Coordination inside centralized radio access networks with limited fronthaul capacity

Jialong Duan

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

UNIVERSITE BRETAGNE LOIRE

THÈSE / IMT Atlantique

sous le sceau de l'Université Bretagne Loire

pour obtenir le grade de

DOCTEUR D'IMT Atlantique

Spécialité : Informatique

École Doctorale Mathématiques et STIC

Présentée par

Jialong Duan

Préparée dans le département Systèmes réseaux, cybersécurité & droit du numérique

Laboratoire Irisa

Coordination inside Centralized Radio Access Networks with Limited Fronthaul Capacity

Thèse soutenue le 27 novembre 2017

devant le jury composé de :

Maryline Hélard
Professeur, Insa - Rennes / présidente

Mohamed Assaad
Professeur, Centrale Supélec – Gif sur Yvette / rapporteur

Christophe Le Martret
Ingénieur expert (HDR), Thales Expert, Thales Communications & Security / rapporteur

Charly Poulliat
Professeur, INP/ENSEEIHT - Toulouse / examinateur

Philippe Chanclou
Ingénieur expert, Orange Labs – Lannion / examinateur

Frédéric Guilloud
Maître de conférences (HDR), IMT Atlantique / examinateur

Xavier Lagrange
Professeur, IMT Atlantique / directeur de thèse

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Jialong DUAN
March 2018

Acknowledgements

On the way to this thesis I received lots of help from different people, I would like to take this opportunity to express my appreciations.

Firstly, I am deeply indebted to my supervisors Prof. Xavier LAGRANGE and Associate Prof. Fédéric GUILLOUD. Their contagious enthusiasm and patience went a long way in bringing this work into its final form. From them, I learnt the value of rigor and motivation in research.

Furthermore, I would like to thank my thesis committee for reviewing my dissertation and for their constructive comments. It is my utmost honor to have all these experts reviewing my work.

Moreover, my sincere thanks also goes to my colleagues in SRCD department: Younes, Xavier, Roberto, Sumy, Saad, Francois. I would never forget the days we worked together. A special thanks to Yangyang CHEN for lots of suggestions on both academic and daily life matters. Also, these years in Rennes, I made some very good friends like Qipeng Song, Yue Li, Yanhuang Li, Wenjing Shuai, Hao Lin, Bihong Huang, Xiaoran Jiang, Xiao Song, Nan Feng and Miaojing Shi. Because of you guys, I had a very colorful life, this would definitely be one of my best memories.

Last but not the least, I would like to thank my family, my parents Hongtao DUAN and Xin ZHAO. No matter what decisions I have made, thank you for your endless love and support.

Abstract

Centralized/Cloud Radio Access Network (C-RAN) is a promising mobile network architecture, which can potentially increase the capacity of mobile networks while reducing operators' cost and energy consumption. However, the feasibility of C-RAN is limited by the large bit rate requirement in the fronthaul. The objective of this thesis is to improve C-RAN performance while considering fronthaul throughput reduction, fronthaul capacity allocation and users scheduling.

We first investigate new functional split architectures between Remote Radio Heads (RRHs) and Baseband Units (BBU) on the uplink to reduce the transmission throughput in fronthaul. Some low layer functions are moved from the BBU to RRHs and a quantitative analysis is provided to illustrate the performance gains.

We then focus on Coordinated Multi-point (CoMP) transmissions on the downlink. CoMP can improve spectral efficiency but needs tight coordination between different cells, which is facilitated by C-RAN only if high fronthaul capacity is available. We compare different transmission strategies without and with multi-cell coordination. Simulation results show that CoMP should be preferred for users located in cell edge areas and when fronthaul capacity is high. We propose a hybrid transmission strategy where users are divided into two parts based on statistical Channel State Informations (CSIs). The users located in cell center areas are served by one transmission point with simple coordinated scheduling and those located in cell edge areas are served with CoMP joint transmission. This proposed hybrid transmission strategy offers a good trade-off between users' transmission rates and fronthaul capacity cost.

Résumé

Le réseau d'accès radio centralisé (C-RAN) peut fortement augmenter la capacité des réseaux mobiles. Cependant, la faisabilité de C-RAN est limitée par le débit considérable engendré sur les liaisons de transport, appelées également *fronthaul*. L'objectif de cette thèse est d'améliorer les performances de C-RAN tout en considérant les limitations du débit sur le frontaul, l'allocation de ressources et l'ordonnancement des utilisateurs.

Nous étudions d'abord les séparations fonctionnelles possibles entre les têtes radios distantes (RRH) et les unités de traitement en bande de base (BBU) sur la liaison montante pour réduire le débit de transmission sur le *fronthaul* : certaines fonctions de couche basse sont déplacées du BBU vers les RRH. Nous fournissons une analyse quantitative des améliorations de performances ainsi obtenues.

Nous nous concentrons ensuite sur la transmission coordonnée Multi-point (CoMP) sur le lien descendant. CoMP peut améliorer l'efficacité spectrale mais nécessite une coordination inter-cellule, ce qui est possible uniquement si une capacité *fronthaul* élevée est disponible. Nous comparons des stratégies de transmission avec et sans coordination inter-cellule. Les résultats de simulation montrent que CoMP doit être préféré pour les utilisateurs situés en bordure de cellule et lorsque la capacité du *fronthaul* est élevée. Nous en déduisons une stratégie hybride pour laquelle Les utilisateurs sont divisés en deux sous-ensembles en fonction de la puissance du signal. Les utilisateurs situés dans les zones centrales sont servis par un seul RRH avec une coordination simple et ceux en bordure de cellule sont servis en mode CoMP. Cette stratégie hybride constitue un bon compromis entre les débits offerts aux utilisateurs et les débits sur le *fronthaul*.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

Mobile data traffic is forecasted by Cisco's Visual Networking Index (VNI) to increase 7-fold between 2016 and 2017, reaching 48.3 EB per month by 2021 [1] as shown in Figure 1.1. Meanwhile, the future 5G system is required to support an increase factor of $10-100$ times of the transmission user data rate and $10-100$ times more devices with low delay (millisecond level) [6]. Therefore, operators need to invest more to increase the mobile network capacity, such as building more base stations (BSs).

Recent studies propose a number of techniques to satisfy the data explosion: Heterogeneous Networks (HetNets) that mix implementation of macro and micro base stations (small cells) [7, 8], Massive MIMO [9, 10] that explore spatial diversity and serve several User Equipments (UEs) simultaneously by a very large number of antennas in the same time-frequency resource. However, these techniques are all focused on hot spots where the density of users is high. We would like also to improve the system performance in a macro cellular network. The inter-cell interference largely limits the data transmission, especially for the UEs located in the cell edge areas. Advanced algorithms such as Coordinated Multiple-point (CoMP) have been introduced to manage inter-cell interference. Nevertheless, they need tight coordination between different cells, which cannot be satisfied by current radio access network (RAN). Furthermore, the current RAN also faces other great challenges.

At first, the building of more BSs and implementation of the above-mentioned new techniques result in high cost for mobile operators. Both capital expenditure (CAPEX) and operating expenditure (OPEX) will largely increase. However, operators will not get a proportionate increase in revenue. With tough price competition, the overall mobile service revenue is expected with only 1.5 percent annually growth from 2016 to 2026, in contrast to a decade ago, with 10-15 percent growth [11]. The Average Revenue Per User (ARPU) is

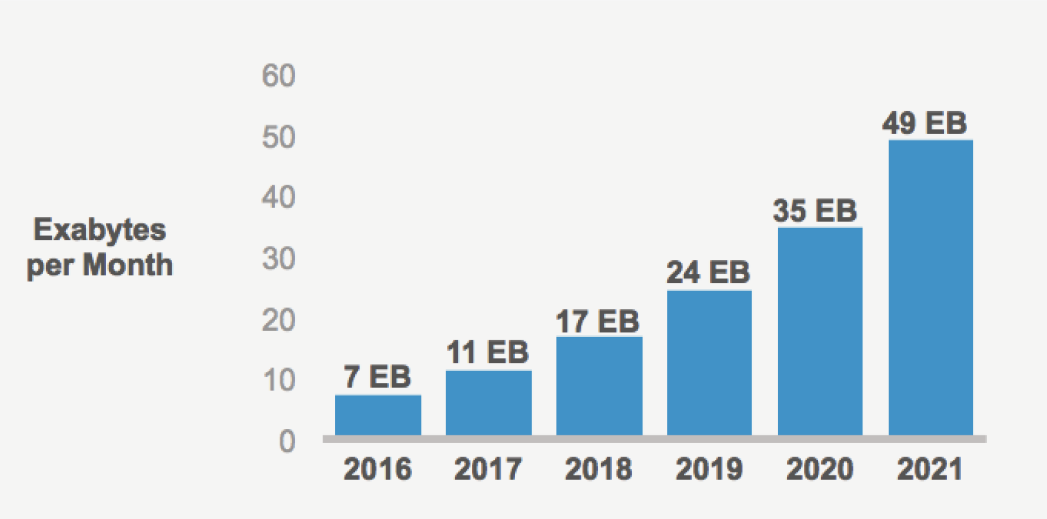

Fig. 1.1 Global mobile data traffic growth [1].

not predicted to increase because a typical user expects to pay less for data usage. Operators are facing the challenges to satisfy the explosion of data usage while reducing the cost.

Secondly, the increasing energy consumption of the current RAN augments both operators' OPEX and green gas emission. The worldwide communication systems swallows up 334 TWh in 2012, counting for 1.5 percent of the total electricity use in the world [12]. The corresponding impact on global warming is not negligible. A major target of future 5G network is to reduce power consumption [13]. Traditional radio access network incurs enormous energy consumption. In [14], it shows that 72% of the power consumption of China mobile originates from the Base Stations (BSs).

All in all, a novel RAN architecture that optimizes cost and energy consumption and benefits network capacity increasing becomes necessary. To address the aforementioned problems, a promising future mobile network architecture, namely Centralized/Cloud Radio Access Network (C-RAN), has been introduced by China Mobile Research Institute in April 2010 [5]. Similar ideas have been proposed several years ago in [15]. The letter C in C-RAN can be interpreted as: Cloud, Centralized processing, Cooperative radio, Collaborative or Clean.

In Long Term Evolution (LTE), most of eNodeBs (eNBs) contain two main parts: Baseband Units (BBUs) and Remote Radio Heads (RRHs). C-RAN remotes BBUs from RRHs and puts them to a BBU pool for centralized processing. The BBU pool connects to the core network through backhaul, and the optical network connecting the BBU pool and RRHs is defined as fronthaul. The centralization of BBUs facilitates the cooperation between different cells. Enormous data exchanging with low delay between different cells becomes possible. It

enables many advanced algorithms that were hard to implement in traditional RAN, such as enhanced Inter-cell Interference Coordination (eICIC), CoMP, network coding to name a few [16]. Furthermore, with clouding and virtualization, the BBUs resources can be easily managed and dynamically allocated on demand. This brings statistical multiplexing gains, energy and resource savings, which reduce the cost. C-RAN can be considered as an instance of Network Function Virtualization [17]. The virtualization of BBU resources can further facilitate scalability and integration of different services [18]. With BBUs centralized in the BBU pool, the deployment of remote sites with light RRHs and antennas reduces the cost in construction and operation and provides flexibility in network upgrades [19]. A quantitative analysis in [20] shows that C-RAN can lead to a $10\% - 15\%$ CAPEX reduction per kilometer comparing with LTE. It also decreases power consumption as less air conditioners are needed to be installed in distributed sites. China Mobile Research Institute forecasts that C-RAN brings 71% powering savings compared with current RAN architecture [5] and ZTE estimates $67\% - 80\%$ depending on the number of cells covered by the BBU pool [21]. A reduction of 50% OPEX is predicted in [5].

The potential remarkable benefits of C-RAN motivate both major mobile operators and equipment vendors regarding it as a competitive realization of mobile network supporting future 5G soft and green technologies. The advocates include mobile operators such as China Mobile, Orange, NTT DoCoMo, Telefonica, etc, and equipment vendors such as Huawei, ZTE, Ericsson, IBM, Nokia, Intel, Texas Instruments, etc [16, 19].

However, the commercial deployment of C-RAN also faces a lot of challenges. A main drawback of C-RAN is the high fronthaul optical transportation network cost [22, 23]. The current widely used interface protocol for IQ data transmission between RRHs and BBUs is Common Public Radio Interface (CPRI) [24]. The estimated IQ data throughput exceeds 10 Gbps for a 3 sector BS with 20 MHz 4×4 MIMO configuration [25, 26]. A BBU pool which is expected to connect $10 - 1000$ RRHs [5] will need vast transmission bandwidth in the fronthaul. The construction cost of optical network is high, e.g. deploying 1 m of optical fiber in urban environments costs up to 100 dollars [22]. The operators who have free/cheap fiber network resources will be more interested in C-RAN.

From a system point of view, a good C-RAN architecture should provide high bit rate with limited cost, in other words, reduced load on the fronthaul. The objective of this thesis is to maximize the user bit rate while taking into account the fronthaul constraints. Moreover, in C-RAN, it is complex to realize user scheduling, resource allocation, advanced coordinated algorithms with centralized BBUs resources and limited fronthaul resources [27]. We also try to solve this problem in this thesis.

1.2 Contributions

This thesis is divided into two parts. In the first part, we propose new functional splits schemes which move part of functions in BBU to RRH in order to reduce fronthaul throughput. As the functional splits are more complex in uplink than downlink, we focus on C-RAN uplink in this study. In the second part, we study the application of advanced coordinated algorithms in C-RAN to improve network capacity. We go deeper into the study of users scheduling and fronthaul resource allocation. A trade-off between load on the fronthaul and throughput on the radio channel is aimed to be optimized. The major contributions of this thesis are summarized as follows.

1.2.1 Performance Analysis of Several Functional Splits in C-RAN Uplink

Various solutions have been proposed to reduce fronthaul throughput. Fronthaul compression techniques, such as applying non-linear quantization [5], distributed source coding [28], compress sensing based compression[29], spatial filtering [30], are the first steps to reduce fronthaul throughput. Another envisaged solution is to change the functional split between RRHs and BBUs [31]. But few papers have done quantitative analyses on the impact of different functional splits.

We propose two new architectures of RRH-BBU functional splits for the C-RAN uplink. In the proposed architectures, the transmission rate between RRHs and BBUs depends on the mobile network load, while that of current architecture is constant. The performance of different functional splits are analyzed quantitatively.

This part of work has been published in VTC 2016 Spring [DLG16b].

1.2.2 Coordinated Transmission Design in C-RAN Downlink with Limited Fronthaul Capacity

C-RAN facilitates the implementation of CoMP and cooperative resource allocation algorithms. These have the potential to largely improve system throughput. However, their high complexity and the corresponding heavy fronthaul load may counteract all the benefits. In this thesis, we study different transmission strategies and resource allocation in a C-RAN downlink system with RRH power constraints and fronthaul capacity constraints. We firstly compare different transmission strategies without and with cooperation between different cells. Then, to improve the performance of distributed MIMO with Zero-forcing, we propose several low complexity user grouping algorithms. In distributed MIMO, several RRHs serve

a number of UEs in parallel. At last, a hybrid transmission strategy is proposed, in which part of UEs are served by only one RRH and the others served by several RRHs.

Comparison of Different Transmission Strategies

At first, we study the maximum throughput of different transmission strategies in a C-RAN cluster with transmission power constraints and fronthaul capacity constraints. Both transmission strategies without cooperation and with cooperation between different cells are considered.

This part of work has been presented in WPMC 2016 [DLG16a].

Improving Zero-Forcing Performance by User Grouping Algorithms

We apply zero-forcing (ZF) for multi-user joint transmission (distributed MIMO) in a C-RAN downlink system. A number of User Equipments (UEs) located in different cells but in the same cluster are assigned to be served into different subframes. The performance of ZF depends on the channel matrix. By appropriately choosing which UEs are served together in the same time frequency resource, the total transmission rate can be improved. With UE data and channel state information shared in BBU pool, C-RAN facilitates the centralized user scheduling. We propose several low complexity user grouping algorithms to maximize the average achievable sum rate. We firstly study the scenario with unlimited fronthaul capacity. This part of work has been published in ICC 2017 [DLG17].

Then we extend the work to the scenario with limited fronthaul capacity and with multiple antennas on the UE and on the RRH.

Hybrid Transmission

CoMP can improve the spectral efficiency but requires also much higher fronthaul capacity than without coordination. In this paper, we propose a hybrid transmission that divides the User Equipments (UEs) into two parts: some UEs are served in single RRH mode (limited coordinated scheduling) whereas the others are served in distributed MIMO mode (CoMP joint processing). The division is based on the UEs' statistical Channel State Informations (CSIs). We also propose a new fronthaul transmission scheme to let the UEs served in distributed MIMO mode exploit the fronthaul capacity not used for UEs served in single RRH mode. This largely improves system performance when the fronthaul capacity is low.

1.3 Outline of The Thesis

The rest of the thesis is organized as follows. We give a survey about RAN architecture evolution, functional splits between BBU and RRH, and CoMP in Chapter 2. The Performances of several functional splits in C-RAN uplink are analyzed in Chapter 3. Then we focus on the studies of transmission strategies and resource allocation in C-RAN downlink. The C-RAN downlink transmission model with RRH transmission power constraints and fronthaul capacity constraints is introduced in Chapter 4. Next, Chapter 5 compares the maximum throughput of different transmission strategies without and with cooperation between different cells in a C-RAN cluster. Chapter 6 presents and analyses several proposed user grouping algorithms which aim to improve the performance of Zero-Forcing. Our proposed hybrid transmission strategy is discussed in Chapter 7. This thesis in concluded in Chapter 8 with summarized results and contributions. In this chapter, some possible future study directions are also presented.

Chapter 2

Evolution towards C-RAN

The RAN deals with all radio-related functionality of the overall mobile network. In this thesis, we focus on the physical layer. Here, we firstly do a brief presentation of the physical layer in LTE. Then the evolution of RAN architecture is introduced, from $1/2$G RAN to C-RAN. C-RAN facilitates the implementation of CoMP. Another main part of our work is to analyze the performance of CoMP in C-RAN downlink with limited fronthaul capacity. In this chapter, we also do a brief survey on different CoMP algorithms.

2.1 LTE Physical Layer

The physical layer in LTE is responsible for coding, modulation, multi-antenna processing, physical time-frequency resources mapping, etc. In 3GPP LTE, Orthogonal Frequency-Division Multiplexing (OFDM) is used as the transmission scheme. In this section, we firstly present OFDM. Then the duplex schemes are introduced. At last, a brief introduction to the physical layer transmission chain is done.

2.1.1 OFDM

OFDM divides frequency-selective wideband channel into overlapping but orthogonal narrow-band subcarriers. As shown in Figure 2.1, a data stream is first separated into N sub-streams and then modulated to N parallel narrowband subcarriers. Each subcarrier has approximately $1/N$ bandwidth of original wideband channel. The transmission on each subcarrier is narrowband. Thus, each subcarrier is non-frequency-selective. OFDM does not need a complex time-domain equalization to adapt to severe time varying channel conditions, such as multi-path channel in an urban area.

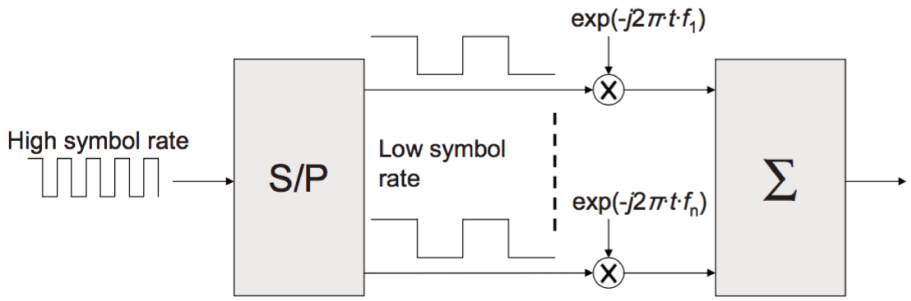

Fig. 2.1 Principle of transmission in OFDM. Source: [2]

In practice, computationally efficient Fast Fourier Transmission is used to implement OFDM with low-complexity. A block chain of corresponding OFDM modulation in LTE downlink is shown in Figure 2.2. N_c users' modulated symbols $a_0,\ a_1,\ldots\ a_{N_c-1}$ are blocked together for transmission. They are extended to size-N with zeros and set as the parallel inputs to a size-N Inverse Discrete Fourier Transform (IDFT). Then the output symbols $x_0,\ x_1,\ldots\ x_{N-1}$ are converted into a stream of data elements transmitted in time sequence, $x(t)$, followed by adding cyclic-prefix (CP) and digital to analog (D/A) conversion. At last, $x(t)$ is shifted to a radio frequency f_{rf}, then be transmitted through the antenna. If N equals to 2^m for some integer m, the size-N IDFT can be realized by radix-2 Inverse Fast Fourier Transform (IFFT) processing.

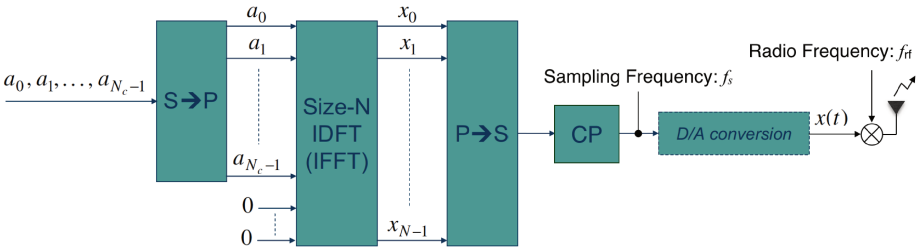

Fig. 2.2 OFDM modulation by means of IFFT processing. Source: [3]

The N_c symbols $a_0,\ a_1,\ldots\ a_{N_c-1}$ are mapped to N_c subcarriers which are called active subcarriers. Meanwhile, the $N-N_c$ zeros are mapped to the subcarriers which are called guard subcarriers or null subcarriers. The guard/null subcarriers are not used to transmit data.

CP insertion is used in both LTE downlink and uplink for OFDM transmission. This is to make OFDM insensitive to a time-dispersive channel. We assume the transmission time for $x_0,\ x_1,\ldots\ x_{N-1}$ in Figure 2.2 has a duration of T_u ($T_u = 1/\Delta f$). As illustrated in Figure 2.3, the last N_{cp} symbols of $x_0,\ x_1,\ldots\ x_{N-1}$ are copied and inserted to the front of the N symbols. This extends the transmission time to $T_u + T_{\text{cp}}$. In a multi-path environment, a UE

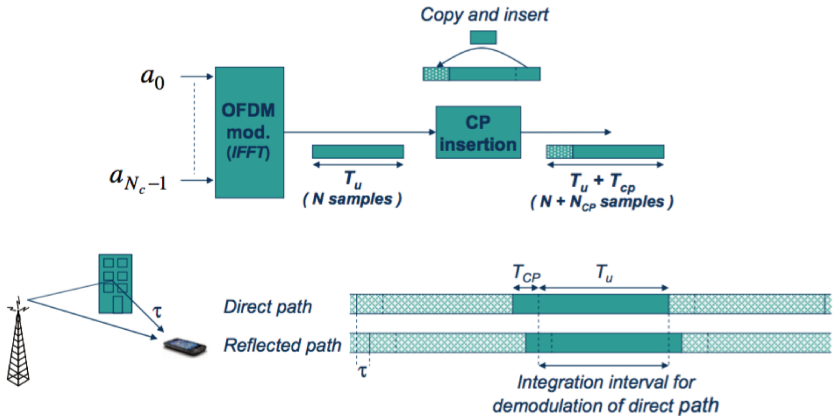

Fig. 2.3 Cyclic-prefix insertion. Source: [3]

may receive signals from a BS through different paths. Figure 2.3 depicts a scenario where a UE receives signals both from a direct path and a reflected path with a time difference of τ. Only if $\tau < T_{\text{cp}}$, after removing CP in the received side, the sum of the signals received from different paths do not lost the orthogonality between different subcarriers.

OFDM has high Peak-to-Average Ratio (PAPR). A high PAPR may lead to excessive cost, size and power consumption of the UE Power Amplifier (PA). The power efficiency is of paramount importance in UEs' side. To avoid high PAPR, 3GPP selects DFT-spread OFDM (DFTS OFDM) for LTE uplink. Nevertheless a high PAPR is not a critical drawback on BSs' side. A multi-user version of OFDM, Orthogonal Frequency-Division Multiplexing Access (OFDMA) is selected for LTE downlink. OFDMA and DFTS OFDM will be presented individually in the following.

OFDMA

OFDMA is based on OFDM. In OFDM, it is assumed that one UE receives data on all the subcarriers when the UE is served. OFDMA distributes the active subcarriers to different UEs at the same time. In LTE downlink, Time Division Multiple Access (TDMA) is also used in combination with OFDMA. The subcarriers and time resources are grouped into blocks. Each block corresponds to a group of subcarriers during a specific duration of time (1 ms in LTE). Each UE is allocated with one or several these resource blocks for serving. Figure 2.4 shows such a strategy as used in LTE.

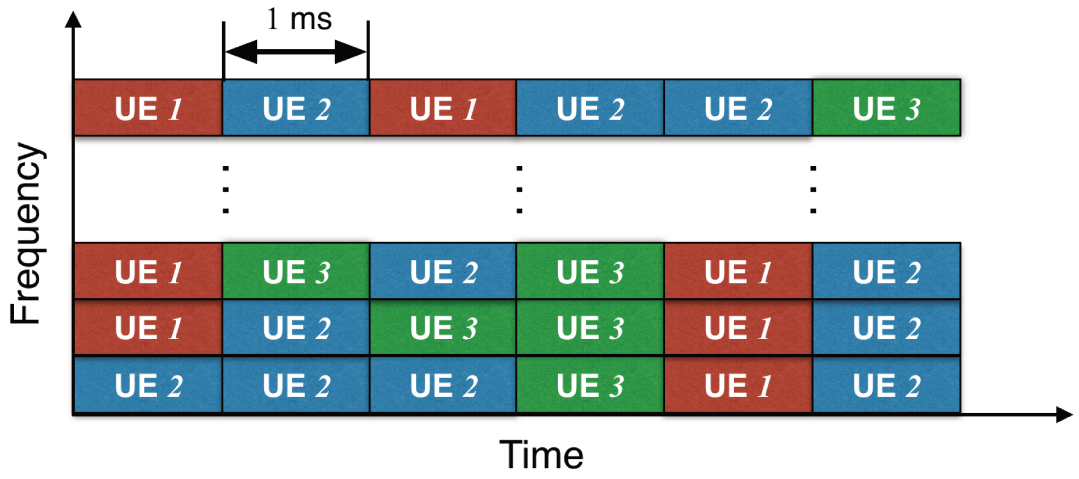

Fig. 2.4 Example of resource allocation in LTE.

DFT-Spread OFDM

DFTS-OFDM can be seen as a normal OFDM with a DFT-based precoding. A simple DFTS-OFDM signal generation block chain is illustrated in Figure 2.5. Different from OFDMA, the size-M symbols $a_0, a_1, \ldots a_{M-1}$ intended to a UE are firstly preformed with a size-M DFT, and then be set as the inputs of the size-N IDFT to be mapped to different subcarriers. M is smaller than N and equals to 2^m for some integer m. The rest of the unused inputs to the IDFT block are set to 0.

The DFT operation gives the output $x(t)$ "single-carrier" properties, which refers to low power variations, low PAPR [32]. The bandwidth of $x(t)$ depends on M. Similar to OFDMA, the time-frequency resources are also divided into blocks and can be allocated to multiple UEs. A more detailed DFTS-OFDM transmission scheme will be presented in Chapter 3.

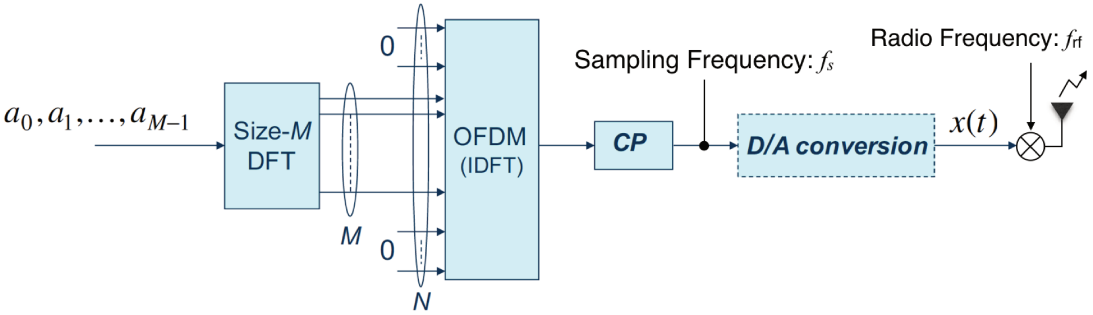

Fig. 2.5 DFTS-OFDM signal generation. Source: [3]

2.1.2 LTE Numerology

In LTE, the subcarrier spacing is $\Delta f = 15$ kHz.The number of subcarriers N ranges from 128 to 2048, depending on channel bandwidth.

Figure 2.6 depicts a basic time-frequency resource structure of LTE with 72 subcarriers. The period of one time slot is 0.5 ms, corresponding to 7 OFDMA/DFTS-OFDM symbol intervals for normal CP duration and 6 for extended CP duration. Two time slots make a subframe and 10 subframes form one radio frame. A Resource Element (RE) is the smallest defined unit. It consists of one subcarrier during one OFDMA/DFTS-OFDM symbol interval.

The Physical Resource Block (PRB) is the smallest chunk of data transmitted in LTE data transmission. Each PRB is composed of 12 subcarriers along one time slot, which results a bandwidth of 15 kHz $\times 12 = 180$ kHz for each PRB. When a normal CP is used, $12 \times 7 = 84$ REs form a PRB. When an extended CP is applied,the number of REs in a PRB is $12 \times 6 = 72$. The number of active subcarriers which are used to transmit data, $N_{\mathrm{c}} = 12 \cdot N_{\mathrm{RB}}$, where N_{RB} is the number of PRBs. During one time slot, one RB is distributed to one single User Equipment (UE) while one UE can be allocated with several PRBs.

Table 2.1 Subcarriers configuration in LTE with different bandwidths.

Channel bandwidth (MHz)	Number of distributable PRBs	Number of distributable active subcarriers	Total number of subcarriers	Number of guard subcarriers	minimum percentage of null subcarriers	Sampling frequency f_{s} (MHz)
1.4	6	72	128	56	43.75%	1.92
3.5	12	144	256	112	43.75%	3.84
5	25	300	512	212	41.41%	7.68
10	50	600	1024	424	41.41%	15.36
15	75	900	1536	636	41.41%	23.04
20	100	1200	2048	848	41.41%	30.72

We use N_{total} to denote the total number of OFDM symbols (including CP) during one subframe. Without oversampling, the sampling frequency for the OFDM symbols is $f_{\mathrm{s}} = N_{\mathrm{total}}/(1\mathrm{ms})$. For example, $N_{\mathrm{total}} = 30720$ for a LTE configuration with 20 MHz bandwidth. Then we get $f_{\mathrm{s}} = 30.72$ MHz.

A summary of the subcarriers configuration in LTE with different bandwidths is shown in Table 2.1. The corresponding sampling frequencies are also noted. The value of the minimum percentage of null subcarriers is obtained in the case when all the active subcarriers are occupied.

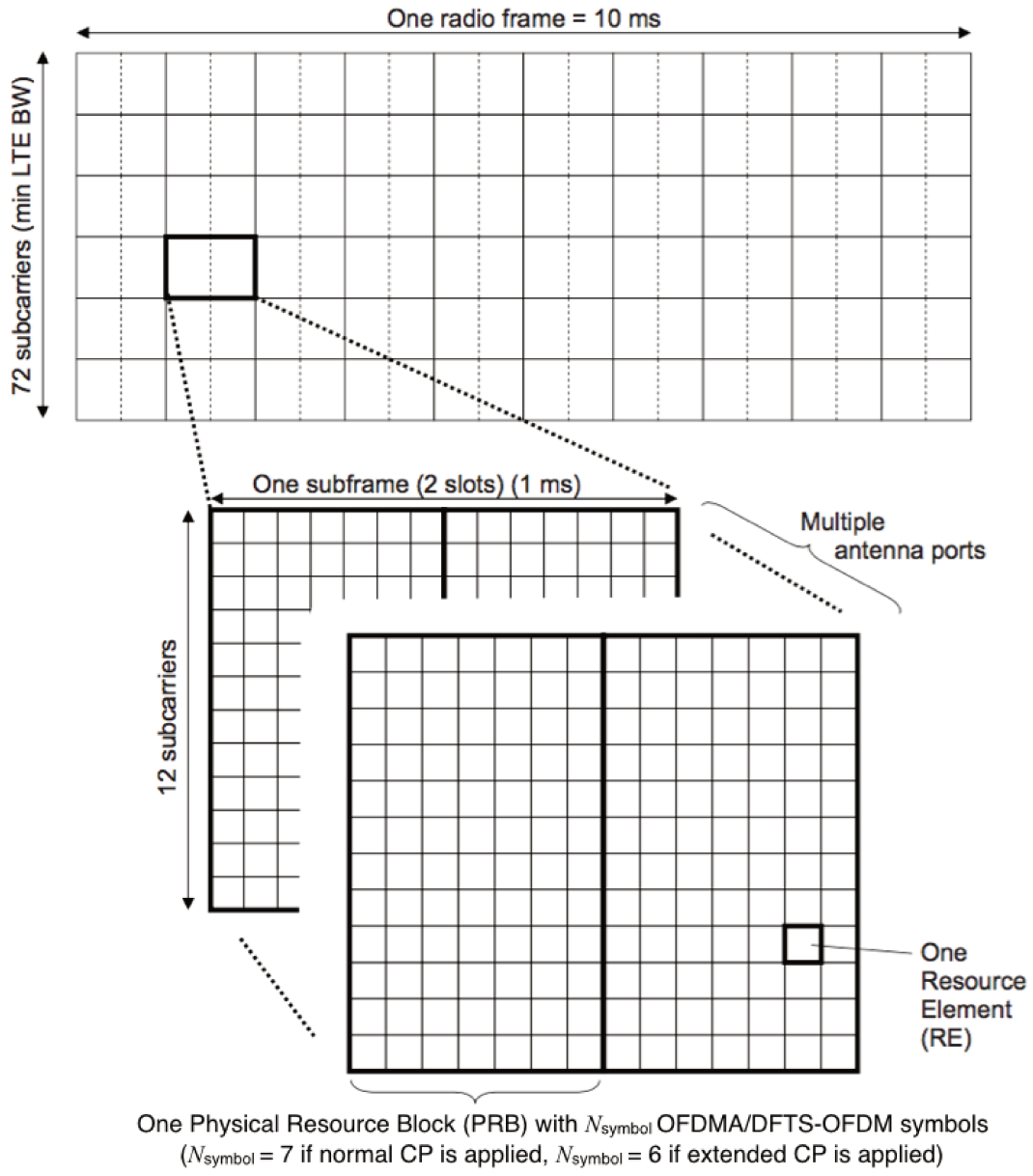

Fig. 2.6 Basic time-frequency resource structure of LTE. Source: [2]

2.1.3 Duplex Schemes

LTE supports both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) [33]. As shown in Figure 2.7, downlink and uplink transmission are divided in frequency domain for FDD mode while in time domain for TDD mode. In FDD, two different carrier frequencies are used, one for uplink transmission (f_{UL}) and the other one for downlink transmission (f_{DL}). In TDD, uplink and downlink transmission take place in the same frequency band, but in non-overlapping time slots.

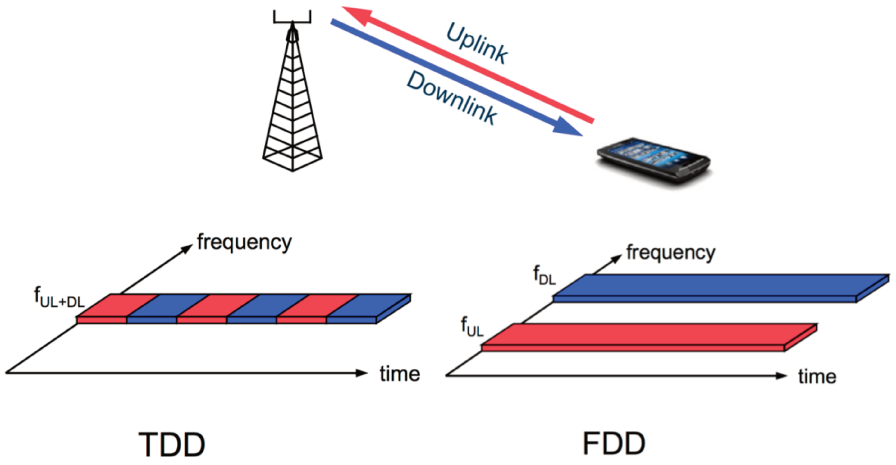

Fig. 2.7 Frequency- and time-division duplex. Source: [3]

2.1.4 Physical layer transmission chain

LTE physical layer translates data into reliable signals for the transmission between eNodeBs and UEs in the radio interface. To be more immune to transmission errors, the data from a higher layer are usually encoded with Cyclic Redundancy Check (CRC) code at first and then channel coded (e.g. Turbo Code). Further processing after channel coding in LTE downlink is illustrated in Figure 2.8. After being scrambled, the coded data are modulated with different modulation schemes which could be QPSK, 16 QAM, 64 QAM etc. Different modulation schemes can be applied for different subcarriers. Next, the modulated data are mapped to different layers and then precoded to be transmitted in different antennas. The "RE mapper" block maps the data to different PRBs. At last, the "OFDM signal generation" blocks convert data to OFDM signals for transmission. In uplink, the physical layer has similar basic structure. The "OFDM signal generation" blocks have to be replaced by "DFTS-OFDM signal generation".

