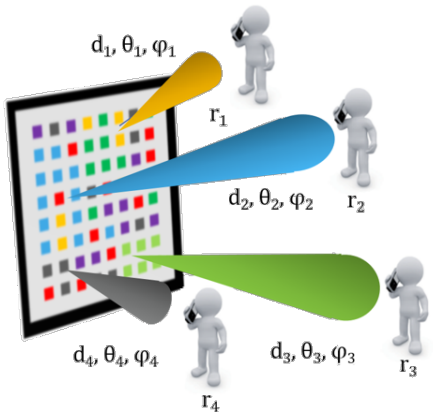

FIGURE 1.2: Sample scenario

1.4 Problem statement

With all the concepts introduced so far, we are now able to define the specific problem this project deals with. In this section this will be done in general terms, while the complete system model and other details can be found in Chapter 3.

We consider an scenario where several MEs are communicating with an AP –or Base Station (BS)– in the mmWave band. Using mmWave instead of THz is a matter of practicality of the final system, bearing in mind that for the moment the THz band is still in its infancy. However, most of the system could be adapted in order to work with THz communications and throughout the present work some of the eventually needed changes will be highlighted.

The BS uses a massive MIMO array to communicate with the users in the system. We focus on the downlink because it is in this case where we will need to share a large array among different receivers. Let us consider then that the BS assigns at a given time instant a certain subset of antenna elements (a subarray) to each one of the users. To do so, it takes into account the QoS requirements of the users, expressed in terms of average s needed. Naturally, the aggregated requirements of the users willing to connect to the BS could be higher than the capacity of the BS. Every user's position is described in spherical coordinates with the distance d_i from the BS, the elevation angle $\Theta_i \in [-\pi/2, \pi/2]$ and the azimuthal angle $\Phi_i \in (-\pi, \pi]$ (see Fig. 1.2).

The BS's array is designed following a hybrid architecture with a switching network that allows any antenna element to be connected to any RF chain. As shown in Fig. 1.2, the subarrays might be formed by non-adjacent antennas: no restrictions forbid irregular subarray shapes. In order to stress the resource allocation problem on the adaptive subarray distribution, we assume that all the users have to share the same temporal and frequency resources. The BS is assumed to have a minimum amount of information about the users's channel, which will be leveraged in order to improve the antenna assignation.

Therefore, the main objective of the Base Station will be to allocate the antenna elements in order to maximize the number of users that are served in a given time slot while meeting their requirements, having a limited number of resources in terms of antennas, power and bandwidth. Further details on the optimization function and the constraints can be found in Chapter 3.

1.4.1 Contributions

Summarizing, the aim of the present work is to propose a cross-layer approach in order to solve a multi-user resource allocation problem in a cellular network environment with resource constraints, when considering the usage of massive MIMO arrays working in high frequency ranges. Our contributions can be summarized as follows:

- We develop a low-complexity analog precoder for a Multi-User MIMO (MU-MIMO) context based on dynamic subarray allocation. Having a switching network allows us to assign each user any antenna in the array, thus optimizing the system and enabling a new perspective in MU-MIMO design by considering the antenna space a resource that is orthogonal to the other existing resources (time, frequency...). This new outlook unleashes new design pathways.
- We have used a reduced amount of Channel State Information (CSI) and assumed throughout the whole work a resource-constrained environment, producing a realistic proposal suitable for cellular network environments.
- A full simulation environment has been designed and implemented in MATLAB in order to simulate the entire system and evaluate the proposal for a wide range of parameters.

Chapter 2

State of the art

2.1 A taxonomy of massive-MIMO techniques in mmWave

Although to the best of our knowledge there is still no other proposal based on a resource allocation approach in order to optimize the user perceived QoS in a multi-user cellular network using massive MIMO arrays in high frequency bands, there has been a lot of work done, specially during the last ten years, in massive MIMO arrays at large. We will divide this section into three parts, each one focused on different aspects relevant to our work: single-user scenarios and the extension to multi-user case, channel information constrain assumption, and adaptive antenna assignment.

2.1.1 A personalized approach

The first works tackling the hybrid MIMO architectures were mainly focused in techniques emulating gains obtained in full-complexity structures with antenna selection. Naturally, the problem considered did not include several users, but started assuming an isolated user communicating with the BS. However, some of this proposals could not long after be extended easily to the multi-user case.

In [22] a novel soft antenna selection is proposed to optimize diversity gain in a single-user multi-stream link, while the work in [23] uses baseband preprocessing to exploit the spatial correlation of the received signals and to perform antenna selection without the need for a selection switch. The work in [34] proposes the optimal analog beamformer in a single-user system when considering interfering signals.

Most of these MIMO techniques developed for carrier frequencies below 6 GHz could illuminate the path for mmWave MIMO as well. Nonetheless, the particular characteristics present in this higher-frequency band must be taken into account. The constraints in hybrid structures and other hardware restrictions, the differences in the channel models to be used, and the size of the arrays lead to different complications when translating techniques not originally designed for mmWave communications [3].

Being one of the first works applying hybrid analog/digital precoding to mmWave MIMO systems, the work in [24] stressed the importance of the analog beamforming for low SNR (power constrained) environments and large distances, proposing an analog precoder based on phase sifters. The authors of [26] proposed to leverage the sparse nature of mmWave channels and the resemblance with the problem of sparse signal recovery with multiple measurement vectors to establish a multiple-stream single-user link based on orthogonal matching pursuit. Similar works based

on signal processing techniques such as block diagonalization and alternative minimization, appeared in the following years [35].

2.1.2 All together now

All these references assume a single-user scenario, but the future importance of mmWave massive MIMO systems in cellular networks involves taking into account multi-user environments, where the Inter-User Interference (IUI) arises as the main problem to solve. Although some single-user proposals could be extended to the multi-user case (e.g. [26] as an extension of [36]), many times this is not possible as the complexity of the analysis increases excessively. It is also important to consider the specific characteristics of the context, such as the increase of multi-path diversity due to the spatial separation of the receivers and the impossibility of having full knowledge of the channel in the BS due to the huge amount of antennas present.

In [27] Joint Spatial Division Multiplexing (JSDM) was applied to mmWave MU-MIMO systems to group users with via multiplexing or orthogonalization. However, there are some practical limitations on the orthogonal grouping of the users[30]. Further work on this scheme was presented in [37], where JSDM was generalized to support a less restrictive grouping. Similar techniques, such as hybrid block diagonalization [29], combination of ZF baseband precoding with Equal Gain Transmission [38] or convex optimization [39], suffer from a lack of flexibility, proposing a fix design for a specific environment.

2.1.3 Let's become real

Having the path cleared with all the previously mentioned works, it was time to tackle some more realistic assumptions. Any theoretical approach usually provides the scientific community with a trustworthy ground where a huge advance can be built, but its analytical nature requires many times to reduce the reality to a set of controlled parameters. Once this work has performed its task, the limits should be removed in order to actually come out with something realistic and functional.

In the case of massive MIMO research, the channel had been traditionally assumed to be perfectly known. Therefore, some works started to think of a limited channel feedback environment (e.g. [28] and [40]). In [28], a two-stage hybrid precoding algorithm was developed reducing the feedback by means of dimensionality reduction and quantization. The work in [40] applied a compressed-sensing technique employing randomization to estimate the channel in multi-user scenarios. Certainly, the theory behind limited channel information goes back many years and is not limited to the findings applied to this specific case, being clearly out of the scope of this work to explain it further.

2.1.4 Flexibility versus complexity

Adaptivity and resilience are not newcomers in the world of wireless communications. However, this has not been the case in the massive MIMO context, traditionally biased towards a signal processing approach. Nonetheless, adaptive antenna subarray techniques can be traced back to the mid-2000s. In [41], an evolutionary method was

proposed to group the antennas into different subarrays to accomodate to the channel at every moment and improve the performance, but without specifying in much detail the hardware structure required. Later on, the work in [32, 33] rationalized the use of switching networks in massive MIMO showing them to be an efficient alternative. The work in [31], apart from providing with a closed-form solution for a coupled design of the analog and digital precoder in both fully-connected and partially-connected hybrid structures, investigated further on the dynamic subarray allocation. Its results show that this adaptive technique provides a performance very close to the fully-connected structures and behave well in a wide range of environments. However, it is focused on a single-user environment. To the best of our knowledge, there is no work using adaptive subarray formation techniques in multi-user cellular networks.

2.2 The mmWave channel

The communication channel, being the medium which is traversed by the signal, determines the quality of the transmission. Its correlation with so many ambient factors makes it usually a quasi-random source of errors in the wireless link. In order to reduce the error rate, we need to know what are the variables that affect the most to the channel, and the way they affect it. If we have sufficient knowledge of the channel, we will be able to revert its effects and improve the quality of the link. That is the reason why we need analytical channel models. In this section we will introduce the most important models developed for environments similar to that we will be facing.

An extensive property analysis of the millimeter wave channel can be found in [42–46]. The millimeter wave band poses new challenges for wireless communications: from higher losses due to molecular absorption [46] to NLoS conditions due to the usage of highly directional beams to overcome the former.

In order to structure the section we will cover the main characteristics of the channel and their modelization following existing literature.

2.2.1 Path loss model

The classical path loss Alpha-Beta-Gamma (ABG) model is still in mmWave frequencies a valid first approach. The work in [43] showed that the model closely resembles the losses in the mmWave band when configuring the parameters accordingly. It uses three parameters to describe the path loss: *Alpha* represents the least square fits of floating intercept, *Beta* stands for the slope over the measured distances (a.k.a. Path Loss Exponent or PLE) and *Gamma* represents the lognormal shadowing variance (see Eq. 2.1).

$$PL(d)[dB] = \alpha + \beta \cdot \log_{10}(d) + \xi, \quad \xi \sim N(0, \sigma^2) \tag{2.1}$$

The numerical values determined for different frequency bands in the existing literature can be looked up in Table 2.1. The reason for the high values in the 60 GHz frequency band is the existing pike in molecular absorption.

On the other hand, authors in [45] propose a Close-In (CI) free space model and claim that, compared with the ABG model, the CI uses fewer parameters while offering

References	Freq. band	LoS PLE	NLoS PLE
[47] and [43]	28 GHz	1.7	4.6
[44] and [48]	38 GHz	2	3.9
[48]	60 GHz	2.25	4
[42]	73 GHz	2	3.4

TABLE 2.1: Numerical values for the PLE in ABG model for different frequency bands, as proposed in existing literature

intuitive physical appeal (see Eq. 2.2). It has been largely used in several applications ([49], [50]).

$$PL^{CI}(f,d)[dB] = FSPL(f,1\text{m})[dB] + 10n\log_{10}(d) + AT[dB] + \chi_{\sigma}^{CI}, \quad d \geq 1\text{m} \quad (2.2)$$

2.2.2 Delay Spread (DS)

The delay spread is found to be considerably smaller in the mmWave band compared to the microwave band. Results in [44] show that the delay spread is inversely proportional with the BS-ME distance. This relationship is sensitive to the propagation environment (indoors vs outdoors) as shown in [51], [43] and [44]. In addition, [51] demonstrates that the beamwidth[1] has also an impact on the delay spread, especially for NLoS conditions.

Regarding numerical values for different bands based on empirical studies, the delay spread was found to be in between 30ns to 80ns [42] in the 28 GHz band, around 12ns [43, 44] in the 38 GHz band and 39 to 47 ns in the 73 GHz band [49]. This results are confirmed in [42], which shows a cumulative distribution function (CDF) for the RMS Delay Spread. Finally, [46] reveals that the average spread in the 60 GHz band is comprised between 15 and 100ns with an average of 20ns.

2.2.3 Multipath components or path Clustering (MPC)

The millimeter wave channel has been largely characterized as a sparse multipath channel ([49], [46] and [52]) and widely adopted by the majority of the mmWave channel models ([51], [52], [53] and [54]). The MPC is characterized altogether by the following parameters: number of clusters (a.k.a. Time Clusters or TC), number of rays per cluster and the inter- and intra- cluster delays. The most important channel models characterizing the MPC are [52]:

(i) Stochastic tapped delay line model: It offers a simplistic cluster model where N clusters are defined, each of them with its own group delay, followed by M intra-cluster rays with equal power at the receiver. The 3GPP/ITU [51] follows this model, with $N = 6$ (typical, with a maximum value of 12) clusters for LoS, $N = 19$ for NLoS, and $M = 20$ equal power rays. [45] extends from the model proposed in [51], modifying the maximum number of clusters for both LoS and

[1]The beamwidth is usually measured as the angle separation between the two points in the space where the power radiated by an antenna first takes a value 3 dB below the maximum.

NLoS to a maximum of 6 and 5 respectively, claiming that such a high number of clusters is not supported by the real-world measurements at mmWave bands.

(ii) **Geometry-based stochastic model:** It provides a more complex channel model by introducing a location-based dependency according to a given probability density function ([52] and [42]). The COST2100 [53] is a extension of this model, with $N = 16$ clusters and $M = 20$ (typical) diffused intra-cluster rays in delay and/or angle. The WLAN model for 60GHz (currently under development) [21] REF!! defines 16 maximum incoming clusters. Each of these clusters contains a central ray, a maximum number of 6 rays preceding it (called pre-cursors) and a maximum number of 8 posterior rays (post-cursors).

(iii) **Semi-deterministic channel model:** It requires a map of the environment, where a reduced number of clusters (representing dominant scattering objects like buildings) can be determined. COST259 [54] was the first one to propose this model to model indoor and microcell environments. The Millimeter Wave Evolution for Backhaul Access model (MiWEBA) configures the number of clusters with a Poisson process with random inter-arrival times and Rayleigh distributed amplitudes [55]. On the other hand, the METIS model uses ray-tracing techniques and measurement-based results to characterize the large- and small-scale fading of the environment.

2.2.4 Angle of Arrival (AoA) and Departure (AoD)

The AoD is fully characterized by the beamforming technique used at the transmitter. As for the AoA, it is strictly related with the MPC model used, which determines the location of the scatters. In [47], the angles for each intra-cluster ray are withdrawn from a wrapped Gaussian distribution consistent with the model in [51]. [42] and [43] provide a comprehensive analysis in an outdoor scenario with NLoS condition. After establishing a -10dB threshold they observed that 2 Spatial Lobes (SL) could be found confined within the main 180 degree direction. Thus, it is reasonable to expect 2 to 3 SL per transmission, justifying the employment of wider beams in the Mobile Stations (MS) in reception to account for it ([51] and [56]). The angle spread at the BS in reception was measured to be around 30 degrees in millimeter wave [44]: that is why a narrower beamwidth is advisable at the BS.

2.2.5 Probability of LoS, NLoS and Outage

The analysis of blockage in mmWave has been thoroughly covered in [57]. Although many papers propose a coverage model, most of them rely on LoS and omnidirectional antennas, something that may not happen in real urban mmWave scenarios. Authors in [51], [58] and [59] propose a blockage model based on random shape theory where blockages are assumed to form a Boolean scheme of rectangles. The impact of blockage on outage and, in turn, the obtained capacity, is studied in [59] and [60] for urban environments. Studies in [47] extend from this theory, where they simplify the probability of being in LoS to a mere exponential function.

Chapter 3

From ideas to reality: design and implementation of a massive MU-MIMO resource allocator

3.1 System Model and Problem Formulation

The present chapter's main aim is to explain the inner workings of the proposal. In this section we will explain in detail the environment assumed in terms of the different parameters taking into account and the channel model used.

3.1.1 System Model

We consider a Base Station (BS) deployed to serve a small cell in the millimeter wave band, equipped with a planar array antenna with N_{BS} total number of antennas and a restricted total available power denoted by P_{BS} (see Fig. 3.1). The antenna elements in the array are considered to be separated by $\lambda/2$, where $lambda$ is the wavelength of the central frequency f_c. The BS serves a limited set of users $U = u_1, u_2, ..., u_M$ using a total bandwidth of W, and employs MU-MIMO to serve users concurrently in a TDMA fashion, i.e. several users are served in a single time slot. Thus, the scope of this work is on the MU-MIMO Downlink (DL).

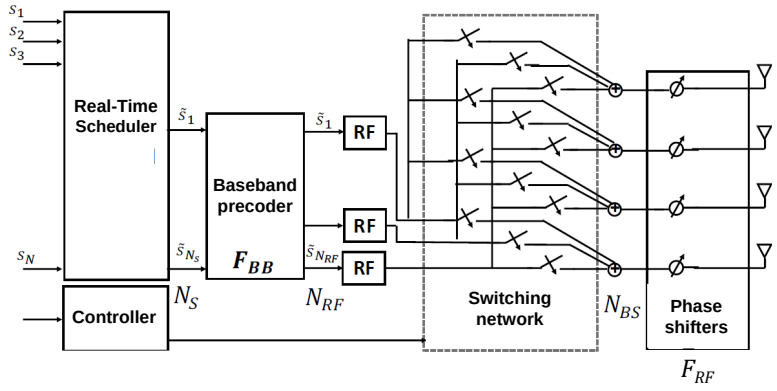

FIGURE 3.1: Block diagram in the Base Station, having separated the different phases (digital processing, switching network and phase shifters network).

The array at the BS is assumed to have limited number of RF chains (N_{RF}), bounded by a power constraint design limitation, that can be connected to a flexible subset of antennas (using a switching network), defining a sub-array per user served. Each antenna element has an analog phase shifter attached, but no amplifier/attenuator, thus having a modulus constraint (all the antenna elements connected to a given RF chain radiate the same signal, keeping the same amplitude but varying the phase).

No more than one bit stream can be allocated per user. Thus, the total number of streams (N_S) or users served concurrently (U_t, where t denotes the time interval) is upper bounded by N_{RF} ($U_t = N_S \leq N_{RF}$).

The Hybrid Beamforming mechanism is the key enabler for efficient directional transmission. The sampled transmitted signal ($\mathbf{x}$) is shown in Eq. 3.1, where $\mathbf{s}$ is a vector of transmitted symbols, with dimensions U_t by 1.

$$\mathbf{x} = \mathbf{F_{BB}}\mathbf{F_{RF}}\mathbf{s} \tag{3.1}$$

The $\mathbf{F_{BB}}$ and $\mathbf{F_{RF}}$ matrices capture the behavior of the transmitter. $\mathbf{F_{BB}}$ represents the baseband precoder at the transmitter, with dimensions N_S by N_{RF}, and controls the allocated power per user. In order to focus on the performance of the dynamic sub-array allocation, no spatial multiplexing is performed. Thus, the $\mathbf{F_{BB}}$ matrix is diagonal, which elements represent the amount of power allocated to every user/stream. $\mathbf{F_{RF}}$ contains the analog beamformer, with dimensions N_{RF} by N_{BS}, and allows for an efficient transmission scheme thanks to the optimal antenna elements allocation.

This analog precoder matrix can be seen as a selection matrix, where the non-zero elements express the connections between RF chains and antenna elements with the added information of the phase shift applied. Therefore, for every row there will be several non-zero elements (number of antenna elements attached to a certain RF chain), while every column is constrained to have only one non-zero element (a single antenna element cannot be connected to more than one RF chain).

Each user $u_i \ \ \forall i \in 1, ..., M$ is equipped with a planar array with N_{MS} total antennas, assumed to be fully digital, which is feasible due to the lower number of antennas expected in the user side. The received signal $\mathbf{r}$, accounting the effects of the channel and the optimal decoder $\mathbf{W}$ is described in Eq. 3.2, where $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_N^2\mathbf{I})$ is the Gaussian noise vector at the receiver.

$$r = \mathbf{W}^*\mathbf{H}\mathbf{x} + \mathbf{W}^*\mathbf{n} = \mathbf{W}^*\mathbf{H}\mathbf{F_{BB}}\mathbf{F_{RF}}\mathbf{s} + \mathbf{W}^*\mathbf{n} \tag{3.2}$$

3.1.2 Channel model

The mmWave channel model used in the present work is a Geometry-based stochastic model (see previous section) and follows the characterization in [42] for a short-range, wide-band, outdoors communication. Results in [47] also validate the robustness of the model. In a typical mmWave communication link, the wide-band property holds, as the measurements on the delay spread (in the order of 80ns for intra-cluster delays and 200ns average delay spread) satisfy $\tau_d \cdot W << 1$.

$$H(t,\Phi_{TX},\Phi_{RX}) = \sqrt{\frac{N_{BS}\cdot N_{MS}}{N_c+M_n}}\sum_{n=1}^{N_c}\sum_{m=1}^{M_n} a_{BS}(\Phi_{TX})\cdot a_{ME}(\Phi_{RX})\cdot\alpha_{m,n}\cdot e^{j\rho_{m,n}}\cdot\delta(t-\tau_{m,n}) \tag{3.3}$$

The directional channel response is shown in Eq. 3.3, where N_c and M_n are the number of clusters and intra-cluster rays respectively and a_{BS}, a_{ME} are the beamforming gains in the Base Station and the Mobile Equipment, which depend on the antenna selection and phases applied to each. $\alpha_{m,n}$ and $\rho_{m,n}$ are the path gains and phase shift respectively, while $\Phi_{TX} = (\phi_{TX}, \theta_{TX})$ and $\Phi_{RX} = (\phi_{RX}, \theta_{RX})$ are the tuples describing the azimuth and elevation angles at the transmitter and receiver respectively.