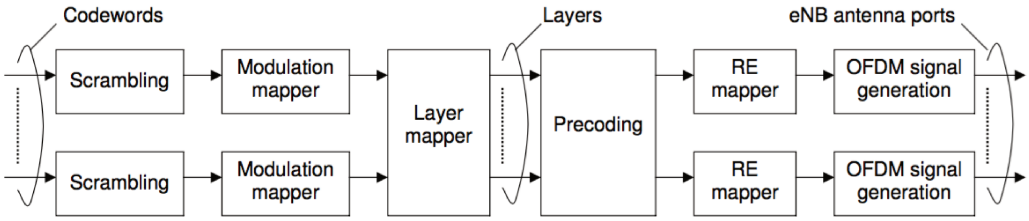

Fig. 2.8 General signal structure for LTE downlink. Source: [2]

2.2 Evolution of RAN Architecture

A mobile network covers an area which is divided into cells. Users located in a cell are served by the corresponding Base Station (BS). The changes of the integration and separation of the functions reflect the evolution of BS architecture. In current deployed LTE RAN architecture, both radio functionalities and baseband processing are all implemented in remote sites, which are called eNodeBs (eNBs). The path towards C-RAN usually are divided into two steps. In the first step, the baseband processing resources of different different remote sites are centralized. In the second step, the centralized baseband processing resources are virtualized and cloudified. This evolution will be introduced in the following.

2.2.1 Functions in a Base Station

In general, the main functions of a BS can be divided into 3 parts: radio functionalities, baseband processing and upper layers.

- The radio functionalities include radio-frequency (RF) application, power amplification, etc. These mainly serve the signal frequency shifting between radio frequency and baseband.
- The baseband processing corresponds to the functions in physical layer: coding, modulation, physical time-frequency resources mapping, etc (see Section 2.1).
- In LTE, the upper layers refer to Medium Access Control (MAC), Radio Link Control (RLC), Packet Data Convergence Control (PDCP) and Radio Resource Control (RRC).

2.2.2 Old Architecture of Base Station

In the old BS architecture, all the functions are integrated into one technical shelter and constitute the BS. The shelter is usually located at the bottom of the tower (see Figure 2.9). The antennas are installed at the top of the tower, and connect to the BS by coaxial cables.

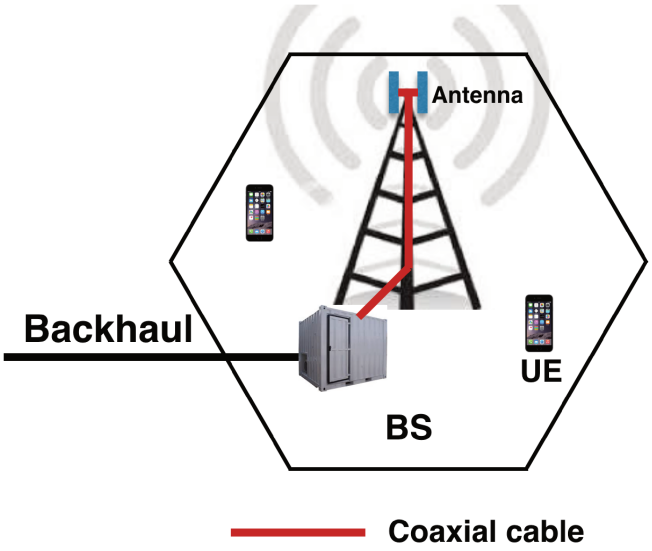

Fig. 2.9 Old architecture of BS.

Due to the high losses of the coaxial cables, there is a constraint on the distance between the BS and the antennas. This architecture is implemented in 1G, 2G and part of 3G mobile network.

2.2.3 RRH-BBU Concept

When the 3G mobile network was being deployed, a new architecture of BS was introduced. In the new architecture, the radio functionalities are integrated in Remote Radio Heads (RRHs). Meanwhile, the baseband processing and upper layers are implemented in Baseband Unit (BBU). One RRH connects to one BBU by optical fiber (mainly), cable, or microwave. This RRH-BBU concept BS architecture is applied in the 3G and 4G mobile network.

An example of the RRH-BBU concept BS architecture is illustrated in Figure 2.10. There are three RRHs serving three different cells corresponding to a 3-sector cellular network. The RRHs are installed near the antennas and connected with coaxial cables. The three RRHs are connected to BBU by Optical fiber (mainly). Usually, there are three BBUs installed together, and each serves one RRH.

Figure 2.11 illustrates a simple 4G LTE RAN architecture. The BBUs are connected to the core network through S1 links. The transport network that connects the core network and remote sites are defined as backhaul. The BBUs of different cells can communicate to each other by the X2 interface.

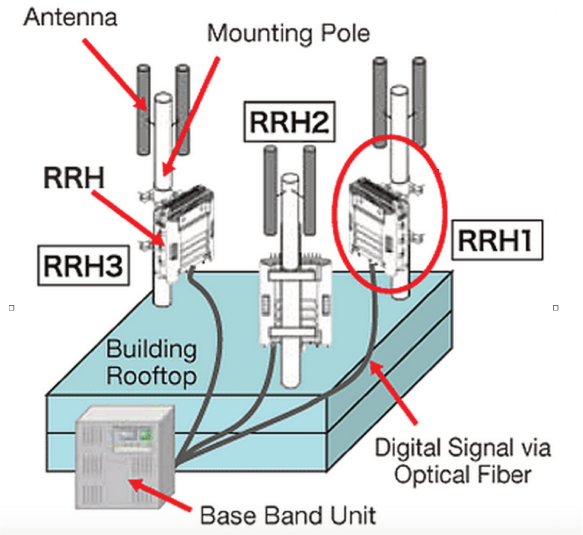

Fig. 2.10 Typical configuration of eNB. Source: [4]

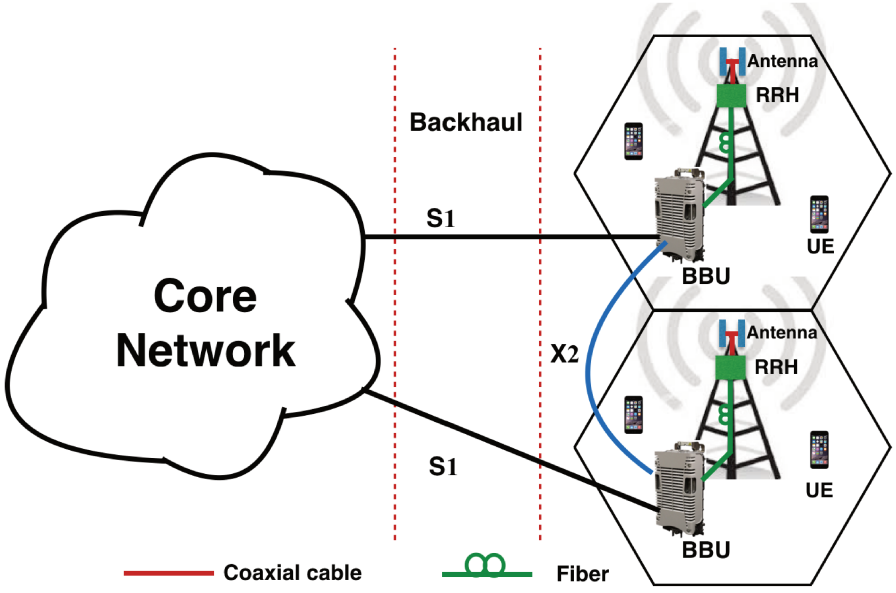

Fig. 2.11 LTE radio access network architecture.

Compared with the old BS architecture, the distance between the RF application module and the antennas is reduced in the RRH-BBU concept. With a shorter coaxial cable, the link budget is improved.

2.2.4 Fronthaul and Required Bit Rate

The connections between between RRHs and BBUs are called fronthaul. The current widely used interface protocol for the data transmission in fronthaul is Common Public Radio Interface (CPRI) [24]. As a constant-bit-rate bidirectional protocol, CPRI requires accurate synchronization and strict latency control. Open Base Station Architecture Initiative (OBSAI) [34] was developed to be a main competing protocol of CPRI. However, it is not widely used. Thus, we take the CPRI case and do the calculation of the required bit rates in the fronthaul.

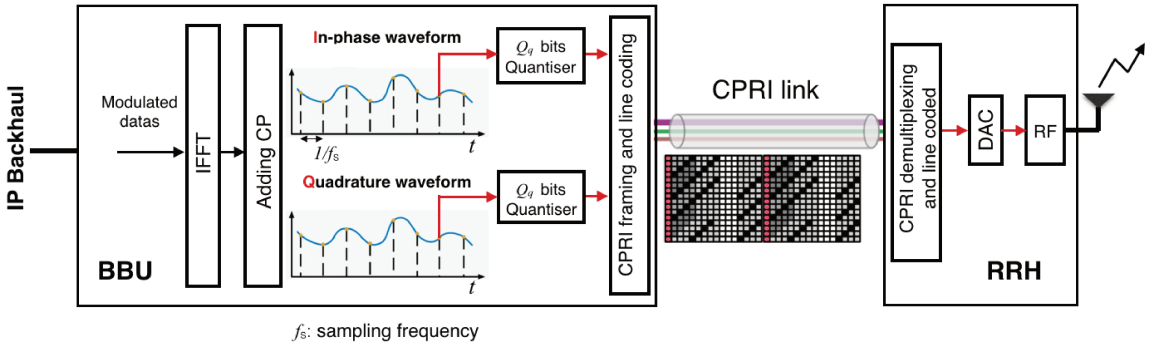

Fig. 2.12 The CPRI interface protocol for IQ data transmission from BBU to RRH in LTE downlink.

The CPRI interface brings enormous overhead to the transmitted data. An example of data transmission from BBU to RRH in LTE downlink applying CPRI is shown in Figure 2.12. The outputs of the block "IFFT" and "adding CP" correspond to OFDM symbols which are complex and analog. The in-phase (I) components and quadrature (Q) components of the analog symbols are sampled with a sampling frequency f_s, separately. The number of quantization bits for each sample is denoted by Q_q. Thus the data transmission rate from one BBU to one RRH is given as

$$D_\mathrm{DL} = 2 \times f_\mathrm{s} \times Q_\mathrm{q} \times N_\mathrm{TA} \tag{2.1}$$

where N_TA is the number of antennas equipped by the RRH and the factor 2 results from that both I and Q signals are quantized and transmitted.

In CPRI, the digitized data are added with extra overheads before being transmitted to RRH:

- one word is added to every 15 data words for control and management [35].
- the digitized data are expanded by $10/8$ when 8B/10B Forward Error Correction (FEC) code is applied. In the CPRI specification, 64B/66B FEC is also an option.

Table 2.2 Subcarriers configuration in LTE with different bandwidths.

LTE downlink configuration	5 MHz bandwidth 2×2 MIMO	10 MHz bandwidth 2×2 MIMO	20 MHz bandwidth 2×2 MIMO	20 MHz bandwidth 4×4 MIMO
Fronthaul rate (Gbits/s)	0.65	1.3	2.6	5.2

Taking into account the CPRI overheads, the data transmission rate from one BBU to one RRH in (2.1) is changed to

$$D_{\mathrm{DL}} = 2 \times f_{\mathrm{s}} \times Q_{\mathrm{q}} \times \frac{16}{15} \times \frac{10}{8} \times N_{\mathrm{TA}}. \tag{2.2}$$

We set $Q_q = 16$. Derived from (2.2) and the last column of Table 2.1, the transmission rates from one BBU to one RRH in different LTE downlink configurations are illustrated in Table 2.2. We can observe that the fronthaul bit rate requirement is huge. The downlink maximum user transmission rate in the radio interface for an LTE configuration with 20 MHz bandwith and 4×4 MIMO is 300 Mbits/s. However, the corresponding required fronthaul bit rate is 5.2 Gbits/s. An LTE network does not always work with full capacity. Assuming a 100 Mbits/s average transmission rate, the required fronthaul bit rate is more than 50 times of the data transmission rate in the radio interface.

2.2.5 Centralized RAN Architecture

The first step towards C-RAN is to move BBUs away from RRHs and centralize them in a BBU pool. This makes a centralized RAN as shown in Figure 2.13. A general Centralized-RAN architecture consists of three main parts: 1) a BBU pool with centralized BBUs, 2) RRHs and antennas located at the remote sites, 3) a transport network which provides a connection between BBU pool and RRHs.

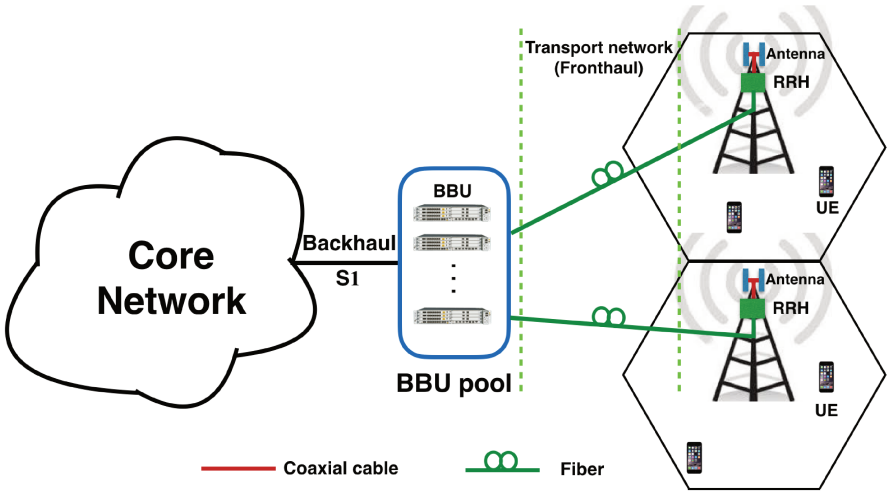

Fig. 2.13 Centralized RAN.

Benefits of Centralized RAN

The centralization of BBUs largely improves the capacity of information exchange between different cells and the delay is also much reduced. This facilitates the implementation of CoMP and other collaborative resource allocation and transmission algorithms.

BBU pool is equipped with a set of "soft" BBU nodes. A BBU is associated to a dedicated RRH or a set of RRHs depending on real-time conditions. All RRHs may connect to a switcher/central device which can flexibly assign the BBU resources to each RRH. This brings advantages in energy saving as explained below.

The density of users in a certain area varies during different periods of a day. People works in an industrial area and goes back to a residential area in the evening. As shown in Figure 2.14, the traffic density is lower during daytime in a residential area, but higher in an industrial area. This is called the tidal effect of users. The current RAN architecture works in an inefficient way in such a nonuniform data traffic scenario. A BBU's capacity is dimensioned to satisfy the peak load time of a cell. Lots of BBU resources are wasted during off-peak load time. Thanks to the centralization of BBUs to BBU pool in C-RAN, the overall utilization rate of the equipments can be improved. The BBU resources can be allocated on demand [36] . It is supposed that less BBU resources are needed in BBU pool than the sum of BBU resources in current BSs. Thus, a statistical multiplexing gain can be achieved. In [37], they have proposed semi-static and adaptive BBU-RRH switching schemes for C-RAN. From the analysis in the Tokyo metropolitan area, the needed number of BBUs is only 25% of that using traditional RAN architecture.

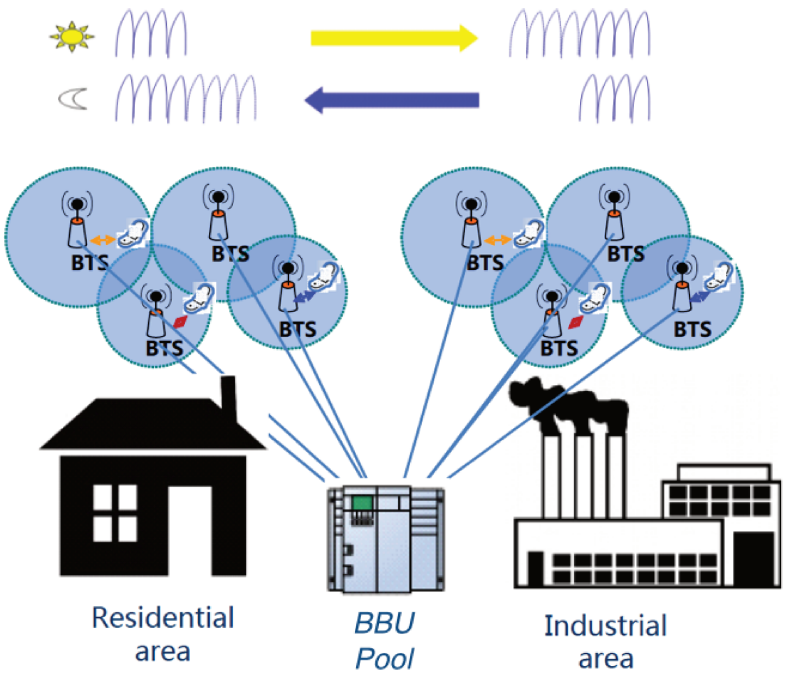

Fig. 2.14 Tidal effect in residential and industrial area. Source: [5].

Fronthaul Compression

The Transport networks connects the RRHs to the BBUs pool. We also call it the fronthaul links in C-RAN. Different technologies, like dark fibre, fibre transport networks or microwave transmission, can be applied depending on the scenarios.

As shown in Table 2.2, with a LTE 20 MHz bandwidth and 4×4 MIMO configuration, the fronthaul link applying CPRI beween one BBU and one single RRH requires 5.2 Gbits/s transmission rate. However, a BBU pool is supposed to connect tens to hundreds RRHs. The fronthaul rate would be enormous. Furthermore, in future 5G network, higher data rates to UEs are predicted, which results in envisioned heavier fronthaul burden. The high bandwidth needed in the fronthaul transport network is a bottleneck of C-RAN. It is important to reduce the fronthaul flow in order to realize C-RAN. The raw IQ samples transmitted between RRHs and BBUs consume too much bandwidth of the transport network. Thus, an efficient compression scheme is needed for the transmission over capacity-constrained transport network between RRHs and BBUs. Several solutions have been proposed in [5], such as reducing signal sampling rate, applying non-linear quantization, frequency subcarrier compression and IQ data compression. A chosen compression scheme can mix several aforementioned techniques. It is a trade-off between achievable compression ratio, power consumption, design complexity, introduced latency and signal distortion [16].

Reducing signal sampling rate improves compression up to 66% with some performance degradation [5]. Mature non-linear quantization algorithm are available to bring up to 50% compression efficiency. Digital Automatic Gain Control (DAGC) can be used for *IQ* data compression by reducing the transmitted signal's dynamic range.

In [38], a CPRI compression algorithm for LTE system is introduced. It's realised by eliminating redundant spectrum bandwidth, performing block scaling and applying non-linear quantizer. The impact of the algorithm to the original data with different compression ratio is analysed. The error vector magnitude (EVM) is less than 1%, when the compression ratio is 2:1, while it has a significant increase in the case of high compression ratio. Similar compression scheme is used in [39]. A drop of 83% transmission data rate is achieved while keeping EVM below 8% (3GPP requirement for 64 QAM).

An alternative method to reduce fronthaul data flow is to move part of functions in BBU to RRH. It is studied in Chapter 3.

2.2.6 Cloud RAN Architecture

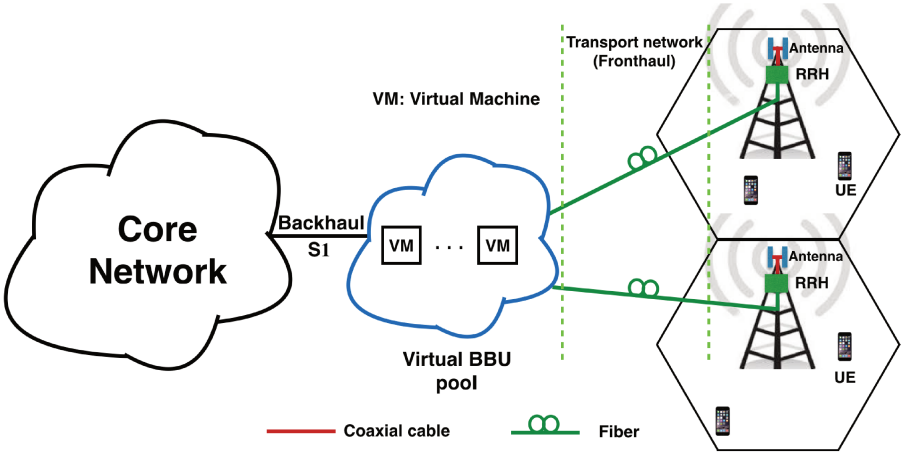

Fig. 2.15 Cloud RAN.

In a high level of view, Cloud-RAN can be seen as an instance of network function virtualization (NFV) techniques [17]. All the baseband resources are virtualized and cloudified in BBU pool. Figure 2.15 depicts an architecture of Cloud-RAN. The BBU pool is virtualized with a number of virtual machines (VMs). When needed, one VM can be built and act as a BBU instance to serve one or a couple of RRHs. The resource cloudification facilitates the dynamic allocation of BBU resources. A virtual BBU pool further eases scalability, cost reduction and integration of different services. The virtualization and cloudification are out of the scope of this thesis. The relative details can be found in [14, 16, 17].

2.3 Cells Coordination

In mobile networks, transmissions on the same time-frequency resource interfere with each other. Modern mobile-broadband system Long Term Evolution (LTE) has opted for a reuse factor 1 to maximize the data rates for users close to the Base Station (BS). This increases the chance of facing low signal-to-interference ratios (SIR), especially in the cell edge area. The overall system efficiency and UE fairness can be improved if the interference from adjacent cells can be avoided. This can be done by coordination between different cells. Initial activities related to coordination between different cells are done by introducing Inter-cell Interference Coordination (ICIC) in LTE release 8. CoMP is a more advanced way of dealing with Inter-Cell Interference (ICI). The 3GPP activities on LTE release 10 firstly discussed CoMP and the main related features are introduced in LTE release 11 [40]. Both of ICIC and CoMP will be introduced in the following.

2.3.1 Inter-cell Interference Coordination

In Inter-cell Interference Coordination (ICIC), a UE can report back to its serving eNodeB (eNB) if it suffers strong interference on certain sub-carriers. Then, the eNB communicates with neighboring eNBs through X2 interface. The neighboring eNBs would try not to schedule their serving cell-edge UEs on these sub-carriers. ICIC carries out a reconfiguration on a time-scale of the order of seconds or longer. This is to make it slow enough permitting eNBs exchange signaling through X2 interface.

2.3.2 Different Downlink CoMP Technologies

The basic idea of Coordinated Multi-point (CoMP) is to jointly avoid ICI or turn ICI into a useful signal through more tight and dynamic coordination between different cells.

In this thesis, we focus on the downlink CoMP aspects in a macro cellular network. One RRH covers a macro cell. We define a number of cooperative RRHs as a cluster of RRHs. The cluster of RRHs jointly serve the UEs located in the corresponding macro cell applying with CoMP algotithms. Here, we do a brief introduction to different downlink CoMP types and different transmission schemes from BBU pool to RRHs in a C-RAN applying CoMP.

The downlink CoMP strategies can be generally divided into three different types: dynamic point selection (DPS), coordinated scheduling/beamforming (CB/CS) and joint transmission (JT) [41, 42]. The three types will be separately presented in the following.

Dynamic Point Selection

Dynamic point selection (DPS) is the first category of downlink CoMP techniques [43]. It can be seen as an extension of ICIC. In DPS, a UE has the opportunity to reselect its serving RRH out of a cluster of cooperative RRHs. The UE reports the index of its preferred RRH to BBU pool. Then, the RRH which has the largest measured SINR will be selected for serving. In the meantime, the corresponding CSIs may also be reported back to RRH or BBU pool for possible MIMO transmission. ICIC does reconfiguration every several seconds or longer. However, the serving RRH in DPS may dynamically change from one subframe to another. The BBU pool mutes the other RRHs in the cluster for the time frequency resource that this UE uses. Different from ICIC, the data of the UE have to be available in each RRH of the cluster.

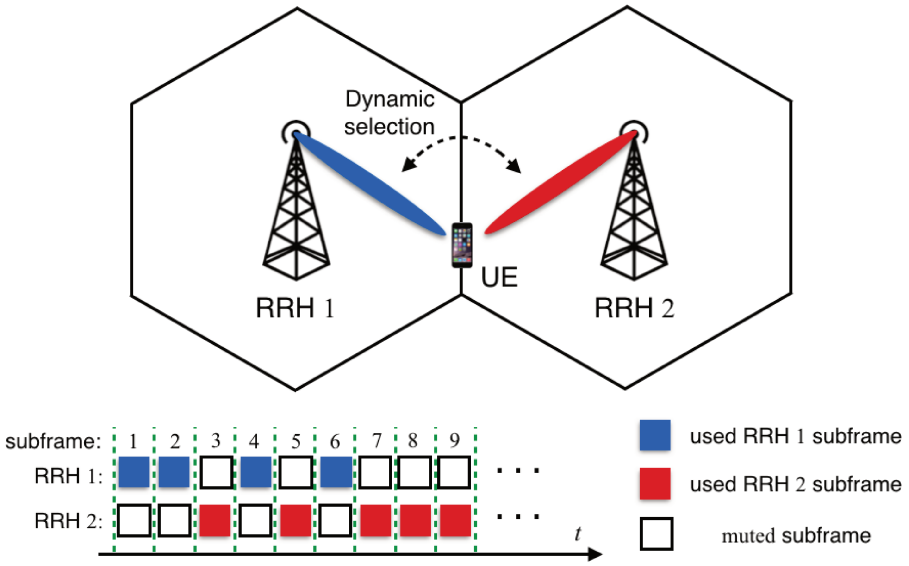

Fig. 2.16 A simple example of dynamic point selection with two cooperative RRHs.

A simple example of DPS is shown in Figure 2.16, in which two RRHs transmit signals alternatively to one UE. We assume that the UE gets higher SINR from RRH 1 than from RRH 2 during subframe 1, 2, 4 and 6. The UE is served by RRH 1 and RRH 2 is muted on the corresponding frequency resources during these subframes, and vice versa during subframe 3, 5, 7, 8 and 9.

Coordinated Scheduling/Coordinated Beamforming

In Coordinated Scheduling/Coordinated Beamforming (CB/CS), each RRH only serves the UEs in its own cell. The CSIs from all the RRHs in the cluster to all their served UEs are

available in the BBU pool. The BBU pool decides the UEs scheduling [44] and beamforming and power allocation [45–47] to minimize the overall interference to the UEs served by a cluster of RRHs [48]. Based on the CSIs of all the UEs served by the cluster of RRHs, the BBU pool can select out the set of UEs which construct the best channel condition [49]. This set of UEs are served at the same time to improve system performance [50]. The transmission beams of different RRHs can also be generated in a coordinated way [51–53]. The above mentioned UEs scheduling and beams selection manages to reduce the overall interference. In general, CS requires less synchronization between different RRHs than CB.

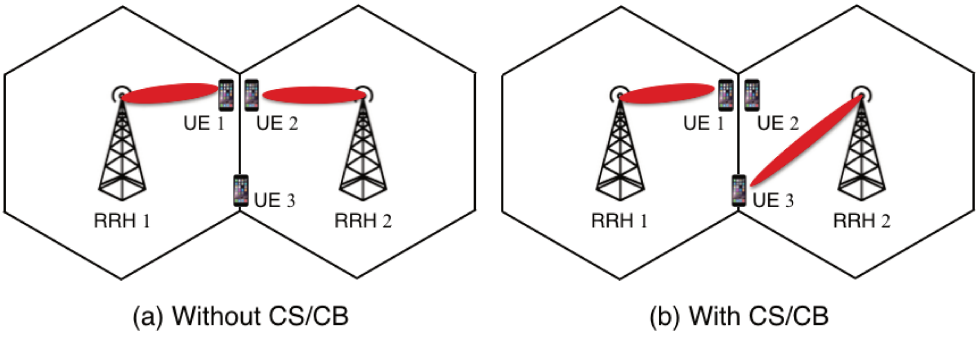

Fig. 2.17 A simple example of cooperative scheduling/cooperative beamforming with two cooperative RRHs.

A simple example of CS/CB is illustrated in Figure 2.17, in which two RRHs transmit signals to two UEs separately. As shown in Figure 2.17 (a), there is beam collision if UE 1 is served by RRH 1 and UE 2 is served by RRH 2 in the same time-frequency resource. Instead of muting one RRH as in ICIC, we can let RRH 2 do a beamforming towards UE 3 as shown in Figure 2.17 (b). In this way, both UE 1 and 3 get small interference from neighboring cell. UE 2 can be served in other time-frequency resources.

Joint Transmission

In CB/CS and DPS, each UE is served by only one RRH. However, in Joint Transmission (JT), each UE is served by all the RRHs in a cluster. The data to be transmitted is available to each RRH in the cluster. In this scenario, several RRHs jointly transmit to one UE or several UEs in the same time frequency resource. We define the case where one UE is served as single user Joint Transmission (SU-JT) and the case where several UEs are served as distributed MIMO (D-MIMO). In D-MIMO, MU-MIMO techniques are applied. In the following, we firstly discuss JT then D-MIMO.

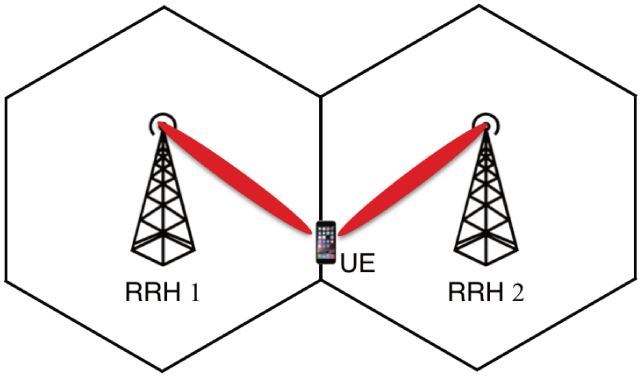

Fig. 2.18 A simple example of Joint Transmission with two cooperative RRHs.

A simple example of SU-JT is illustrated in Figure 2.18, in which two RRHs transmit signals to the same UE. Different from CB/CS and DPS, the interference from neighboring cells are not simply reduced or eliminated, but turned into useful signal to the UE. However, SU-JT takes up the resources of several RRHs to serve only one UE. This is not suitable for a high loaded network.

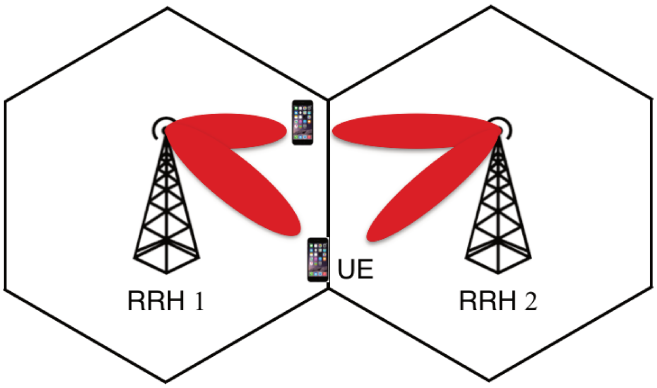

Fig. 2.19 A simple example of distributed MIMO with two cooperative RRHs.

To further increase system capacity, we can expand SU-JT to D-MIMO which combines SU-JT and MU-MIMO. A simple example of D-MIMO is illustrated in Figure 2.19, in which two RRHs transmit signals to two UEs in parallel.

Another categorization distinguishes JT into non-coherent JT and coherent JT, depending on whether the transmission is non-coherent or coherent [54]. Downlink coherent transmission are based on CSIs feedback to BBU pool, while CSIs are not used in downlink non-coherent transmission. Non-coherent JT can use open-loop MIMO techniques like cyclic delay diversity (CDD) or Space Time/Frequency block coded (STBC/SFBC) schemes to exploit diversity gain and also increase total transmit power to the served UEs [55–57]. On

the other hand, with shared CSIs, coherent JT can apply closed-loop MIMO techniques, such as Zero-Forcing (ZF), minimum mean square error (MMSE), dirty paper coding (DPC), etc [58]. This provides coherent JT theoretically a high gain over non-coherent JT as compared in [59]. However, coherent JT requires tighter RRHs synchronization than non-coherent JT. The needing for full CSIs available in BBU pool also makes coherent JT more complex than non-coherent JT.

Compared with CB/CS and DPS, coherent JT needs a tighter synchronization between different RRHs. Nevertheless, coherent JT can explore more degree of diversity than CB/CS and DPS to improve system throughput.

Different CoMP downlink algorithms for a cluster of cooperative RRHs are summarized in Table 2.3.

Table 2.3 A summary of CoMP downlink algorithms.

Downlink CoMP algorithms	dynamic point selection	coordinated scheduling	coordinated beamforming	non-coherent joint transmission	coherent joint transmission
Acronym	DPS	CS	CB	Non-coherent JT	Coherent JT
User data	all data available in each RRH	a RRH only has the data for its serving UEs	a RRH only has the data for its serving UEs	all data available in each RRH	all data available in each RRH
CSIs available in BBU pool	yes	yes	yes	no	yes
Sychonization requirements between different RRHs	0.05 ppm frequency and 3 μs timing accuracy [16]	0.05 ppm frequency and 3 μs timing accuracy [16]; 5 μs timing accuracy [60]	0.05 ppm frequency and 3 μs timing accuracy [16]; 1.5 μs timing accuracy [60]	5 μs timing accuracy [60]	0.02 ppm frequency and 0.5 μs timing accuracy [16]

2.3.3 Transmission Schemes from BBU Pool to RRHs

In recent studies, in downlink, two main strategies of forwarding signals from BBU pool to RRHs to realize CoMP are proposed: data-sharing strategy and compression-based strategy [61]. The two transmission schemes and their comparison are introduced in the following.

Data-sharing

In data-sharing strategy, the channel state informations (CSIs) from all RRHs in a cluster to each served UEs are available in each RRH in the cluster. The user data flow of a given UE is transmitted to the corresponding RRH by the BBU pool. The beamformed signals are locally generated in each RRH.

For CB/CS, an RRH only needs the user flow data of its served UEs. By contrast, all the user flow data of the UEs served by the cluster of RRHs need to be available at each RRH. Thus, JT consumes much more fronthaul capacity than CB/CS in data-sharing strategy.

Compression based

We will introduce both Compression after precoding and compression before precoding. But only compression after precoding is applied in this thesis.

In compression-based strategy, each RRH collects only the CSIs from itself to all the served UEs and forwards them to the BBU pool. The joint processing is done in the BBU pool, e.g. the precoding matrices are calculated in the BBU pool [62]. The compression-based strategy can be further divided to two sub-strategies: compression-after-precoding (CAP) and compression-before-precoding (CBP) [62]. In CAP (studied in [62]), the user flow data are firstly precoded in BBU pool then quantized before transmitted to the corresponding RRHs. In CBP (studied in [62–64]), the user flow data and quantized precoding matrices are transmitted to the corresponding RRHs and then the user flow data are precoded in RRHs. A RRH only need the user flow data of its served UEs in CB/CS, while the user flow data of all the UEs served by the cluster of RRHs in JT. Furthermore, BBU pool transmits a smaller precoding matrix to each RRH in CB/CS than in JT. Therefore, JT consumes much more fronthaul capacity than CB/CS in CBP, however not significant in CAP.

A simple example of CAP and CBP where 3 UEs are jointly served by 3 RRHs applying D-MIMO is shown in Figure 2.20. For CAP, BBU pool need to transmit quantized corresponding precoded signals to different RRHs. On the contrary, for CBP as illustrated in Figure 2.21, all discrete user data of UE 1, 2 and 3 together with corresponding quantized precoding matrix will be transmitted from BBU pool to different RRHs.

CBP consumes less fronthaul capacity than CAP in CB, however not in JT. Furthermore, when the size of a cluster is large, CBP consume more fronthaul capacity than CAP in JT. In this thesis, we focus on JT instead of CB. Thus, CAP is chosen as the transmission scheme from the BBU pool to RRHs for CoMP in this thesis.

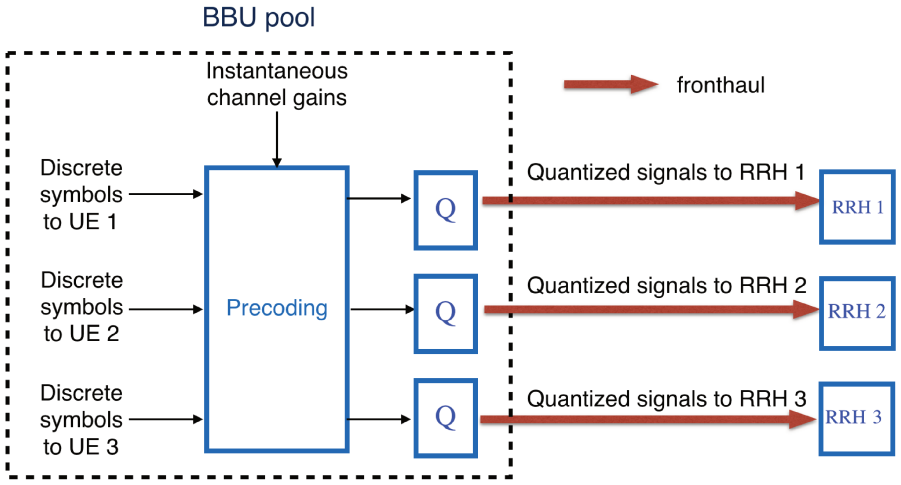

Fig. 2.20 Downlink compression after precoding (CAP) transmission scheme ("Q" represents compression).

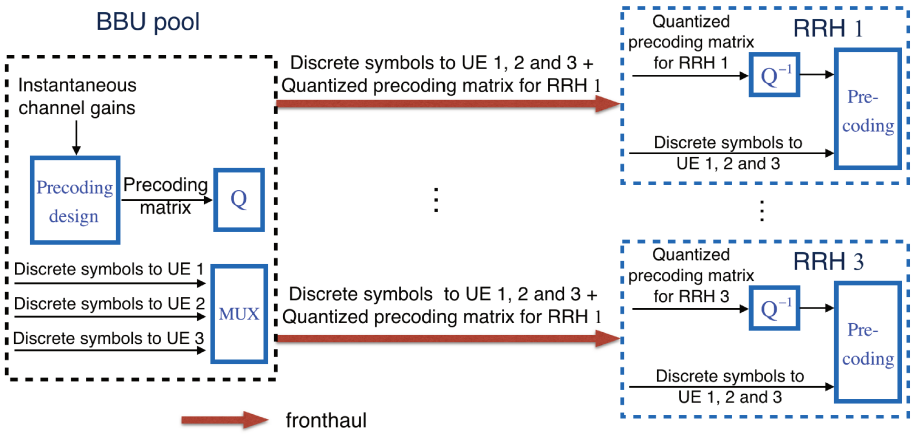

Fig. 2.21 Downlink compression before precoding (CBP) transmission scheme ("Q" represents compression and "Q^{-1}" represents decompression).

Comparison of the Two Transmission Schemes

Despite the introduction of extra quantization noise, compression-based strategy is more practical than data-sharing strategy in C-RAN. At first, the RRH should be as light as possible in C-RAN to reduce cost. The joint processing is too complex to be done in RRHs. Secondly, the CSIs from all RRHs in a cluster to their served UEs should be available in each RRH. This costs too much resources.

For a traditional transmission strategy without applying CoMP (non-CoMP), each RRH only serves the UEs in its cell and no complex cooperation exists between different RRHs. An RRH only needs to collect the CSIs from itself to its serving UEs and the precoding can also be done locally in the RRH by applying simple linear precoding algorithm. In this way, BBU pool only needs to transmit the user flow data of the UEs served by this RRH to it.

Chapter 3

Performance Analysis of Several Functional Splits in C-RAN

3.1 Introduction

In a C-RAN architecture where all baseband processing functions in traditional BBU are put into BBU pool, the burden on fronthaul is high. Several solutions have been proposed in [5], such as reducing signal sampling rate, applying non-linear quantization, frequency subcarrier compression and IQ data compression. Another envisioned method to reduce the data throughput between RRHs and BBUs is to change the current functional split architecture between RRH and BBU [31]. Part of baseband processing functions in BBU are moved to RRH. This will largely reduce the overheads of the signal transmitted through fronthaul.

This chapter investigates on the performance of different functional splits with a mathematical model. Two new functional split architectures are proposed, which move part of the functions in the physical layer of LTE transmission system from BBU to RRH.

For the downlink, the base station is concerned by the transmission side. The baseband signal is defined in a finite set of discrete symbols. Developing compression scheme is easy as illustrated in Figure 2.21 of Section 2.3.3. For the uplink, the base station is concerned by the receiver side. As the signal received is affected by noise, it is fundamentally analog. The analog received signal should be quantized before being transmitted from the RRH to the BBU. The quantization has an impact both on the quality of the reception and the throughput between the RRH and the BBU. Therefore, we focus on LTE uplink in this chapter. This work has been published in [DLG16b].

This chapter is organized as follows. In Section 3.2, we firstly give a state of the art about functional splits between RRHs and BBUs. In Section 3.3, the Discrete Fourier Transform

Spread Orthogonal Frequency-Division Multiplexing (DFTS-OFDM) transmission system in LTE uplink is briefly introduced. The current and proposed functional split architectures between RRH and BBU are modeled and presented in Section 3.4. The algorithms applied and numerical configurations in the simulation for numerical estimation of the distortion are proposed in Section 3.5. The simulation results are shown in Section 3.6. At last, this chapter is concluded in Section 3.7.

3.2 State of the Art

To reduce the throughput on the fronthaul, some baseband functionalities at the physical layer can be moved from the BBU pool to the RRHs (functional splits in the uppers layers are beyond the scope of this thesis). The fronthaul transmission rates, advantages and disadvantages have been discussed in [5, 19, 22, 31, 65]. With more functionalities in the RRHs, it requires more storage and larger calculation capacity available in the remote sites. The multiplexing gain in energy consumption will be reduced. In the same time, this also lowers the coordination level between different cells.

The authors of [22, 31] studied the implementation of the FFT/IFFT blocks in the RRHs. They found that this can bring 40% reduction of the fronthaul bit rate when combined with the omission of guard carriers and of the cyclic prefix (CP). These studies are based on very simple models that consider a given bit rate budget for each layer.

In [66], a further reduction of fronthaul bit rate is obtained by moving resource demapping for the uplink and resource mapping for the downlink from the BBU pool to the RRHs. The fronthaul bit rate becomes user-traffic dependent. Only the occupied resource blocks need to be transmitted on the fronthaul. This allows the system to exploit the statistical multiplexing gain based on the real time traffic. Furthermore, different physical channels can be distinguished. The authors of [67] applied higher quantization resolution to channels carrying higher-order modulation and lower quantization resolution to channels carrying lower-order modulation. In [68], the authors gave quantitative analysis of multiplexing gains of traffic-dependent C-RAN functional splits. They conclude that the implementation of such functional splits is much beneficial when the traffic is variable bit rate, has low load or is bursty. The requirements on the fronthaul link can be largely relaxed.

All the above studies did not consider the impact of the splits on the overall signal quality. Our objective is to analyse the fronthaul bit rate of different functional splits while maintaining the same quality expressed with the error vector magnitude (EVM).

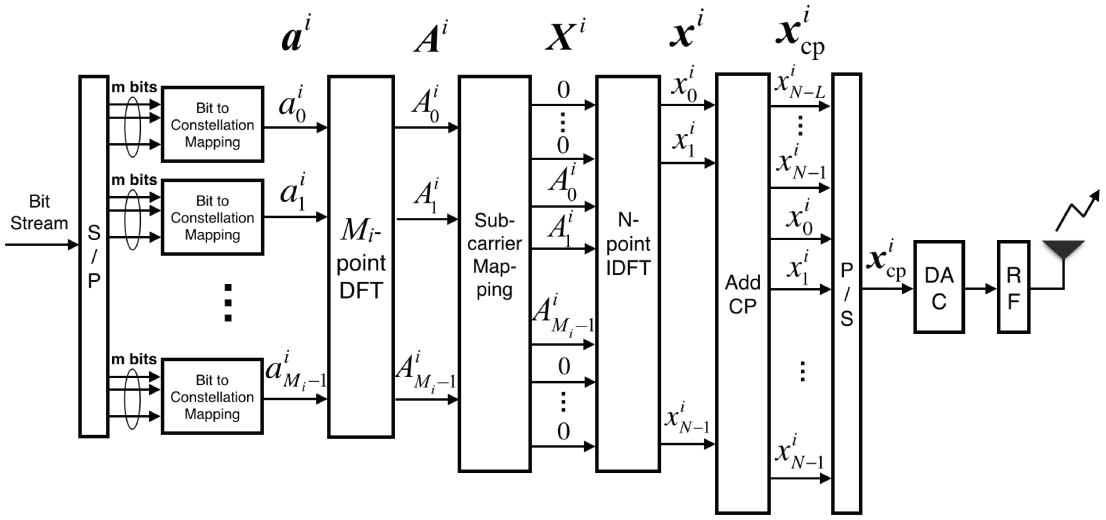

Fig. 3.1 Localized DFTS-OFDM transmitter.

3.3 Overview of DFTS OFDM Transmission System

LTE uplink transmission is based on DFTS-OFDM rather than Orthogonal Frequency-Division Multiplexing Access (OFDMA), which is used for LTE downlink transmission. OFDMA is less power efficient than DFTS-OFDM due to its high Peak to Average Power Ratio (PAPR) [69]. The lower PAPR feature of DFTS-OFDM greatly benefits the User Equipment (UE) in terms of transmit power efficiency and reduced cost of the power amplifier. Time Division Duplex (TDD) and Frequency Division Duplex (FDD) are both supported in LTE with flexible bandwidth i.e. 1.4, 3, 5, 10, 15, 20 MHz. Modulation schemes QPSK, 16 QAM, 64 QAM etc can be adapted.

3.3.1 Transmitter

Let us consider a transmitter i of a DFTS-OFDM system. It converts input bit stream signal to a set of modulated sub-carriers. It uses M_i sub-carriers over a total of N sub-carriers. The related signal processing operations are shown in Figure 3.1. A set of input data bits is firstly collected and then modulated by using QPSK, 16 QAM or 64 QAM, which results in a complex vector $\mathbf{a}^i = \{a_0^i, a_1^i, ..., a_{M_i-1}^i\}$ with dimension M_i. After an M_i-point unitary Discrete Fourier Transform (DFT), the output complex row vector $\mathbf{A}^i = \{A_0^i, A_1^i, ..., A_{M_i-1}^i\}$ is obtained.

Then $\mathbf{A}^i$ is mapped to a group of M_i adjacent sub-carriers of $\mathbf{X}^i$. The remaining sub-carriers of $\mathbf{X}^i$ are allocated with zeros. The aforementioned mapping is called localized

sub-carrier mapping. Next, a DFTS-OFDM symbol $\mathbf{x}^i = \{x_0^i, x_1^i, ..., x_{N-1}^i\}$ is generated by performing an N-point unitary inverse DFT to $\mathbf{X}^i$.

To combat multi-path fading, a Cyclic Prefix (CP) is added at the beginning of each symbol $\mathbf{x}^i$ (see Section 2.1 in Chapter 2). A CP is created by copying the last L samples of symbol $\mathbf{x}^i$:

$$\mathbf{x}_{\text{cp}}^i = \{x_{N-L}^i, ..., x_{N-1}^i, x_0^i, x_1^i, ..., x_{N-1}^i\}. \tag{3.1}$$

After a parallel-to-serial operation, $\mathbf{x}_{\text{cp}}^i$ is converted to analog signal and modulated to radio frequency, then finally transmitted.

The average symbol energy is defined as

$$E_s = \mathbb{E}(|a_k^i|^2), \tag{3.2}$$

where $\mathbb{E}$ denotes expectation.

For a given transmission (i.e. a transmission of a random symbol sequence $\{a_0^i, a_1^i, ..., a_{M-1}^i\}$), the energy of $\mathbf{a}^i$, $\mathbf{A}^i$ and $\mathbf{x}^i$ is then

$$\begin{cases} E(\mathbf{a}^i) = \mathbb{E}\left(\mathbf{a}^i(\mathbf{a}^i)^{\text{H}}\right) = \sum_{n=0}^{M-1} \mathbb{E}(|a_n^i|^2) \\ E(\mathbf{A}^i) = \mathbb{E}\left(\mathbf{A}^i(\mathbf{A}^i)^{\text{H}}\right) = \sum_{n=0}^{M-1} \mathbb{E}(|A_n^i|^2) \\ E(\mathbf{x}^i) = \mathbb{E}\left(\mathbf{x}^i(\mathbf{x}^i)^{\text{H}}\right) = \sum_{n=0}^{N-1} \mathbb{E}(|x_n^i|^2). \end{cases} \tag{3.3}$$

As we apply unitary DFT and unitary IDFT, $E(\mathbf{a}^i) = E(\mathbf{A}^i) = E(\mathbf{x}^i) = ME_s$.

3.3.2 White gaussian noise channel model

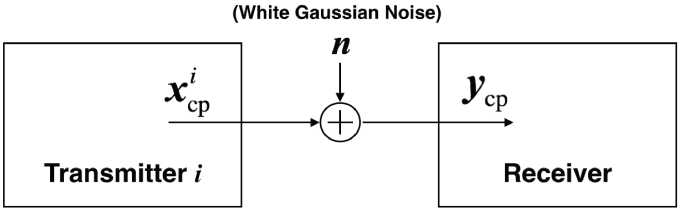

Fig. 3.2 Model of adding white gaussian noise in the transmission channel between transmitter i and receiver.

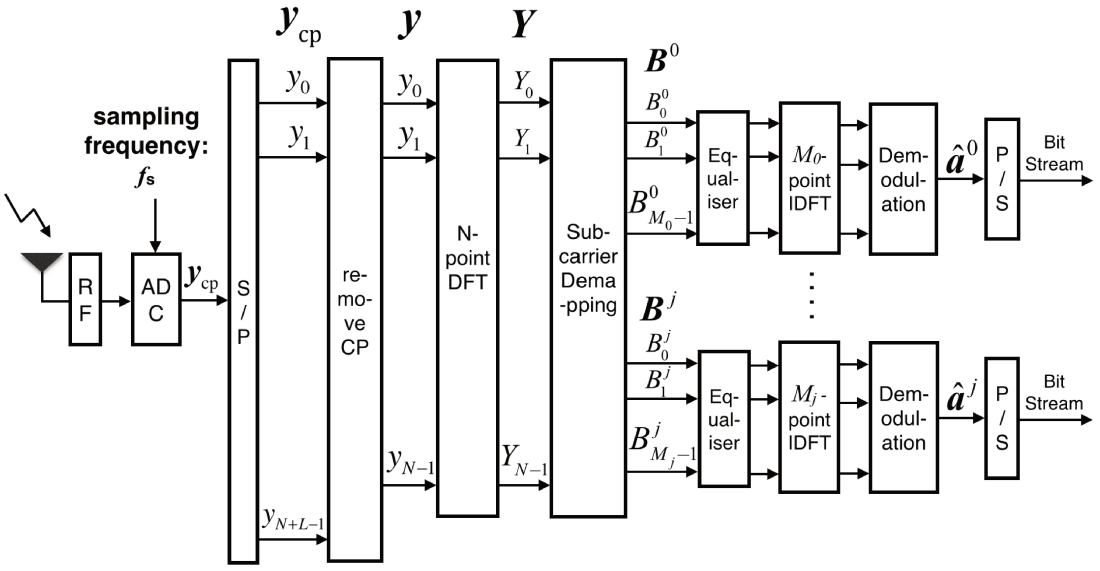

Fig. 3.3 Localized DFTS-OFDM receiver.

We model the transmission channel with noise between transmitter i and receiver by adding white gaussian noise as shown in Figure 3.2. The white gaussian noise is given by

$$\mathbf{n} = \frac{\sigma}{\sqrt{2}}(\mathbf{n}_{\mathrm{I}} + j \times \mathbf{n}_{\mathrm{Q}}), \tag{3.4}$$

where $j^2 = -1$, both $\mathbf{n}_{\mathrm{I}}$ and $\mathbf{n}_{\mathrm{Q}}$ follow a standard normal distribution and σ is the standard deviation of the adding white gaussian noise.

The variance of the adding white gaussian noise is

$$\sigma^2 = E_s \times 10^{-(\frac{E_s}{N_0})_{dB}/10}, \tag{3.5}$$

where E_s is the defined average symbol energy mentioned in Subsection 3.3.1, and N_0 is the double-sided power spectral density of the white gaussian noise.

The received symbols $\mathbf{y}_{\mathrm{cp}} = \mathbf{x}^i_{\mathrm{cp}} + \mathbf{n}$.

3.3.3 Receiver

Figure 3.3 includes a block diagram of a DFTS-OFDM receiver for the recovery of signals transmitted from multiple transmitters. The signal processing process is inverse to that of the aforementioned transmitter. The sub-carriers distributed to Transmitter j are selected out by the Sub-carrier Demapping block. The signal corresponding to Transmitter j is denoted by $\mathbf{B}^j$. A frequency-domain-equalization (FDE) can be performed after the Sub-carrier

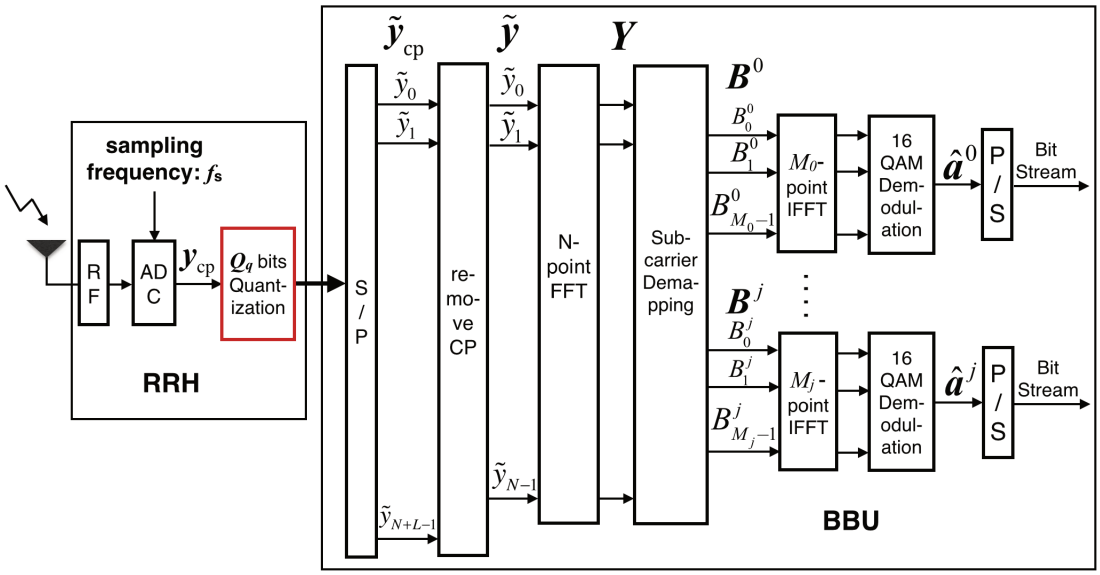

Fig. 3.4 Method 1 of functional split between RRH and BBU (used in CPRI).

Demapping to combat frequency-selective fading and phase distortion. We do not consider equalization in this work.

3.4 Different Functional Splits

3.4.1 Current Functional Split Between RRH and BBU

A model of receiver applying typical functional split is shown in Figure 3.4, which we denote as Method 1. A Q_q-bits linear quantization is applied to both the real part and the imaginary part of $\mathbf{y}_{cp}$ separately before it is transmitted from RRH to BBU. The RRH in Method 1 cannot identify which and how many Physical Resource Blocks (PRBs) are occupied, thus it has to transmit all the PRBs even if all of them are vacant. Therefore, the transmission rate between RRHs and BBUs is constant, independent of the mobile network load.

3.4.2 Proposed Functional Splits Between RRH and BBU

In LTE transmission, the number of guard sub-carriers is about 40 percent of all the sub-carriers transmitted and the cyclic prefix (CP) also forms 7 to 25 percents of the data transmitted [31]. We consider to remove CP and the guard sub-carriers before the transmission between RRH and BBU. Two methods of functional split which move “remove CP” and

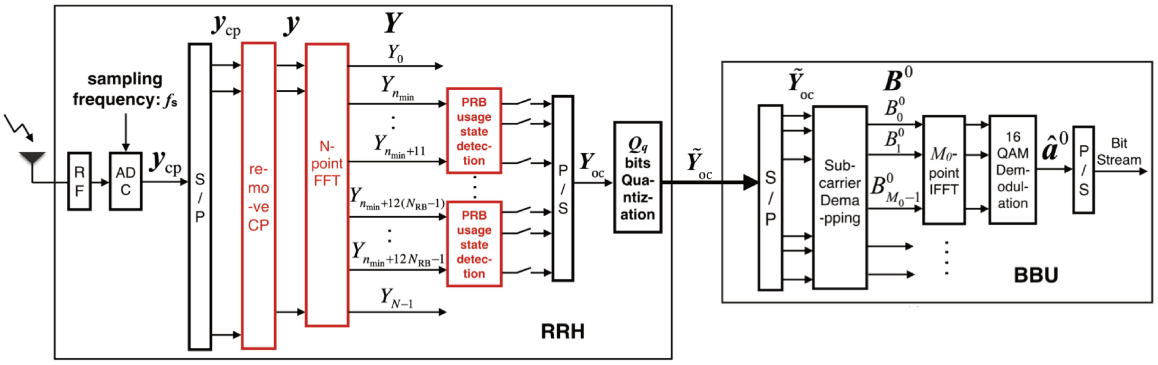

Fig. 3.5 Method 2 of functional split between RRHs and BBUs.

"FFT" blocks from BBU to RRH are proposed. We denote them Method 2 and 3, and detail them hereafter.

Method 2 of Functional Split Between RRH and BBU

The model of Method 2 is shown in Figure 3.5. The difference between Method 1 mentioned in Section 3.4.1 and Method 2 is that we move the "remove CP" block and "N-point-FFT" block from BBU to RRH. The new block "PRB usage state detection" detects which sets of 12 sub-carriers corresponding to one PRB, are occupied in $\mathbf{Y}$.

Usually, not all the PRBs are occupied in LTE transmission. If we can detect which PRBs are occupied or vacant, then we can transmit just the PRBs occupied to further reduce the data transmission rate between RRH and BBU.

For uplinks, one possible method is to calculate the sum of the modulus of 12 adjacent elements in $\mathbf{Y}$, which correspond to 12 sub-carriers in one PRB. Assume that $\{Y_{12k+n_{\min}}, Y_{12k+n_{\min}+1}, \ldots, Y_{12k+n_{\min}+11}\}$ in $\mathbf{Y}$ belong to the same PRB k, where $n_{\min}$ is the index of the first active sub-carrier. The sum of the modulus of the 12 elements is given as

$$W_k = \sum_{n=0}^{11} |Y_{12k+n}|^2 \tag{3.6}$$

If W_k is less than a threshold, we can decide that the 12 sub-carriers in PRB k are vacant.

Method 3 of Functional Split Between RRH and BBU

The model of Method 3 is shown in Figure 3.6. The dynamic of the signal $\mathbf{y}$ is higher in the frequency domain than $\mathbf{Y}$ in the time domain. In method 2, the signal transmitted from RRH to BBU is in frequency domain. We consider to transform it to time domain before its

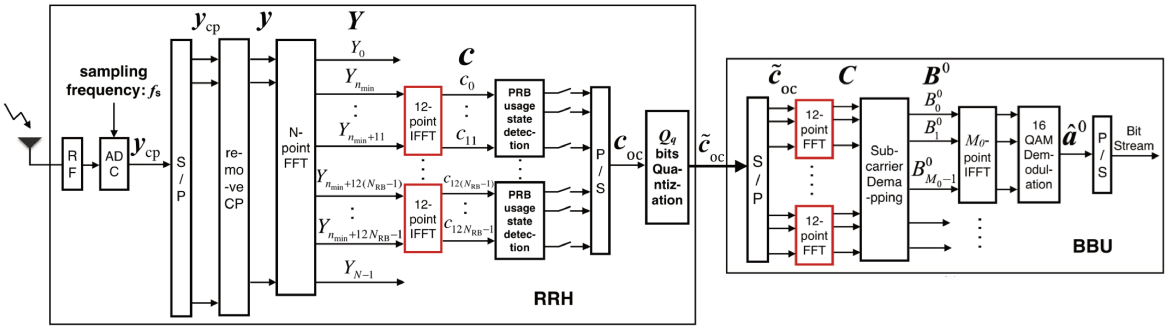

Fig. 3.6 Method 3 of functional split between RRHs and BBUs.

transmission to BBU, aiming to reduce the dynamic. To realize this, different from method 2, we do a 12-point unitary IFFT to $\{Y_{12k+n_{\min}}, Y_{12k+n_{\min}+1}, ..., Y_{12k+n_{\min}+11}\}$ in $\mathbf{Y}$ in Method 3. Then we obtain a vector $\{c_{12k}, c_{12k+1}, ..., c_{12k+11}\}$, where $k = 0, 1, ..., N_{\mathrm{RB}} - 1$. The total number of PRBs during one LTE time slot (0.5 ms) is denoted by N_{RB}. Next, the block "PRB usage state detection" decides if this vector will be quantized and transmitted to BBUs.

In BBUs, the serial "12-point FFT" blocks are used to reverse the 12-point unitary IFFT operation in the RRHs.

Method 2 and 3 can satisfy the demand for statistical multiplexing in Next Generation Fronthaul Interface (NGFI) defined in [70], by transmitting only the occupied PRBs. Here, statistical multiplexing means that transmission rate can adapt to the varying network load.

Table 3.1 Summary of different methods of functional splits between RRHs and BBUs.

Methods of functional splits	Features
Method 1	CPRI like, FFT in the BBU
Method 2	FFT in the RRH
Method 3	FFT in the RRH, additional IFFT in the RRH and additional FFT in the BBU

The different methods of functional splits between RRHs and BBUs are summarized in Table 3.1.

3.5 Algorithms and Numerical Configurations

This section will present the simulation of DFT-Spread OFDM in LTE Uplink. Three different methods of functional split between RRH and BBU are applied and the performances will be compared.

We simulate a LTE Uplink physical layer with a bandwidth of 20 MHz, which includes 2048 sub-carriers where 1200 of them are active. There are 100 PRBs within one LTE time slot. Modulation scheme 16 QAM is considered. "DAC" and "RF" blocks in the transmitter are not considered as well as "ADC" and "RF" in the receiver. In our simulation, the number of sub-carriers $N = 2048$ and the length of CP $L = 160$. As 16 QAM modulation scheme is adapted, there are $m = 4$ bits per symbol.

3.5.1 Quantization and Frame Arrangement

In order to minimize the quantization error, a digital automatic gain control is applied. A scaling factor F_{s} is determined for each block of N_s received I/Q samples. Then the N_s received I/Q samples are quantized linearly with Q_q bits resolution per complex component based on F_{s} . The scaling factor is sent together with the N_s quantized I/Q samples from RRHs to BBUs. In order not to introduce a large extra delay to the transmission system, the value of N_s should be a factor of $N+L$ for Method 1, and of total number of sub-carriers occupied during each LTE time slot for Method 2 and 3.

We define the largest absolute value as

$$V_{\max} = \max_{k=0,\ldots,N_{\mathrm{s}}-1} \{|\Re(y_{\mathrm{cp}_k})|, |\Im(y_{\mathrm{cp}_k})|\} \tag{3.7}$$

where $\Re(y_{\mathrm{cp}_k})$ and $\Im(y_{\mathrm{cp}_k})$ denotes the real and imaginary part of y_{cp_k}, respectively.

The corresponding scaling factor is determined as

$$F_{\mathrm{s}} = \begin{cases} \lceil \frac{V_{\max}}{p} \rceil & \text{for } \lceil \frac{V_{\max}}{p} \rceil \leq 2^{Q_{\mathrm{s}}} - 1 \\ 2^{Q_{\mathrm{s}}} - 1 & \text{for } \lceil \frac{V_{\max}}{p} \rceil > 2^{Q_{\mathrm{s}}} - 1 \end{cases} \tag{3.8}$$

where p is the quantization step for $V_{\max}$ and Q_{s} is the number of bits used to represent F_{s}.

Then the I/Q samples are linearly quantized to $2^{Q_{\mathrm{q}}}$ levels ranging from $-F_{\mathrm{s}}$ to F_{s}. The n-th quantization level is given by

$$q_n = -F_{\mathrm{s}} p + (n + \frac{1}{2}) \frac{F_{\mathrm{s}} p}{2^{Q_{\mathrm{q}}-1}} \tag{3.9}$$

where $n = 0, 1, ..., 2^{Q_{\mathrm{q}}} - 1$.

Applying the quantization levels obtained, the I/Q samples $\hat{\boldsymbol{s}}$ is linearly quantized as

$$u_I(k) = \arg\min_n |q_n - \Re(y_{\mathrm{cp}_k})|, \tag{3.10}$$

$$u_Q(k) = \arg\min_n |q_n - \Im(y_{\mathrm{cp}_k})| \tag{3.11}$$

where $k = 0, \ldots, N_{\mathrm{s}} - 1$.

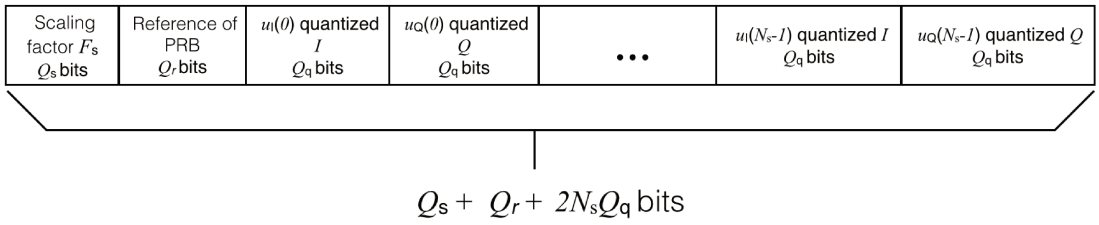

Fig. 3.7 Possible frame structure.

Inspired by [39], the symbols transmitted from RRH to BBU may be organized as a frame given in Figure 3.7. The first Q_s bits are used to represent the scaling factor F_{s}. Moreover, the references of the corresponding occupied PRBs are also represented by Q_{r} bits. This is necessary for Method 2 and 3 mentioned in Section 3.4.2 to identify which PRBs the samples received belong to. For Method 1, $Q_{\mathrm{r}} = 0$.

Taking into account the transmission of the scaling factor and the references of PRB, the average number of bits used to transport a complex element (I or Q) for Method k is:

$$Q_{\mathrm{eff}k} = \frac{2N_{\mathrm{s}_k}Q_{\mathrm{q}_k} + Q_{\mathrm{s}} + Q_{\mathrm{r}}}{2N_{\mathrm{s}_k}} \tag{3.12}$$

where $k = 1, 2, 3$, N_{s_k} is the number of I/Q samples in each frame and Q_{q_k} is the number of quantization bits per complex component.

3.5.2 Data Transmission Rate From RRH to BBU

In the following, we will present the formula of data transmission rate calculation from RRH to BBU separately.

Data Transmission Rate for Method 1

The data transmission rate between RRH and BBU for Method 1 is given by

$$D_1 = F_{\mathrm{os}} \times f_{\mathrm{s}} \times 2 \times Q_{\mathrm{eff}1} \times \frac{1}{r_{\mathrm{c}}} \tag{3.13}$$

where F_{os} is the oversampling factor, f_s is the minimum sampling frequency, and $Q_{\mathrm{eff}1}$ is the number of effective quantization bits for each complex element (I or Q) as mentioned in Section 3.5.1. The factor 2 in Equation 3.13 results from that both I and Q signals should be

transmitted. Due to the Forward Error Correction (FEC) code applied in CPRI, a code rate r_c needs to be considered.

In LTE uplink with 20 MHz bandwidth, there are totally 2048 sub-carriers including 1200 active ones used for transmission. Therefore, the minimum sampling frequency $f_s = 2048 \times \Delta f = 2048 \times 15$ kHz $= 30.72$ MHz. Oversampling is sampling a signal with a frequency higher than the Nyquist rate (minimum sampling rate). Oversampling is applied to improve the performance of system.

Table 3.2 Simulation parameters for method 1.

PAPAMETER	SYMBOL	VALUE
Oversampling factor	F_{os}	1
Minimum sampling frequency	f_s	30.72 MHz
Number of samples for each frame	N_s	138
Number of bits for scaling factor	Q_s	16
Number of bits for reference of PRB	Q_r	0
Code rate of FEC applied	r_c	$8/10$

Table 3.2 illustrates the values of the parameters for Method 1 in our simulation.

Data Transmission Rate for Method 2 **and** 3

The data transmission rate from RRH to BBU for Method 2 is denoted by D_2, and D_3 for Method 3. The data transmission rate is given by

$$D_k = N_{Sc} \times \frac{1}{T_s} \times 2 \times Q_{\text{eff}k} \times \frac{1}{r_c} \times \eta \tag{3.14}$$

where N_{Sc} is the number of active sub-carriers for data transmission, T_s is the symbol duration, $k = 2,3$, and η is the assumed PRB utilization ratio. The factor 2 in Equation 3.14 results from that both I and Q signals should be transmitted. We also apply FEC code in the transmission, thus the same code rate r_c as in Method 1 is included in Equation 3.14.

Table 3.3 shows the value assignments of the parameters in our simulation for method 2 and 3.

3.5.3 Error Vector Magnitude

The error vector magnitude (EVM) is used to quantify the performance of aforementioned different methods of receiver. EVM is defined as

Table 3.3 Simulation parameters for method 2 and 3.

PAPAMETER	SYMBOL	VALUE
Number of active subcarriers	N_{Sc}	1200
Symbol duration (for normal CP)	T_s	71.43 μs
Number of samples for each frame	$N_{\mathrm{s}_k}(k=2,3)$	12
Number of bits for scaling factor	Q_{s}	16
Number of bits for reference of PRB	Q_{r}	7
Code rate of FEC applied	r_{c}	$8/10$
Assumed PRB utilization ratio	η	0.5or1

$$\mathrm{EVM} = \sqrt{\frac{\mathbb{E}[|a_k^i - \hat{a}_k^i|^2]}{\mathbb{E}[|a_k^i|^2]}} \tag{3.15}$$

where a_k^i are the 16 QAM modulated symbols to be sent in transmitters as shown in Figure 3.1, and $\hat{a}_k^i$ are the received symbols before 16 QAM demodulation as shown in Figure 3.3.

We have also considered the cases with AWGN in the transmission channel between transmitters and receivers (Section 3.3.2).

If there is no quantization noise,

$$\mathrm{EVM} = \sqrt{\frac{\mathbb{E}[|a_k^i - \hat{a}_k^i|^2]}{\mathbb{E}[|a_k^i|^2]}} = \sqrt{\frac{N_0}{E_s}}. \tag{3.16}$$

From Equation 3.16, we can get

$$\mathrm{EVM} = 10^{-(\frac{E_s}{N_0})_{dB}/20} \tag{3.17}$$

when the number of quantization bits $Q_q \to \infty$.

3.6 Simulation Results

In our simulation, all the three methods of receiver applied (Method 1 mentioned in Section 3.4.1, Method 2 and 3 mentioned in Section 3.4.2) are tested. The performances in the three different scenarios with and without AWGN have been simulated. Using Equation 3.13 and 3.14, we can get the data transmission rate from RRH to BBU for different methods of receiver applied and quantization resolution Q_{q}.

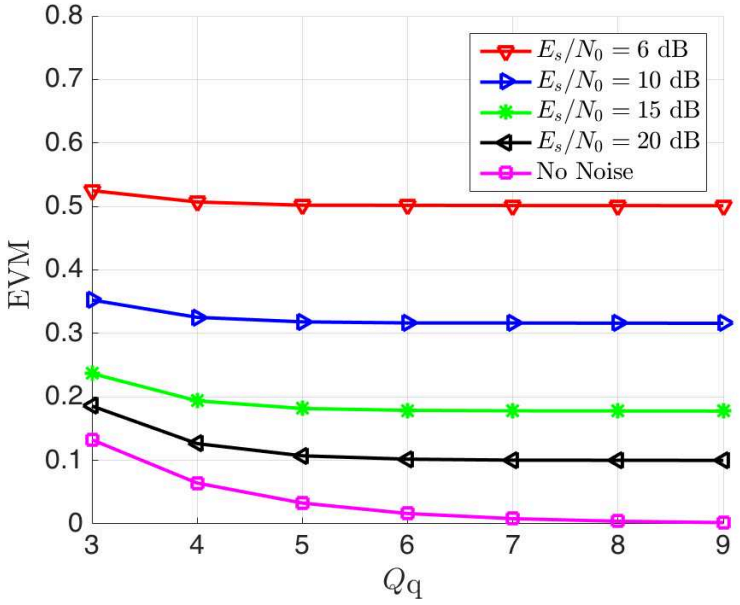

Fig. 3.8 EVMs for applying method 1 of receiver with different $(\frac{E_s}{N_0})_{\text{dB}}$ and quantization resolution Q_{q}.

3.6.1 Limitation of EVM On Function of AWGN.

At first, simulations have been done to verify Equation 3.17. We have measured the values of EVM applying Method 1 of receiver with different values of $(\frac{E_s}{N_0})_{\text{dB}}$ and quantization resolution. The simulation results are illustrated in Figure 3.8.

Table 3.4 EVM on function of $(\frac{E_s}{N_0})_{dB}$.

$(\frac{E_s}{N_0})_{dB}$	6	10	15	20	No AWGN
EVM ($Q_q \to \infty$)	0.5012	0.3162	0.1778	0.1000	0

Using Equation 3.17, we can obtain the values of EVM for different levels of $(\frac{E_s}{N_0})_{\text{dB}}$, when $Q_{\text{q}} \to \infty$. Thus we get the results as shown in Table 3.4, which are coherent with what we observe from Figure 3.8.

3.6.2 Performance Comparison Between Method 1 and 2.

Figure 3.9 shows the relative transmission data rate between Method 2 and 1 ($\frac{D_2}{D_1}$) with respect to the same value of EVM. The assumed PRB occupation ratio $\eta = 0.5$ and $\eta = 1$ are considered in both cases. Scenarios without AWGN, $\frac{E_s}{N_0} = 10$ dB and $\frac{E_s}{N_0} = 6$ dB are simulated. It can be observed that, similar results are obtained with adding different levels

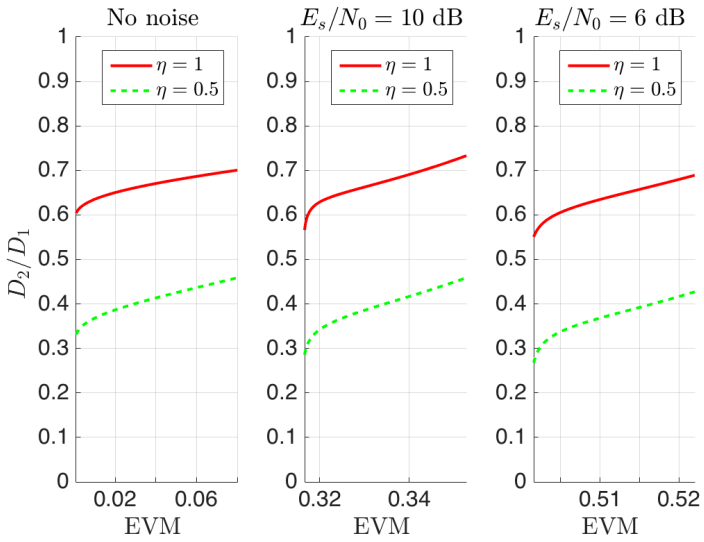

Fig. 3.9 The relative transmission data rate between Method 1 and 2 with respect to the same value of EVM.

of white Gaussian noise. Compared with Method 1, Method 2 can save 30% to 40% of the bandwidth between RRH and BBU when all the PRBs are occupied, and 55% to 70% when only half of the PRBs are occupied.