3.1.3 Problem formulation

Having explained in detail the system model, this subsection will be devoted to briefly state the problem in a mathematical formulation so as to clarify what the main objective and the existing constraints of our algorithm are.

The aim of this project is to maximize the number of users that can be allocated within one time slot while guaranteeing their required QoS. This problem can be formulated using the previously introduced notation as:

$$\begin{aligned} \{\mathbf{F}^*_{\mathbf{RF}}, \mathbf{F}^*_{\mathbf{BB}}\} \quad &= \quad \underset{\mathbf{F_{RF}},\mathbf{F_{BB}}}{\arg\min} \quad \sum_{\forall u\in U} \tilde{r}_u - r_u \\ s.t. \qquad & \quad \textstyle\sum_{i=j} \mathbf{F}^{*(\mathbf{i},\mathbf{j})}_{\mathbf{RF}} = 1 \\ & \quad \textstyle\sum_{\forall j} \mathbf{F}^{*(\mathbf{i},\mathbf{j})}_{\mathbf{RF}} = 1, \quad \forall i \in [1, N_{RF}] \\ & \quad \textstyle\sum_{i=j} \mathbf{F}^{*(\mathbf{i},\mathbf{j})}_{\mathbf{BB}} = p \\ & \quad \textstyle\sum_{i\neq j} \mathbf{F}^{*(\mathbf{i},\mathbf{j})}_{\mathbf{BB}} = 0 \end{aligned} \tag{3.4}$$

That is, the objective function is the difference between the required (r) and the actual ($\tilde{r}$) throughput, summed for all the users in the system. The constraints basically refer to the constant modulus in $\mathbf{F}^*_{\mathbf{RF}}$, the fact that an antenna can be connected to one and only one RF chain, the maximum power p in $\mathbf{F}^*_{\mathbf{BB}}$ and the fact that we are not using spatial multiplexing to mix streams into different RF chains.

3.2 Optimization Mechanism

The problem previously stated is NP-complete, thus having no algorithm to solve it in polynomial time. The amount of combinations of antenna selections allocated to each user that moreover satisfy the constraints makes it absolutely impossible to try an exhaustive serach in the solutions space.

For these reasons, we need to design a solver that is able to find suboptimal solutions with the best quality and in the least time. In this work we have followed two different approaches: the first aims at solving several subproblems that *discretize* the solution

space and then combine the solutions searching for the best. The second solution speeds up the process by means of a greedy approach.

3.2.1 First approach: decentralized heuristics

Divide and conquer. This sentence, used by the roman emperor Julius Cæsar and the french conqueror Napoleon Bonaparte, refers to the political strategy that divides the greatest parts of an entity in order to rule over them without difficulty, by reducing their individual power to the minimum. It has also been widely used in different algorithms, as an strategy that allows solving big problems with a simple plan of action.

The problem to be solved in this work is clearly a perfect match for this type of method, applying a decentralization of the scheduling policy. Instead of having a centralized brain optimizing the global problem in order to obtain the best solution to a highly complex problem, each user could *propose* several solutions taking into account some partial information from the other users. With the set of sub-solutions, the centralized authority could ultimately combine them and decide over the final allocation.

Agreeing on the number of proposals submitted by each user is not a trivial problem. A low number would eventually mean that no global solution could be found. On the other hand, a very high number would increase dramatically the execution time, rendering totally useless the application of the *divide and conquer* strategy. Therefore, this number is in fact one of the parameters of the optimization algorithm.

Briefly, the algorithm can be described with the following procedure: each user receives the basic data from the centralized entity of the problem to solve: the position and CSI of the other users and the configuration of the array. The CSI is limited, needing only the Angles of Departure (AoD) and the path gains of the channel observed per each user, i.e. based on the channel models presented in the previous section, the algorithm takes into account a reduced amount of information of the environment.

Knowing this, the user searches for the best subset of antennas that maximizes its received power and minimizes the received interference to the other users. In this way, when these partial solutions are combined with other sub-solutions, the joint interference will be minimized and thus the perceived SNR to all users will be maximized. The objective function to minimize expresses this idea as the weighted sum of two factors defining the quality of the antenna selection: the ratio between the resulting interference in the other users and the received power, and the width of the beam (which serves as a way to numerically formulate the directivity of the selected subarray) (see Eq. 3.5).

$$Q_u(\mathbf{F}^*_{\mathbf{RF}}, \mathbf{F}^*_{\mathbf{BB}}) = w_1 \cdot \frac{\sum_{\forall u' \neq u} I_u^{u'}}{P_{RX}^u} + w_2 \cdot BW_u \tag{3.5}$$

Observe that this optimization is equivalent to an analog beamforming optimization with the same main goal, looking for thin and highly directive beams and minimizing the secondary lobes in the directions where may affect the most to other users.

Each user solves this minimization sub-problem several times. The parameter that is modified for each different proposal is the number of antennas to use, called N_{MAX}. Therefore, each user provides the centralized solver with several solutions with different sizes in terms of percentage of resources (in this case antennas) used. Usually this

strategy will allow a uniform *discretization* of the solution space where the solutions with a higher number of antennas will provide a better solution in terms of the objective function Q, but consuming a lot of the constrained resources, while the solutions using only a few antennas might be a low-cost low-quality version.

The centralized entity simply needs to do a combinatorial search in the now very reduced solutions space. When combining sub-solutions it will be able to predict more or less accurately the final SNR per user, thus being able to determine whether the solution meets the requirements or not. The aforementioned discretization will be of use when looking for an optimal combination: the low-cost low-quality solutions will be useful for scenarios with low-demanding requirements or with good channel characteristics. On the other hand, high quality solutions will be required in highly-constrained problems.

The combinatorial search will maximize the objective function shown in Eq. 3.4, i.e. aims at maximizing the number of users that are allocated sufficient resources to meet their requirements. This, however, does not mean that the users will always receive the amount of resources needed, because the channel conditions and/or the resources available are insufficient to satisfy all the users' requirements. In this case, the algorithm will choose a solution including only some of the users, as a way to relax the constrains. In an extension of this work, an global scheduler working in an upper layer will be the one to decide which user is to be *evicted* of the system in a certain time slot by means of e.g. user priorities.

3.2.2 Second approach: greedy version

The presented approach takes the basic idea of *divide and conquer* strategy to reduce the complexity of the problem we are facing. However, the discretization of the solution space is performed in an arbitrary way and usually the resolution will not be enough to provide good results. In some cases, problems without extreme unbalance on the users' requirements or under good channel conditions will cause difficulties to the algorithm. Moreover, computing all the different sub-problems per user is still time- and resource-consuming. Therefore, we need to find a way to simplify the approach while increasing the adaptive capacity and the performance in terms of execution time.

The adaptive capacity of the previously presented algorithm is basically lost by the fact that every user has to execute their sub-problem for a set of *fixed* number of antennas. Moreover, the sub-solutions provided might not be compatible with the sub-solutions from other users, because of spatial constrains, as in the case where two sub-solutions have selected two subsets of antennas with a non-null intersection. Finally, the combinations could result in a combined selection where some antennas are not used, thus causing a lost in resource utilization that dramatically decreases the algorithm's efficiency.

Thus, a greedy strategy is proposed, where the users will try to get the most antennas they can, although keeping proportionality among them via their requirements. The algorithm is described in the diagram shown in Fig. 3.2. The number of antennas assigned to all users will sum up to the total number of antennas in the array, in order to maximize the resource utilization. Having assigned this fixed number of antennas, they will solve the same sub-problem as in the previous approach, only with a difference: the users do not solve it in parallel, but sequentially. In this way, the sub-solutions proposed will always be feasible to be combined. The sequential order

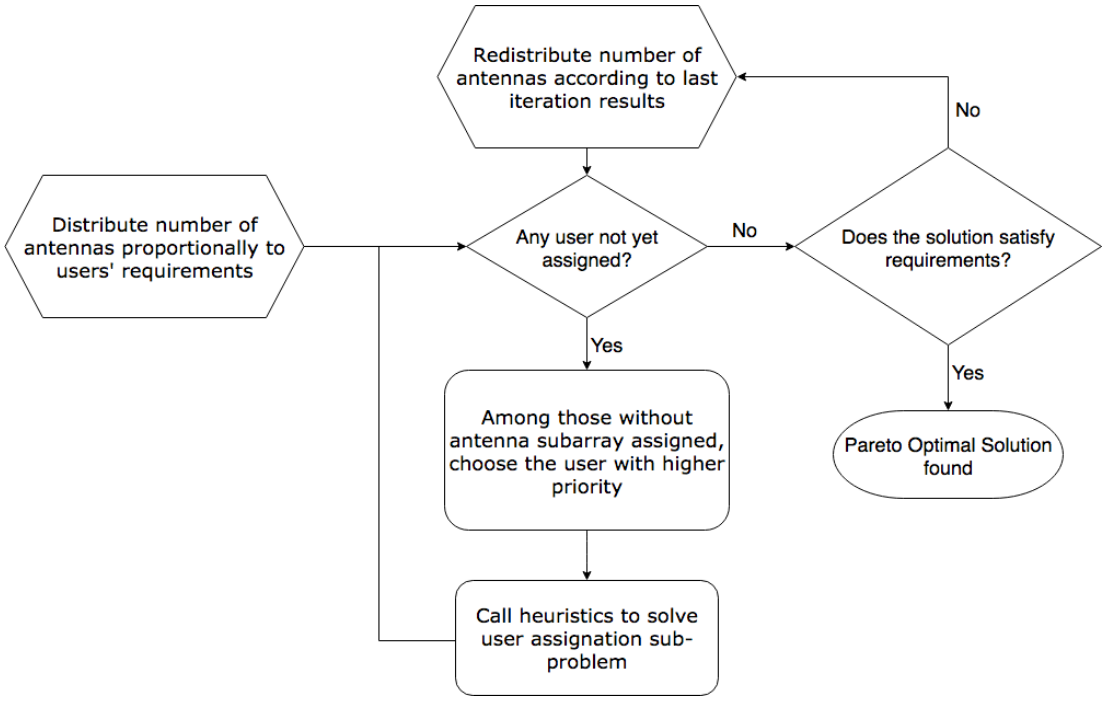

FIGURE 3.2: Greedy algorithm for antenna assignment

is not trivial, as the first users to choose will be able to obtain better fitted solutions, while the last user will have no choice.