3.6.3 Performance Comparison Between Method 2 and 3.

Figure 3.10 shows the relative transmission data rate between Method 3 and 2 ($\frac{D_3}{D_2}$) with respect to the same value of EVM. The number of PRBs allocated to each transmitter is denoted by $\mathrm{NUM_{RB}}$. Both the scenarios without AWGN and $\frac{E_s}{N_0} = 6$ dB are simulated. It can be observed that, compared with Method 2, Method 3 can save 5% to 10% of the bandwidth between RRH and BBU when each transmitter is only allocated with one PRB without AWGN, and 12% to 15% when $\frac{E_s}{N_0} = 6$ dB. The more PRBs are distributed to each transmitter, the less bandwidth gain Method 3 can be obtained compared with Method 2. This is because the more PRBs are allocated to one transmitter, the more asymmetrical are the serial 12-point IFFT process in receiver and the M-point FFT process in the transmitter.

3.7 Conclusion

This chapter introduces two new architectures of functional split between RRH and BBU: Method 2 and 3, which have been modeled and simulated. Digital automatic gain control and linear quantization have been applied. In method 2, removing CP, FFT and resource

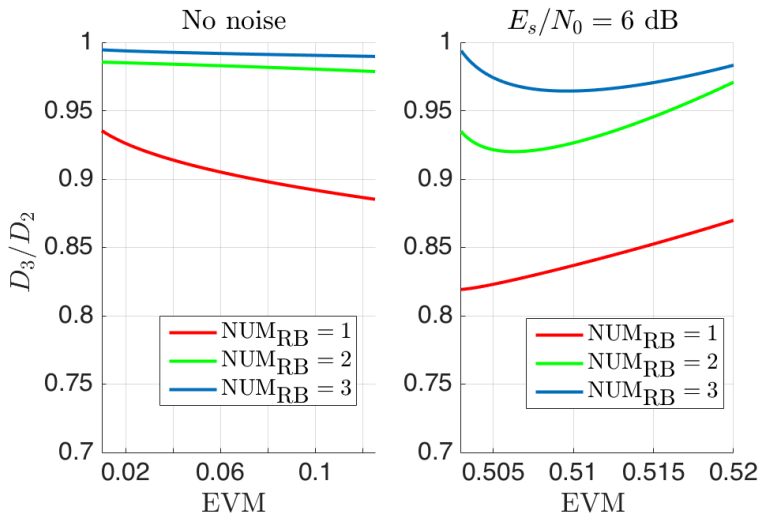

Fig. 3.10 The relative transmission data rate between method 2 and 3 with respect to the same value of EVM.

usage detection are processed in RRH. In method 3, besides the modules in Method 2, a serial 12-point IFFT is made in RRH after FFT and a serial 12-point FFT in BBU. Simulation results illustrate that, Method 2 brings a drop of 30% to 40% of the transmission rate between RRHs and BBUs compared with current functional split architecture (Method 1) when all the PRBs are occupied, and up to 70% when half of the PRBs are occupied. Method 3 can further reduce the transmission rate when UEs are allocated with few PRBs.

Applying non-linear quantization algorithms have the potential to further reduce the throughput. Non-linear quantization algorithms are compatible with the different proposed architectures, and will be investigated in a future work.

Chapter 4

C-RAN Downlink Model

4.1 Introduction

In the following of this thesis, we concentrate on a C-RAN downlink transmission scenario. This chapter deals with the C-RAN downlink system model. At first, the mathematical model of the system is introduced. Then, we present the transmission model for single RRH mode and CoMP mode. In single RRH mode, each UE is served by only one RRH. There is only limited coordinated scheduling between different RRHs. In CoMP mode, both CS/CB and JT can be applied. We rather apply CAP than "data-sharing" in data transmission scheme from a BBU pool to RRHs for CoMP. This is more realistic. Our system configuration is adapted to a 3GPP 3-sector cellular network [71].

We also propose a new RRH clustering mechanism in this chapter. A cluster with large size of RRHs will cause numerous fronthaul overheads and high complexity for CoMP transmission. In general, CoMP is considered to be implemented in a cluster with limited number of RRHs [72, 73]. The clustering schemes can be categorized to two types[74]: disjoint clustering [75] and user-centric clustering [73, 76–78]. In disjoint clustering, we divide the whole network into non-overlapping clusters. The RRHs in each cluster jointly serve the UEs inside their coverage area. This can effectively mitigate inter-cell interference. Nevertheless, the UEs located in cluster-edge still suffer great inter-cluster interference from neighbouring clusters. This disadvantage can be overcome by user-centric clustering [79]. In this clustering scheme, each UE selects the nearest RRHs to form the cluster which jointly serves it. The clusters for different UEs may be overlapped. In this way, there is no explicit cluster edges. Thus, the UEs fairness can be improved. However, the dynamic selection of clusters and the overlapping of clusters make user-centric clustering much complex than disjoint clustering. [80]. Here we propose a new RRH clustering scheme. It keeps the

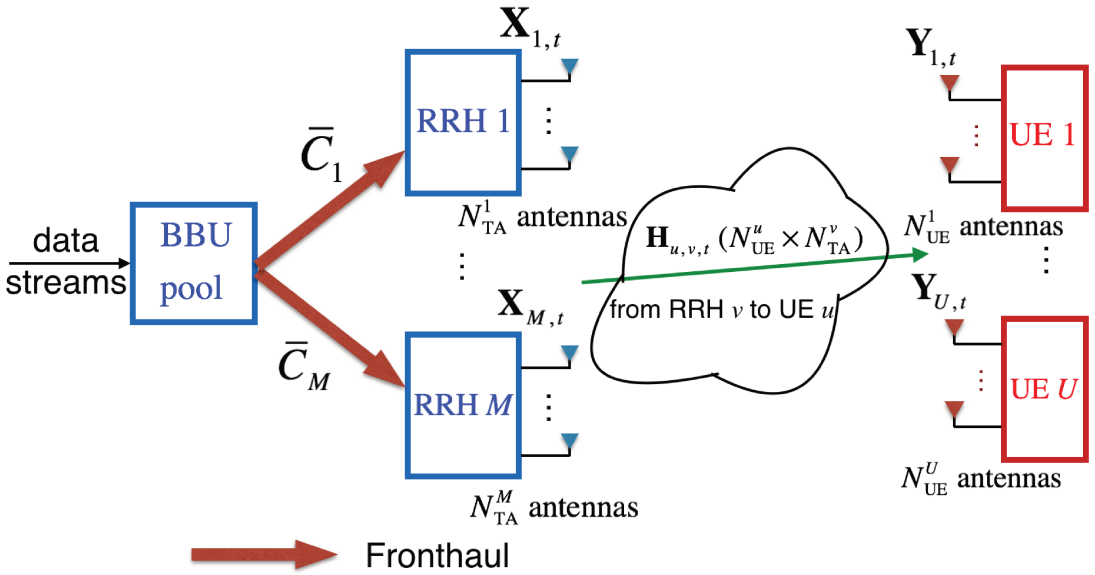

Fig. 4.1 Downlink of a cluster of RRHs which connect to a BBU pool via finite-capacity fronthaul links in C-RAN.

simplicity of disjoint clustering. Meanwhile, it reduces or eliminates the interference from other clusters to the UEs located in cluster-edge areas.

This chapter is structured as follows. At first, the system model is presented in Section 4.2. Then the problem formulation for single RRH mode and CoMP mode is separately introduced in Section 4.3 and Section 4.4. Next, Section 4.5 presents system configurations. Our proposed RRH clustering scheme is discussed in Section 4.6. At last, this chapter is concluded in Section 4.7.

4.2 System Model

4.2.1 Overall Architecture

We consider that a BBU pool controls and is connected to M_{tot} RRHs via finite-capacity fronthaul links. This thesis investigates the downlink of a cluster $\mathcal{C}$ of M RRHs out of the M_{tot} RRHs serving U UEs ($M \leq M_{\mathrm{tot}}$) as shown in Figure 4.1. The set of all the RRHs is denoted as $\mathcal{N}_{\mathrm{TP}} = \{1, 2, ..., M\}$, and the set of all the UEs is denoted as $\mathcal{N}_{\mathrm{UE}} = \{1, 2, ..., U\}$. Each v-th RRH is equipped with N_{TA}^v antennas and each u-th UE is equipped with N_{UA}^u antennas. The maximum power of RRH v is $\overline{P}_v$. The v-th fronthaul link connects the BBU pool with RRH v. The capacity of the v-th fronthaul link is denoted by $\overline{C}_v$.

4.2.2 Transmission Chain Model

We consider ideal OFDM transmission, which is equivalent to a set of independent narrow-band transmissions. Like in most cellular systems, the time is divided into subframes (1-ms subframes for LTE). Each transmission is thus made on a sub-carrier f and during a subframe t. In this thesis, we restrict ourselves on one single sub-carrier. For subframe t, the downlink propagation channel inside the coordinated cluster is defined by a set of matrix $\mathbf{H}_{u,v,t} \in \mathbb{C}^{[N^u_{\mathrm{UA}} \times N_{\mathrm{TA}}]}$ where $\mathbf{H}_{u,v,t}$ is the channel matrix from RRH v to UE u. The channel matrix from M RRHs to UE u is $\mathbf{H}_{u,t} = [\mathbf{H}_{1,u,t},\ \mathbf{H}_{2,u,t},\ ...,\ \mathbf{H}_{M,u,t}]$. The whole channel matrix from M RRHs to all UEs is $\mathbf{H}_t = [\mathbf{H}_{1,t};\ \mathbf{H}_{2,t};\ ...;\ \mathbf{H}_{U,t}]$

During subframe t, RRH v transmits the signal denoted by $\mathbf{X}_{v,t} \in \mathbb{C}^{[N_{\mathrm{TA}} \times N_s]}$ following the power constraints:

$$\frac{1}{N_s}\ \mathbb{E}(\|\mathbf{X}_{v,t}\|^2) \leq \overline{P},\ \forall\ v \in \mathcal{N}_{\mathrm{TP}}, \tag{4.1}$$

where N_s is the number of OFDM symbols transmitted from an antenna during one subframe.

The $N^u_{\mathrm{UA}} \times N_s$ signal $\mathbf{Y}_{u,t}$ received by UE u during subframe t is given by

$$\mathbf{Y}_{u,t} = \sum_{v=1}^{M} \mathbf{H}_{u,v,t} \cdot \mathbf{X}_{v,t} + \mathbf{N}_{u,t}, \tag{4.2}$$

where $\mathbf{N}_{u,t} \in \mathbb{C}^{[N^u_{\mathrm{UA}} \times N_s]}$ is the noise matrix, which consists of i.i.d $\mathcal{CN}(0,1)$ entries.

In this thesis, a set of successive N_T subframes is called a frame and is denoted by $\mathcal{T} = \{1,\ 2, \ldots\ N_T\}$. Within a frame, all active UEs are served. The U UEs are thus partitioned into N_T disjoint groups $\mathcal{N}_{\mathrm{UE},t}$ with $t \in \mathcal{T}$ and the UEs belonging to the same group are served simultaneously in the same subframe. Thus, $\bigcup_{t \in \mathcal{T}} \mathcal{N}_{\mathrm{UE},t} = \mathcal{N}_{\mathrm{UE}}$ and $\mathcal{N}_{\mathrm{UE},i} \bigcap \mathcal{N}_{\mathrm{UE},j} = \emptyset$, where $i \neq j\ \forall i \in \mathcal{T}$ and $\forall j \in \mathcal{T}$. The group of UEs served in subframe $t \in \mathcal{T}$ is denoted as $\mathcal{N}_{\mathrm{UE},t} = \{u^t_1,\ u^t_2, \ldots\ u^t_{N_{\mathrm{g},t}}\}$, where $N_{\mathrm{g},t}$ is the size of the group during subframe t.

The UEs in $\mathcal{N}_{\mathrm{UE},t}$ are served together by the M RRHs under the following constraints:

$$N^u_{\mathrm{RL}}\ \leq\ N^u_{\mathrm{UA}}, \forall u \in \mathcal{N}_{\mathrm{UE},t}, \tag{4.3}$$

$$\sum_{u \in \mathcal{N}_{\mathrm{UE},t}} N^u_{\mathrm{RL}}\ \leq\ \sum_{v \in \mathcal{N}_{\mathrm{TP}}} N^v_{\mathrm{TP}}. \tag{4.4}$$

where N^u_{RL} is the number of parallel symbols (layers) transmitted to UE u.

4.2.3 Channel model

We assume the channel to be constant within a coherence time, while they vary in an ergodic way across a large number of coherence periods. We assume the coherence time is longer than a frame. Thus, we omit index t in the channel matrix for simplicity within a frame.

The instantaneous channel between RRH v and UE u is modeled as:

$$\mathbf{H}_{u,v} = \sqrt{A(\theta_{u,v})\alpha_{u,v}\rho_{u,v}} \cdot \tilde{\mathbf{H}}_{u,v} \tag{4.5}$$

where the small-scale multipath fading matrix $\tilde{\mathbf{H}}_{u,v} \in \mathbb{C}^{[N_{\mathrm{UA}}^{u} \times N_{\mathrm{TA}}]}$ has i.i.d $\mathcal{CN}(0,1)$ entries, $\rho_{u,v}$ is the shadow fading coefficient, $\alpha_{u,v}$ is the path loss coefficient for downlink from RRH v to UE u and $A(\theta_{u,v})$ is the antenna gain, which depends on angle $\theta_{u,v}$ between the line (RRH v, UE u) and on the antenna orientation of RRH v.

The path loss coefficient is

$$\alpha_{u,v} = \frac{1}{1 + \left(\frac{d_{u,v}}{d_0}\right)^{\eta}}, \tag{4.6}$$

where $d_{u,v}$ denotes the distance between RRH v and UE u, d_0 is a reference distance and η is the path loss exponent.

The shadow fading coefficients between different RRHs and one UE are correlated. This is made by defining two different independent variables: β_u and $\beta_{u,v}$ with

$$\rho_{u,v} = \beta_u \beta_{u,v}, \tag{4.7}$$

where $10\log_{10}\beta_u \sim \mathcal{N}(0, \sigma_u^2)$ and $10\log_{10}\beta_{u,v} \sim \mathcal{N}(0, \sigma_{u,v}^2)$.

4.2.4 Discussion on CSI

Channel State Information (CSI) refers to the measured channel properties. There are generally two levels of CSI: instantaneous CSI and statistical CSI [3].

Instantaneous CSI represents the current channel conditions ($\mathbf{H}_{u,v}$). It corresponds to both small-scale multipath fading part and slow fading part in (4.5).

In this thesis, Statistical CSI refers to the average channel gain. It corresponds to only the term $\sqrt{A(\theta_{u,v})\alpha_{u,v}\rho_{u,v}}$ in (4.5). The statistical CSI varies slower than the instantaneous CSI.

When a system is based on FDD, the channel conditions on the uplink and on the downlink are different. The channel can be only measured in the receiver side. For the downlink, a reference signal with constant power is transmitted from the base station to the

UEs to estimate the instantaneous CSIs. Then, the UEs quantize the CSIs and report them back to the base station.

In the case of a system with TDD, the reverse-link estimation is possible. This is due to the reciprocity of the uplink channel and the downlink channel [3]. The downlink instantaneous CSIs can be directly estimated in the base station through uplink. In this way, compared with FDD, there will be less CSI errors caused by quantization and delay. Meanwhile, less transmission resources on the radio interface are consumed. In this thesis, we assume that the system is TDD and all the available CSIs are perfect.

4.3 Single RRH Mode

The single RRH mode is defined as when each UE is served by one and only one RRH and each RRH serves only one UE during a subframe. Hence, the size of the set of served UEs in a subframe equals to the number of RRHs, $|\mathcal{N}_{\mathrm{UE},t}| = M$.

For beamforming, we apply Maximum Ratio Combining (MRC). The precoding is done in RRH. We assume the statistical CSIs from each RRH to each UE in $\mathcal{C}$ are available in the BBU pool. In each RRH, only the instantaneous CSI between itself and its serving UE is available. The corresponding downlink transmission scheme is shown in Figure 4.2.

After channel coding, the symbols to be transmitted to UE $u \in \mathcal{N}_{\mathrm{UE},t}$ during subframe t are denoted by $\mathbf{S}_{u,t} \in \mathbb{C}^{[1\times N_s]}$, which are assumed to consist of i.i.d. $\mathcal{CN}(0,1)$ entries. The "UEs mapping" block decides which RRH serves which UE based on the statistical CSIs and distributes the coded symbols to the corresponding RRHs through the fronthaul links. In this mode, no quantization is needed. The mapping strategy is that an UE is served by the RRH corresponding to the highest statistical channel gain.

In RRH v, the output of the precoding is denoted by $\mathbf{X}_{v,t} = \mathbf{V}_{v,t}\mathbf{S}_{k_{v,t}}$, where $\mathbf{V}_{v,t}$ is the precoding matrix for the symbols transmitted by RRH v during subframe t and $k_{v,t}$ is the index of the UE served by RRH v during subframe t. The precoding matrix is given as

$$\mathbf{V}_{v,t} = \gamma_{v,t}\mathbf{H}^{\mathrm{H}}_{k_{v,t},v} \tag{4.8}$$

where $\gamma_{v,t}$ is the regulation factor which determines the transmission power of RRH v during subframe t.

In Single RRH mode, no coordinated power control is considered. The beamforming is individually done in each RRH and there is no coordinated beamforming between different RRHs. Each RRH tries its best to do the transmission. The power transmitted by RRH v

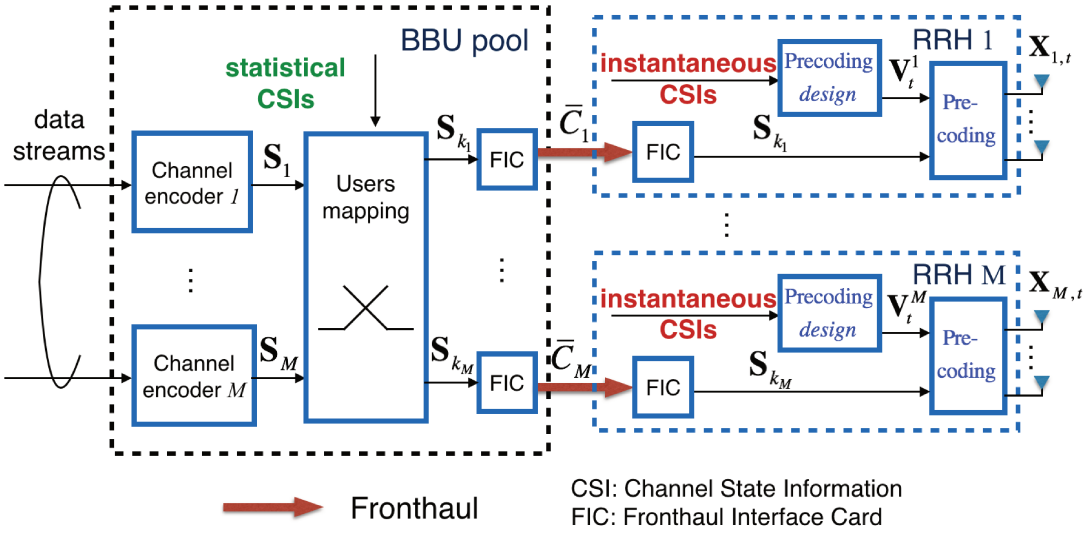

Fig. 4.2 Downlink transmission scheme from BBU pool to RRHs (single RRH mode).

during subframe t is

$$P_{v,t}(\mathbf{V}_{v,t}) = \frac{1}{N_s}\,\mathbb{E}(\|\mathbf{X}_{v,t}\|^2) = \gamma_{v,t}^2\|\mathbf{H}_{k_{v,t},v}\|^2. \tag{4.9}$$

With full power transmission, the regulation factor is

$$\gamma_{v,t} = \sqrt{\frac{\overline{P}}{\|\mathbf{H}_{k_{v,t},v}\|^2}} \tag{4.10}$$

, where $\overline{P}$ is defined in (4.1).

The achievable transmission rate for UE $k_{v,t}$ on the radio interface is limited by the interference from all the RRHs not serving it [48]:

$$\overline{R}_{k_{v,t}} = \log\left(1 + \frac{\gamma_{v,t}^2\|\mathbf{H}_{k_{v,t},v}\|^4}{1+\sum_{i\neq v,\ i\in\mathcal{N}_{\mathrm{TP}}}\gamma_{i,t}^2\|\mathbf{H}_{k_{i,t},v}\mathbf{H}_{k_{i,t},i}^{\mathrm{H}}\|^2}\right) \tag{4.11}$$

We now determine the transmission rate $R_{k_{v,t}}$ during subframe t with the fronthaul capacity constraint. As only the discrete symbols for UE $k_{v,t}$ are transmitted to RRH v, we get a lower bound of $R_{k_{v,t}}$:

$$R_{k_{v,t}} = \min\{\overline{R}_{k_{v,t}}, \overline{C}_v\}, \tag{4.12}$$

where $\overline{C}_v$ is the capacity constraint of the v-th fronthaul link.

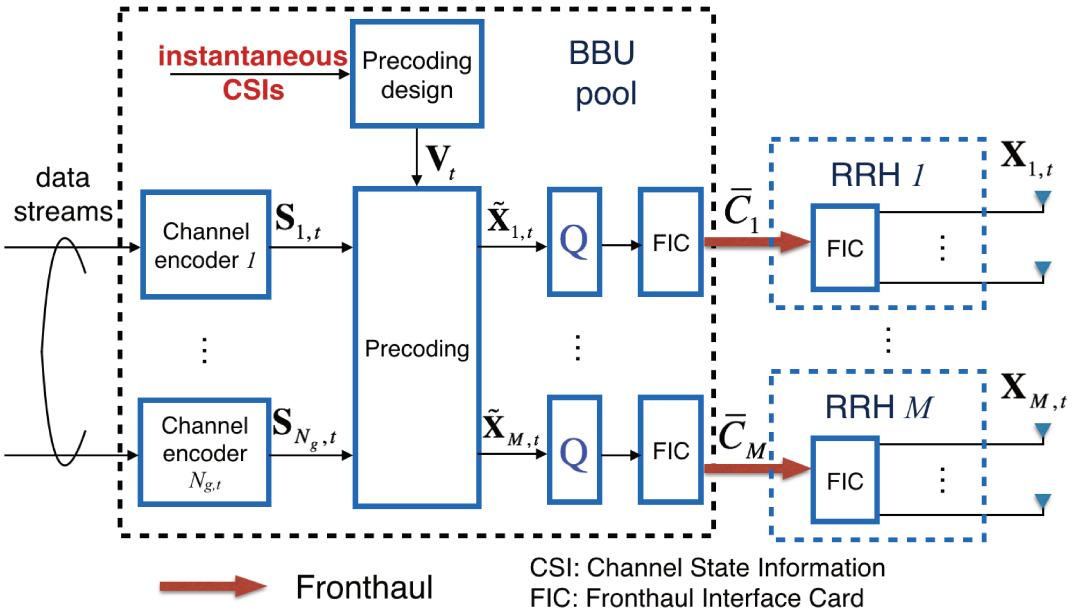

Fig. 4.3 Compression-After-Precoding (CAP) scheme for CoMP ("Q" represents fronthaul compression).

4.4 CoMP Mode

In a CoMP Mode, the U UEs are served the cluster of M RRHs by different CoMP techniques. Instead of applying a simple data-sharing strategy to forward data from BBU pool to RRHs in [81, 82], a compression-based strategy is used in this thesis. A block scheme of CAP transmission for CoMP is illustrated in Figure 4.3.

The set of symbols for all data streams during subframe t is given by

$$\mathbf{S}_t^{\mathrm{T}} = [\mathbf{S}_{\mathcal{N}_{\mathrm{UE},t}[1]},\ \mathbf{S}_{\mathcal{N}_{\mathrm{UE},t}[2]},\dots\ \mathbf{S}_{\mathcal{N}_{\mathrm{UE},t}[N_{\mathrm{g},t}]}]. \tag{4.13}$$

The output $\tilde{\mathbf{X}}_t$ of the precoding is:

$$\tilde{\mathbf{X}}_t = \mathbf{V}_t \mathbf{S}_t. \tag{4.14}$$

The output of the precoder is split into flows $\tilde{\mathbf{X}}_{1,t},\ \tilde{\mathbf{X}}_{2,t},\dots\ \tilde{\mathbf{X}}_{M,t}$, each one being sent to a given RRH. We introduce matrix $\mathbf{D}^r_{t,v}$ to extract the flow $\tilde{\mathbf{X}}_{v,t}$ from $\tilde{\mathbf{X}}_t$, which is sent to RRH v: $\mathbf{V}^r_{t,v} = (\mathbf{D}^r_{t,v})^{\mathrm{T}}\,\mathbf{V}_t$ and $\tilde{\mathbf{X}}_{v,t} = \mathbf{V}^r_{t,v}\mathbf{S}_t$. The matrix $\mathbf{D}^r_{t,v} \in \mathbb{1}^{[(\sum_{v\in\mathcal{N}_{\mathrm{TA}}} N^v_{\mathrm{TP}})\times N^v_{\mathrm{TA}}]}$ contains an $N^v_{\mathrm{TA}} \times N^v_{\mathrm{TA}}$ identity matrix in the rows from $\sum_{i=1}^{v} N^i_{\mathrm{TA}} - N^v_{\mathrm{TA}} + 1$ to $\sum_{i=1}^{v} N^i_{\mathrm{TA}}$ and all zero elements in the other rows. Note that $\mathbf{V}^r_{t,v}$ is obtained by selecting the rows of precoding matrix $\mathbf{V}_t$ for RRH v (upper index r stands for row).

Before being transmitted via the v-th fronthaul, the baseband signal sequence $\tilde{\mathbf{X}}_{v,t}$ is quantized. The compressed signals $\mathbf{X}_{v,t}$ (after quantization) is:

$$\mathbf{X}_{v,t} = \tilde{\mathbf{X}}_{v,t} + \mathbf{Q}_{v,t},\ v \in \mathcal{N}_{\mathrm{UE},t}, \tag{4.15}$$

where $\mathbf{Q}_{v,t}$ is the quantization noise matrix on the v-th fronthaul during t. We assume that the random entries of $\mathbf{Q}_{v,t}$ are complex gaussian with variance $\sigma^2_{v,t}$ and mean 0 (i.i.d $\mathcal{CN}(0,\sigma^2_{v,t})$), where $\sigma^2_{v,t}$ is an optimization variable. A fine quantization corresponds to a low value of $\sigma^2_{v,t}$, but leads to a high bit rate on the fronthaul.

Power constraint. The power transmitted by RRH v is given by [62]

$$P_{v,t}(\mathbf{V}_t, \sigma^2_{v,t}) = \frac{1}{N_s}\,\mathbb{E}(\|\mathbf{X}_{v,t}\|^2) = \mathrm{tr}\big((\mathbf{D}^r_{t,v})^{\mathrm{T}}\,\mathbf{V}_t\mathbf{V}_t^{\mathrm{H}}\,\mathbf{D}^r_{t,v} + \sigma^2_{v,t}\mathbf{I}_{N^v_{\mathrm{TA}}}\big) \tag{4.16}$$

which should respect the power constraint

$$P_{v,t}(\mathbf{V}_t, \sigma^2_{v,t}) \leq \overline{P}. \tag{4.17}$$

Fronthaul capacity constraint. The rate required on the fronthaul between RRH v and BBUs pool during subframe t can be quantified by [62]

$$C_{v,t}(\mathbf{V}_t, \sigma^2_{v,t}) = \frac{1}{N_s} I(\tilde{\mathbf{X}}_{v,t}; \mathbf{X}_{v,t}) = \log\det\big((\mathbf{D}^r_{t,v})^{\mathrm{T}}\,\mathbf{V}_t\mathbf{V}_t^{\mathrm{H}}\,\mathbf{D}^r_{t,v} + \sigma^2_{v,t}\mathbf{I}_{N_{\mathrm{TA}}}) - N_{\mathrm{TA}}\log(\sigma^2_{v,t}\big) \tag{4.18}$$

which should respect the fronthaul capacity constraint

$$C_{v,t}(\mathbf{V}_t, \sigma^2_{v,t}) \leq \overline{C}. \tag{4.19}$$

Achievable rate for UE u. The precoding matrix $\mathbf{V}^c_{t,u}$ for the data stream specific to UE u can be obtained by selecting the corresponding columns of $\mathbf{V}_t$. The index l_u of UE u in $\mathcal{N}_{\mathrm{UE},t}$ is given as $l_u = \{i | \mathcal{N}_{\mathrm{UE},t}[i] = u\}$. The matrix $\mathbf{V}^c_{t,u}$ is defined as $\mathbf{V}^c_{t,u} = \mathbf{V}_t\mathbf{D}^c_{t,u}$, where matrix $\mathbf{D}^c_{t,u} \in \mathbb{1}^{[(\sum_{u\in\mathcal{N}_{\mathrm{UE},t}} N^u_{\mathrm{RL}})\times N^u_{\mathrm{RL}}]}$ contains an $N^u_{\mathrm{RL}} \times N^u_{\mathrm{RL}}$ identity matrix in the rows from $\sum_{i=1}^{g_u} N_{\mathrm{RL}}^{\mathcal{N}_{\mathrm{UE},t}[i]} - N^u_{\mathrm{RL}} + 1$ to $\sum_{i=1}^{g_u} N_{\mathrm{RL}}^{\mathcal{N}_{\mathrm{UE},t}[i]}$ and all zero elements in the other rows. The corresponding covariance precoding matrix for UE u is defined as

$$\mathbf{G}_u = \mathbf{V}^c_{t,u}(\mathbf{V}^c_{t,u})^{\mathrm{H}},\ \forall u \in \mathcal{N}_{\mathrm{UE},t}. \tag{4.20}$$

The achievable rate for UE $u \in \mathcal{N}_{\mathrm{UE},t}$ is [62][83]

$$R_u = \log \frac{\det\left(\mathbf{I}_{N_{\mathrm{UA}}^u} + \mathbf{H}_u(\sum_{k\in\mathcal{N}_{\mathrm{UE},t}} \mathbf{G}_k + \Omega_t)\mathbf{H}_u^{\mathrm{H}}\right)}{\det\left(\mathbf{I}_{N_{\mathrm{UA}}^u} + \mathbf{H}_u(\sum_{k\in\mathcal{N}_{\mathrm{UE},t},k\neq u} \mathbf{G}_k + \Omega_t)\mathbf{H}_u^{\mathrm{H}}\right)} \tag{4.21}$$

where $\mathbf{H}_u = [\mathbf{H}_{1,u},\ \mathbf{H}_{2,u}, \dots\ \mathbf{H}_{M,u}]$ and covariance matrix $\Omega_t = \mathrm{diag}([\sigma_{1,t}^2 \mathbf{I}_{N_{\mathrm{TA}}^1}, \dots\ \sigma_{M,t}^2 \mathbf{I}_{N_{\mathrm{TA}}^M}])$.

Fundamental optimization problem. Our objective is to find the quantizations and the precoding matrix that maximize the global transmission rate. The problem of optimizing $R(\mathcal{N}_{\mathrm{UE},t})$ can be formulated as:

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{u\in\mathcal{N}_{\mathrm{UE},t}} R_u \\ \text{over} \quad & \mathbf{V}_t, \quad \Omega_t \\ \text{s.t.} \quad & C_{v,t}(\mathbf{V}_t, \sigma_{v,t}^2) \leq \overline{C},\ \forall v \in \mathcal{N}_{\mathrm{TP}} \\ & P_{v,t}(\mathbf{V}_t, \sigma_{v,t}^2) \leq \overline{P},\ \forall v \in \mathcal{N}_{\mathrm{TP}}. \end{aligned} \tag{4.22}$$

4.5 System Configuration

4.5.1 Network geometry

We consider a hexagonal three-sectored cellular wireless network as shown in Figure 4.4. The cell radius is denoted by r (the cell range is $2r$). Each site is equipped with three RRHs and the distance between two neighbor sites is $3r$. Each black arrow represents the antenna orientation for a corresponding RRH. The Cartesian coordinates of RRHs 1, 2 and 3 in the (x,y) plane are $Z_{\mathrm{RRH}}^1 = [-\frac{3}{2}r, 0]$, $Z_{\mathrm{RRH}}^2 = [\frac{3}{2}r, 0]$ and $Z_{\mathrm{RRH}}^3 = [0, \frac{3\sqrt{3}}{2}r]$. Each RRH and UE is supposed to be equipped with only one antenna.

4.5.2 Simulation parameters

In the following of this thesis, we let each cluster of RRHs have a size of 3. Therefore, $M = 3$. We consider a simple cluster including RRH 1, 2 and 3 serving a number of UEs. The set of RRHs is $\mathcal{N}_{\mathrm{TP}} = \{1,2,3\}$. The served UEs are uniformly randomly distributed in the area enclosed by the triangle whose three vertices are the positions of the three RRHs.

When possible, we take all parameters from 3GPP reference scenarios defined in [71]. Simulations are performed in an urban area LTE downlink scenario where the cell radius r is 500 m, the carrier frequency is 2000 MHz and the base station antenna height is 15 m above average rooftop level.

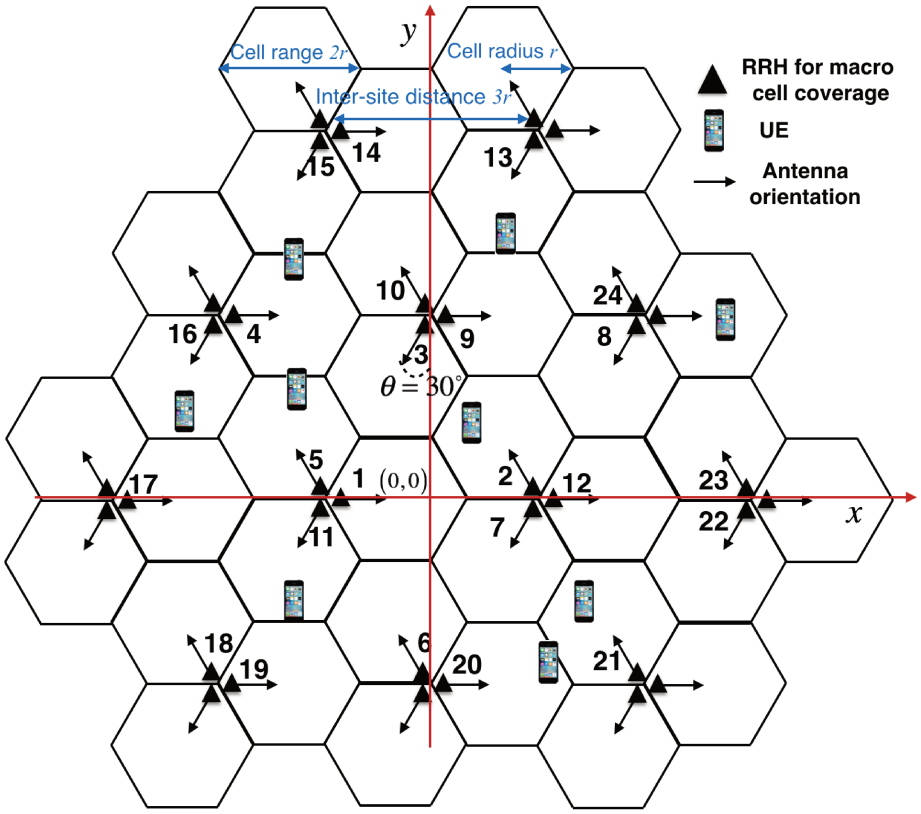

Fig. 4.4 Geometry distribution and clustering of RRHs.

Antenna radiation pattern

We apply the Base Station antenna radiation pattern in [71] where antenna gain is defined as:

$$10\log_{10}A(\theta_{u,v}) = -\min\left[12\left(\frac{\theta_{u,v}}{\theta_{3dB}}\right)^2, A_m\right] + A_0 \quad \text{where } -180^\circ \leq \theta_{u,v} \leq 180^\circ. \quad (4.23)$$

where $\theta_{3dB} = 65^\circ$ is the 3 dB beam width, $A_m = 20$ dB is the maximum attenuation and A_0 is the forward antenna gain (in the bore-sight direction of the antenna beam).

In our simulation, we take $A_0 = 15$ dBi for an urban area and a 2000 MHz carrier frequency [71].

Path loss

The propagation from RRH to UE attenuation model for our considered simulation scenario given in [71] is

$$L = 128.1 + 37.6\log_{10}(d_{u,v}) \text{ dB}, \quad (4.24)$$

where $d_{u,v}$ is in kilometers. This model is designed mainly for a value of $d_{u,v}$ varying from few hundreds meters to kilometers.

Comparing Equation 4.24 with the following formula

$$-10\log_{10}(\alpha_{u,v}) = 10\log_{10}\left(1 + \left(\frac{d_{u,v}}{d_0}\right)^{\eta}\right) \quad (4.25)$$

for a comparably large value of $d_{u,v}$, we can get $\eta \approx 3.76$ and $d_0 \approx 3.92 \times 10^{-4}$ km.

We will take $\eta = 3.76$ and $d_0 = 3.92 \times 10^{-4}$ km for the following simulation.

Shadow fading

As in [71], we take $10\log_{10}\rho_{u,v} \sim \mathcal{N}(0, 10 \text{ dB})$ in (4.7) and a shadowing correlation factor of 0.5 for the shadowing between sites. As RRH 1, 2 and 3 belong to different sites, we take $10\log_{10}\rho_u \sim \mathcal{N}(0, 5 \text{ dB})$ and $10\log_{10}\rho_{u,v} \sim \mathcal{N}(0, 5 \text{ dB})$.

RRH transmission power

In this study, we assume the noise has a variance equal to 1 (noise power): the value of each RRH power constraint $\overline{P}$ is a ratio to the noise power. To make the simulations results more intuitive and comparable with practical system configuration, we convert the unit of power constraint to dBm.

At the UE side, noise figure is 9 dB and the white Gaussian noise power is -101 dBm for a noise temperature of 300 K and a 20 MHz bandwidth. A noise power equal to 1 in previous study corresponds to -101 dBm $+ 9$ dB $= 92$ dBm. The corresponding RRH power constraint in dBm has a ratio of $\overline{P}$ to the noise power in dBm and is given by

$$P_{\text{dBm}} = \overline{P} - 92 \text{ dBm}. \tag{4.26}$$

As defined in [71], for Universal Terrestrial Radio Access (E-UTRA) downlink with a bandwidth of $W = 20$ MHz, the maximum RRH transmission power is 46 dBm. In our simulation, each RRH power constraint is set as $P_{\text{dBm}} = 46$ dBm.

Fronthaul capacity

In the system model, the unit of fronthaul capacity is bit/channel. For the same reason that of converting the unit of RRH transmission power, we convert the unit of fronthaul capacity to Mbits/s. In LTE downlink with 20 MHz bandwidth, there are $1200 \times 14 = 16800$ Resource Elements (REs) during one subframe (1 ms), where 1200 refers to 1200 subcarriers and 14 refers to 14 REs for one subcarrier during one subframe. One RE is the equivalent of 1 modulation symbol on a subcarrier. Here we consider one RE as a channel and there are 16800 channels during one subframe. Thus, a transmission rate r bits/channel equals to $16800r$ bits/ 1 ms $= 16.8r$ Mbits/s in LTE downlink with 20 MHz bandwidth.

In Section 5.5, we do simulations with $r = 8$ and $r = 2$. These correspond to a fronthaul capacity of $16.8 \times 8 = 134.4$ Mbits/s and a fronthaul capacity of $16.8 \times 2 = 33.6$ Mbits/s, respectively.

4.6 RRH Clustering

In this section, we introduce our proposed RRH clustering mechanism. As shown in Figure 4.5, the network is divided into regions of triangular form. Each 3 RRHs on neighbour sites located at the three vertices of a triangle region constitute a cluster, and mainly jointly serve the UEs located inside this region. Meanwhile, the antenna direction of one RRH of a cluster should point to one of the other two RRHs in the cluster. For example, RRH 1, 2 and 3 form a cluster but RRH 1, 2 and 9 do not belong to the same cluster.

Each RRH belongs to two different clusters. For example, RRH 1, 2 and 3 form a cluster while RRH 1, 6 and 7 form another different cluster. To guarantee the coverage of the whole network, each RRH works in each of its two corresponding clusters half of the time. Thus the sum transmission rate got from (4.12) and (4.22) should be divided by 2. During one half

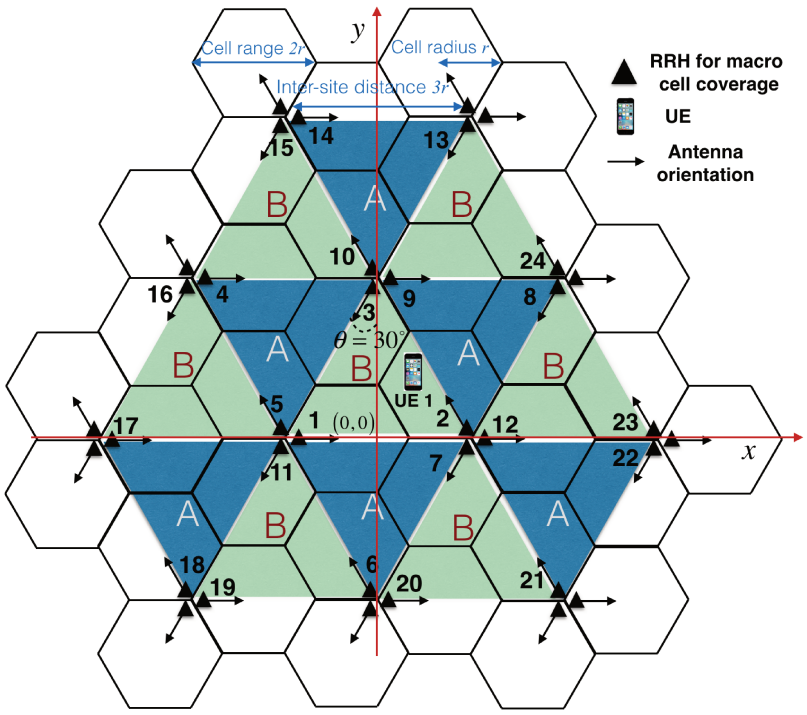

Fig. 4.5 Clustering of RRHs.

of the time, each 3 RRHs located at the three vertices of a "A" type triangle region as shown in Figure 4.4 constitute a cluster. we call this mode "A". During the second half of the time, the system works on mode "B" where each 3 RRHs located at the three vertices of a "B" type triangle region constitute a cluster. For each two N_T subframes, the system works on mode "A" to mainly serve the UEs in "A" type regions during N_T subframes, and on mode "B" to mainly cover the UEs in "B" type regions during the other N_T subframes. The alternating change between mode "A" and mode "B" can largely reduce the inter-clustering interference.

To assign a UE to a cluster of RRHs, we search for the RRH having the largest large-scale (average) channel gain to this UE. Then, we will make a choice between the two clusters associated to this RRH. For example, UE 1 get the largest average channel gain from RRH 2. Then, we should make a choice between the cluster of RRH 1, 2 and 3 and the cluster of RRH 2, 8 and 9. This choice depends on which RRH among RRH 1, 3, 8 and 9 gives the maximum average channel gain to UE 1. If it is RRH 1 or 3, we choose UE 1 to be served by the cluster of RRH 1, 2 and 3, otherwise by the cluster of RRH 2, 8 and 9.

4.7 Conclusion

The C-RAN downlink system model and configurations for the following of this thesis is presented. We propose a new RRH clustering scheme. It is a trade-off between disjoint clustering and user-centric clustering.

Chapter 5

Comparison of Different Transmission Strategies

5.1 Introduction

In mobile networks, transmissions on the same time-frequency resource interfere with each other. Early cellular technologies apply static frequency separation between neighbor cells to avoid the interference. This corresponds to a frequency reuse factor higher than 1, which is the number of cells that use different frequencies. However, modern mobile-broadband system Long Term Evolution (LTE) has opted for a reuse factor 1, to maximize the data rates for users close to the Base Station (BS).

In reuse-one deployment, low signal-to-interference ratios (SIR) may occur, especially in the cell edge area, where the power of the useful signal has the same order of magnitude as the interference. C-RAN can facilitate the cooperation among different cells and allows advanced algorithms (e.g. CoMP) to manage interference.

However, the digitized baseband signals exchanged between BBU pool and RRHs require a large bit rate. This is a main limitation of the feasibility of C-RAN. Therefore, it is important to include fronthaul capacity constraint when evaluating the performance of different advanced cooperation algorithms in C-RAN [84, 85]. In this chapter, we study the performance of different transmission strategies with RRH power constraints and fronthaul capacity constraints.

This study was published in WPMC 2016 [DLG16a]. It was made during the first part of the thesis. Hence, only one antenna per RRH and per UE is considered. This limitation is relaxed in the following chapters.

The rest of this chapter is organized as follows. Firstly, different transmission strategies are presented in Section 5.2. The system model is introduced in Section 5.3. Then the problem formulation for different transmission strategies is discussed in Section 5.4. Next, Section 5.5 presents the simulation results. In the end, Section 5.6 concludes this chapter.

5.2 Different Transmission Strategies

In order to understand the performance of different transmission strategies (e.g. single transmission vs distributed MIMO), we consider the cooperation cluster including 3 RRHs (RRH 1, 2 and 3) which serves 3 UEs (UE 1, 2 and 3). Each RRH and UE is supposed to be equipped with only one antenna. We assume the functional module 'FFT' is in BBU pool. The statistical CSIs are assumed to be known by the BBU pool. In some transmission schemes, we assume the instantaneous CSIs are also known.

In the following, five different transmission schemes in downlink for the cooperation cluster are introduced: 1) single RRH mode (see Section 4.3); 2) dynamic point selection (DPS) (see Subsection 2.3.2); 3) round robin selection (RR); 4) single user joint transmission (SU-JT) (see Subsection 2.3.2); 5) distributed Multiple Input Multiple Output mode (D-MIMO). All the transmissions are supposed to be done at the same frequency resource.

5.2.1 Single RRH Mode

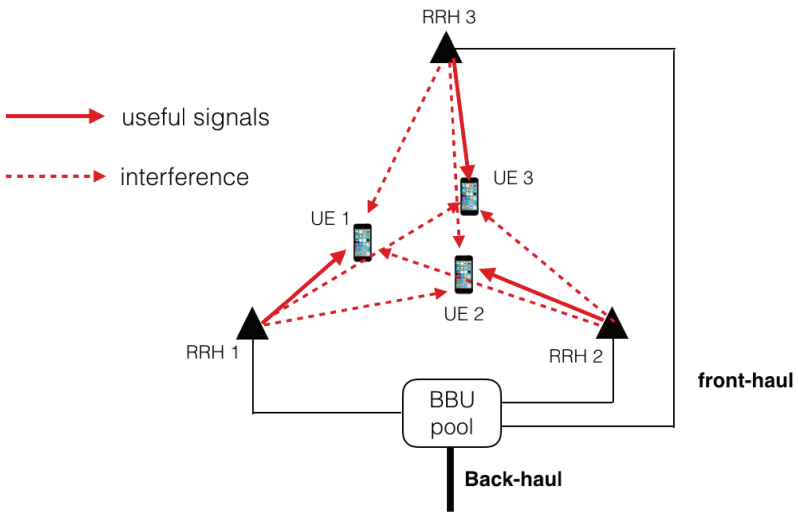

Fig. 5.1 Downlink transmission scheme of single RRH mode.

In single RRH mode, the statistical CSIs are shared among different cells, meanwhile the instantaneous CSIs are not. The association of which RRH to which UE depends on statistical CSIs. A simple example of downlink in single RRH mode is shown in Figure 5.1. RRH 1 transmits signals to UE 1, RRH 2 to UE 2 and RRH 3 to UE 3 at the same time. No power control is applied and each RRH transmits signals with maximum power. BBU pool need to transmit discrete user data for UE u only to RRH u through fronthaul links, where $u = 1,\ 2,\ 3$.

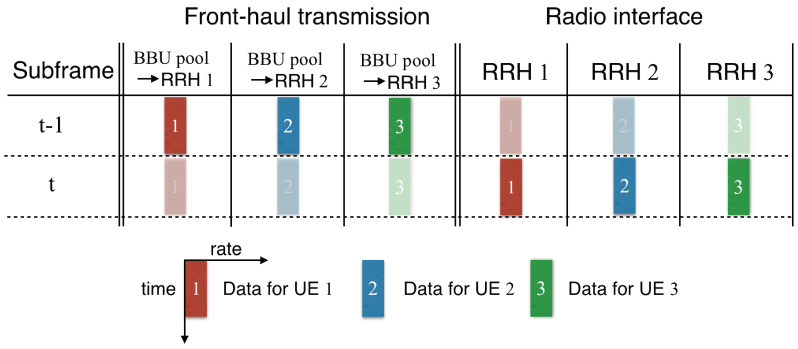

Fig. 5.2 Downlink fronthaul and radio interface transmission scheme of single RRH mode.

Figure 5.2 shows the fronthaul and radio interface transmission scheme of single RRH mode. During subframe t, each RRH transmits the data received during subframe $t-1$ from the BBU pool to its serving UE.

As all the transmissions are done at the same time and frequency resource, each UE will suffer interference from its non serving RRHs. The interference will become huge when a UE is located at the edge of cells. Applying advanced coordinated algorithms can eliminate the interference or turn it to useful signals. several CoMP algorithms are presented in the following sections.

5.2.2 Dynamic Point Selection (DPS)

In dynamic point selection (DPS), only one UE is served during a given subframe and the signal to the UE is transmitted only from a single RRH selected in the cooperation cluster. We assume that the BBU pool knows the instantaneous channel information between each RRH and each served UE in the cooperation cluster. The BBU pool selects the RRH that gives the highest instantaneous channel gain to the served UE. The selected RRH is used for transmission with maximum power, while the other two RRHs are muted. Different UEs are served sequentially (TDMA mode). As only one UE is served at each time and frequency resource, the served UE will no more suffer interference from the other two RRHs.

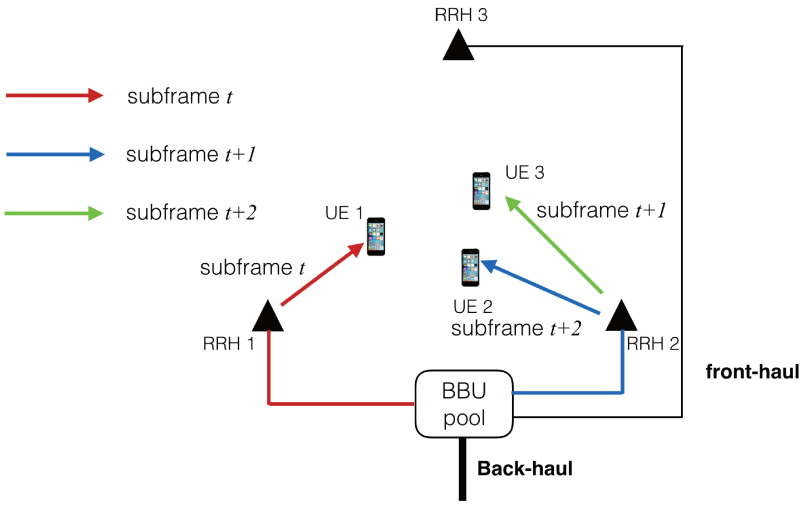

Fig. 5.3 Downlink transmission scheme of dynamic point selection.

A simple dynamic point selection downlink transmission scheme for the cooperation cluster of 3 RRHs and 3 served UEs is illustrated in Figure 5.3. In this example, the channel gain is supposed to be constant during subframe t, $t+1$ and $t+2$. The highest instantaneous channel gain is given by RRH 1 for UE 1, and RRH 2 for both UE 2 and UE 3. Hence, RRH 3 is not used but RRH 2 is transmitting during 2 subframes.

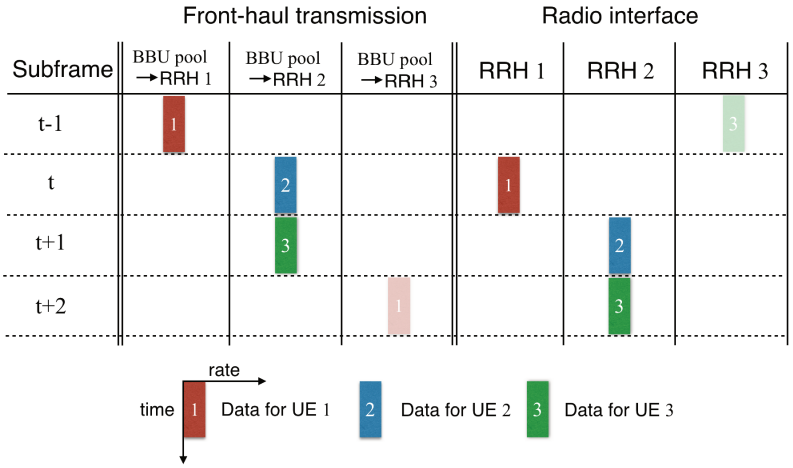

Fig. 5.4 Downlink fronthaul and radio interface transmission scheme of dynamic point selection with TDMA mode.

Figure 5.4 shows the fronthaul and radio interface transmission scheme of dynamic point selection. Once a RRH is selected to do transmission during a subframe t, the corresponding discrete user data are transmitted to the RRH from BBU pool during the previous subframe

$t-1$. The fronthaul link between the selected RRH and the BBU pool is occupied during subframe $t-1$ while the other two are idle. The selection of RRH for transmission is dynamic for each period of 3 subframes. BBU pool does not know which RRHs will do transmission 3 subframes later. Thus the vacant fronthaul capacity cannot be used for predicted transmission: only one third of the whole fronthaul capacity is used. This shortage can be compensated by round robin transmission scheme which is introduced in the following section.

5.2.3 Round Robin Selection (RR)

We assume here that, only average channel gains are known at BBU pool (no instantaneous channel knowledge). The association of a RRH and its serving UE is fixed during a longer time than 3 subframes depending on the average channel gains. In round robin selection (RR), each RRH transmits in a round robin way and during one subframe for each turn. Like for dynamic point selection, only one RRH transmits at the same time while the other two are muted. However, RR cannot achieve diversity gain compared with DPS.

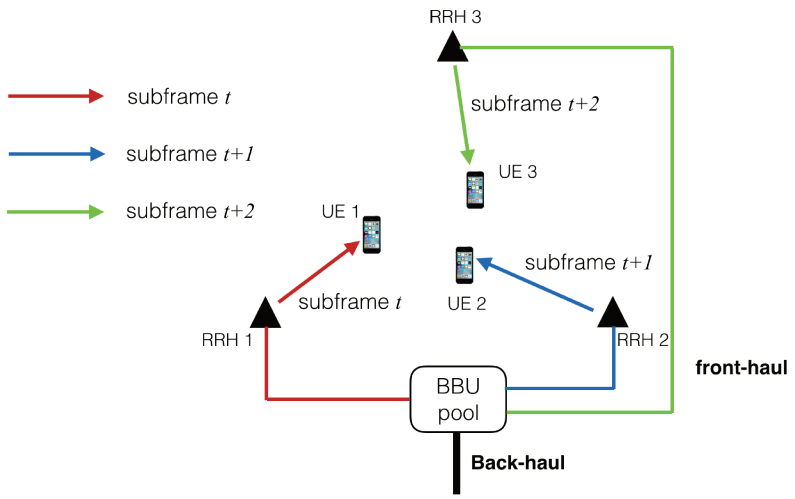

Fig. 5.5 Downlink transmission scheme of round robin selection.

An example of round robin selection is shown in Figure 5.5. In this example, UE 1 is served by RRH 1, UE 2 by RRH 2 and UE 3 by RRH 3 in a round robin way, with each UE being served during one subframe each turn.

Figure 5.6 shows the fronthaul and radio interface transmission scheme of round robin selection. As the association between each RRH and each UE is fixed during a rather longer time than 3 subframes, BBU pool can do predicted transmission to each RRH. User data to be transmitted is stored in a buffer for a period of 3 subframes or less. When one RRH does transmission during subframe t, it can transmit all the user data received during subframe

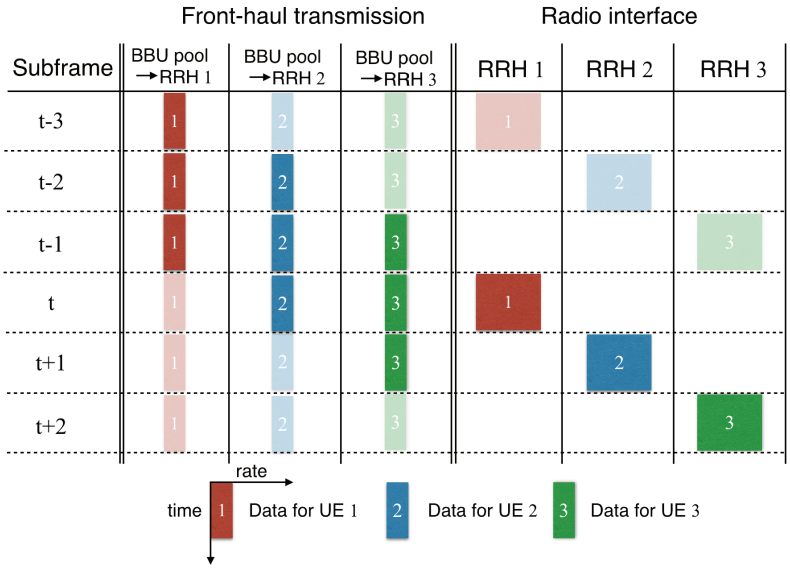

Fig. 5.6 Downlink fronthaul and radio interface transmission scheme of round robin selection with TDMA mode.

$t-1$, $t-2$ and $t-3$. In this way, the useful fronthaul capacity is three times higher than for dynamic point selection.

5.2.4 Single User Joint Transmission (SU-JT)

Dynamic point selection can eliminate interference. However, two RRHs are muted during each subframe. Different from dynamic point selection, in single user joint transmission (SU-JT), all the 3 RRHs transmit signals coherently to one and only one UE during each subframe.

The downlink transmission scheme of joint transmission is shown in Figure 5.7. During subframe $t+u-1$, all the three RRHs transmit the same signals to UE u with maximum power, where $u = 1,\ 2,\ 3$. BBU pool needs to transmit all discrete user data for UE 1, 2 and 3 to each of the three RRHs.

Figure 5.8 shows the fronthaul and radio interface transmission scheme of SU-JT. During subframe $t+u-1$, the BBU pool transmits the same data of UE u to all the 3 RRHs. And the RRHs transmit the received data to UE u during the next subframe.

5.2.5 Distributed MIMO Mode (D-MIMO)

Joint transmission brings a power gain compared with dynamic point selection. If all channel informations including both slow fading and fast fading are shared among different cells in

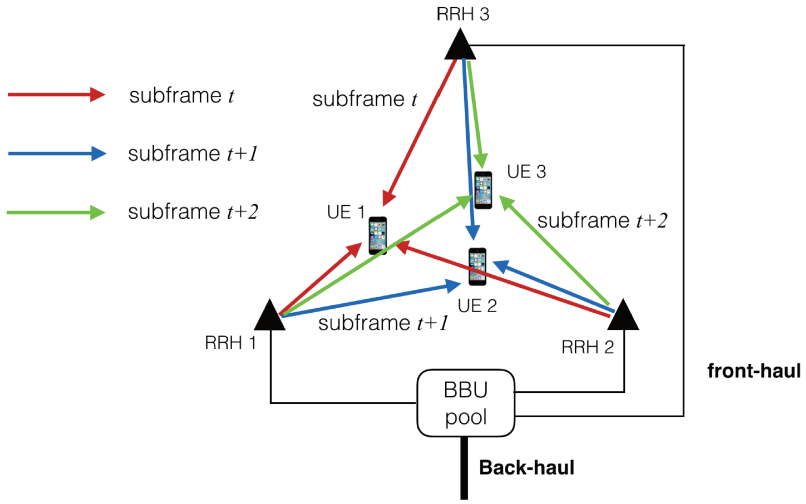

Fig. 5.7 Downlink transmission scheme of single user joint transmission.

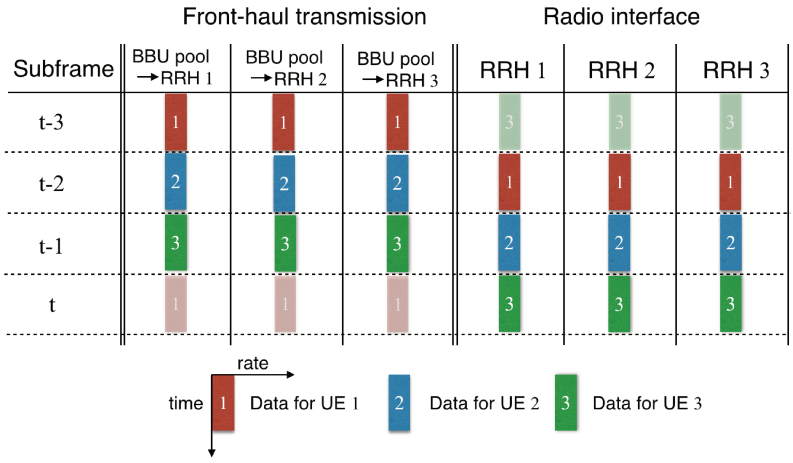

Fig. 5.8 Downlink fronthaul and radio interface transmission scheme of SU-JT with TDMA mode.

the cooperation cluster, all the 3 RRHs can transmit parallel data to the 3 UEs using different Multi-User MIMO technologies. In this study, we consider Zero-Forcing (ZF) algorithm for the parallel data transmission. Here, we denote this transmission scheme as distributed MIMO mode (D-MIMO). D-MIMO can bring a multiplexing gain compared with dynamic point selection.

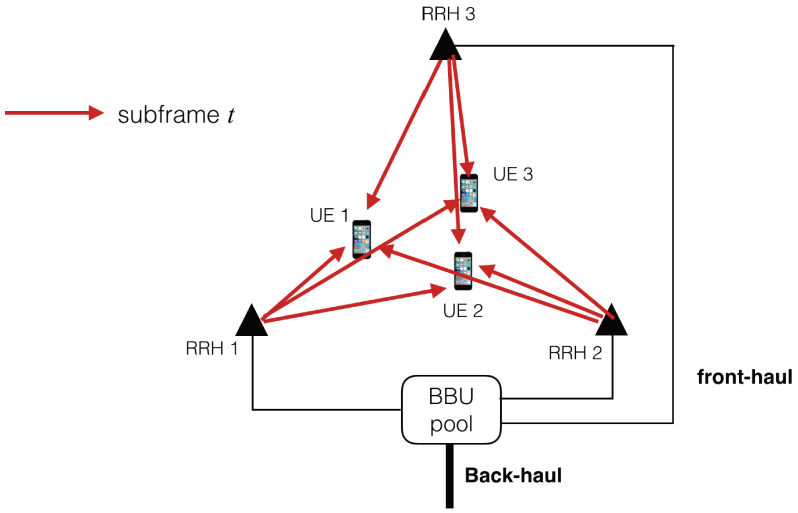

Fig. 5.9 Downlink transmission scheme of distributed MIMO.

As shown in Figure 5.9, during each time period, RRH 1, 2 and 3 transmit parallel data to UE 1, 2 and 3.

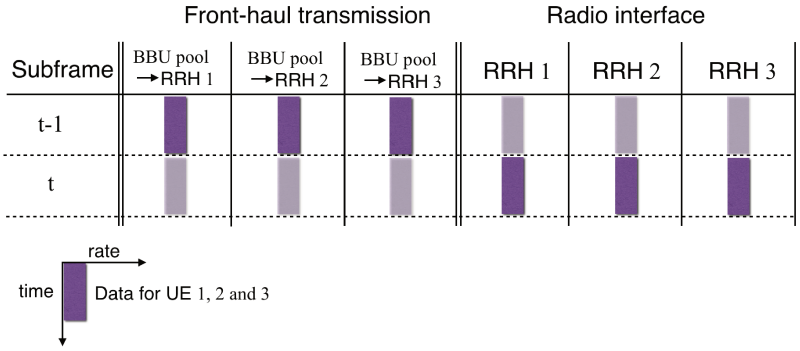

Fig. 5.10 Downlink fronthaul and radio interface transmission scheme of D-MIMO.

Figure 5.10 shows the fronthaul and radio interface transmission scheme of D-MIMO. During subframe t, each RRH transmits the data received during subframe $t-1$ from the BBU pool to all the UEs.

5.3 System Model

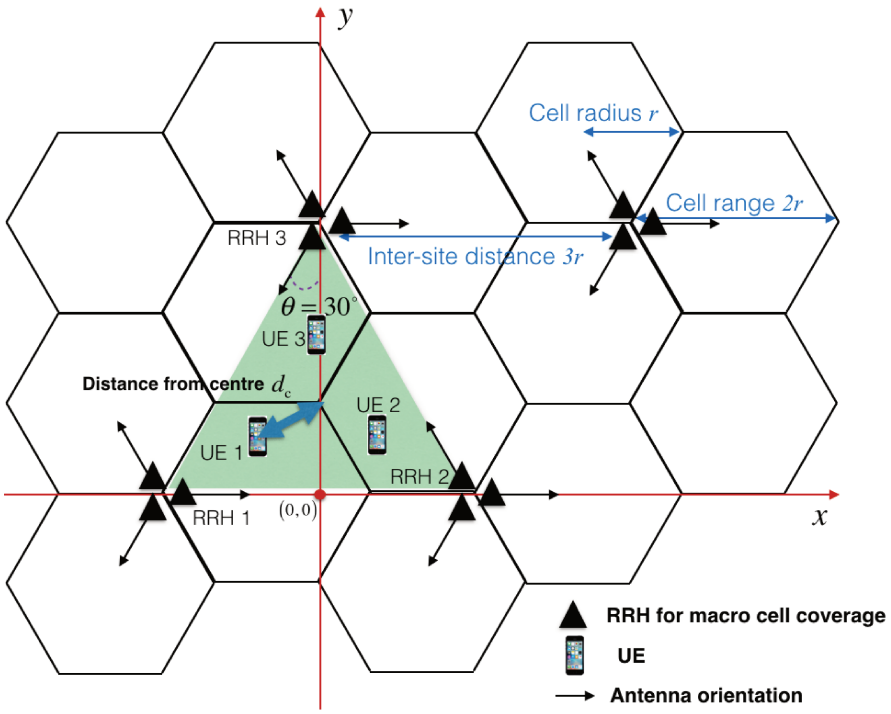

Fig. 5.11 Geometry distribution of RRHs and UEs.

We apply the system model and configuration presented in Chapter 4. We consider a simple cluster including RRH 1, 2 and 3 serving three UEs as shown in Figure 5.11. Thus, $U = 3$. The set of RRHs $\mathcal{N}_{\text{TP}} = \{1,2,3\}$ and the set of UEs $\mathcal{N}_{\text{UE}} = \{1,2,3\}$. Each RRH and each UE is assumed to be equipped with one antenna: $N_{\text{TA},v} = 1$, $N_{\text{UE},u} = 1$, $\forall v \in \mathcal{N}_{\text{TP}}$, $\forall u \in \mathcal{N}_{\text{UE}}$. The position of the center of RRH 1, 2 and 3 is $[0, \frac{\sqrt{3}}{2}r]$. UE u is located on the line connecting RRH u and the center of RRH 1, 2 and 3, $\forall u \in \mathcal{N}_{\text{UE}}$. The distances between each UE and the center of RRH 1, 2 and 3 are the same which is denoted by d_{c}.

We assume each RRH is subject to the same fronthaul constraint $\overline{C}$ and has the same power constraint $\overline{P}$. Thus, $\overline{C}_v = \overline{C}$ and $\overline{P}_v = \overline{P}$, $\forall v \in \mathcal{N}_{\text{TP}}$. Each RRH and UE is supposed to be equipped with only one antenna. Thus $\mathbf{H}_{u,v}$ becomes a scalar. To simplify the notation, let $h_{u,v} = \mathbf{H}_{u,v}$, $\forall u \in \mathcal{N}_{\text{UE}},\ v \in \mathcal{N}_{\text{TP}}$.

5.4 Transmission Rates for Different Transmission Strategies

In this section, we apply the mathematic model in Chapter 4 to the five transmission strategies introduced in Section 5.2.

5.4.1 Single RRH mode

For single RRH mode, we assume $T = 1$, and $\mathcal{T} = \{1\}$. Without losing generality, the index t will be omitted in the following of this subsection 5.4.1.

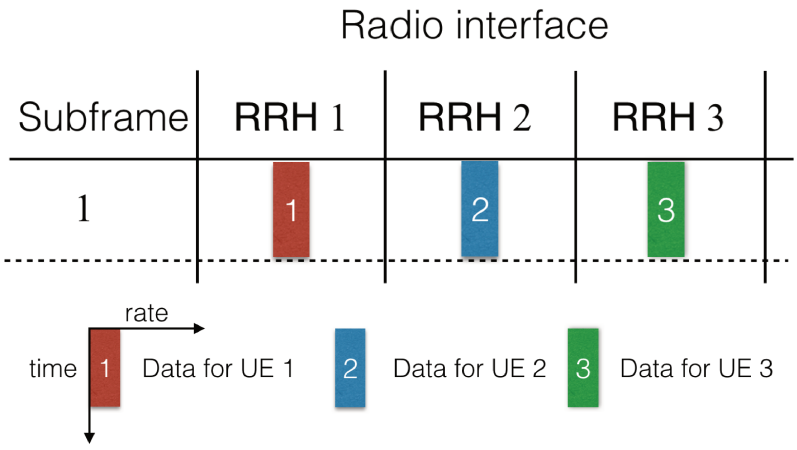

Fig. 5.12 RRH - UE serving association for single RRH mode.

Each UE is served only by one RRH during each subframe. We suppose the serving association in Figure 5.12 is the best choice depending on statistical CSIs during subframe 1, where UE u is served by RRH u and $u = 1,\ 2,\ 3$.

Applying (4.11) in Section 4.3, the achievable transmission rate without fronthaul capacity limitation for UE u is

$$\overline{R}_u = \log\left(\frac{\overline{P}\|h_{u,u}\|^2}{1+\overline{P}\sum_{i=1,\ i\neq u}^{3}\|h_{u,i}\|^2}+1\right)\ \forall\ u \in \mathcal{N}_{\mathrm{UE}} \tag{5.1}$$

The achievable transmission rate with fronthaul capacity constraint for UE u during subframe 1 is

$$R_u = \min\left(\overline{R}_u,\ \overline{C}\right) \tag{5.2}$$

where $u = 1,\ 2,\ 3$.

The sum of achievable transmission rates during subframe 1 with fronthaul capacity constraint is

$$R_{\text{tot}} = R_1 + R_2 + R_3. \tag{5.3}$$

5.4.2 Dynamic Point Selection (DPS)

For dynamic point selection (DPS), we assume $T = 3$, and $\mathcal{T} = \{1,\ 2,\ 3\}$. UE u is supposed to be the only one being served during subframe u, where $u = 1,\ 2,\ 3$. And only the RRH bringing the highest channel gain for UE u does transmission during subframe u.

As the serving RRH transmits signals with maximum power, the achievable transmission rate without fronthaul capacity limitation during subframe u for UE u is

$$\overline{R}_u = \log\left(1 + \overline{P} \times \max\left(\|h_{u,1}\|^2, \|h_{u,2}\|^2, \|h_{u,3}\|^2\right)\right)\ \forall\ u \in \mathcal{N}_{\text{UE}} \tag{5.4}$$

The achievable transmission rate with fronthaul capacity constraint for UE u during subframe u is

$$R_u = \min\left(\overline{R}_u,\ \overline{C}\right) \tag{5.5}$$

where $u = 1,\ 2,\ 3$.

The sum achievable transmission rate during subframe 1, 2 and 3 with fronthaul capacity constraint is

$$R_{\text{tot}} = \frac{1}{3}(R_1 + R_2 + R_3). \tag{5.6}$$

5.4.3 Round Robin Transmission (RR)

For round robin transmission (RR), we assume $T = 3$, and $\mathcal{T} = \{1,\ 2,\ 3\}$. UE u is supposed to be the only one being served during subframe u, where $u = 1,\ 2,\ 3$. Only one different RRH transmits during each subframe. Depending on slow fading channel information during subframes 1, 2 and 3, we suppose the association that RRH u serves UE u where $u = 1,\ 2,\ 3$ is the best choice.

As the serving RRH transmits signals with maximum power, the achievable transmission rate without fronthaul capacity limitation during subframe u for UE u is

$$\overline{R}_u = \log\left(1 + \overline{P} \times \|h_{u,u}\|^2\right)\ \forall\ u \in \mathcal{N}_{\text{UE}} \tag{5.7}$$

The achievable transmission rate with fronthaul capacity constraint for UE u during subframe u is

$$R_u = \min\left(\overline{R}_u,\ 3 \times \overline{C}\right) \tag{5.8}$$

where $u = 1,\ 2,\ 3$.

The sum achievable transmission rate during subframe 1, 2 and 3 with fronthaul capacity constraint is

$$R_{\text{tot}} = \frac{1}{3}(R_1 + R_2 + R_3). \tag{5.9}$$

5.4.4 Single User Joint Transmission (SU-JT)

For single user joint transmission (SU-JT), we assume $T = 3$, and $\mathcal{T} = \{1,\ 2,\ 3\}$. UE u is supposed to be the only one being served during subframe u, where $u = 1,\ 2,\ 3$. And all the 3 RRH transmit the same signals to UE u during subframe u.

As the serving RRHs transmit signals with maximum power, the achievable transmission rate without fronthaul capacity limitation during subframe u for UE u is

$$\overline{R}_u = \log\left(1 + \overline{P} \times \sum_{v=1}^{3} \|h_{u,v}\|^2\right)\ \forall\ u \in \mathcal{N}_{\text{UE}} \tag{5.10}$$

The achievable transmission rate with fronthaul capacity constraint for UE u during subframe u is

$$R_u = \min\left(\overline{R}_u,\ \overline{C}\right) \tag{5.11}$$

where $u = 1,\ 2,\ 3$.

The sum achievable transmission rate during subframe 1, 2 and 3 with fronthaul capacity constraint is

$$R_{\text{tot}} = \frac{1}{3}(R_1 + R_2 + R_3). \tag{5.12}$$

5.4.5 Distributed MIMO Mode (D-MIMO)

For distributed MIMO mode (D-MIMO), we assume $T = 1$, and $\mathcal{T} = \{1\}$. Without losing generality, index t will be omitted in the following of this section. All the 3 RRHs transmit parallel data together to all the 3 UEs.