It is easy to observe that such approach is not optimal in terms of the objective function to optimize in this resource allocation problem (review Eq. 3.4): we are trying to serve all users, without taking into account whether the system has enough capacity for all. This information is of course unknown, but leaving the algorithm as has been explained so far would be a clearly non-optimal. To avoid that, the algorithm revises the solution found in order to determine if there is *room for improvement*: we may found that while the users requirements have not been met some users have been assigned a number of antennas that has allowed them to have a predicted throughput much higher than what was required. What if we assign the excess antenna elements to the users that have not met their requirements? To do so, an approximation of the number of excess antennas per user is needed –obviously, the capacity achieved per antenna depends on the user, the channel state and the antenna within the whole assignment. This approximation can be computed as the ratio between the excess throughput and the average capacity delivered per antenna (computed separately per each user).

3.2.3 The big picture: a scheduler under a resource limited system

The presented approaches tackle only the resource allocation in a given time instant: having a fixed set of users, their requirements, and a configuration of the array, along with an static channel state information, they compute a near-optimal antenna assignment in order to meet all the requirements.

However, the real problem needs to extend this resource allocation over time, having a dynamic system, with traffic variations, users entering and leaving and constantly changing channel conditions. Therefore, an upper layer scheduler, which is in charge

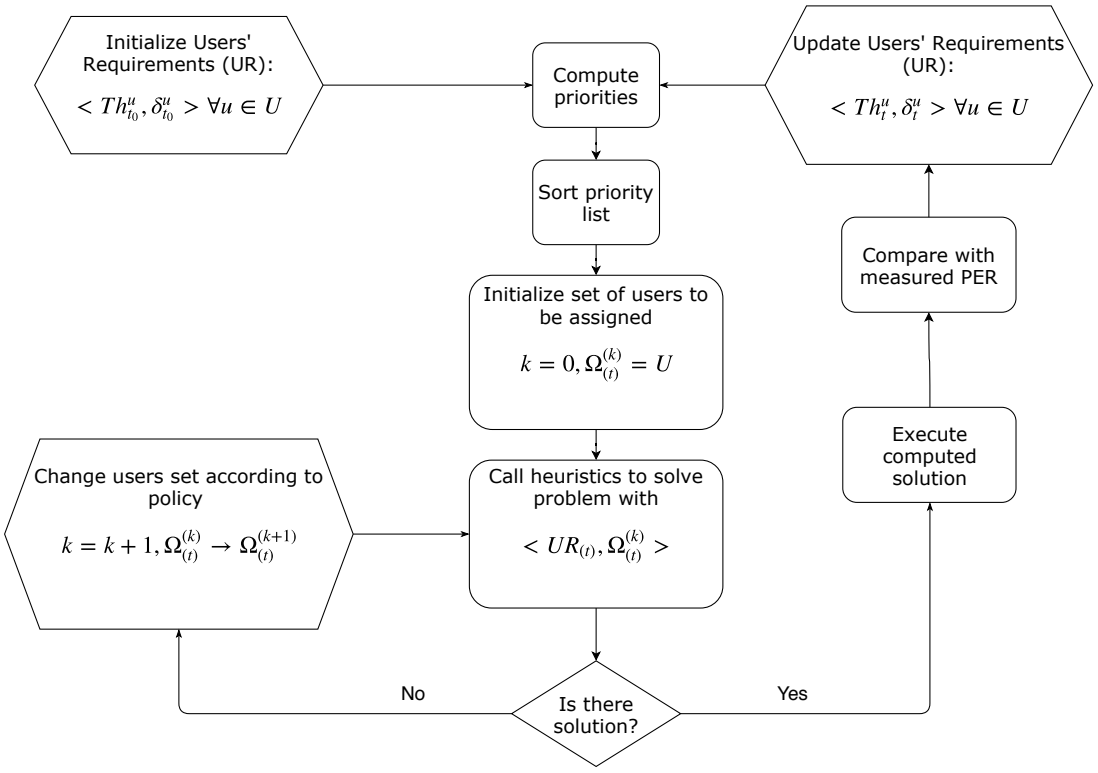

FIGURE 3.3: Upper layer orchestrator for a multi-user real time antenna subarray allocator

of running the allocation algorithm every time slot, is needed. A solution for this could take the form shown in Fig. 3.3.

The proposal, to be developed as future work, would be in charge of selecting the set of users to allocate in each time slot, and to distribute the throughput needs over time. That is to say, this scheduler would try to maximize the success rate (minimize the complexity) of each time slot's allocation.

3.2.4 A short study on some available heuristics

Along this section the resource allocation problem has been further explained and divided into several sub-problems, the most important of which is the resource allocation performed for a single user under the constraints of the channel and other users in the system. However, no specific description on the solution of this problem itself has been provided.

As the general all-users antenna allocation, the single-user case is an optimization problem. Both are non-linear and thus no simple approach can be applied. Solving this very complex problems usually requires either of the following two strategies: a highly-specific solution that leverages all the information known about the properties of the objective function, or a generic solution using heuristic algorithms.

The first possibility is clearly better, as the more we know about what we are looking for, the simpler the search is. However, the amount of information we can extract from a certain problem depends on many factors regarding the problem itself. In our case, we could intuitively guess that some antenna sub-array arrangements provide

a thinner beam or that the antenna phases can follow a certain distribution in order to aim the beam to a certain direction. However, irregular shapes when configuring antenna arrangements are known to perform better for some cases. For that reason, the problem is sufficiently complex so as to render infeasible any way to obtain a near-to-closed form for the solution.

Therefore, we have to explore the second alternative. In this case, we only need to describe the problem in a way that the chosen heuristic can *read* it and explore the solution space following its own strategy. Of course, the better we describe the problem and adapt it to the type of problems that a certain heuristic is designed to solve, the more quality and performance we will obtain.

Optimization heuristics exist since long time ago, being a clean and good performing strategy to solve problems that in other way would be infeasible. They are a big group of different algorithms that aim at traversing the solution space in an efficient way, both in terms of execution time and solution optimality. Therefore, they are mainly search algorithms. Evolutionary strategies, genetic algorithms and simulated annealing are different examples of search heuristics. In our problem we needed to determine which of these algorithms could perform better.

In order to make this decision, three different algorithms were implemented: Genetic Algorithms (GA), Particle Search Optimization (PSO) and Generalized Pattern Search (GPS). A brief description on each algorithm an its implementation details for our problem follows:

- **Genetic Algorithm [61]:** Genetic Algorithms may be considered a class of evolutionary algorithms, where the solution space is described using a biological counterpart. Each possible solution is formatted as a chromosome, a structure that contains a number of genes. At the beginning of the execution, there is an initial group of chromosomes chosen at random or following an specific rule. Every iteration, also called generation, the chromosomes evolve by means of a mutation or a crossover process (See Alg. 1). The chromosomes (solutions) with lower objective function score are called elite, and are the ones that will be chosen for the generation of the next group of chromosomes. In this way, at each iteration the algorithm will monotonically decrease the score of the minimum value found so far. Usually this technique is very useful and adaptive for highly complex, non-linearly bounded problems like this, accepting both discrete and continuous solution spaces. Although it usually obtains high-quality solutions, it is very computationally expensive.
- **Particle Swarm Optimization [62]:** This algorithm is a relatively recent heuristic invented in the mid 1990s and has given very promising results in all sorts of optimization problems. The basic idea is to mimic the behavior of the physical particles, flock of birds, herds of cattle or swarms of bees, which represent the possible solutions, and propagate their movement throughout the solution space taking into account their position and velocity at each iteration, as well as their best position so far (See Alg. 2). These parameters are influenced by the score of the different points in the solution space visited over time. Consequently, all the solutions in the set are aiming together at an optimum value. PSO is also enabled to solve non-linearly bounded problems obtaining high efectiveness without being too computationally demanding.
- **Generalized Pattern Search [63]:** Also known as direct search, this algorithm is in fact a family of numerical optimization techniques that do not require a

Algorithm 1 Genetic Algorithm (GA)

$Population \leftarrow$ InitializePopulation($Population_{size}$)
EvaluatePopulation($Population$)
$Sol_{best} \leftarrow$ GetBestSolution($Population$)
while ¬StopCondition() **do**
    $Parents \leftarrow$ SelectParents($Population, P_{elite}$)
    $Children \leftarrow \emptyset$
    for all $\langle p_1, p_2 \rangle \in Parents$ **do**
        $c_1, c_2 \leftarrow Crossover(p_1, p_2, P_{crossover})$
        $Children \leftarrow Children \cup$ Mutate($c_1, p_{mutation}$)
        $Children \leftarrow Children \cup$ Mutate($c_2, p_{mutation}$)
    EvaluatePopulation($Population$)
    $Sol_{best} \leftarrow$ GetBestSolution($Population$)
    $Population \leftarrow$ Replace($Population, Children$)
return Sol_{best}

gradient function. For the general case, the execution starts at a given point and explores moving a certain distance in all directions and comparing the scores for each. The point with the lowest score is used in the next iteration as the starting point. The distance used (also called mesh size) is reduced on the long term as we move toward the minimum (See Alg. 3). Although it has been used for a great range of different problems, it is usually a good option for local searches after applying other heuristics for the global search.

The three options presented are equally valid for our aim, as they are sufficiently flexible to be able to solve complex problems as the one we are tackling now. They have been shown to obtain high quality solutions while ensuring convergence in a reasonable time, although PSO is more computationally efficient [64] and GPS may obtain poor performance in functions with many discontinuities [65]. Encoding our problem for these heuristics is very similar in the three cases: for GA, each chromosome represents the antennas chosen for the subarray, the complex weights applied to them (amplitude and phase) and the values for the $\mathbf{F}^*_{\mathbf{BB}}$ matrix, while the other two see this exact same information as an N-dimensional point, where N is the total number of variables involved.

After implementing these three alternatives, we found nearly no remarkable difference among them, and we used most of the time GA for its configurability and easy-to-understand inner workings.