With perfect Channel State Informations (CSI) available at BBU pool, both linear procoding and dirty paper coding (DPC) techniques can be applied for the joint transmission [19]. DPC can achieve a better performance but is difficult to implement in practical systems. This

is due to the high complexity of successive encodings and decodings involved, especially when the number of UEs is large [86]. In this study, we consider Zero-Forcing (ZF) algorithm [87] for the parallel data transmission. Applying Zero Forcing algorithm, the precoding matrix is given as:

$$\mathbf{V} = \gamma \mathbf{H}^{\mathrm{H}} (\mathbf{H}\mathbf{H}^{\mathrm{H}})^{-1} \tag{5.13}$$

where $\mathbf{H}$ is the channel matrix between the 3 RRHs and the 3 UEs, and γ is a normalization factor which is selected to satisfy the RRH power constraint and fronthaul capacity constraint. Water-filling is not considered in this thesis.

We apply precoding matrix $\mathbf{V}$ to problem (4.22) in Section 4.4 to calculate the maximum sum transmission rate. Problem 4.22 is a non-convex problem. We apply an adapted Majorization Minimization scheme proposed in [88] to solve this problem.

5.5 Simulation Results

In this section, we compare the performance of different transmission strategies discussed in the previous section.

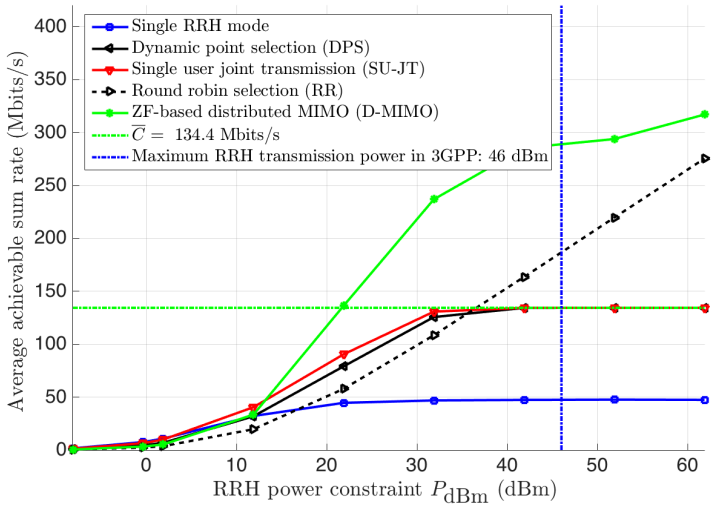

Fig. 5.13 Average achievable sum rate vs. RRH power constraint P_{dBm} ($\overline{C} = 134.4$ Mbits/s, $d_{\mathrm{c}} = 0$ m).

We start by investigating the effect of RRH power limitation on the average achievable sum rate with $d_{\mathrm{c}} = 0$ m. In other words, the 3 UEs are at the same location: at the common conner of Cell 1, 2 and 3, where RRH u locates at Cell u and $u = 1,\ 2,\ 3$.

At first, we consider a fronthaul capacity $\overline{C} = 8$ bits/channel which corresponds to $\overline{C} = 134.4$ Mbits/s (see Section 4.5.2). The simulation results are shown in Figure 5.13. ZF-based D-MIMO can achieve the highest average sum rate and single RRH mode can achieve the lowest in RRH high-power regime. However, ZF-based D-MIMO is less preferred in RRH low-power regime. RR has a better performance than other transmission strategies except for ZF-based D-MIMO in RRH high-power regime. For DPS and SU-JT, each fronthaul link has to guarantee enough capacity to convey all the symbols transmitted to the RRHs during each subframe. Therefore, the sum transmission rates for these transmission schemes are always less than $\overline{C}$.

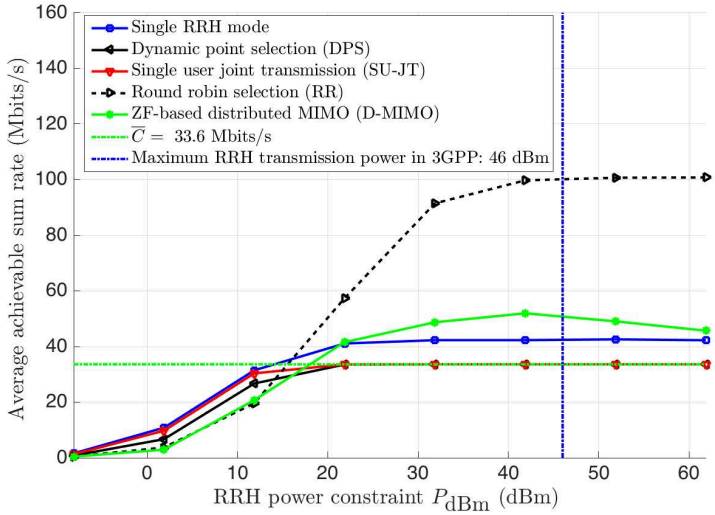

Fig. 5.14 Average achievable sum rate vs. RRH power constraint P_{dBm} with small fronthaul capacity ($\overline{C} = 33.6$ Mbits/s, $d_{\mathrm{c}} = 0$ m).

Then, we consider a smaller fronthaul capacity $\overline{C} = 2$ bits/channel which corresponds to $\overline{C} = 33.6$ Mbits/s (see Section 4.5.2). The simulation results are illustrated in Figure 5.14. It shows that RR has the best performance in RRH high-power regime. The performance of single RRH mode is better than SU-JT and DPS. The performance difference between single RRH mode and ZF-based D-MIMO becomes smaller in RRH high-power regime.

Then, the effect of UE distance from the center of RRH 1, 2 and 3 on the achievable sum rate is tested, with $P_{\mathrm{dBm}} = 41.4$ dBm. The simulation results with $\overline{C} = 134.4$ Mbits/s are illustrated in Figure 5.15. ZF-based D-MIMO has a better performance than the others. However, the performance difference between it and single RRH mode becomes smaller and smaller with the increasing of UE distance from center. Single RRH mode has a worse performance than RR in low UE distance from center regime, but better in high regime. The average achievable sum rates of DPS and SU-JT are limited to be less than $\overline{C}$.

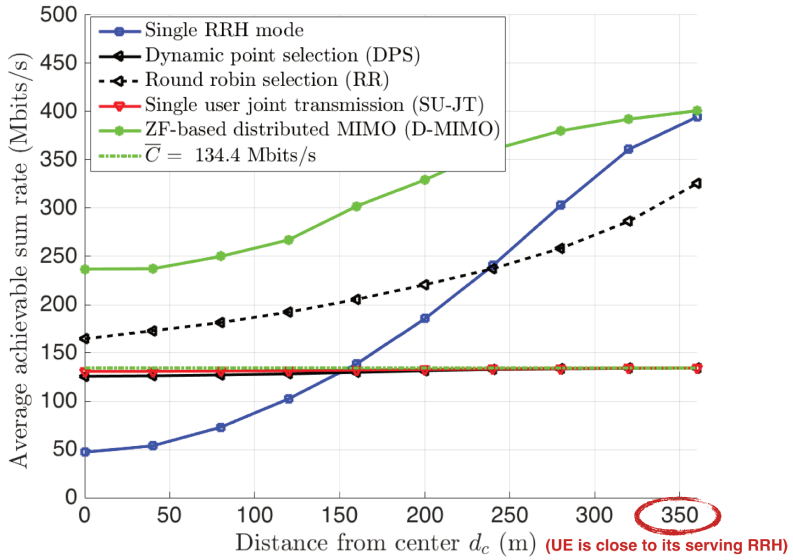

Fig. 5.15 Average achievable sum rate vs. UE distance from the center of RRH 1, 2 and 3 d_c ($P_{\text{dBm}} = 41.4$ dBm, $\overline{C} = 134.4$ Mbits/s).

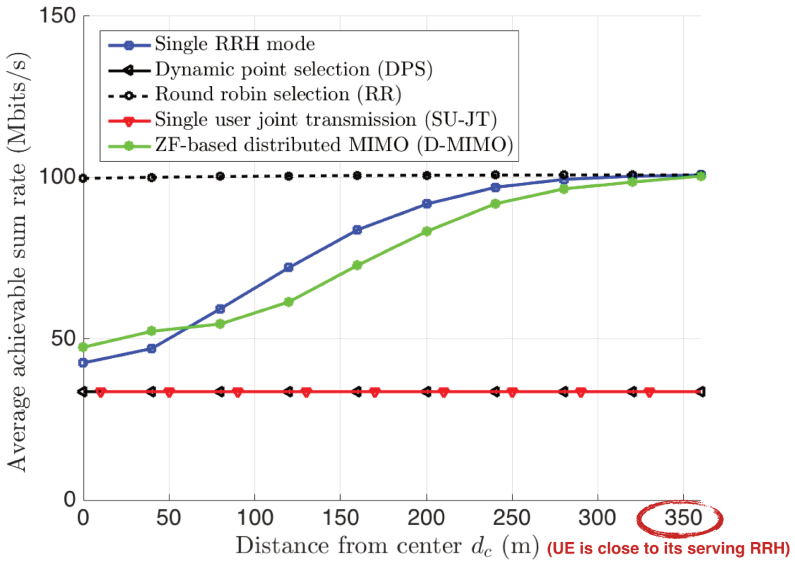

Fig. 5.16 Average achievable sum rate vs. UE distance from the center of RRH 1, 2 and 3 d_c with small fronthaul capacity ($P_{\text{dBm}} = 41.4$ dBm, $\overline{C} = 33.6$ Mbits/s).

When the fronthaul capacity is small ($\overline{C} = 33.6$ Mbits/s, see Figure 5.16), RR have the best performance in small UE distance from center regime. The average achievable sum rates of DPS and SU-JT are limited to be no more than 33.6 Mbits/s. The performances of single RRH mode and ZF-based D-MIMO are similar to each other with different values of UE distance from center.

5.6 Conclusion

We have studied the performance of different coordinated transmission strategies for a cooperation cluster of 3 RRHs serving 3 UEs with RRH power constraints and fronthaul capacity constraints. Each RRH and UE is assumed to be equipped with only one antenna.

With each fronthaul link capacity $\overline{C} = 134.4$ Mbits/s, ZF-based distributed MIMO (D-MIMO) is preferred in RRH high-power regime. When each UE is close to RRH, the performance difference between single RRH mode and ZF-based D-MIMO is small. However, ZF-based D-MIMO needs much more calculation resources, precise channel information feedback from UEs and requires high level of synchronization among coordinated RRHs. It is not interesting to apply ZF-based D-MIMO for only a negligible rate improvement compared with single RRH mode.

When we reduce the fronthaul link capacity to $\overline{C} = 33.6$ Mbits/s, round robin selection (RR) has the best performance in RRH high-power regime and in the case that each UE is close to one different RRH. There is no much performance difference between single RRH mode and ZF-based D-MIMO.

Chapter 6

Analysis of Several User Grouping Algorithms

6.1 Introduction

In Chapter 5, we compare different transmission strategies without and with cooperations between different cells. Simulation results show that distributed MIMO mode achieves the highest sum transmission rates when the fronthaul capacity is high. We apply ZF in D-MIMO for parallel data transmission. In a C-RAN architecture, the CSIs of the UEs served by a cluster of RRHs can be easily shared. Based on the CSIs, the UEs could be appropriately scheduled in order to further improve the performance of D-MIMO.

ZF is known to have low performance when the ratio of the greatest to the smallest singular value of the channel matrix is large [89]. The more linearly-dependent the channel gain vectors are, the larger this ratio is. We assume the U UEs have a total number of antennas larger than the M RRHs. The U UEs are assumed to be randomly distributed at different positions with different shadow fading. We divide these UEs into several groups with equal size, and each group of UEs is served during separate subframe. By appropriately grouping the UEs, the condition of the channel matrix during each subframe can be improved. However, to determine the best grouping choice to get the maximum sum transmission rate requires an exhaustive search over all possible grouping possibilities.

The method to improve the performance of ZF by selecting a subset of UEs among a number of UEs is widely studied in single-cell multi-user Multiple-Input Multiple-Output (MIMO) [90, 91]. In [89], the sub-optimal greedy user selection algorithms are classified into two categories: 1) capacity-based algorithms and 2) Frobenius norm-based algorithms. The capacity based algorithms choose UEs greedily based on the sum rate variation: such as

a UE is added if this increases the sum rate [92][90]. The Frobenius norm-based algorithms choose UEs greedily based on the variation of the condition of channel matrix, such as maximizing the determinant of the composite channel matrix in [93]. Another representative of Frobenius norm-based algorithms is the semi-orthogonal user selection (SUS) algorithm proposed in [94], where the UEs are greedily selected by choosing the UE whose CSI has the largest projected norm on the channel matrices of the already selected UEs. Capacity based algorithms usually achieve higher sum rate than Frobenius norm-based algorithms, because they can guarantee sum rate increment in each step of user selection, while Frobenius norm-based algorithms can not. However, Forbenius norm-based algorithms are less complex as no calculations of the sum rate is done [89].

In this chapter, we propose several user grouping algorithms to improve the performance of ZF in a C-RAN downlink system with limited fronthaul capacity. This corresponds to a multi-cell MIMO scenario. We apply a practical uniform scalar quantization for the analog data transmitted on fronthaul links as in [95], rather than getting results based on information-theoretical quantization methods as in [62, 88, 96]. In D-MIMO mode, we study a cluster of RRHs serving a group of UEs and we apply linear precoding to do joint transmission. Maximizing the sum transmission rate with per RRH power constraint and per fronthaul link capacity constraint is a non-convex optimization problem. Different from [62] and [88] which adopt the Majorization Minimization (MM) scheme to convert the non-convex problem to a series of convex optimization problems, we relax the above mentioned non-convex problem by applying uniform scalar quantization and ZF precoding algorithm and get a closed-form solution.

The remainder of this chapter is organized as follows. In Section 6.2, we introduce the system model and problem formulation. Next our proposed UEs grouping algorithms are described in Section 6.3. The simulation results are presented in Section 6.4. Finally we conclude in Section 6.5.

6.2 System Model and Problem Formulation

We apply the system model presented in Section 4.2. Each UE and each RRH are assumed to be equipped with equal number of antennas N_{UA} and N_{TA}, respectively. Thus $N_{\mathrm{UA}}^{u} = N_{\mathrm{UA}}$, $\forall u \in \mathcal{N}_{\mathrm{UE}}$ and $N_{\mathrm{TA}}^{v} = N_{\mathrm{TA}}$, $\forall v \in \mathcal{N}_{\mathrm{TP}}$. As shown in Figure 6.1, the U UE are disjointly served by the M RRHs during N_T subframes. To simplify the analysis, the number of UEs served during each subframe is the same, $N_{g,t} = U/N_T = N_{g,\mathrm{d}},\ \forall t \in \mathcal{T}$. We assume the M RRHs serve the maximum number of UEs in parallel that they can during each subframe, thus $N_{g,\mathrm{d}} = MN_{\mathrm{TA}}/N_{\mathrm{UA}}$.

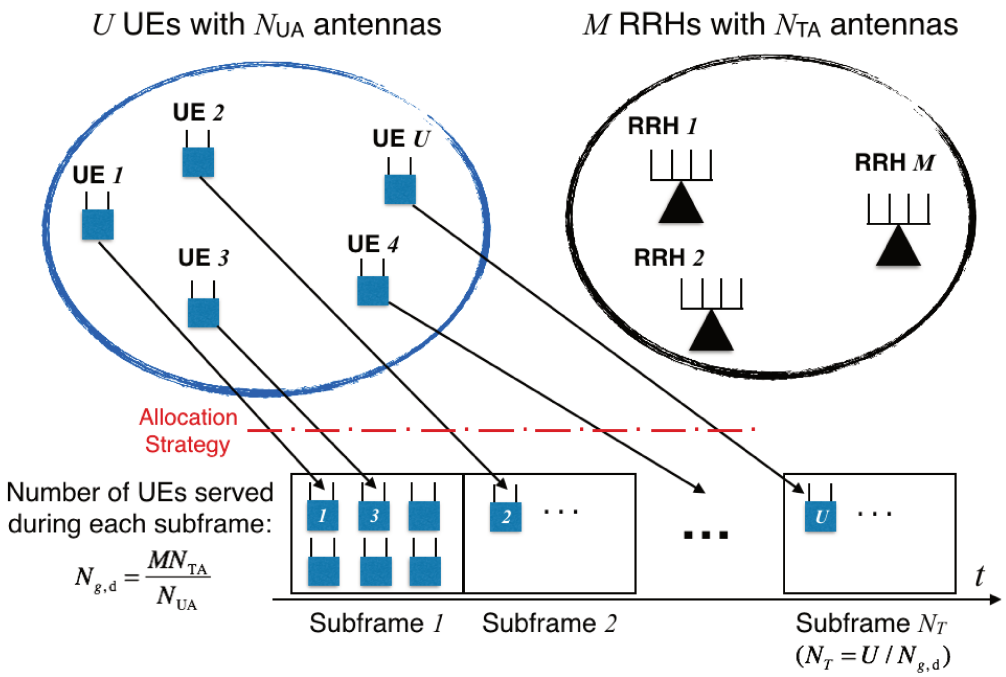

Fig. 6.1 M RRHs serve U UEs during N_T subframes.

Let $\mathcal{G} = \{\mathcal{N}_{\text{UE},1}, \mathcal{N}_{\text{UE},2}, \ldots, \mathcal{N}_{\text{UE},N_T}\}$ be a set of any arbitrary grouping of UEs. For example, $\mathcal{G} = \{\{1, 3\}, \{2, 4\}\}$ is one possible grouping of 4 UEs. As the channel fading is the same during the N_T subframes, the set $\mathcal{G}$ is unordered. Therefore, there is no difference between $\mathcal{G} = \{\{1, 3\}, \{2, 4\}\}$ and $\mathcal{G} = \{\{2, 4\}, \{1, 3\}\}$. We denote all possible groupings of dividing U UEs into disjoint groups with same size $N_{g,\text{d}}$ by an unordered set $\mathcal{K}_U^{N_{g,\text{d}}}$. With $U = 4$, $N_T = 2$ and $N_{g,\text{d}} = 2$, we have $\mathcal{K}_4^2 = \Big\{\{\{1, 2\}, \{3, 4\}\}, \{\{1, 3\}, \{2, 4\}\}, \{\{1, 4\}, \{2, 3\}\}\Big\}$. The number of all possible groupings is given by

$$|\mathcal{K}_U^{N_{g,\text{d}}}| = \prod_{k=1}^{N_T} \begin{pmatrix} N_{g,\text{d}} - 1 \\ kN_{g,\text{d}} - 1 \end{pmatrix} = \frac{\prod_{k=1}^{N_T}(kN_{g,\text{d}} - 1)!}{((N_{g,\text{d}} - 1)!)^{N_T} \prod_{k=1}^{N_T}((k-1)N_{g,\text{d}})!}. \tag{6.1}$$

The UEs in $\mathcal{N}_{\text{UE},t}$ are served together by the RRHs in $\mathcal{N}_{\text{TP}}$ on respecting to the constraints:

$$N_{\text{RL}}^u \leq N_{\text{UA}}, \forall u \in \mathcal{N}_{\text{UE},t}, \tag{6.2}$$

$$\sum_{u \in \mathcal{N}_{\text{UE},t}} N_{\text{RL}}^u \leq MN_{\text{TA}}. \tag{6.3}$$

where N_{RL}^{u} is the number of parallel symbols (layers) transmitted to UE u. In this paper, we assume $N_{\mathrm{RL}}^{u} = N_{\mathrm{UA}}$ and the M RRHs serve the maximum number of UEs that they can during each subframe.

We define the achievable sum rate for a certain set of UEs served during subframe t by

$$R(\mathcal{N}_{\mathrm{UE},t}) = \sum_{u \in \mathcal{N}_{\mathrm{UE},t}} R_u, \tag{6.4}$$

where R_u is the achievable transmission rate for UE u.

Here we consider optimizing the achievable sum-rate for all UEs during each subframe which is defined as

$$R_{\mathrm{tot}} = \frac{1}{N_T} \sum_{t=1}^{N_T} R(\mathcal{N}_{\mathrm{UE},t}). \tag{6.5}$$

The sum-rate can be optimized over subframes allocation for different UEs, the precoding matrix and the compression noise under fronthaul capacity and RRH power constraints. In the following, we formulate the sum-rate optimization problem.

For a certain set of UEs served during subframe t, problem (4.22) is a non-convex optimization problem. In the following, we apply uniform scalar quantization and ZF algorithm to relax this problem. After the relaxation, we get a closed form expression for the maximum achievable transmission rate.

6.2.1 Uniform Scalar Quantization

In practice, to find the quantization codebooks to achieve the sum rate given in (4.21) with the fronthaul capacity constraints given in (4.18) is very difficult. Inspired by [95], we apply uniform scalar quantization technique for the signals to be transmitted from BBUs pool to RRHs and derive the corresponding achievable sum rate. We generalize the case where each RRH is equipped with only one antenna in [95] to multiple antennas in each RRH.

The interface protocol Common Public Radio Interface (CPRI) is currently used for the data transmission between RRHs and BBUs. It applies separate Inphase (I) and Quadrature (Q) quantization. The baseband signal for the v-th RRH can be expressed as

$$\tilde{\mathbf{X}}_{v,t} = \tilde{\mathbf{X}}_{v,t}^{I} + j\tilde{\mathbf{X}}_{v,t}^{Q}. \tag{6.6}$$

Then the I/Q samples are linearly quantized to $2^{Q_{\mathrm{q},v}}$ levels ranging from $-\eta_{v,t}$ to $\eta_{v,t}$, $\forall v \in \mathcal{N}_{\mathrm{TP}}$. The n-th quantization level in RRH v is given by

$$q_{n,v} = \left(-1 + \frac{1}{2^{Q_{\mathrm{q},v}}} + n\frac{2}{2^{Q_{\mathrm{q},v}}}\right)\eta_{v,t}, \tag{6.7}$$

where $n = 0,\ 1,\dots\ 2^{Q_{\mathrm{q},v}} - 1$ and $\eta_{v,t}$ is the scaling factor for the quantized signals transmitted from BBU pool to RRH v during subframe t. With $Q_{\mathrm{q},v}$ bits resolution per complex component, the transmission rate in the v-th fronthaul link is $C_v = 2N_{\mathrm{TA}}Q_{\mathrm{q},v}$ bits/channel.

The lower the quantization noise is, the higher transmission rate can be achieved for the UEs on the radio channel. A higher quantization resolution reduces the quantization noise. Taking the fronthaul capacity constraint $2N_{\mathrm{TA}}Q_{\mathrm{q},v} \leq \overline{C}$ into consideration, to minimize the quantization noise, we take $Q_{\mathrm{q},v} = \lfloor \frac{\overline{C}}{2N_{\mathrm{TA}}} \rfloor$ as the value of the quantization resolution for the v-th fronthaul link, $\forall v \in \mathcal{N}_{\mathrm{TP}}$. In other words, the fronthaul is fully used.

We choose the scaling factor $\eta_{v,t}$ to control the probability of overflow of I/Q samples in a low level. We adapt the "three-sigma rule" applied in [95]. The average power of both Inphase (I) and Quadrature (Q) parts of each element in the k-th row of $\tilde{\mathbf{X}}_{v,t}$ is $\mathbf{V}^r_{t,v}[k](\mathbf{V}^r_{t,v}[k])^{\mathrm{H}}/2$, where $\mathbf{V}^r_{t,v}[k]$ denotes the k-th row of $\mathbf{V}^r_{t,v}$.

Let $p = \max\{\frac{\mathbf{V}^r_{t,v}[1](\mathbf{V}^r_{t,v}[1])^{\mathrm{H}}}{2}, \dots\ \frac{\mathbf{V}^r_{t,v}[N_{\mathrm{TA}}](\mathbf{V}^r_{t,v}[N_{\mathrm{TA}}])^{\mathrm{H}}}{2}\}$. We set

$$\eta_{v,t} = 3\sqrt{p}. \tag{6.8}$$

The probability of overflow for both I-branch and Q-branch samples is expressed as

$$P(|\tilde{\mathbf{X}}^I_{v,t}[i,j]| > \eta_{v,t}) = P(|\tilde{\mathbf{X}}^Q_{v,t}[i,j]| > \eta_{v,t}) \leq 0.0027 \tag{6.9}$$

where $i = 1,\ 2,\dots\ N_T$ and $j = 1,\ 2,\dots\ N_s$.

Each element in $\tilde{\mathbf{X}}_{v,t}$ is linearly quantized as

$$\mathbf{X}^I_{v,t}[i,j] = \underset{n}{\operatorname{argmin}} |q_n - \tilde{\mathbf{X}}^I_{v,t}[i,j]|, \tag{6.10}$$

$$\mathbf{X}^Q_{v,t}[i,j] = \underset{n}{\operatorname{argmin}} |q_n - \tilde{\mathbf{X}}^Q_{v,t}[i,j]| \tag{6.11}$$

where $i = 1,\ 2,\dots\ N_T$ and $j = 1,\ 2,\dots\ N_s$.

As mentioned in [95], according to Widrow Theorem, if the number of quantization levels (i.e. $2^{Q_{\mathrm{q},v}}$) is large, and the signal varies by at least some quantization levels from sample to sample, the quantization noise can be assumed to be uniformly distributed. The

quantization step size is $\frac{2}{2^{Q_{\mathrm{q},v}}} = 2^{1-Q_{\mathrm{q},v}}$. The corresponding quantization noise variance is $\sigma_{v,t}^2 = \frac{(\eta_{v,t})^2 2^{2-2Q_{\mathrm{q},v}}}{6}$ according to Appendix C in [97].

Applying linear quantization with a quantization resolution $Q_{\mathrm{q},v} = \lfloor \frac{\overline{C}}{2N_{\mathrm{TA}}} \rfloor$, problem (4.22) can be modified to

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{u\in\mathcal{N}_{\mathrm{UE},t}} R_u \\ \text{over} \quad & \mathbf{V}_t,\ \forall t \in \mathcal{T} \\ \text{s.t.} \quad & P_{v,t}(\mathbf{V}_{t,v}^r, \sigma_{v,t}^2) \le \overline{P}_v,\ \forall v \in \mathcal{N}_{\mathrm{TP}} \end{aligned} \tag{6.12}$$

where

$$\sigma_{v,t}^2 = \frac{(\eta_{v,t})^2 2^{2-2Q_{\mathrm{q},v}}}{6},\ \forall v \in \mathcal{N}_{\mathrm{TP}} \tag{6.13}$$

$$\eta_{v,t} = 3\sqrt{\max\{\frac{\mathbf{V}_{t,v}^r[1](\mathbf{V}_{t,v}^r[1])^{\mathrm{H}}}{2}, \dots \frac{\mathbf{V}_{t,v}^r[N_{\mathrm{TA}}](\mathbf{V}_{t,v}^r[N_T])^{\mathrm{H}}}{2}\}} \tag{6.14}$$

$$Q_{\mathrm{q},v} = \lfloor \frac{\overline{C}}{2N_{\mathrm{TA}}} \rfloor. \tag{6.15}$$

6.2.2 Zero Forcing

Applying ZF algorithm, the precoding matrix during subframe t is given as:

$$\mathbf{V}_t = \gamma_t \mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}}^{\mathrm{H}} (\mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}} \mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}}^{\mathrm{H}})^{-1} \tag{6.16}$$

where $\mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}} = [\mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}[1]};\ \mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}[2]};\ \dots;\ \mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}[N_{g,\mathrm{d}}]}]$ and γ_t is a normalization factor which is selected to satisfy the RRH power constraint during subframe t.

Applying $\tilde{\mathbf{V}}_t = \mathbf{V}_t/\gamma_t$, $\tilde{\sigma}_{v,t} = \sigma_{v,t}/\gamma_t$, $N_{\mathrm{UA}}^u = N_{\mathrm{UA}}\ \forall u \in \mathcal{N}_{\mathrm{UE}}$ and (6.16) to (4.21), we get

$$R_u = \log \frac{\det(\mathbf{I}_{N_{\mathrm{UA}}} + \gamma_t^2 \mathbf{I}_{N_{\mathrm{UA}}} + \gamma_t^2 \mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\mathrm{H}})}{\det(\mathbf{I}_{N_{\mathrm{UA}}} + \gamma_t^2 \mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\mathrm{H}})},\ \forall u \in \mathcal{N}_{\mathrm{UE},t} \tag{6.17}$$

where

$$\sigma_{v,t}^2 = \frac{\gamma_t^2 (\tilde{\eta}_{v,t})^2 2^{2-2Q_{\mathrm{q},v}}}{6} = \gamma_t^2 \tilde{\sigma}_{v,t}^2,\ \forall v \in \mathcal{N}_{\mathrm{TP}} \tag{6.18}$$

$$\tilde{\eta}_{v,t} = 3\sqrt{\max\{\frac{\tilde{\mathbf{V}}_{t,v}^r[1](\tilde{\mathbf{V}}_{t,v}^r[1])^{\mathrm{H}}}{2}, \dots \frac{\tilde{\mathbf{V}}_{t,v}^r[N_{\mathrm{TA}}](\tilde{\mathbf{V}}_{t,v}^r[N_{\mathrm{TA}}])^{\mathrm{H}}}{2}\}} \tag{6.19}$$

$$\tilde{\mathbf{V}}_t = \mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}}^{\mathrm{H}} (\mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}} \mathbf{H}_{\mathcal{N}_{\mathrm{UE},t}}^{\mathrm{H}})^{-1} \tag{6.20}$$

$$\tilde{\Omega} = \Omega/\gamma_t^2 \tag{6.21}$$

$$Q_{\mathrm{q},v} = \lfloor \frac{\overline{C}}{2N_{\mathrm{TA}}} \rfloor. \tag{6.22}$$

Applying $\tilde{\mathbf{V}}_t$ and $\tilde{\sigma}_{v,t}$ to (4.16), we get

$$P_{v,t}(\mathbf{V}_t, \sigma_{v,t}^2) = \gamma_t^2 \mathrm{tr}\big(\tilde{\mathbf{V}}_{t,v}^r (\tilde{\mathbf{V}}_{t,v}^r)^{\mathrm{H}} + \tilde{\sigma}_{v,t}^2 \mathbf{I}_{N_{\mathrm{TA}}}\big). \tag{6.23}$$

Then problem (4.22) can be rewritten as

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{u\in\mathcal{N}_{\mathrm{UE},t}} R_u = \log \frac{\det(\mathbf{I}_{N_{\mathrm{UA}}} + \gamma_t^2 \mathbf{I}_{N_{\mathrm{UA}}} + \gamma_t^2 \mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\mathrm{H}})}{\det(\mathbf{I}_{N_{\mathrm{UA}}} + \gamma_t^2 \mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\mathrm{H}})} \\ \text{over} \quad & \gamma_t^2 \\ \text{s.t.} \quad & \gamma_t^2 \mathrm{tr}\big(\tilde{\mathbf{V}}_{t,v}^r (\tilde{\mathbf{V}}_{t,v}^r)^{\mathrm{H}} + \tilde{\sigma}_{v,t}^2 \mathbf{I}_{N_{\mathrm{TA}}}\big) \le \overline{P} \ \forall v \in \mathcal{N}_{\mathrm{TP}} \end{aligned} \tag{6.24}$$

The multiple constraints in problem (6.24)

$$\gamma_t^2 \le \frac{\overline{P}}{\mathrm{tr}\big(\tilde{\mathbf{V}}_{t,v}^r (\tilde{\mathbf{V}}_{t,v}^r)^{\mathrm{H}} + \tilde{\sigma}_{v,t}^2 \mathbf{I}_{N_{\mathrm{TA}}}\big)} \ \forall v \in \mathcal{N}_{\mathrm{TP}} \tag{6.25}$$

can be adapted to one constraint

$$\gamma_t^2 \le \min\{\frac{\overline{P}}{\mathrm{tr}\big(\tilde{\mathbf{V}}_{t,1}^r (\tilde{\mathbf{V}}_{t,1}^r)^{\mathrm{H}} + \tilde{\sigma}_{1,t}^2 \mathbf{I}_{N_{\mathrm{TA}}}\big)}, \dots \frac{\overline{P}}{\mathrm{tr}\big(\tilde{\mathbf{V}}_{t,M}^r (\tilde{\mathbf{V}}_{t,M}^r)^{\mathrm{H}} + \tilde{\sigma}_{M,t}^2 \mathbf{I}_{N_{\mathrm{TA}}}\big)}\}. \tag{6.26}$$

Theorem 1. *Problem (6.24) can be solved by*

$$\gamma_t^2 = \min\{\frac{\overline{P}}{tr\big(\tilde{\mathbf{V}}_{t,1}^r (\tilde{\mathbf{V}}_{t,1}^r)^{H} + \tilde{\sigma}_{1,t}^2 \mathbf{I}_{N_{TA}}\big)}, \dots \frac{\overline{P}}{tr\big(\tilde{\mathbf{V}}_{t,M}^r (\tilde{\mathbf{V}}_{t,M}^r)^{H} + \tilde{\sigma}_{M,t}^2 \mathbf{I}_{N_{TA}}\big)}\}. \tag{6.27}$$

Proof: See Appendix A.

Therefore, the problem of maximizing R_{tot} can be formulated as below:

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{t=1}^{N_T} \sum_{u\in\mathcal{N}_{\mathrm{UE},t}} \log\Big(1 + \frac{\gamma_t^2}{1+\gamma_t^2 \mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\mathrm{H}}}\Big) \\ \text{over} \quad & \mathcal{G} = \{\mathcal{N}_{\mathrm{UE},1}, \mathcal{N}_{\mathrm{UE},2}, \dots, \mathcal{N}_{\mathrm{UE},N_{g,\mathrm{d}}}\} \in \mathcal{K}_U^{N_{g,\mathrm{d}}} \end{aligned} \tag{6.28}$$

where

$$\gamma_t^2 = \min\{\frac{\overline{P}}{\mathrm{tr}\big(\tilde{\mathbf{V}}_{t,1}^r (\tilde{\mathbf{V}}_{t,1}^r)^{\mathrm{H}} + \tilde{\sigma}_{1,t}^2 \mathbf{I}_{N_{\mathrm{TA}}}\big)}, \dots \frac{\overline{P}}{\mathrm{tr}\big(\tilde{\mathbf{V}}_{t,M}^r (\tilde{\mathbf{V}}_{t,M}^r)^{\mathrm{H}} + \tilde{\sigma}_{M,t}^2 \mathbf{I}_{N_{\mathrm{TA}}}\big)}\} \tag{6.29}$$

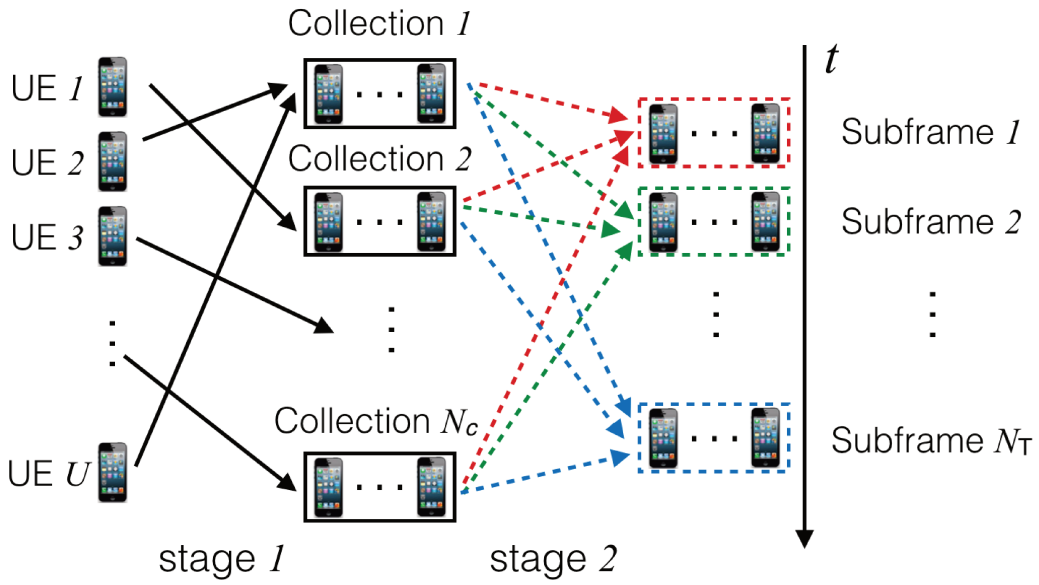

Fig. 6.2 Process of allocating UEs to be served in different subframes with UDA.

6.3 User Grouping Algorithm

Problem (6.28) can be solved by an exhaustive search over all possible UEs groupings. This brute force algorithm is simple with a small number of UEs U. Nevertheless, the number of all possible UEs groupings increases exponentially with U for a fixed $N_{g,\mathrm{d}} > 1$. To reduce the complexity, we propose different algorithms that solve these optimization problems approximately.

We adapt the iterative capacity based greedy algorithm in [98] to a C-RAN system, and name it "Greedy User Grouping Algorithm" (GUGA). We present it in Section 6.3.1. In Section 6.3.2, we adapt the Semi-orthogonal User Selection (SUS) algorithm proposed in [94] to an UEs grouping algorithm named "Semi-orthogonal User Grouping Algorithm" (SUGA).