Algorithm 2 Particle Swarm Optimization (PSO)

$Population \leftarrow \emptyset$
$P_{g,best} \leftarrow \emptyset$
for $i = 1 : Population_{size}$ **do**
    $P_{vel} \leftarrow$ RandomVelocity()
    $P_{pos} \leftarrow$ RandomPosition($Population_{size}$)
    $P_{p,best} \leftarrow P_{pos}$
    if ObjFunc($P_{p,best}$) $\leq$ ObjFunc($P_{g,best}$) **then**
        $P_{g,best} \leftarrow P_{p,best}$
while ¬StopCondition() **do**
    for all $p \in Population$ **do**
        $P_{vel} \leftarrow$ UpdateVelocity($P_{vel}, P_{g,best}, P_{p,best}$)
        $P_{pos} \leftarrow$ UpdatePosition(P_{pos}, P_{vel})
        if ObjFunc(P_{pos}) $\leq$ ObjFunc($P_{p,best}$) **then**
            $P_{p,best} \leftarrow P_{pos}$
            if ObjFunc($P_{p,best}$) $\leq$ ObjFunc($P_{g,best}$) **then**
                $P_{g,best} \leftarrow P_{p,best}$
return $P_{g,best}$

Algorithm 3 Generalized Pattern Search (GPS)

$x_0 \leftarrow$ RandomValue()
$\Delta_0 \leftarrow$ InputValue(Δ)
$k \leftarrow 0$
$Sol_{best} \leftarrow \emptyset$
while ¬StopCondition() **do**
    $D_k \leftarrow$ GetPositiveSpanningDirections(x_k, Δ_k)
    for all $d \in D_k$ **do**
        $x_{k+1} = x_k + \Delta_k d$
        if ObjFunc(x_{k+1}) $<$ ObjFunc(x_{k+1}) **then**
            $\Delta_{k+1} \geq \Delta_k$
            $k \leftarrow k + 1$
            $Sol_{best} \leftarrow x_{k+1}$
            break
    $x_{k+1} = x_k$
    $\Delta_{k+1} < \Delta_k$
    $k \leftarrow k + 1$
return Sol_{best}

Chapter 4

Results

4.1 Evaluating complex problems: always an invaluable task

In the previous section all the details of the system and its implementation have been explained. In this section we aim at presenting some performance results and a brief analysis before entering into the last section, which concludes the present work.

When dealing with highly complex problems such as the one we are trying to solve, it is of the utmost importance to characterize and somehow reduce the input problem space in order to effectively evaluate and analyze the performance of a given solver. Without this simplification task, there is no way to come up with any conclusion regarding the efficiency and even the correctness of the solution proposed.

In our present problem we can clearly identify the most important variables affecting the *difficulty* of a given input problem:

- **Number of users.** As the number of users (i.e. candidates to be scheduled) increases, the complexity of the problem increases, as we have a higher number of allocation sub-problems to solve. Moreover, it will be more difficult to find a feasible solution that satisfies the requirements of all the users.
- **Number of antennas.** For the same number of users, decreasing the number of antennas available decreases the problem complexity (less number of possible antenna subsets), but it gets more difficult to find a solution meeting all users' requirements.
- **Users' requirements.** The influence of this variable on the problem's complexity is even more involved: it is directly related to the function we are trying to minimize, but in such a way that not only the absolute values of the requirements are important, but their statistical dispersion over the different users. That is, if we have a high sum of users' requirement, it is more likely to end up having no feasible solution serving all users, but obtaining a solution for a large subset of users will depend on whether the users have similar requirements or not. If, for instance, all the users have the same requirements we will probably find solutions serving less users than in the case where we have an unbalanced requirements situation: we could simply remove the most demanding user(s).
- **Users' relative position and the channel.** Of course, this is the most complex variable to control and the one that probably introduces the most intricate impact on the behavior of the system and the performance of the algorithm. It is

important to take into account that the channel's information used by the algorithm is limited, due to the impossibility of having a perfect knowledge of the channel, and thus, its effect is not completely cancelled.

- **Other minor variables.** The system can also be affected in varying nature and intensity when some parameters are changed: bandwidth and working frequency, geometrical disposition of the antennas in the array, and other physical characteristics of the Base Station and the Mobile Equipment.

Sweeping these parameters will show the behavior of the algorithm and its input limits (i.e. the limits on the input's complexity for the algorithm to be able to solve it in a reasonable period of time and spending a reasonable amount of memory). This previous analysis of the impact of each parameter shall be confirmed by the experimental simulations presented in this section.

4.2 Performance Tests

Now that we have briefly characterized the input variables that are more likely to affect the problem and how they might impact its complexity, we should describe the tests that have been performed in order to evaluate the behavior of the proposed solution.

However, in order to draw correct conclusions, the role of this algorithm in a bigger picture should be recalled. The measurements here performed are not capable to capture the system including the scheduler at a higher layer which is in charge of controlling the allocation algorithm presented here. This scheduler is out of the scope of this work, and is still being developed. Therefore, the results presented in this work, though useful for the analysis of the allocation problem, shall be considered only a partial and hence insufficient view of the performance of the whole system proposed.

A simple tester application has been implemented, in order to have a suitable testing platform to perform extensive simulations and sweeps of the critical variables to capture statistically valid measures. This application works as follows: given a set of input variables (number of antennas in the array, distance between them, working frequency, number of users, users' requirements, users' positions, channel parameters per user...), including sweeping variables with a range of values, it executes the antenna allocation algorithm for the static scenario defined. The default scenario considers a rectangular array with antenna elements along both axis, and is fully defined by the number and position of the users that are to be assigned, the antenna array characteristics (number and distribution of the antennas), the channel measurements and other parameters such as the working frequency and the Noise figure.

Some minor modifications and improvements to the algorithm can also be tested, such the solution refiner that aims at balancing the number of antennas assigned to the user explained in the previous section. Each simulation is performed several times in order to extract an statistically valid set of results. MATLAB programming language and some of its toolboxes have been used in the implementation: the antenna arrays are simulated using the Phased Array System Toolbox, the optimization algorithms use some functionalities of the Global Optimization Toolbox, and the execution is parallelized using the Parallel Computing Toolbox.

Four main test sets have been performed: three tests in which one out of the main critical variables (i.e. number of antennas in the array, number of users to be allocated and complexity of the channel in terms of the number of multi-path components) has been swept in a given range; and a test comparing our dynamic subarray proposal with several configurations that make use of statically conformed subarrays[1].

In the three main tests two different versions of the algorithm were run: the first one tries to refine the solution if it does not satisfy the constraint by removing the users with stronger requirements, one at a time, till it finds a solution or there are no users left; the second version only searches for a solution including all users, whether the requirements are met or not. This two versions allow us to benchmark the algorithm in two extreme cases, which will be useful in order to know its limitations. In all cases we have measured the capacity (bits/s/Hz) delivered by the array and the number of users being served:

1. In the first benchmark, the number of antennas is swept from 10 to 100, having fixed the number of users to 4 and the number of multi-path components to 6 (typical figure, as shown in section 2).
2. In the second benchmark, the number of users is swept from 2 to 10, while the number of antennas is set to 64 (a middle sized massive MIMO array) and the number of multi-path components is kept to 6.
3. In the third benchmark, the number of users is again fixed to 4 and we keep working with a 64-element array, while the multi-path components are swept from 3 to 12.
4. The fourth benchmark sets an environment equivalent to the first test, but for each problem created, four different antenna elements allocation policies are used: our dynamic proposal (any user can be assigned any subset of antennas in the array), localized (the user will receive only a subset of contiguous antennas in a square-like arrangement), interleaved (the possible subsets of antennas are formed by antennas interleaved in an vertical-horizontal fashion), and diagonally interleaved (same as before, but interleaving the antenna elements diagonally). See Fig. 4.1 for a graphical representation of these policies.

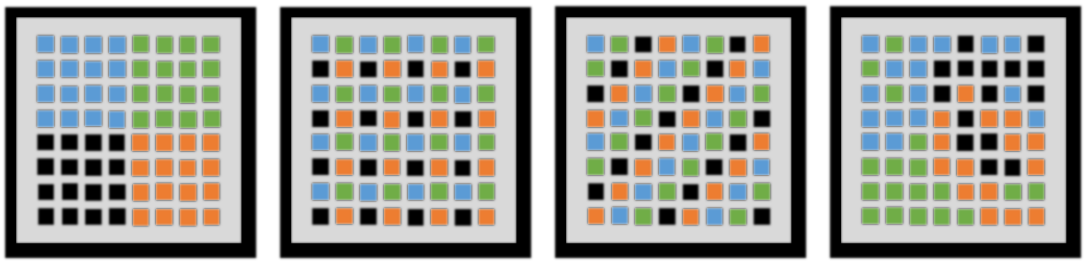

FIGURE 4.1: Main subarray arrangement policies (examples): localized (upper-left), interleaved (upper-right), diagonally interleaved (bottom-left) and dynamic (bottom-right)

All the tests were performed working at 60 GHz, with a 2 GHz channel bandwidth and a separation among antennas of $\lambda/2$, arranged in a Uniform Rectangular Array. The users requirements were uniformly distributed in the range 0.02 - 4 Gbps (very

[1]The sweeping ranges and the fixed values of the main input variables are chosen from the most representative cases.

highly demanding users), and were spatially distributed following a truncated normal distribution in the interval $\phi = [-45^o, 45^o]$, $\theta = [-45^o, 45^o]$, $d = [2, 12]$ (azimuth, elevation and distance respectively).

For a better comprehension and analysis of each one of the simulations performed, the results will be shown in three different sections: the solutions quality analysis (in terms of delivered capacity and number of users assigned), the algorithm performance analysis (execution time in the machine used for the experiments) and the behavior of the dynamic subarray allocation when compared to fixed subarrays preallocation.

4.2.1 Solutions quality analysis

The complexity of the problem to be solved in this work prevent us, as we have thoroughly explained, from being able to find the optimal solution in a reasonable amount of time. A sub-optimal algorithm is required, which performance is necessarily bounded by time and memory requirements as well as by its simplified assumptions and approach. However, there is no easy way to measure its performance but grading the quality of the solutions obtained for a given set of problems. In this subsection we will try to do so by showing and analyzing the results obtained by the algorithm proposed in terms of delivered capacity and number of users assigned. The first value will give us a benchmark of the utilization of the antenna array, while the second figure will give us an idea of the ability of the physical array in combination with the allocation method proposed to satisfy the input requirements.

With regard to the capacity of the system (see Figs. 4.2 and 4.3), both the average capacity and the total capacity is represented. We can see a clear downgrade in performance when trying to allocate all users: the total capacity obtained is around two times larger when optimizing the users finally allocated, due to the Inter-User Interference reduction. Conversely, the average capacity accounting all users is always larger as well, which means that applying this algorithm in a long-term basis would benefit the average throughput obtained.