We also propose two new two-stage UEs grouping algorithms. An exhaustive search of all UEs groupings to find the best one is complex. We thus decide to first restrict the selection of UEs grouping into a certain region. The process of these two-stage UEs grouping algorithms is shown in Figure 6.2. In the first stage, the U UEs are partitioned into N_c *collections* with the same size N_{d}. Then, in the second stage, UEs in the same *collection* are assigned to different subframes. Thus, a collection of UEs is defined as the UEs that are not served during the same subframe. To get a larger achievable sum rate, we should let the channel matrices between all the RRHs and each UE served in the same subframe be less linearly

dependent. We put the UEs with similar channel gain to the same collection and avoid that they are served on the same time.

During a given subframe, there is exactly one UE in each collection that is served. Thus the number of UEs in each collection is $N_{\mathrm{d}} = N_T = U/N_{g,\mathrm{d}}$. As the group of UEs served during the same subframe contains one UE from each collection, the number of collections equals to the number of UEs served during each subframe is $N_c = N_{g,\mathrm{d}} = \frac{MN_{\mathrm{TA}}}{N_{\mathrm{UA}}}$. To simplify the analysis, we assume that the value of $\frac{N_{\mathrm{TA}}}{N_{\mathrm{UA}}}$ is always an integer.

We denote the i-th collection as $\mathcal{M}_i$. Thus, $\bigcup_{i=1}^{N_c} \mathcal{M}_i = \mathcal{N}_{\mathrm{UE}}$ and $\mathcal{M}_i \bigcap \mathcal{M}_j = \emptyset, \forall i \neq j$. Let $\mathcal{M}_U^{N_{\mathrm{c}}}$ be the unordered set of all possible groupings of partitioning U UEs belonging to N_{c} collections into groups with same size $N_{g,\mathrm{d}}$ with the above mentioned constraint. The cardinal of $\mathcal{M}_U^{N_{\mathrm{c}}}$ is given by

$$|\mathcal{M}_U^{N_{\mathrm{c}}}| = \left(\frac{U}{N_{\mathrm{c}}}!\right)^{N_{\mathrm{c}}-1} \tag{6.30}$$

For example, with $M = 3$, $U = 6$, $N_c = 3$, we could have $\mathcal{M}_1 = \{1,\ 4\}$, $\mathcal{M}_2 = \{2,\ 6\}$ and $\mathcal{M}_3 = \{3,\ 5\}$. In this case, UE 1 and UE 4 are in the same collection, they cannot be served during the same subframe. The same applies also to the UEs in $\mathcal{M}_2$ and $\mathcal{M}_3$. With this partition of collections,

$$\begin{aligned}\mathcal{M}_6^3 = \Big\{ &\big\{\{1,\ 2,\ 3\},\{4,\ 6,\ 5\}\big\}, \big\{\{1,\ 2,\ 5\},\{4,\ 6,\ 3\}\big\}, \\ &\big\{\{1,\ 6,\ 3\},\{4,\ 2,\ 5\}\big\}, \big\{\{1,\ 6,\ 5\},\{4,\ 2,\ 3\}\big\}\Big\}.\end{aligned} \tag{6.31}$$

And we have $|\mathcal{M}_U^{N_{\mathrm{c}}}| = (\frac{6}{3}!)^{3-1} = 4$.

We propose an algorithm called "User Division Algorithm" (UDA) to do the first stage process which partitions the UEs to different collections. UDA and the motivation to apply UDA are introduced in Section 6.3.3. For the second stage, we apply separately random selection and GUGA for the two different proposed two-stage UEs grouping algorithms. We call the one which combines UDA and random selection by " Pre Partitioned - Random User Grouping Algorithm" (PP-RUGA); the one which combines UDA and GUGA by "Pre Partitioned - Greedy User Grouping Algorithm"(PP-GUGA). We present the two proposed two-stage UEs grouping algorithms in Section 6.3.4. Among the above mentioned UEs grouping algorithms, PP-RUGA and SUGA are Frobenius norm-based algorithms, GUGA and PP-GUGA are capacity-based algorithms. All the four algorithms are summarized in Table 6.1.

Table 6.1 Summary of different UEs grouping algorithms.

Acronym	Meaning	Number of stages	First stage	Second stage
GUGA	Greedy user grouping algorithm	1	GUGA	
SUGA	Semi-orthogonal user grouping algorithm		SUGA	
PP-RUGA	Pre partitioned - random user grouping algorithm	2	UDA	random selection
PP-GUGA	Pre partitioned - greedy user grouping algorithm			GUGA

6.3.1 Greedy User Grouping Algorithm (GUGA)

In this section, we describe GUGA algorithm. For each turn, the U UEs are assigned separately to N_T different subframes to be served. The GUGA algorithm has a number of iterations which is equal to N_T. In each iteration, we assign $N_{g,\mathrm{d}}$ UEs to one subframe. Let $\mathcal{J}$ be the set of all possible unordered subsets of size $N_{g,\mathrm{d}}$ of $\mathcal{N}_{\mathrm{UE}}$. The size of $\mathcal{J}$ is

$$|\mathcal{J}| = \begin{pmatrix} N_{g,\mathrm{d}} \\ U \end{pmatrix} = \frac{U!}{(U-N_{g,\mathrm{d}})!N_{g,\mathrm{d}}!}. \tag{6.32}$$

Meanwhile, we denote the set of all possible unordered subsets of size $N_{g,\mathrm{d}}$ of the still unassigned UEs at the beginning of the k-th iteration by $\mathcal{J}^{(k)}$. Thus, we have $\mathcal{J}^{(1)} = \mathcal{J}$.

Before starting the iterations, we first calculate the value of $R(\mathcal{S})$ for each $\mathcal{S} \in \mathcal{J}$ by (6.4), (6.17) and (6.27), and save them.

In the k-th iteration, we find the set of UEs $\mathcal{S} \in \mathcal{J}^{(k)}$ that has the largest value of $R(\mathcal{S})$ (achievable sum rate of $\mathcal{S}$). This set of UEs is assigned to be served during subframe k and is denoted by $\mathcal{S}_k$. Thus, we have

$$\mathcal{S}_k = \underset{\mathcal{S}\in\mathcal{J}^{(k)}}{\operatorname{argmax}} R(\mathcal{S}). \tag{6.33}$$

Then we remove the sets in $\mathcal{J}^{(k)}$ which contain any element in $\mathcal{S}_k$ to get $\mathcal{J}^{(k+1)}$, i.e.

$$\mathcal{J}^{(k+1)} = \left\{ \mathcal{S} | \mathcal{S} \in \mathcal{J}^{(k)} \text{ and } i \notin \mathcal{S}\ \forall i \in \mathcal{S}_k \right\}. \tag{6.34}$$

Next we move onto the $(k+1)$-th iteration and terminate at the end of the N_T-th iteration. At last, the chosen grouping of UEs is given by $\tilde{G} = \{\mathcal{S}_1,\ \mathcal{S}_2,\ ...,\ \mathcal{S}_{N_T}\}$.

6.3.2 Semi-orthogonal User Grouping Algorithm (SUGA)

Semi-orthogonal User Grouping Algorithm (SUGA) is an extension of the user grouping algorithm in [94] to the scenario where both RRHs and UEs are equipped with multiple antennas. The basic idea of SUGA is to greedily construct a channel matrix whose rows are as orthogonal as possible to each other.

SUGA is an iterative algorithm which assigns the U UEs to N_T contiguous subframes. We denote the set of unallocated UEs as $\mathcal{K}$ and the set of allocated UEs during subframe k as $\mathcal{S}_k$.

Let $\mathbf{H}_u = [\mathbf{h}_{u,1}^{\mathrm{T}},\ \mathbf{h}_{u,2}^{\mathrm{T}},\ ...,\ \mathbf{h}_{u,N_{\mathrm{UA}}}^{\mathrm{T}}]$, where $\mathbf{h}_{u,i}$ is the channel vector from the RRHs to the i-th antenna of UE u. We initialize $n = 1$.

For each i-th antenna of each UE $u \in \mathcal{K}$, calculate the component of $\mathbf{h}_{u,i}$ orthogonal to the subspace spanned by $\{\tilde{\mathbf{g}}_1,\ ...,\ \tilde{\mathbf{g}}_{(n-1)N_{\mathrm{UA}}+i-1}\}$:

$$\mathbf{g}_{(u-1)N_{\mathrm{UA}}+i} = \mathbf{h}_{u,i} - \sum_{j=1}^{(n-1)N_{\mathrm{UA}}+i-1} \frac{\mathbf{h}_{u,i}\tilde{\mathbf{g}}_j^{\mathrm{H}}}{\|\tilde{\mathbf{g}}_j\|^2}\tilde{\mathbf{g}}_j \tag{6.35}$$

$$\tilde{\mathbf{g}}_{(n-1)N_{\mathrm{UA}}+i} = \mathbf{g}_{(u-1)N_{\mathrm{UA}}+i}. \tag{6.36}$$

When $n = 1$, this implies $\mathbf{g}_{(u-1)N_{\mathrm{UA}}+i} = \mathbf{h}_{u,i}$, where $i = 1,\ 2,\ ...,\ N_{\mathrm{UA}}$.

Let $\tilde{u}$ be the UE in $\mathcal{K}$ that maximizes $\sum_{i=1}^{N_{\mathrm{UA}}} \|\mathbf{g}_{(u-1)N_{\mathrm{UA}}+i}\|^2$. We remove UE $\tilde{u}$ from $\mathcal{K}$, put it in $\mathcal{S}_k$ and increment n. Next, we select another UE with the same process until $N_{g,\mathrm{d}}$ UEs are selected.

Once $N_{g,\mathrm{d}}$ UEs are assigned to subframe k, we increment k, reset $n = 1$, $\mathcal{S}_k = \emptyset$ and allocate another $N_{g,\mathrm{d}}$ UEs to the next subframe with the same procedure. After allocating all the U UEs, we get $\tilde{G} = \{\mathcal{S}_1,\ \mathcal{S}_2,\ ...,\ \mathcal{S}_{N_T}\}$, which is the grouping of the UEs.

SUGA is summarized in Algorithm 1.

6.3.3 User Division Algorithm (UDA) and Motivation to Apply UDA

UDA

The objective of UDA is to generate $N_c = N_{g,\mathrm{d}}$ collections of same size N_T. In a collection, the UEs are served in different subframes. UDA allocates the UEs that get the highest sum channel gain from N_{UA} antennas of a RRH into the same collection. It firstly calculates a reference matrix $\mathbf{Z} \in \mathbb{R}^{[U \times N_c]}$. Let $r = N_{\mathrm{TA}}/N_{\mathrm{UA}}$. Each element of the matrix is

$$\mathbf{Z}_{u,l} = \sum_{i=(u-1)N_{\mathrm{UA}}+1}^{uN_{\mathrm{UA}}} \sum_{j=(l-1)r+1}^{lr} \|\mathbf{H}(i,j)\|^2 \tag{6.37}$$

Algorithm 1 SUGA: Semi-orthogonal User Grouping Algorithm

Input: $\mathbf{H}$ (Channel matrix), M (No. of RRHs), U (No. of UEs), $N_{g,\mathrm{d}}$ (No. of UEs served during each subframe), N_{UA} (No. of antennas of each UE), $\mathcal{N}_{\mathrm{UE}}$ (Set of UEs to be served)

Output: $\tilde{\mathcal{G}}$ (UEs grouping)

Initialisation : Let $\mathcal{S}_k = \emptyset$ (Set of UEs to be served during subframe k), where $k = 1,\ 2,\ ...,\ U/N_{g,\mathrm{d}}$

$\mathbf{h}_{u,i} = [\mathbf{H}_{(u-1)N_{\mathrm{UA}}+i,1},\ ...,\ \mathbf{H}_{(u-1)N_{\mathrm{UA}}+i,MN_{\mathrm{TA}}}]$, where $u = 1,\ 2,\ ...,\ U$ and $i = 1,\ 2,\ ...,\ N_{\mathrm{UA}}$
$\mathcal{K} = \mathcal{N}_{\mathrm{UE}}$ (Set of UEs still unallocated)
for $k = 1$ to $U/N_{g,\mathrm{d}}$ **do**
  for $l = 1$ to $N_{g,\mathrm{d}}$ **do**
    if $l = 1$ **then**
      for $m = 1$ to $|\mathcal{K}|$ **do**
        for $n = 1$ to N_{UA} **do**
          $\mathbf{g}_{(m-1)N_{\mathrm{UA}}+n} = \mathbf{h}_{\mathcal{K}[m],n}$
        end for
      end for
    else
      for $m = 1$ to $|\mathcal{K}|$ **do**
        for $n = 1$ to N_{UA} **do**
          $\mathbf{g}_{(m-1)N_{\mathrm{UA}}+n} = \mathbf{h}_{\mathcal{K}[m],n} - \sum_{i=1}^{(l-1)N_{\mathrm{UA}}+n-1} \frac{\mathbf{h}_{\mathcal{K}[m],n}\tilde{\mathbf{g}}_i^{\mathrm{H}}}{\|\tilde{\mathbf{g}}_i\|^2}\tilde{\mathbf{g}}_i$
          $\tilde{\mathbf{g}}_{(l-1)N_{\mathrm{UA}}+n} = \mathbf{g}_{(m-1)N_{\mathrm{UA}}+n}$
        end for
      end for
    end if
    Let $\mathcal{F} = \{1,\ 2,\ ...,\ |\mathcal{K}|\}$
    $\tilde{m} = \mathrm{argmax}_{m\in\mathcal{F}} \sum_{i=1}^{N_{\mathrm{UA}}} \|\mathbf{g}_{(m-1)N_{\mathrm{UA}}+i}\|^2$ /* Select the UE. */
    for $i = 1$ to N_{UA} **do**
      $\tilde{\mathbf{g}}_{(l-1)N_{\mathrm{UA}}+i} = \mathbf{g}_{(\tilde{m}-1)N_{\mathrm{UA}}+i}$
    end for
    $\mathcal{S}_k \leftarrow \mathcal{S}_k \cup \{\mathcal{K}[\tilde{m}]\}$ /* Add $\mathcal{K}[\tilde{m}]$ to $\mathcal{S}_k$. */
    $\mathcal{K} \leftarrow \mathcal{K}/\mathcal{K}[\tilde{m}]$
  end for
end for
return $\tilde{G} = \{\mathcal{S}_1,\ \mathcal{S}_2,\ ...,\ \mathcal{S}_{N_T}\}$

where $u = 1,\ 2,\ \ldots,\ U$ and $l = 1,\ 2,\ \ldots,\ N_c$.

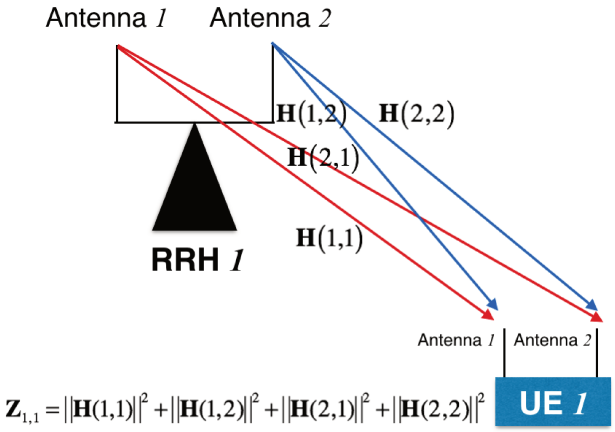

Fig. 6.3 An example of the calculation of one element of the reference matrix.

Note that each element in $\mathbf{Z}$ is the sum of the channel gain from each of N_{UA} antennas in a RRH to each antenna in a UE. Figure 6.3 shows an example of the calculation of one element in the reference matrix. In this example, each RRH and each UE is equipped with 2 antennas ($N_{\mathrm{TA}} = N_{\mathrm{UA}} = 2$).

Secondly, UDA initializes $\mathcal{N}_{\mathrm{c}} = \{1,\ 2,\ \ldots,\ N_c\}$ and the N_c collections all to be empty sets: $\mathcal{M}_i = \emptyset\ \forall i \in \mathcal{N}_{\mathrm{c}}$. Then UDA does U iterations to assign each of the U UEs to a collection.

In each iteration, UDA finds the index of the maximum value in $\mathbf{Z}$:

$$\{m,n\} = \underset{i\in\mathcal{N}_{\mathrm{UE}}, j\in\mathcal{N}_{\mathrm{TP}}}{\operatorname{argmax}} \mathbf{Z}_{i,j}. \tag{6.38}$$

Then UE m is added to the set $\mathcal{M}_n$ and removed from $\mathcal{N}_{\mathrm{UE}}$. When the n-th collection is full ($\mathcal{M}_n = U/N_c$), we delete collection n from $\mathcal{N}_{\mathrm{c}}$.

UDA is summarized in Algorithm 2.

Motivation to Apply UDA

In the following, we present several examples to show the sensitivity of the average available sum rate w.r.t. the chosen UEs grouping.

Here we let $M = 3$, $U = 9$, $N_{\mathrm{TA}} = 1$ and $N_{\mathrm{UA}} = 1$. Thus, we have $N_{g,\mathrm{d}} = MN_{\mathrm{TA}}/N_{\mathrm{UA}} = 3$ and $N_T = U/N_{g,\mathrm{d}} = 3$. In this case, applying Equation (6.1), the number of all possible combinations $|\mathcal{K}_U^{N_{g,\mathrm{d}}}| = |\mathcal{K}_9^3| = 280$.

Algorithm 2 UDA: UEs Division Algorithm

Input: $\mathbf{H}$ (Channel matrix), M (No. of RRHs), U (No. of UEs), N_{TA} (No. of antennas of each RRH, N_{UA} (No. of antennas of each UE

Output: $\{\mathcal{M}_1,\ \mathcal{M}_2,\ ...,\ \mathcal{M}_{N_c}\}$ (UEs collections)

Initialisation : Let the set of all UEs $\mathcal{N}_{\mathrm{UE}} = \{1,\ 2,\ ...,\ U\}$

Let $N_c = MN_{\mathrm{TA}}/N_{\mathrm{UA}}$ /∗ Calculate the number of collections ∗/

Let $\mathcal{N}_{\mathrm{c}} = \{1,\ 2,\ ...,\ N_c\}$, $\mathcal{M}_i = \emptyset$ (UEs collections), where $i = 1,\ 2,\ ...,\ N_c$

Let $r = N_{\mathrm{TA}}/N_{\mathrm{UA}}$

Let $\mathbf{Z} \in \mathbb{R}^{[U \times N_c]}$ be a reference matrix,
where $\mathbf{Z}_{u,l} = \sum_{i=(u-1)N_{\mathrm{UA}}+1}^{uN_{\mathrm{UA}}} \sum_{j=(l-1)r+1}^{lr} \|\mathbf{H}(i,j)\|^2$

for $k = 1$ to U **do**

  $\{m,n\} = \mathrm{argmax}_{i \in \mathcal{N}_{\mathrm{UE}}, j \in \mathcal{N}_{\mathrm{C}}} \mathbf{Z}_{i,j}$

  Assign UE m to the set $\mathcal{M}_n$

  $\mathcal{N}_{\mathrm{UE}} \leftarrow \mathcal{N}_{\mathrm{UE}}/\{m\}$

  if $|\mathcal{M}_n| = U/N_c$ **then**

    $\mathcal{N}_{\mathrm{c}} \leftarrow \mathcal{N}_{\mathrm{c}}/\{n\}$ /∗ When a collection already has U/N_c elements, it should not be assigned with new UEs any more.∗/

  end if

end for

return $\{\mathcal{M}_1,\ \mathcal{M}_2,\ ...,\ \mathcal{M}_{N_c}\}$

UDA partitions the U UEs into $N_c = N_{g,\mathrm{d}} = 3$ collections with same size $N_{\mathrm{d}} = N_T = 3$. The number of all possible UEs groupings with the constraint that UEs in the same collection cannot be served during the same subframe is $|\mathcal{M}_U^{N_c}| = |\mathcal{M}_9^3| = 36$ (applying (6.30)).

We apply the simulation configuration presented in Section 6.2 and set the RRH power constraint to $\overline{P} = 41.86$ dBm. Each fronthaul is assumed to have unlimited transmission capacity. In each simulation, we randomly generate $U = 9$ UEs located at different positions which are served by the cluster of the 3 RRHs with a random channel matrix. For the given channel matrix in this simulation, we calculate the sum rate with each of the $|\mathcal{K}_9^3| = 280$ UEs groupings using (6.17) and (6.27). Then we order the 280 UEs groupings $\mathcal{K}_9^3$ in increasing order of the average achievable sum rates achieved by them. Next, we do the same process to the $|\mathcal{M}_9^3| = 36$ UEs groupings. After the reordering, each i-th element of $\mathcal{K}_U^{N_{g,\mathrm{d}}}$ is labeled with the number i. $\mathcal{M}_9^3$ is a subset of $\mathcal{K}_9^3$. Next, we let each element in $\mathcal{M}_9^3$ have the same label as it has in $\mathcal{K}_9^3$.

We perform 5 different simulations, and plot the average achievable sum rate versus the labels of each element in $\mathcal{K}_9^3$ and $\mathcal{M}_9^3$ of each simulation in Figure 6.4. We can observe that there is a wide variation of average achievable sum rate for different UEs groupings in each simulation. The highest average achievable sum rate can be twice as the lowest. Therefore, it is interesting to find a good UE grouping to increase the sum rate. $\mathcal{M}_9^3$ is a set of UEs

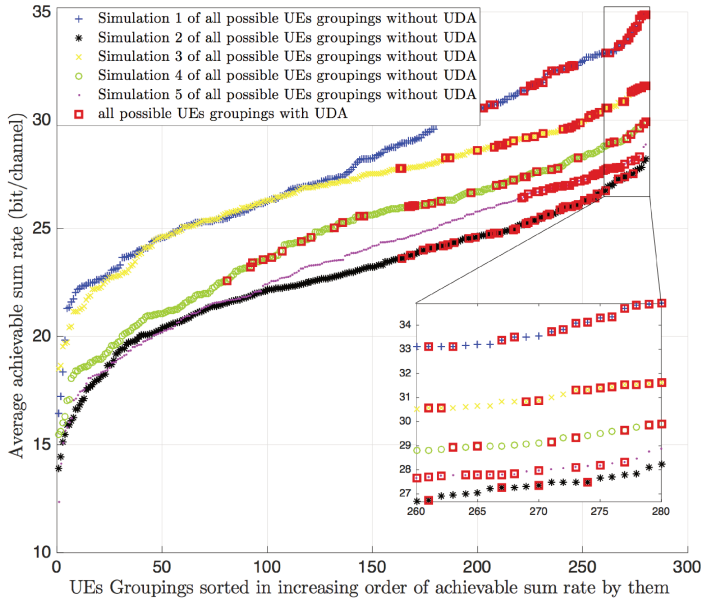

Fig. 6.4 Sensitivity of the achievable sum rate towards the chosen grouping of UEs for the channel matrix.

groupings selected from $\mathcal{K}_9^3$ by UDA. We can observe that the UEs groupings in $\mathcal{M}_9^3$ mainly locate in the high average achievable sum rate regime of $\mathcal{K}_9^3$. To select one UEs grouping from $\mathcal{M}_9^3$ gives a larger chance to get a high sum rate than from $\mathcal{K}_9^3$ and also gives a much lower chance to get a very low sum rate. Moreover, we have large possibility to get the maximum or near maximum sum rate by selecting one UEs grouping from the $|\mathcal{M}_9^3| = 36$ UEs groupings rather than from the $|\mathcal{K}_9^3| = 280$ UEs groupings.

6.3.4 Two-Stage UEs Grouping Algorithms

In the following, we present the two proposed pre partitioned UEs grouping algorithms separately.

Pre Partitioned - Random User Grouping Algorithm (PP-RUGA)

In Section 6.3.3, it is shown that, by applying UDA, the selecting of one UE from each collection increases the chance to form a group of UEs which achieve a high sum rate. A simply random selection can already increase the system performance. PP-RUGA applies a pseudo-random way to select one UE from each collection to be served during each subframe. After the assignment for each collection by UDA, we sort $\mathcal{M}_i$ into increasing order, where

$i = 1,\ 2, \ldots\ N_c$. Then we select the k-th UE from each collection to be served during the k-th subframe, $\forall k \in \mathcal{T}$. PP-RUGA is summarized in Algorithm 3.

Algorithm 3 PP-RUGA: Pre Partitioned - Random User Grouping Algorithm

Input: $\mathbf{H}$ (Channel matrix), M (No. of RRHs), U (No. of UEs), $N_{g,\mathrm{d}}$ (No. of UEs served during each subframe), N_{UA} (No. of antennas of each UE), N_{TA} (No. of antennas of each RRH)

Output: $\tilde{\mathcal{G}}$ (UEs grouping)

Initialisation : $\mathcal{S}_k = \emptyset$ (Set of UEs to be served during subframe k), where $k = 1,\ 2,\ ...,\ U/N_{g,\mathrm{d}}$

1: $N_c = MN_{\mathrm{TA}}/N_{\mathrm{UA}}$ /* Number of collections */
2: $\{\mathcal{M}_1,\ \mathcal{M}_2,\ ...,\ \mathcal{M}_{N_c}\} = \mathrm{UDA}(\mathbf{H}, M, U, N_{\mathrm{UA}}, N_{\mathrm{TA}})$
3: Sort $\mathcal{M}_i$ into increasing order, $\forall i \in \mathcal{N}_{\mathrm{TP}}$
4: **for** $k = 1$ to $U/N_{g,\mathrm{d}}$ **do**
5: **for** $l = 1$ to N_c **do**
6: $\mathcal{S}_k = \mathcal{S}_k \bigcup \left(\bigcup_{i=(k-1)N_{g,\mathrm{d}}/M+1}^{kN_{g,\mathrm{d}}/M} \{\mathcal{M}_l[i]\} \right)$ /* Assign one UE from each collection to be served during subframe k. */
7: **end for**
8: **end for**
9: **return** $\tilde{G} = \{\mathcal{S}_1,\ \mathcal{S}_2,\ ...,\ \mathcal{S}_{N_T}\}$

Pre Partitioned - Greedy User Grouping Algorithm (PP-GUGA)

Pre Partitioned - Greedy User Grouping Algorithm (PP-GUGA) is a combination of UDA and GUGA. At first, PP-GUGA uses UDA to divide the U UEs into $N_c = MN_{\mathrm{TA}}/N_{\mathrm{UA}}$ collections. Then, it gets the set of all possible unordered subsets of size $N_{g,\mathrm{d}}$ of $\mathcal{N}_{\mathrm{UE}}$ with the constraint that each subset should contain at least and only one UE from each collection: $\tilde{\mathcal{J}}$. To construct one subset of $\tilde{\mathcal{J}}$, we just need to select one UE from each collection. Thus, the size of $\tilde{\mathcal{J}}$ is

$$|\tilde{\mathcal{J}}| = \left(\frac{U}{N_c}\right)^{N_c}. \tag{6.39}$$

The second stage of PP-GUGA has the same process as GUGA, expect that we need to replace $\mathcal{J}$ in GUGA (see Section 6.3.1) by $\tilde{\mathcal{J}}$. PP-GUGA is summarized in Algorithm 4.

6.3.5 Complexity Analysis

In this subsection, we discuss the complexity of different user grouping algorithms. Let $\Theta(\cdot)$ be the complexity of one algorithm.

Algorithm 4 PP-GUGA: Pre Partitioned Greedy User Grouping Algorithm

Input: $\mathbf{H}$ (Channel matrix), M (No. of RRHs), $\overline{P}$ (RRH power constraint), U (No. of UEs), $N_{g,\mathrm{d}}$ (No. of UEs served during each subframe), N_{UA} (No. of antennas of each UE), N_{TA} (No. of antennas of each RRH)

Output: $\tilde{\mathcal{G}}$ (UEs grouping)

Initialisation : Let the set of all UEs $\mathcal{N}_{\mathrm{UE}} = \{1,\ 2,\ ...,\ U\}$, $\mathcal{S}_k = \emptyset$ (Set of UEs to be served during subframe k), where $k = 1,\ 2,\ ...,\ U/N_{g,\mathrm{d}}$

1: $N_c = MN_{\mathrm{TA}}/N_{\mathrm{UA}}$ /* Number of collections */

2: $\{\mathcal{M}_1,\ \mathcal{M}_2,\ ...,\ \mathcal{M}_{N_c}\} = \mathrm{UDA}(\mathbf{H}, M, U, N_{\mathrm{UA}}, N_{\mathrm{TA}})$

3: Get the set of all possible unordered subsets of size $N_{g,\mathrm{d}}$ of $\mathcal{N}_{\mathrm{UE}}$ with the constraint that each subset should contain at least and only one UE from $\mathcal{M}_i$ where $i = 1,\ 2,\ ...,\ N_c : \tilde{\mathcal{J}}$

4: Calculate the value of $R(\mathcal{S})$ for each $\mathcal{S} \in \tilde{\mathcal{J}}$ by (6.17) and (6.27)

5: Let $\tilde{\mathcal{J}}^{(1)} = \tilde{\mathcal{J}}$

6: **for** $k = 1$ to $U/N_{g,\mathrm{d}}$ **do**

7: $\quad \mathcal{S}_k = \mathrm{argmax}_{\mathcal{S} \in \tilde{\mathcal{J}}^{(k)}} R(\mathcal{S})$ /* Set of UEs $\mathcal{S}_k$ having the highest achievable sum rate is chosen to be served during subframe k */

8: $\quad \tilde{\mathcal{J}}^{(k+1)} = \left\{ \mathcal{S} | \mathcal{S} \in \tilde{\mathcal{J}}^{(k)} \text{ and } i \notin \mathcal{S}\ \forall i \in \mathcal{S}_k \right\}$

9: **end for**

10: **return** $\tilde{G} = \{\mathcal{S}_1,\ \mathcal{S}_2,\ ...,\ \mathcal{S}_{N_T}\}$

GUGA

The main cost of GUGA is the computation of the sum transmission rate of each subset of UEs $\mathcal{S}$ in $\mathcal{J}$, where $\mathcal{J}$ is the set of all possible unordered subsets of size $N_{g,\mathrm{d}}$ of $\mathcal{N}_{\mathrm{UE}}$. In this computation, the operation $\left(\mathbf{H}_{\mathcal{S}}(\mathbf{H}_{\mathcal{S}})^{\mathrm{H}}\right)^{-1}$ costs the most. Each calculation of $\left(\mathbf{H}_{\mathcal{S}}(\mathbf{H}_{\mathcal{S}})^{\mathrm{H}}\right)^{-1}$ has a complexity of $O(N_{g,\mathrm{d}}^2 N_{\mathrm{UA}}^2 MN_{\mathrm{TA}})$ [98]. GUGA needs to do $|\mathcal{J}| = \frac{U!}{(U-N_{g,\mathrm{d}})!N_{g,\mathrm{d}}!}$ times this operation. Thus, the complexity of GUGA algorithm is $O(N_{g,\mathrm{d}}^2 N_{\mathrm{UA}}^2 MN_{\mathrm{TA}} \cdot U^{N_{g,\mathrm{d}}})$. Applying $N_{g,\mathrm{d}} = MN_{\mathrm{TA}}/N_{\mathrm{UA}}$, we get

$$\Theta(\mathrm{GUGA}) = O(M^3 N_{\mathrm{TA}}^3 U^{(MN_{\mathrm{TA}}/N_{\mathrm{UA}})}). \tag{6.40}$$

SUGA

In (6.35), we need at most $MN_{\mathrm{TA}} - 1$ times $(1 \times MN_{\mathrm{TA}}) \times (MN_{\mathrm{TA}} \times 1) \times (1 \times MN_{\mathrm{TA}})$ (column vector)-(row vector)-(column vector) multiplication whose complexity is $O(MN_{\mathrm{PA}})$ for each antenna of each unallocated UE. To allocate a UE, we need to calculate (6.35) for at most UN_{UA} antennas in UEs' side and choose the one who has the maximum module. Thus, the complexity of allocating one UE is $O(UN_{\mathrm{UA}}M^2N_{\mathrm{TA}}^2) + O(UN_{\mathrm{UA}}) = O(UN_{\mathrm{UA}}M^2N_{\mathrm{TA}}^2)$.

Therefore, the total complexity of SUGA is

$$\Theta(\text{SUGA}) = O(M^2 N_{\text{TA}}^2 N_{\text{UA}} U^2). \tag{6.41}$$

PP-RUGA

In UDA, we need to do $MUN_{\text{UA}}N_{\text{TA}}$ multiplications and a sorting to MU elements. The complexity for UDA is $O(MU \log MU) + O(MUN_{\text{UA}}N_{\text{TA}})$. After applying UDA, the second stage of PP-RUGA is just a randomly allocation of different UEs to be served during different subframes. Thus, the complexity of PP-RUGA is

$$\Theta(\text{PP-RUGA}) = O(MU \log MU) + O(MUN_{\text{UA}}N_{\text{TA}}). \tag{6.42}$$

PP-GUGA

The first stage of PP-GUGA has the same complexity as UDA. Similar to GUGA, the computation of the sum transmission rate of each possible subset of UEs $\mathcal{S}$ in $\tilde{\mathcal{J}}$ (see Section 6.3.4) is the main cost of the second stage of PP-GUGA. However, the size of $|\tilde{\mathcal{J}}| = (U/N_c)^{N_c}$ is much less than $|\tilde{\mathcal{J}}|$. Note that $N_c = N_{g,\text{d}}$. The complexity of the second stage of PP-GUGA is $O(N_{g,\text{d}}^2 N_{\text{UA}}^2 MN_{\text{TA}} \cdot (U/N_{g,\text{d}})^{N_{g,\text{d}}})$. This is higher than that of UDA. Thus, the total complexity of PP-GUGA is $O(N_{g,\text{d}}^2 N_{\text{UA}}^2 MN_{\text{TA}} \cdot (U/N_{g,\text{d}})^{N_{g,\text{d}}})$. Applying $N_{g,\text{d}} = MN_{\text{TA}}/N_{\text{UA}}$, we get

$$\Theta(\text{PP-GUGA}) = O\big(M^3 N_{\text{TA}}^3 \left(\frac{UN_{\text{UA}}}{MN_{\text{TA}}}\right)^{(MN_{\text{TA}}/N_{\text{UA}})}\big). \tag{6.43}$$

6.4 Simulation Results

We apply the system configuration presented in Chapter 4. We consider a simple cluster of $M = 3$ RRHs serving U UEs. The set of RRHs is $\mathcal{N}_{\text{TP}} = \{1,\ 2,\ 3\}$. Each RRH and each UE is equipped with $N_{\text{TA}} = N_{\text{UA}} = 2$ antennas. During each subframe, $N_{g,\text{d}} = 3$ UEs are served. Each fronthaul link is assumed to apply the linear quantization with the same number of bits: $Q_{\text{q},v} = Q_{\text{q}}\ \forall v \in \mathcal{N}_{\text{TP}}$. For each numerical result, an average over 100 times randomly generated U UEs allocated to the studied cluster of RRHs with different large-scale channel gain is performed. For each fixed large-scale channel gain, we take average over 100 times small-scale multi-path fading.

We denote the exhaustive enumeration of all possible UEs grouping to solve optimization (6.28) by "Exhaustive algorithm". We call the transmission without applying any user

grouping algorithm by “Random UEs grouping”. In this section, we compare the average achievable sum rate of applying different user grouping algorithms GUGA, PP-GUGA, SUGA and PP-RUGA with applying “Exhaustive algorithm” and “Random UEs grouping”.

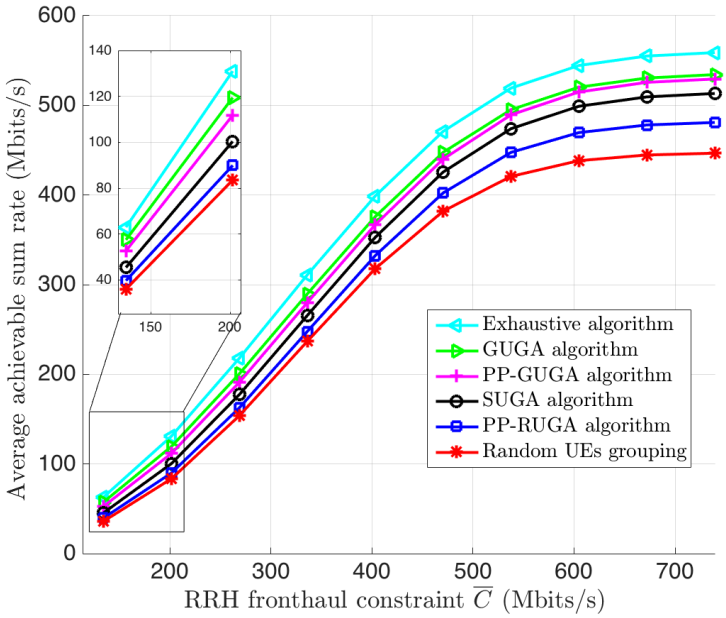

Fig. 6.5 Average achievable sum rate vs. RRH power constraint P_W ($U = 9$, $Q_\mathrm{q} = 4$).

At first, we fix the number of UEs $U = 9$ and each RRH power $P_\mathrm{dBm} = 46$ dBm. The average achievable sum rate of the six algorithms with respect to different fronthaul constraints is shown in Figure 6.5. Secondly, we fix the number of UEs $U = 9$ and each fronhaul quantization bits $Q_\mathrm{q} = 4$. The average achievable sum rate of the six algorithms with respect to different RRH power constraints is shown in Figure 6.5.

It is observed that applying the algorithms which selects an optimal or suboptimal UEs grouping increases the system performance compared to randomly selecting a UEs grouping. Applying GUGA has a similar performance to that of applying “Exhaustive algorithm”. Meanwhile applying PP-GUGA achieves a smaller average sum rate than applying GUGA, but higher than SUGA. PP-RUGA also achieves a higher average sum rate than “Random UEs grouping”, however smaller than all the other user grouping algorithms.