The effect of increasing the number of users is worse than that observed when the multi-path components in the channel are multiplied: in fact, having a larger number of channel paths does not necessarily mean a bad-conditioned channel, as we could, in principle, take advantage of potentially orthogonal components to increase the performance.

Increasing the number of antenna elements has an interesting outcome: when assigning all users, we observe a non-monotonic increase in the capacity offered, which is consistent with the changes introduced in the antenna array. However, when optimizing the subset of users to be allocated, in order to obtain a better result, a second effect is present: for large antenna arrays, the algorithm gets very complex and is not able to converge to a good solution, not being able to obtain good results. This is the main reason why the capacity decreases at first when increasing the number of antennas. This effect is overcome later on and a slight increase is observed. Finally, for very large arrays, the delivered capacity decreases again. Therefore, we could affirm that there is an optimum value of number of antennas per combination of channel characteristics and number of users / data requirements to be allocated. In this case, we observe this clearly around 50-60 antenna elements. Obtaining this *golden figure* could imply having an easy-to-tweak design parameter in order to adapt the base station to the channel characteristics and traffic predicted.

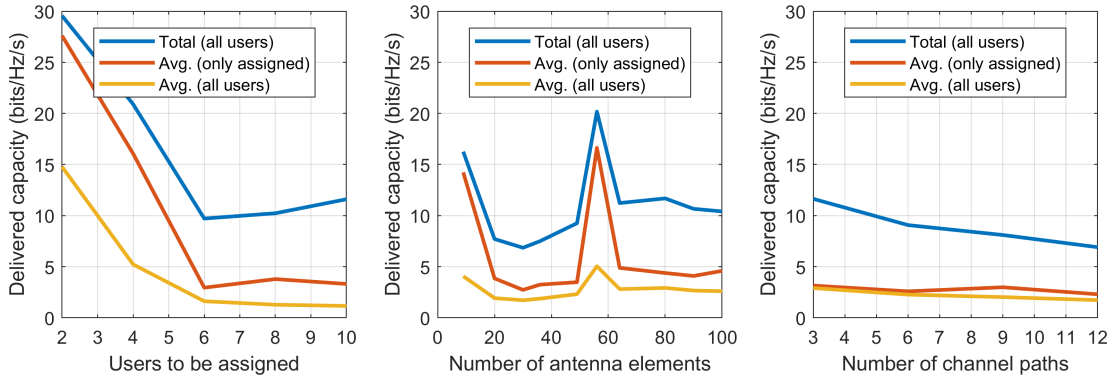

FIGURE 4.2: Delivered capacity when optimizing the subset of users to be allocated

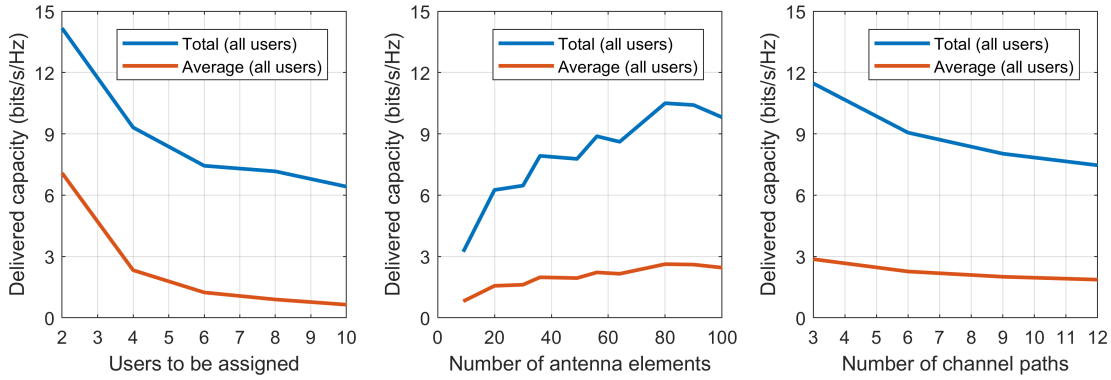

FIGURE 4.3: Delivered capacity for an all-users allocation

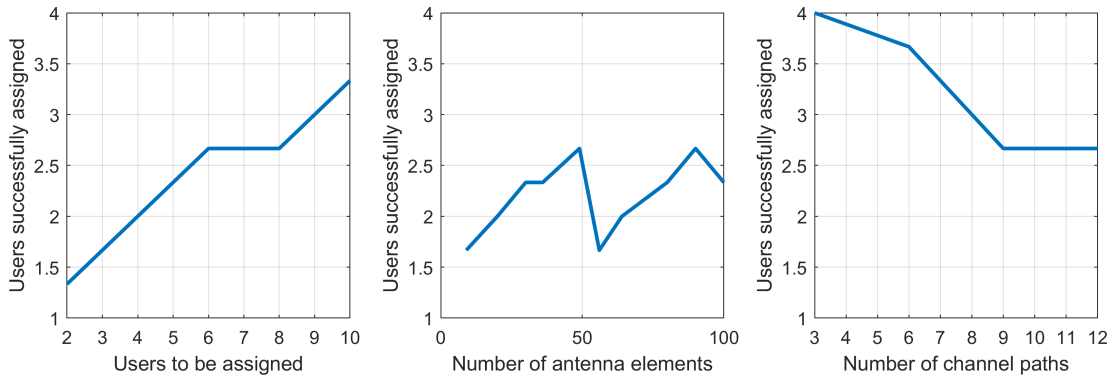

FIGURE 4.4: Users assigned when varying separately the number of antennas, the number of users and the multi-path components

The delivered capacity is not the only parameter giving out information about the quality of the solutions, we need to know as well the number of users successfully assigned. In Fig. 4.4 the results for the case where we optimize the subset of users to be allocated is shown (the case where all users are assigned is meaningless in this analysis). The Inter-User Interference (highly correlated to the number of users in the system), as well as the number of multi-path components, affects the environment introducing a downgrade in the performance. The fact that the users' requirements selected for these tests are very demanding is reflected on the results: it is not possible in almost all cases to meet all the existing demands. On the other hand, increasing the number of antenna elements in the array, which affects positively to the capacity of the system, is not a sufficient condition to overcome the Inter-User Interference.

This is related to the fact that the algorithm complexity increases with the number of antennas and hence the solutions' quality gets worse.

4.2.2 Algorithm performance analysis

The quality of the solutions obtained by the algorithm is of the utmost importance in order to analyze the optimality of the approach used. However, its performance depends also in the amount of resources used, specially the most critical ones in online / real-time applications such as the one we are facing: a resource allocation algorithm to be run continuously in a highly dynamic cellular network. We will devote this section to analyze the preliminary results obtained on the algorithm performance.

The absolute measurements of execution time could render useless if we did not provide along with them the characteristics of the testbed used. In this case, we have used the last available version of MATLAB (2017b) running on an Intel® Core™i5-3210M CPU @ 2.50 GHz (capable of running 4 threads in parallel) with 8 GB of RAM. We could expect a much more powerful machine in a realistic scenario, thus having better results than those shown here.

In Fig. 4.5 and 4.6 the execution time for the three tests performed is shown. The absolute values are in general unacceptable for a real case, taking in some extreme cases up to half an hour. This should be tackled by means of a optimization of the code and an upgrade of the computing power of the machine in charge of executing the algorithm. In any case, the relative values (comparison among the results obtained) remain perfectly valid for our analysis.

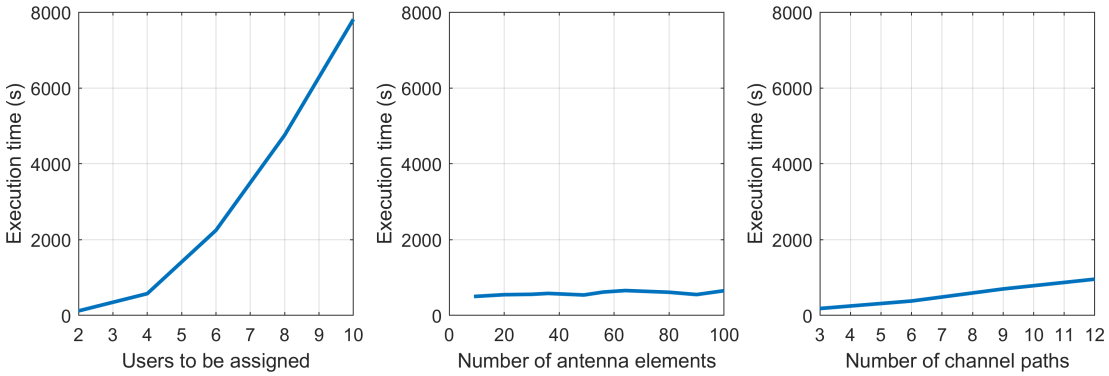

FIGURE 4.5: Execution time when optimizing the subset of users to be allocated

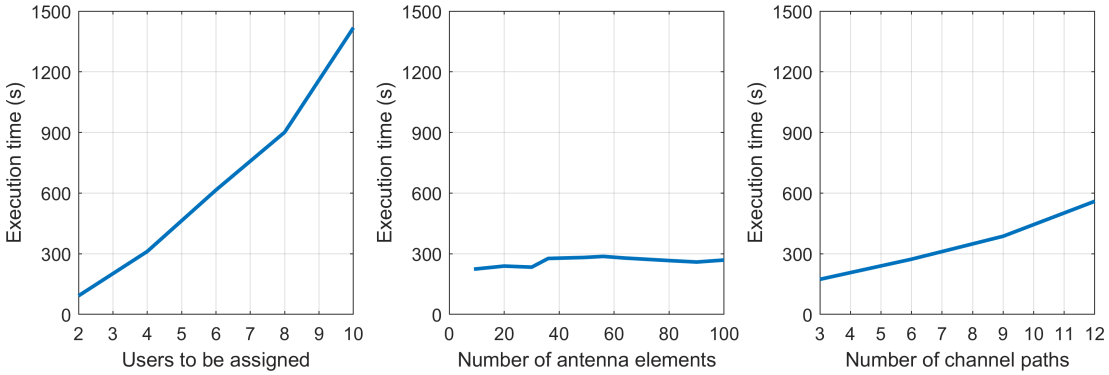

FIGURE 4.6: Execution time for an all-users allocation

As one could expect, the solution that optimizes the users to be assigned takes much longer, around twice the time it takes to do the same task in the version that only tries to assign all users. The *divide and conquer* approach leads to a clear exponential increase of the execution time when increasing the number of users. We can also see that the correlation between the number of antennas and the execution time is quite low: there is a very low increase along the whole range tested. Finally, the change in multi-path components affects only modestly to the execution time, which presents an almost constant slope increment.

4.2.3 Dynamic vs fixed subarrays

Although already appearing in [31], the concept of dynamically arranged subarrays is one of our most novel proposals. Using software-controlled switches or any similar device to dynamically divide a massive array into several subarrays enables the base station to adapt to the current channel conditions and traffic. The most frequently used fixed techniques, represented in Fig. 4.1, provide only a limited degree of freedom for allocating several users/streams of information in the same time/frequency slot, although they reduce the production cost and complexity of the array.

The average capacity delivered by a system with varying number of antennas, four users and a channel with 6 multi-path components when using four different subarray allocation techniques is shown in Fig. 4.7. For small antenna arrays, the advantage of dynamic subarray allocation over any fixed subarray technique is clear. The fact that this difference disappears for large antenna arrays is probably due to the time constraint given to the Genetic Algorithm that optimizes the selection, which reduces the quality of the solutions obtained for the dynamic subarray allocation algorithm.

This result supports clearly this technique as a highly performance technique that is able to squeeze the capacity of large antenna arrays adapting to the channel conditions and users requirements.

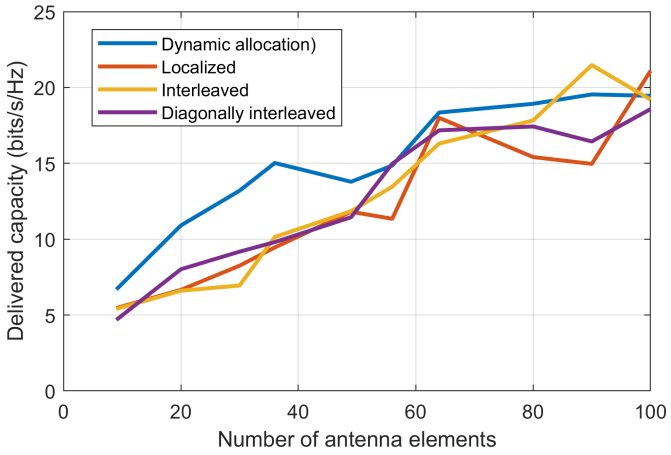

FIGURE 4.7: Performance comparison between dynamic and fixed subarray allocation techniques

Chapter 5

Conclusions

Multi-user systems using Massive MIMO arrays in mmWave are only starting to appear in the preliminary and more theoretical research. These systems leverage the previously no man's land of high frequency communications and the possibility of working with antenna arrays containing dozens or even hundreds of antenna elements, instead of the traditional MIMO devices with only up to 16 antennas. They respond to the call for high capacity systems that enable multiple ultra fast communications in mobile scenarios as one of the primary goals for 5G. However, being in a very early stage of development the challenges presented by these systems are countless. Dynamic antenna allocation has been proposed as a game changer to tackle them and push forward the capacity of such a promising technology.

In this work a dynamic subarray allocation algorithm has been fully designed and implemented, in a first approach using a simplified scenario and part of a bigger picture in which a real-time scheduler could be managing the users' demands and the available resources while using this algorithm to allocate antennas every certain period of time. The present implementation has been tested in a huge variety of scenarios, rigorously defining the parameters for large testing benchmarks. These tests have measured quantitatively the performance of the algorithm and a thorough analysis has been done in order to extract qualitatively conclusions. In the present section we will sum up all the experience gained with this work and outline some next steps to be done in the future.

5.1 Analysis of the results and validity of the proposal

In chapter 4 the most significant results have already been shown, but the analysis carried out there has been focused only on those aspects of the behavior of the algorithm that could be extracted from the figures obtained. Therefore, it is important to complete that information with a further description of some other things that should be taken into account in order to derive correct conclusions out of the present work. Most of them come from enlightening the qualitative analysis with the full knowledge of the actual flaws and simplifications of the present implementation, hence enabling us to clearly understand the reasons beneath every shadow and every light found in the results.

The most important thing to consider is probably the complexity of the simulation of an scenario that fulfills the requirements in order to accurately represent the real thing. The channel conditions, among others, are recreated in a very simplistic way for, as we mentioned before, considering a full knowledge of the channel is impractical in a real world system. Moreover, in our tests we have only changed the multi-path

characteristic of our channel model, leaving untouched the path gains, the statistical model for the AoD and the path loss model.

In second place, the scope of our tests for the parameters under study is reduced because of basic time constraints. Although in order to extract more meaningful conclusions and explore other scenarios we would need to do many more tests, for the moment the amount of time needed to execute an instance of the algorithm renders this task impossible.

Finally, the dummy scheduler used to control the algorithm (which in cases where it is not possible to allocate all users it simply decides to expel the user with the most demanding requirement) and the use of random user positions and requirements are responsible from having some amount of uncertainty of the extension of the results.

Nonetheless, the results presented in this work are very promising. Although there is need of doing a much larger and deep benchmark of the algorithm, and there is much room for improvement, the behavior of the algorithm in terms of delivered capacity fulfills the theoretical expectations. Despite the fact of requiring extra hardware and computational power, the algorithm presented could be a great alternative for future mobile communications to gain one dimension more in the resource allocation problem, while providing a solution that leverages the advantages and takes into account the limits of the mmWave channel.

5.2 Future work

The present work is an unfinished or even an ongoing project for several reasons: first and foremost, because it is a research work. The basis of the research is the *unsatisfiability*, the never-ending thirst for knowing better the world we live in and for improving it without limit. However, this project is an ongoing project in the full sense of the word, as it is the core element of an ambitious resource scheduler/planner for mobile communications in mmWave.

In any case, many different parts within the present work could be further improved in the future. From the results obtained we can identify at least the following pathways for the next steps to be done:

- The results shown in the previous section report the low performance of the algorithm in terms of execution time. Even though the computational power used could be increased, there is a clear pending task of improvement on the heuristics performance, which is the bottleneck of the whole execution. This improvement would be focused not only on the time reduction, but also on the solution quality and the efficiency of the solution space search process. In principle, we could use either of the already implemented algorithms (PSO, GA, GPS), tweaking their input parameters (generations, mesh size, inertia and acceleration...) to find the best fitting values.
- Within the algorithm's logic, the most important flaw is clearly the assignation of a certain number of antennas to each user, which depends (proportionally) on their requirements. In homogeneous cases this is not an issue. However, when solving a problem where the users' requirements are very heterogeneous, the proportional assignation could lead to high interference caused by users with low requirements reducing the overall thorughput of the system.

- Although the channel model already takes into account a realistic reduction of the information, it would be very interesting both for the improvement on the algorithm and the testing of more realistic cases to have a more accurate users' positions model and to use existing statistical models for AoD and other channel characteristics.
- The algorithm presented in this work is designed to be used by an upper layer scheduler, which would be in charge of deciding on which users need to be allocated and specifying their requirements, in order to satisfy a certain traffic demand. All this operation would be performed in real time. Therefore, the implementation of the whole system will give us a much more accurate vision of the behavior of this proposal when working in scenarios that are closer to the initial design.
- In this proposal all the users are assumed to transmit only one stream of information at a time. However, including several streams per user could eventually benefit the performance of the system by leveraging the multiplexing gain. That being said, the fact that the system is already quite complex and the scarcity in terms of multi-path components in mmWave could render spatial multiplexing almost useless.
- Although most of the work and the tests are focused in the mmWave band, the whole proposal has been designed having in mind any high frequency bands, including the THz band. Most of its characteristics are shared with mmWave, and the present proposal leverages them all (need of directive communications, high path loss, scarce number of multi-path components...).

Bibliography

[1] Cisco Visual Networking Index. "The zettabyte era–trends and analysis". In: *Cisco white paper* (2017).

[2] Qian Clara Li et al. "5G network capacity: Key elements and technologies". In: *IEEE Vehicular Technology Magazine* 9.1 (2014), pp. 71–78.

[3] 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.

[4] Carlos Cordeiro, Dmitry Akhmetov, and Minyoung Park. "Ieee 802.11Ad: Introduction and Performance Evaluation of the First Multi-gbps Wifi Technology". In: *Proceedings of the 2010 ACM International Workshop on mmWave Communications: From Circuits to Networks*. Chicago, Illinois, USA, 2010, pp. 3–8.

[5] Theodore S Rappaport, James N Murdock, and Felix 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.

[6] T. 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.

[7] Thomas Kürner and Sebastian Priebe. "Towards THz Communications - Status in Research, Standardization and Regulation". In: *Journal of Infrared, Millimeter, and Terahertz Waves* 35.1 (2014), pp. 53–62.

[8] IF Akyildiz, JM Jornet, and C Han. "TeraNets: Ultra-Broadband Communication Networks in the Terahertz Band". In: *Wireless Communications, IEEE* (2014), pp. 130–135. ISSN: 1536-1284. DOI: 10.1109/MWC.2014.6882305.

[9] Ian F. Akyildiz, Josep Miquel Jornet, and Chong Han. "Terahertz band: Next frontier for wireless communications". In: *Physical Communication* 12 (2014), pp. 16–32. ISSN: 18744907. DOI: 10.1016/j.phycom.2014.01.006.

[10] Darrel T Emerson. "The work of Jagadis Chandra Bose: 100 years of millimeter-wave research". In: *IEEE transactions on microwave theory and techniques* 45.12 (1997), pp. 2267–2273.

[11] Prasanna Adhikari. "Understanding millimeter wave wireless communication". In: (2008).

[12] "FCC ONLINE TABLE OF FREQUENCY ALLOCATIONS". In: (2017).

[13] Ian F. Akyildiz and Josep Miquel Jornet. "The Internet of Nano-Things". In: *Ieee Wireless Communications* December (2010), pp. 2–10. ISSN: 1607-551X. DOI: 10.1227/01.NEU.0000297013.35469.37.

[14] Ian F. Akyildiz and Josep Miquel Jornet. "Realizing Ultra-Massive MIMO (1024x1024) communication in the (0.06-10) Terahertz band". In: *Nano Communication Networks* 8 (2016), pp. 46–54. ISSN: 18787789. DOI: 10.1016/j.nancom.2016.02.001.

[15] Chong Han, A. Ozan Bicen, and Ian F. Akyildiz. "Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band". In: *IEEE Transactions on Wireless Communications* 14.5 (2015), pp. 2402–2412. ISSN: 15361276. DOI: 10.1109/TWC.2014.2386335.

[16] Chong Han, Wenqian Tong, and Xin Wei Yao. "MA-ADM: A memory-assisted angular-division-multiplexing MAC protocol in Terahertz communication networks". In: *Nano Communication Networks* 13 (2017), pp. 51–59. ISSN: 18787789. DOI: 10.1016/j.nancom.2017.08.001.

[17] J M Jornet and I F Akyildiz. "Channel Capacity of Electromagnetic Nanonetworks in the Terahertz Band". In: *Communications (ICC), 2010 IEEE International Conference on* (2010), pp. 1–6. ISSN: 1550-3607. DOI: 10.1109/ICC.2010.5501885.

[18] Josep Miquel Jornet and Ian F. Akyildiz. "Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the terahertz band". In: *IEEE Transactions on Wireless Communications* 10.10 (2011), pp. 3211–3221. ISSN: 15361276. DOI: 10.1109/TWC.2011.081011.100545.

[19] Josep Miquel Jornet and Ian F. Akyildiz. "Information capacity of pulse-based Wireless Nanosensor Networks". In: *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, SECON 2011* (2011), pp. 80–88. ISSN: 2155-5486. DOI: 10.1109/SAHCN.2011.5984951.

[20] Joao Vieira et al. "A flexible 100-antenna testbed for massive MIMO". In: *Globecom Workshops (GC Wkshps), 2014*. IEEE. 2014, pp. 287–293.

[21] Shu Sun et al. "MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both?" In: *IEEE Communications Magazine* 52.12 (2014), pp. 110–121.

[22] Xinying Zhang, Andreas F Molisch, and Sun-Yuan Kung. "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection". In: *IEEE Transactions on Signal Processing* 53.11 (2005), pp. 4091–4103.

[23] Pallav Sudarshan et al. "Channel statistics-based RF pre-processing with antenna selection". In: *IEEE Transactions on Wireless Communications* 5.12 (2006).

[24] Omar El Ayach et al. "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays". In: *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE. 2013, pp. 3476–3480.

[25] Ahmed Alkhateeb et al. "MIMO precoding and combining solutions for millimeter-wave systems". In: *IEEE Communications Magazine* 52.12 (2014), pp. 122–131.

[26] Omar El Ayach et al. "Spatially sparse precoding in millimeter wave MIMO systems". In: *IEEE transactions on wireless communications* 13.3 (2014), pp. 1499–1513.

[27] Ansuman Adhikary et al. "Joint spatial division and multiplexing for mm-wave channels". In: *IEEE Journal on Selected Areas in Communications* 32.6 (2014), pp. 1239–1255.

[28] Ahmed Alkhateeb, Geert Leus, and Robert W Heath. "Limited feedback hybrid precoding for multi-user millimeter wave systems". In: *IEEE transactions on wireless communications* 14.11 (2015), pp. 6481–6494.

[29] Weiheng Ni and Xiaodai Dong. "Hybrid block diagonalization for massive multiuser MIMO systems". In: *IEEE transactions on communications* 64.1 (2016), pp. 201–211.

[30] Andreas F Molisch et al. "Hybrid beamforming for massive MIMO: A survey". In: *IEEE Communications Magazine* 55.9 (2017), pp. 134–141.

[31] Sungwoo Park, Ahmed Alkhateeb, and Robert W Heath. "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems". In: *IEEE Transactions on Wireless Communications* 16.5 (2017), pp. 2907–2920.

[32] Roi Méndez-Rial et al. "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" In: *IEEE Access* 4 (2016), pp. 247–267.

[33] Ahmed Alkhateeb et al. "Massive MIMO combining with switches". In: *IEEE Wireless Communications Letters* 5.3 (2016), pp. 232–235.

[34] Vijay Venkateswaran and Alle-Jan van der Veen. "Analog beamforming in MIMO communications with phase shift networks and online channel estimation". In: *IEEE Transactions on Signal Processing* 58.8 (2010), pp. 4131–4143.

[35] Xianghao Yu et al. "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems". In: *IEEE Journal of Selected Topics in Signal Processing* 10.3 (2016), pp. 485–500.

[36] Omar El Ayach et al. "The capacity optimality of beam steering in large millimeter wave MIMO systems". In: *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*. IEEE. 2012, pp. 100–104.

[37] Zheda Li, Shengqian Han, and Andreas F Molisch. "Hybrid beamforming design for millimeter-wave multi-user massive MIMO downlink". In: *Communications (ICC), 2016 IEEE International Conference on*. IEEE. 2016, pp. 1–6.

[38] Le Liang, Wei Xu, and Xiaodai Dong. "Low-complexity hybrid precoding in massive multiuser MIMO systems". In: *IEEE Wireless Communications Letters* 3.6 (2014), pp. 653–656.

[39] An Liu and Vincent Lau. "Phase only RF precoding for massive MIMO systems with limited RF chains". In: *IEEE Transactions on Signal Processing* 62.17 (2014), pp. 4505–4515.

[40] Ahmed Alkhateeby, Geert Leusz, and Robert W Heath. "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 2909–2913.

[41] Panagiotis D Karamalis, Nikolaos D Skentos, and Athanasios G Kanatas. "Adaptive antenna subarray formation for MIMO systems". In: *IEEE Transactions on Wireless Communications* 5.11 (2006).

[42] M. K. Samimi and T. S. Rappaport. "3-D Millimeter-Wave Statistical Channel Model for 5G Wireless System Design". In: *IEEE Transactions on Microwave Theory and Techniques* 64.7 (2016), pp. 2207–2225. ISSN: 0018-9480.

[43] Theodore S Rappaport et al. "Millimeter wave mobile communications for 5G cellular: It will work!" In: *IEEE Access* 1 (2013), pp. 335–349.

[44] T. S. Rappaport et al. "Broadband Millimeter-Wave Propagation Measurements and Models Using Adaptive-Beam Antennas for Outdoor Urban Cellular Communications". In: *IEEE Transactions on Antennas and Propagation* 61.4 (2013), pp. 1850–1859. ISSN: 0018-926X.

[45] G. R. MacCartney, M. K. Samimi, and T. S. Rappaport. "Omnidirectional path loss models in New York City at 28 GHz and 73 GHz". In: *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*. 2014, pp. 227–231.

[46] P. F. M. Smulders and L. M. Correia. "Characterisation of propagation in 60 GHz radio channels". In: *Electronics Communication Engineering Journal* 9.2 (1997), pp. 73–80. ISSN: 0954-0695.

[47] M. R. Akdeniz et al. "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation". In: *IEEE Journal on Selected Areas in Communications* 32.6 (2014), pp. 1164–1179. ISSN: 0733-8716.

[48] T. S. Rappaport et al. "38 GHz and 60 GHz angle-dependent propagation for cellular amp; peer-to-peer wireless communications". In: *2012 IEEE International Conference on Communications (ICC)*. 2012, pp. 4568–4573.

[49] T. S. Rappaport et al. “Wideband Millimeter-Wave Propagation Measurements and Channel Models for Future Wireless Communication System Design”. In: *IEEE Transactions on Communications* 63.9 (2015), pp. 3029–3056. ISSN: 0090-6778.

[50] S. Sun et al. “Investigation of Prediction Accuracy, Sensitivity, and Parameter Stability of Large-Scale Propagation Path Loss Models for 5G Wireless Communications”. In: *IEEE Transactions on Vehicular Technology* 65.5 (2016), pp. 2843–2860. ISSN: 0018-9545.

[51] T. Bai, V. Desai, and R. W. Heath. “Millimeter wave cellular channel models for system evaluation”. In: *2014 International Conference on Computing, Networking and Communications (ICNC)*. 2014, pp. 178–182.

[52] A. F. Molisch et al. “Millimeter-wave channels in urban environments”. In: *2016 10th European Conference on Antennas and Propagation (EuCAP)*. 2016, pp. 1–5.

[53] L. Liu et al. “The COST 2100 MIMO channel model”. In: *IEEE Wireless Communications* 19.6 (2012), pp. 92–99. ISSN: 1536-1284.

[54] A. F. Molisch et al. “The COST259 Directional Channel Model-Part I: Overview and Methodology”. In: *IEEE Transactions on Wireless Communications* 5.12 (2006), pp. 3421–3433. ISSN: 1536-1276.

[55] Richard J Weiler et al. “Quasi-deterministic millimeter-wave channel models in MiWEBA”. In: *EURASIP Journal on Wireless Communications and Networking* 2016.1 (2016), p. 84.

[56] S. Rajagopal, S. Abu-Surra, and M. Malmirchegini. “Channel Feasibility for Outdoor Non-Line-of-Sight mmWave Mobile Communication”. In: *2012 IEEE Vehicular Technology Conference (VTC Fall)*. 2012, pp. 1–6.

[57] S. Singh et al. “Blockage and directivity in 60 GHz wireless personal area networks: from cross-layer model to multihop MAC design”. In: *IEEE Journal on Selected Areas in Communications* 27.8 (2009), pp. 1400–1413. ISSN: 0733-8716.

[58] T. Bai, R. Vaze, and R. W. Heath. “Using random shape theory to model blockage in random cellular networks”. In: *2012 International Conference on Signal Processing and Communications (SPCOM)*. 2012, pp. 1–5.

[59] T. Bai, R. Vaze, and R. W. Heath. “Analysis of Blockage Effects on Urban Cellular Networks”. In: *IEEE Transactions on Wireless Communications* 13.9 (2014), pp. 5070–5083. ISSN: 1536-1276.

[60] T. Bai and R. W. Heath. “Coverage and Rate Analysis for Millimeter-Wave Cellular Networks”. In: *IEEE Transactions on Wireless Communications* 14.2 (2015), pp. 1100–1114. ISSN: 1536-1276.

[61] David B Fogel. *Evolutionary computation: the fossil record*. Wiley-IEEE Press, 1998.

[62] James Kennedy. “Particle swarm optimization”. In: *Encyclopedia of machine learning*. Springer, 2011, pp. 760–766.

[63] Robert Hooke and T. A. Jeeves. ““ Direct Search” Solution of Numerical and Statistical Problems”. In: *J. ACM* 8.2 (Apr. 1961), pp. 212–229. ISSN: 0004-5411. DOI: 10.1145/321062.321069.

[64] Rania Hassan et al. “A comparison of particle swarm optimization and the genetic algorithm”. In: *46th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference*, p. 1897.

[65] Michael Wetter and Jonathan Wright. “Comparison of a generalized pattern search and a genetic algorithm optimization method”. In: *Proceedings of the 8th International IBPSA Conference, Eindhoven, Netherlands*. 2003, pp. 1401–1408.