Then, we fix the RRH power constraint $P_\mathrm{dBm} = 46$ dBm and each fronhaul quantization bits $Q_\mathrm{q} = 4$. We vary the number of UEs U. The corresponding ratios of increasing average achievable sum rate compared with applying “Random user grouping” realized by applying GUGA, PP-GUGA, SUGA and UDA are illustrated in Figure 6.7. With the increasing number of UEs for grouping, the performances of all the proposed user grouping algorithms

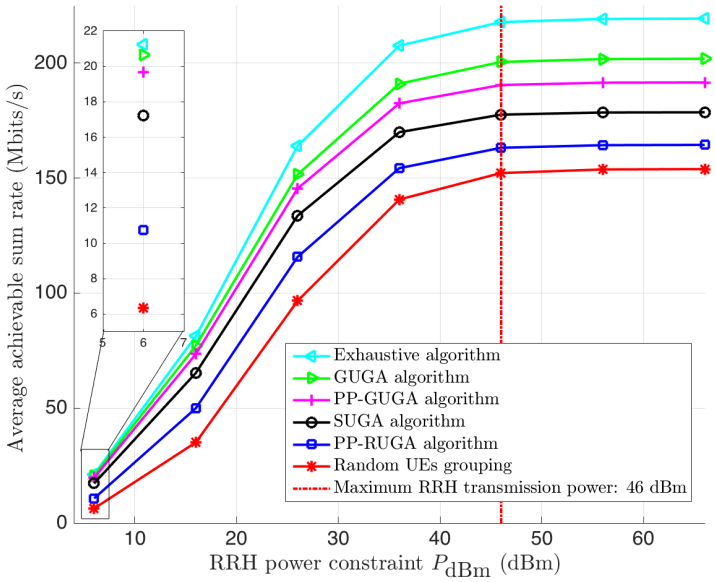

Fig. 6.6 Average achievable sum rate vs. RRH power constraint P_W ($U = 9$, $Q_{\mathrm{q}} = 4$).

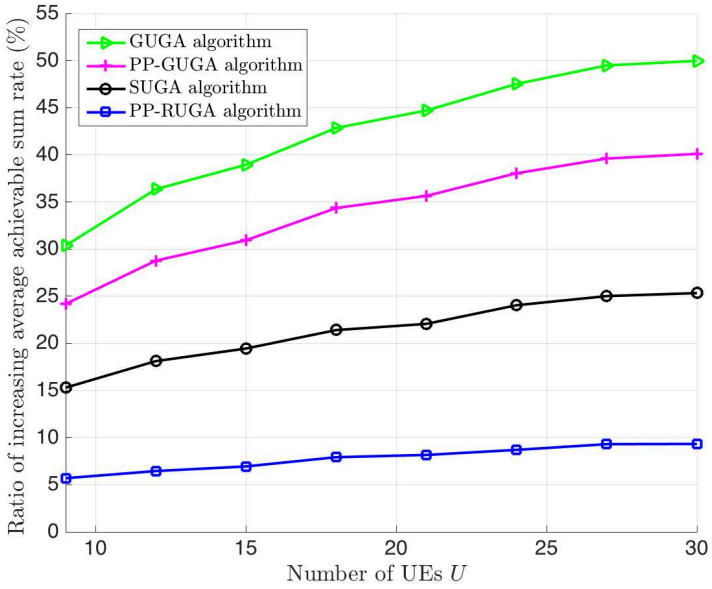

Fig. 6.7 Ratio of increasing average achievable sum rate compared with applying "Random user grouping" vs. number of UEs U ($P_{\mathrm{dBm}} = 46$ dBm, $Q_{\mathrm{q}} = 4$).

increase. Applying GUGA increase 50% the average achievable sum rate when $U = 30$, 40% with PP-GUGA, 25% with PP-SUGA and around 10% with PP-RUGA. However, the increase rate of the average achievable sum rate is gelling lower as U increases, because most of the spatial diversity has been exploited.

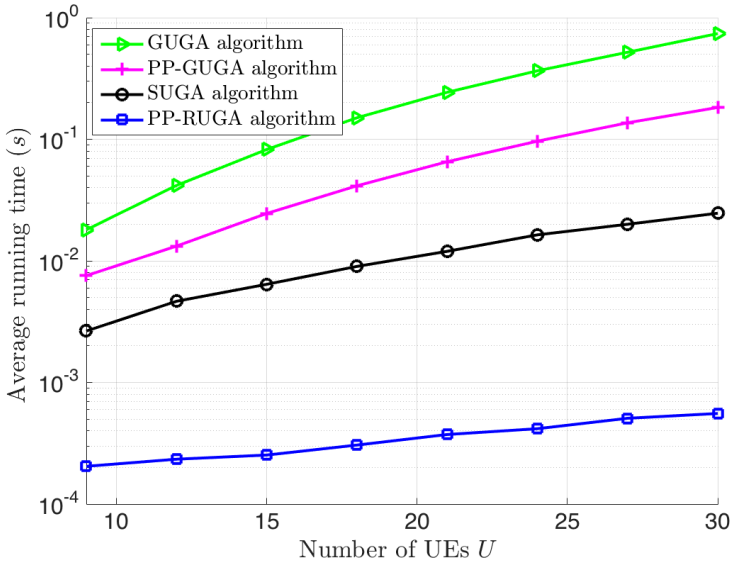

Fig. 6.8 Running time vs. number of UEs U ($P_{\text{dBm}} = 46$ dBm, $Q_{\text{q}} = 4$).

The simulation is done by Matlab in a MacBook Air with 1.4 GHz Intel Core i5. The average running time for GUGA, PP-GUGA, SUGA and PP-RUGA with respect to different number of UEs U is shown in Figure 6.8. It illustrates that, considering running time, GUGA $\gg$ PP-GUGA $\gg$ SUGA $\gg$ PP-RUGA.

When we fix the values of M, N_{TA} and N_{UA}, the complexities of the four user grouping algorithms (6.40), (6.43), (6.41) and (6.42) become:

$$\begin{aligned} \Theta(\text{GUGA}) &= O(U^3) \\ \Theta(\text{PP-GUGA}) &= O(U^3) \\ \Theta(\text{SUGA}) &= O(U^2) \\ \Theta(\text{PP-RUGA}) &= O(U \log U). \end{aligned} \tag{6.44}$$

These correspond to the increase of average running time with respect to the number of UEs U as depicted in Figure 6.8. Note that, although the increasing rate of GUGA is similar to that of PP-GUGA, (6.40) and (6.42) show that GUGA is more complex than PP-GUGA. This is consistent with Figure 6.8.

6.5 Conclusion

In this chapter, we proposed four user grouping algorithms denoted GUGA, SUGA, PP-RUGA and PP-SUGA. ZF precoding is used for joint transmission in a C-RAN downlink system with limited fronthaul capacity. For a number of UEs, the algorithms select which UEs are served simultaneously during the same subframe to increase the average achievable sum rate.

GUGA can achieve a similar performance of that of exhaustive enumeration of all possible UEs groupings with a lower complexity. However GUGA is still too complex when the number of UEs U becomes large for a fixed number of UEs served during each subframe. PP-SUGA and SUGA exhibit less complexity with a small loss in performance. PP-GUGA has a better performance than SUGA but a higher complexity. The complexity can be further significantly reduced with PP-RUGA. However, PP-RUGA achieves less average sum rate than all the other three algorithms.

Chapter 7

Hybrid Joint Transmission

7.1 Introduction

In Chapter 5, simulation results show that CoMP should be preferred for UEs located in cell edge areas and when fronthaul capacity is high. The UEs in cell edge area benefit more from CoMP than UEs in cell center area.

CoMP can largely improve system performance but needs more fronthaul capacity [27], especially with JT: data of all served UEs have to be shared between different RRHs [51]. The heavy burden on fronthaul links is the major bottleneck of the feasibility of C-RAN [22]. The extra fronthaul load and CSIs requirement in CoMP may counteract the throughput gain for UEs in cell center area [99].

To improve system performance with limited fronthaul capacity, we propose a hybrid transmission strategy. An example of hybrid transmission is shown in Figure 7.1. The UEs in near RRH area are served in single RRH mode (served by only one RRH), and UEs in cell edge area served in distributed MIMO mode (jointly served by all the RRHs in the cluster).

Single RRH mode corresponds to non-CoMP with limited coordinated Scheduling (CB). The BBU pool just needs to collect the statistical CSIs to decide which RRH serves which UE and which UEs are served at the same time. Distributed MIMO mode is one coherent JT algorithm where we apply Zero-Forcing (ZF) precoding to let a cluster of RRHs jointly transmit signals to a group of UEs.

The performances of CB/CS, JT and non-CoMP with limited fronthaul capacity are compared in a number of studies. Simulation results in Section 5.5 and [100] show that non-CoMP outperforms CoMP and CB/CS outperforms JP when the fronthaul capacity is low. To make a trade-off between system performance and fronthaul load, a number of studies has been done to switch between CB/CS, JP and non-CoMP. In [81], a semi-dynamic hard switching scheme is proposed to decide which UEs are to be served by CB or JT by

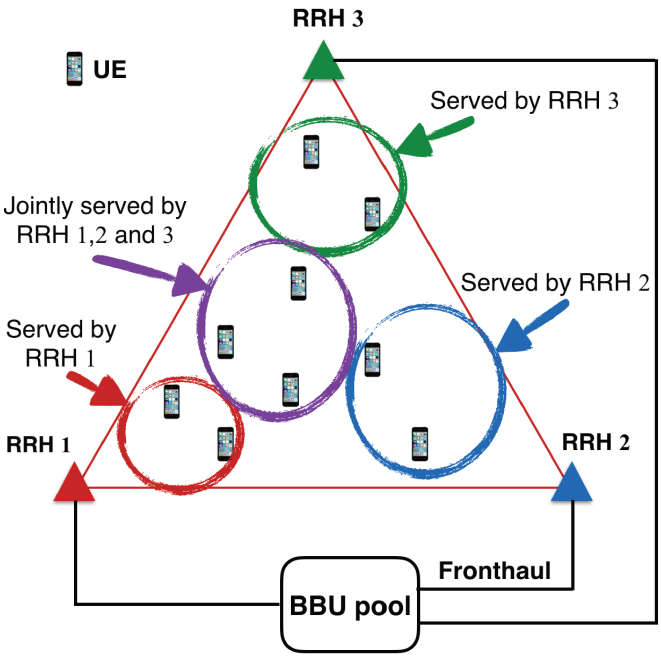

Fig. 7.1 An example of hybrid transmission.

comparing their achievable sum rates under CB and JP with limited fronthaul capacity. In [82, 101], the UE data are divided into two part: private data and common data. The private data are not shared among the RRHs and are transmitted by CB. By contrast, the common data are shared among the RRHs are transmitted by JT. The authors in [82] show how the limited fronthaul capacity determines the best amount of data worth sharing across the RRHs. Both [81] and [82] use a data-sharing based strategy forwarding data from BBU pool to RRHs. However, a more practical compression-based strategy is applied in our work for D-MIMO to transmit data from BBU pool to RRHs.

The remainder of this chapter is organized as follows. In Section 7.2, we introduce the system model. Next our proposed UEs scheduling algorithms are described in Section 7.3. The simulation results are presented in Section 7.4. Finally we conclude in Section 7.5.

7.2 System Model

In this chapter, we compare three different transmission strategies: 1) Fully single RRH transmission; 2) Fully distributed MIMO; 3) Hybrid transmission. In Fully single RRH transmission, all the U UEs are served in a single RRH mode, in which each of the U UEs is served by one and only one RRH. Note that all the RRHs of a cluster transmit simultaneously, and several UEs can be served at the same time in single RRH mode. In Fully distributed

MIMO, all the U UEs are served in a distributed MIMO mode, in which several UEs are jointly served by all the M RRHs at the same time. In hybrid transmission, part of the U UEs are served in single RRH mode and the rest in Distributed MIMO mode.

For each mode, we have to determine which UEs are served at the same time (during the same subframe). The U UEs are thus partitioned into N_T disjoint groups $\mathcal{N}_{\mathrm{UE},t}$ with $t \in \mathcal{T}$ and each group is served in one subframe. Thus, $\bigcup_{t\in\mathcal{T}} \mathcal{N}_{\mathrm{UE},t} = \mathcal{N}_{\mathrm{UE}}$ and $\mathcal{N}_{\mathrm{UE},i} \bigcap \mathcal{N}_{\mathrm{UE},j} = \emptyset$, where $i \neq j \; \forall i \in \mathcal{T}$ and $\forall j \in \mathcal{T}$. The group of UEs served in subframe $t \in \mathcal{T}$ is denoted as $\mathcal{N}_{\mathrm{UE},t} = \{u_1^t, u_2^t, \dots u_{N_{g,t}}^t\}$, where $N_{g,t}$ is the size of the group during subframe t.

Let $\mathcal{G} = \{\mathcal{N}_{\mathrm{UE},1}, \mathcal{N}_{\mathrm{UE},2}, \dots \mathcal{N}_{\mathrm{UE},N_T}\}$ be a set of any arbitrary grouping of UEs.

UEs in $\mathcal{N}_{\mathrm{UE},t}$ are served together or seperately by the RRHs in $\mathcal{N}_{\mathrm{TP}}$ on respecting to the constraints:

$$N_{g,t} \leq MN_{\mathrm{TA}}. \tag{7.1}$$

7.3 UEs Scheduling

We denote the number of UEs served in single RRH mode as U_{s}, and the number of UEs served in distributed MIMO mode by U_{d}. Then we have $U_{\mathrm{s}} + U_{\mathrm{d}} = U$. Let $\mathcal{N}_{\mathrm{UE,s}} = \{u_{\mathrm{s},1}, u_{\mathrm{s},2}, \dots u_{\mathrm{s},U_{\mathrm{s}}}\}$ be the set of UEs served in single RRH mode. Meanwhile, let $\mathcal{N}_{\mathrm{UE,d}} = \{u_{\mathrm{d},1}, u_{\mathrm{d},2}, \dots u_{\mathrm{d},U_{\mathrm{d}}}\}$ be the set of UEs served in distributed MIMO mode. In Fully Single RRH transmission strategy, $U_{\mathrm{s}} = U$, while $U_{\mathrm{d}} = U$ in Fully Distributed MIMO transmission strategy. Let $\mathcal{T}_{\mathrm{s}} = \{t_{\mathrm{s},1}, t_{\mathrm{s},2}, \dots t_{\mathrm{s},N_{T,\mathrm{s}}}\}$ be the set of subframes out of $\mathcal{T}$ during which the UEs are served in single RRH mode, where $N_{T,\mathrm{s}}$ is the length of $\mathcal{T}_{\mathrm{s}}$. Let $\mathcal{T}_{\mathrm{d}} = \{t_{\mathrm{d},1}, t_{\mathrm{d},2}, \dots t_{\mathrm{d},N_{T,\mathrm{d}}}\}$ be the set of subframes out of $\mathcal{T}$ during which the UEs are served in distributed MIMO mode, where $N_{T,\mathrm{d}}$ is the length of $\mathcal{T}_{\mathrm{d}}$. And we have $N_{T,\mathrm{s}} + N_{T,\mathrm{d}} = N_T$.

As mentioned in Section 4.3 and Section 6.2, $N_{g,\mathrm{s}} = M$ UEs are served during one subframe in single RRH mode and $N_{g,\mathrm{d}} = MN_{\mathrm{TA}}$ in distributed MIMO mode.

7.3.1 UEs Scheduling for Distributed MIMO Mode

We partition the U_{d} UEs served in distributed MIMO mode to $N_{T,\mathrm{d}}$ groups with same size $N_{g,\mathrm{d}} = MN_{\mathrm{TA}}$. Each group of UEs is served during one subframe in $\mathcal{T}_{\mathrm{d}}$. We need an heuristic to decide which UEs are served at the same time in order to get a high sum transmission rate.

Here, we apply Semi-orthogonal User Grouping Algorithm (SUGA) adapted from [94] and presented in Section 6.3.2 to do the UEs grouping. SUGA is an iterative algorithm which assigns the U_{d} UEs disjointly to $N_{T,\mathrm{d}}$ subframes.

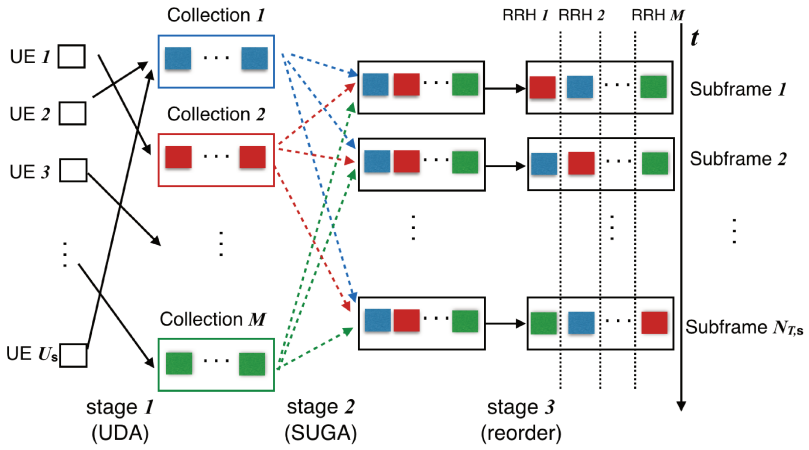

Fig. 7.2 Process of PP-SUGA.

7.3.2 UEs Scheduling for Single RRH Mode

We partition the U_s UEs served in single RRH mode to $N_{T,s}$ groups with same size $N_{g,s} = M$. Let $\mathcal{G}_s = \{\mathcal{N}_{UE,t_{s,1}}, \mathcal{N}_{UE,t_{s,2}}, ..., \mathcal{N}_{UE,t_{s,N_{T,s}}}\}$ be the grouping of UEs served in single RRH mode. The i-th UE in the set $\mathcal{N}_{UE,t_{s,k}}$ is served by the i-th RRH in $\mathcal{N}_{TP}$, where $k = 1, 2, \ldots N_{T,s}$.

In single RRH mode, only statistical CSIs are available in BBU pool. The statistical channel matrix $\mathbf{Q}$ is used for UEs scheduling. We propose Pre Partitioned - Semi-orthogonal User Grouping Algorithm (PP-SUGA) to do the UEs scheduling.

PP-SUGA is a three-stage algorithm as shown in Figure 7.2. In stage 1, it applies an adapted version of Users Division Algorithm (UDA) (see Section 6.3.3) to partition the U_s UEs into M collections with same size $U_s/M = N_{T,s}$. The i-th collection is allocated with the UEs which locates nearest to RRH i. Then in stage 2, UEs in the same collection are assigned to different subframes by applying SUGA. To remind, a collection of UEs is defined as the UEs that are not served during the same subframe. At last, in stage 3, each group of UEs is reordered. The RRH that brings the highest channel gain to the UE is chosen to serve it.

Stage 1

The adapted version of UDA initializes $\mathcal{K} = \mathcal{N}_{UE,s}$, $\mathcal{S} = \mathcal{N}_{TP,s}$. Let $\mathbf{Q}_{\mathcal{K},\mathcal{D}}$ be the statistical channel matrix between the set of RHHs $\mathcal{K}$ and the set of UEs $\mathcal{S}$.

In each iteration, UDA finds the index of the maximum value in $\mathbf{Q}_{\mathcal{K},\mathcal{S}}$:

$$\{m,n\} = \underset{i,j}{\operatorname{argmax}} \mathbf{Q}_{\mathcal{K},\mathcal{S}}(i,j). \tag{7.2}$$

Then UE $\mathcal{K}[m]$ is added to $\mathcal{M}_{\mathcal{S}[n]}$ and removed from $\mathcal{K}$. When the $\mathcal{S}[n]$-th collection is full, RRH $\mathcal{S}[n]$ is deleted from $\mathcal{S}$.

This adapted version of UDA for Chapter 7 is detailed in Algorithm 5.

Algorithm 5 UDA: Users Division Algorithm

Input: $\mathbf{Q}$ (Statistical channel matrix), M (No. of RRHs), U_s (No. of UEs), $\mathcal{N}_{\mathrm{UE,s}}$ (Set of UEs).

Output: $\{\mathcal{M}_1,\ \mathcal{M}_2,\ ...,\ \mathcal{M}_M\}$ (UEs collections)

Initialisation : $\mathcal{K} = \mathcal{N}_{\mathrm{UE,s}}$, $\mathcal{S} = \{1,\ 2,\ ...,\ M\}$ (set of collections), $\mathcal{M}_i = \emptyset$ (UEs collection), where $i = 1,\ 2,\ ...,\ M$

1: **for** $k = 1$ to U_s **do**
2: $\{m,n\} = \mathrm{argmax}_{i,j}\mathbf{Q}_{\mathcal{K},\mathcal{S}}(i,j)$
3: $\mathcal{M}_{\mathcal{S}[n]} \leftarrow \mathcal{M}_{\mathcal{S}[n]} \cup \{\mathcal{K}[m]\}$, $\mathcal{K} \leftarrow \mathcal{K}/\{\mathcal{K}[m]\}$
4: **if** $|\mathcal{M}_{\mathcal{S}[n]}| = U_\mathrm{s}/M$ **then**
5: $\mathcal{S} \leftarrow \mathcal{S}/\{\mathcal{S}[n]\}$ /* When a collection already has U_s/M elements, it should not be assigned with new UEs any more.*/
6: **end if**
7: **end for**
8: **return** $\{\mathcal{M}_1,\ \mathcal{M}_2,\ ...,\ \mathcal{M}_M\}$

Stage 2

PP-SUGA realizes stage 2 by using SUGA with the constraint that UEs in the same collection cannot be served during the same subframe. The input to SUGA in Section 7.3.1 is replaced by Q, M, $N_{g,\mathrm{d}}$ and $\mathcal{N}_{\mathrm{UE,d}}$. At the end of stage 2, we get $\tilde{\mathcal{G}}_\mathrm{d} = \{\tilde{\mathcal{S}}_1,\ \tilde{\mathcal{S}}_2,\ ...,\ \tilde{\mathcal{S}}_{N_{T,\mathrm{s}}}\}$.

Stage 3

In stage 3, each set is reordered in $\tilde{\mathcal{G}}_\mathrm{d}$ to decide which RRH serves which UE. A RRH will serve the nearest UE if no shadow fading is considered. It is as dividing M UEs into M collections with size 1. This can be realized by applying UDA: $\tilde{\mathcal{S}}_k = \mathrm{UDA}(Q,M,M,\tilde{\mathcal{S}}_k)$, where $k = 1,\ 2,\ldots\ |\mathcal{T}_\mathrm{s}|$.

At last, the chosen grouping of UEs is given by $\mathcal{G}_\mathrm{s} = \{\mathcal{S}_1,\ \mathcal{S}_2,\ ...,\ \mathcal{S}_{|\mathcal{T}_\mathrm{s}|}\}$.

In the following, we discuss the complexity of PP-SUGA. UDA in stage 1 does a sorting to $U_\mathrm{s}M$ elements. Thus the complexity of stage 1 is $O(U_\mathrm{s}M\log U_\mathrm{s}M)$. In stage 2, SUGA is applied with statistical channel matrix and M UEs served during each subframe. Referring to Subsection 6.3.5, the complexity of stage 2 is $O(U_\mathrm{s}^2M^2)$. In stage 3, we do a sorting of M^2 elements for each subframe in $\mathcal{T}_\mathrm{s}$. Thus the complexity of stage 3 is $O(N_{T,\mathrm{s}}M^2\log M^2) =$

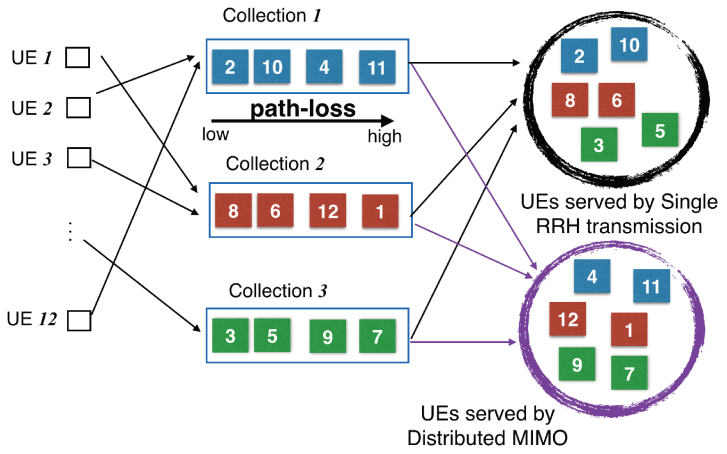

Fig. 7.3 Process of separating UEs to twos parts, some served in single RRH mode and the others in distributed MIMO mode.

$O(U_{\mathrm{s}}M\log M^2)$. Finally, the total complexity of PP-SUGA is $O(U_{\mathrm{s}}M\log U_{\mathrm{s}}M)+O(U_{\mathrm{s}}^2M^2)+O(U_{\mathrm{s}}M\log M^2)=O(U_{\mathrm{s}}^2M^2)$.

7.3.3 UEs Scheduling for Hybrid Transmission

In Hybrid transmission, based on statistical CSIs, the set of UEs $\mathcal{N}_{\mathrm{UE}}$ is separated to two parts: $\mathcal{N}_{\mathrm{UE,s}}$ with size U_{s} and $\mathcal{N}_{\mathrm{UE,d}}$ with size U_{d}. The UEs in $\mathcal{N}_{\mathrm{UE},s}$ are served on Single transmission mode while the UEs in $\mathcal{N}_{\mathrm{UE,d}}$ in distributed MIMO mode. Our strategy is to allocate the UEs close to any one of the M RRHs to $\mathcal{N}_{\mathrm{UE,s}}$ and the UEs located in the center of the M RRHs to $\mathcal{N}_{\mathrm{UE,d}}$.

At first, UDA is applied to divide the U UEs into M collections. The lower the index of a UE in the i-th collection is, the higher statistical channel gain from RRH i to the UE is. Then the first U_{s}/M elements in each collection are allocated to $\mathcal{N}_{\mathrm{UE},s}$ and the rest to $\mathcal{N}_{\mathrm{UE,d}}$. The complexity of this process is $O(UM\log UM)$. An example of this process with $U=12$, $M=3$ and $U_{\mathrm{s}}=U_{\mathrm{d}}=6$ is illustrated in Figure 7.3.

Next, the set of UEs $\mathcal{N}_{\mathrm{UE,s}}$ are grouped to $N_{T,\mathrm{s}}$ groups with equal size M by applying SUGA in Subsection 7.3.1. Each group is served during one subframe in $\mathcal{T}_{\mathrm{s}}$. Meanwhile, the set of UEs $\mathcal{N}_{\mathrm{UE,d}}$ are grouped to $N_{T,\mathrm{d}}$ groups with equal size MN_{TA} by applying PP-SUGA in Subsection 7.3.2. Each group is served during one subframe in $\mathcal{T}_{\mathrm{d}}$. PP-SUGA for UEs served in single RRH mode is based on statistical CSIs and SUGA for UEs served in distributed MIMO mode is based on instantaneous CSIs. Therefore, PP-SUGA need to be applied once during a number of frames while SUGA should be applied once during each frame.

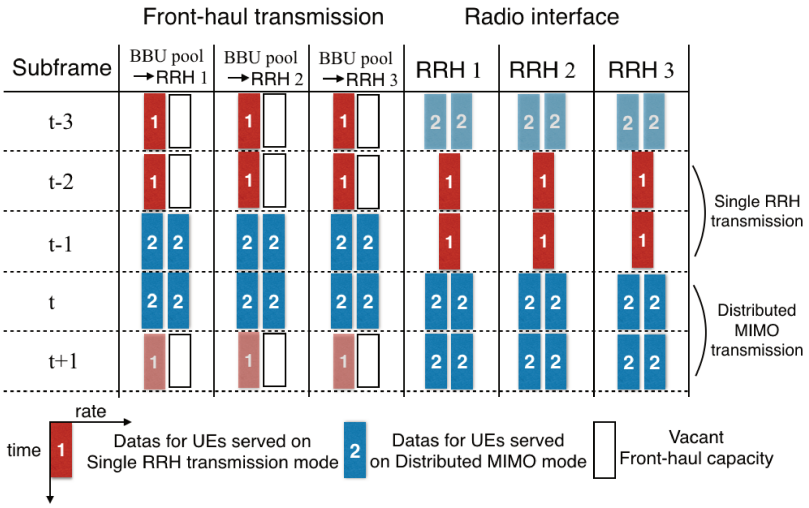

Fig. 7.4 Equal fronthaul transmission capacity constraint for the UEs served on Single transmission mode and in distributed MIMO mode.

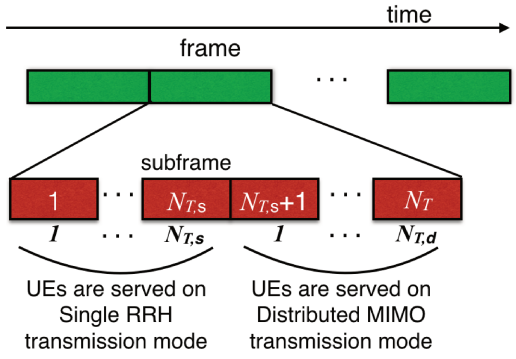

Fig. 7.5 Subframes scheduling of B-FTS.

7.3.4 Transmission Strategies on the Fronthaul Links

In Synchronous Fronthaul Transmission Scheme (S-FTS), the data transmitted from BBU pool to RRHs during subframe t are transmitted from RRHs to UEs on radio interface during subframe $t+1$. In single RRH mode, discrete symbols are transmitted from BBU pool to RRHs. By contrast, in distributed MIMO mode, the discrete symbols are first precoded and quantized and then transmitted from BBU pool to RRHs. Thus, Distributed MIMO transmission mode consumes more fronthaul capacity than Single RRH transmission mode.

In S-FTS, the fronthaul capacity for Single RRH transmission mode can be not fully used, while the transmission rate for Distributed MIMO transmission mode is limited by the fronthaul capacity constraint. Such an example with 3 RRHs is shown in Figure 7.4.

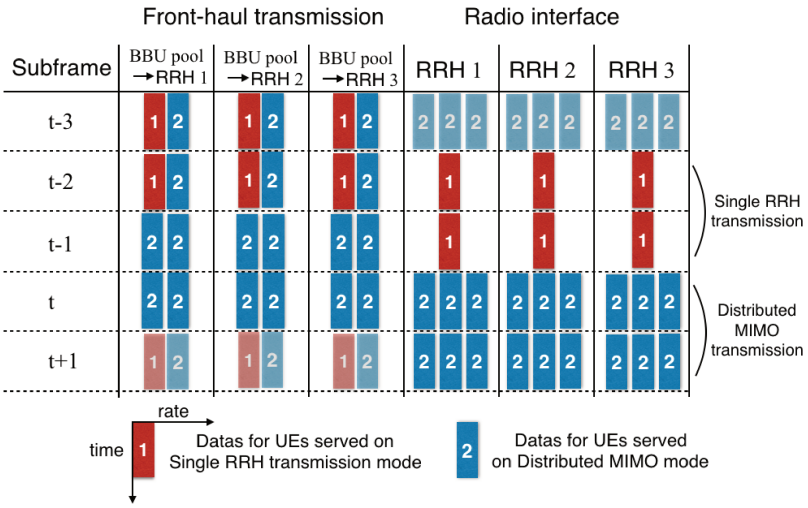

Fig. 7.6 Fronthaul transmission with previous data transmission for the UEs served in distributed MIMO mode.

To exploit the vacant fronthaul capacity for Single RRH transmission mode, we propose Buffered Fronthaul Transmission Scheme (B-FTS). In B-FTS, during each frame, the UEs in $\mathcal{N}_{\mathrm{UE,s}}$, which are served in single RRH mode, are served before the UEs in $\mathcal{N}_{\mathrm{UE,d}}$ during the first $N_{T,\mathrm{s}}$ successive subframes in $\mathcal{T}$ (see Figure 7.5). Like in S-FTS, in subframe $t-1$, the fronthaul transmits the data for the UEs served in single RRH mode which are transmitted on radio interface during subframe t. However, part of the precoded symbols for the UEs in $\mathcal{N}_{\mathrm{UE,d}}$ are transmitted in advance in the vacant fronthaul capacity available during subframes $0,\ 1,\ldots\ N_{T,\mathrm{s}}-1$. The precoded symbols are then buffered in the corresponding RRHs and transmitted when necessary on the radio interface. In this way, extra fronthaul capacity is got for Distributed MIMO transmission mode.

As the instantaneous CSIs from BBU pool to the UEs in $\mathcal{N}_{\mathrm{UE,d}}$ are constant during all the subframes in $\mathcal{T}$, BBU pool can calculate what precoded symbols to transmit to each RRH for all UEs in $\mathcal{N}_{\mathrm{UE,d}}$ from the beginning of subframe 1. An example of B-FTS with 3 RRHs is illustrated in Figure 7.6.

We denote the set of UEs served in single RRH mode by RRH v as $\mathcal{U}_v$. The sum vacant capacity of the fronthaul link from BBU pool to RRH v during subframes $0,\ 1,\ldots\ N_{T,\mathrm{s}}-1$ is

$$\tilde{C}_v = N_{T,\mathrm{s}}\overline{C} - \sum_{u\in\mathcal{U}_v} R_u. \tag{7.3}$$

We equally divide $\tilde{C}_v$ to $N_{T,\mathrm{d}}$ parts, and each part is used to transmit the precoded symbols from BBU pool to RRH v for the UEs served in distributed MIMO mode during one subframe in $\mathcal{T}_{\mathrm{d}}$ on radio interface. Then the equivalent capacity of the v-th fronthaul link for the UEs

Table 7.1 Comparison of CSI signaling cost of different transmission strategies

Transmission strategy	CSI signaling cost
Fully single RRH transmission	statistical CSIs between all the M RRHs and all the U UEs every N_L frames
Hybrid transmission	statistical CSIs between all the M RRHs and all the U UEs every N_L frames instantaneous CSIs between all the M RRHs and all the U_d UEs in $\mathcal{N}_{\mathrm{UE,d}}$ each frame
Fully distributed MIMO transmission	instantaneous CSIs between all the M RRHs and all the U UEs each frame

served on distributed MIMO mode becomes

$$\overline{C}_{\mathrm{eq},v} = \overline{C} + \frac{\tilde{C}_v}{N_{T,\mathrm{d}}} \quad v \in \mathcal{N}_{\mathrm{TP}}, \forall t \in \mathcal{T}_\mathrm{d}. \tag{7.4}$$

To calculate the corresponding transmission rates for the UEs served in distributed MIMO mode when B-FTS is activated, we replace $\overline{C}$ with $\overline{C}_{\mathrm{eq},v}$ in (4.19).

7.3.5 Cost of CSIs

In this subsection, we discuss the cost of CSIs of different transmission strategies. We assume the statistical CSIs to be constant within a period of N_L frames, while they vary in an ergodic way across a large number of periods of N_L frames. The instantaneous CSIs are updated each frame, and the statistical CSIs are updated every N_L frames. The CSI signaling cost of the three transmission strategies is shown in Table 7.1. As N_L is usually in a scale of hundreds and statistical CSIs have a small dimension than instantaneous CSIs, the update of statistical CSIs costs much less resources than instantaneous CSIs. Thus, fully single RRH transmission needs the least CSIs while fully distributed MIMO transmission needs the most. Hybrid transmission consumes $\frac{100U_\mathrm{d}}{U}$ percent instantaneous CSIs of that consumed by fully distributed MIMO transmission.

7.4 Simulation Results

In this section, we first evaluate the performance of SUGA and PP-SUGA. Then, we compare the average achievable sum rate of applying fully single RRH transmission, fully distributed MIMO transmission and hybrid transmission. Both S-FTS and B-FTS are evaluated for hybrid transmission.

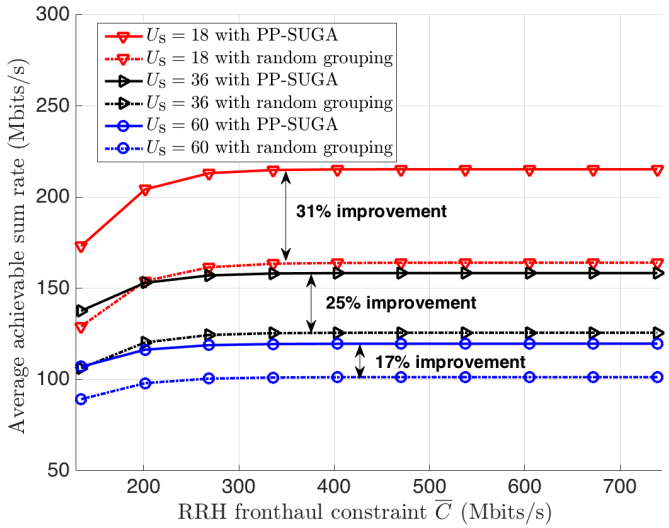

Fig. 7.7 Average achievable sum rate vs. fronthaul capacity constraint $\overline{C}$ for the UEs served in single RRH mode.

We apply the system configuration presented in Chapter 4. We consider a simple cluster of $M = 3$ RRHs serving U UEs. For each numerical result, an average over 100 times randomly generated U UEs allocated to the studied cluster of RRHs with different large-scale channel gain is performed. For each fixed large-scale channel gain, we take average over 100 times small-scale multi-path fading.

The average achievable sum rate during one subframe for a set of UEs $\tilde{\mathcal{N}}$ served during a set of subframes $\tilde{\mathcal{T}}$ is

$$\overline{R} = \frac{1}{|\tilde{\mathcal{T}}|} \sum_{u \in \mathcal{N}} R_u. \tag{7.5}$$

7.4.1 Performance of Users Grouping Algorithms

The 60 UEs are separated into two parts: U_s UEs served in single RRH mode and U_d UEs served in distributed MIMO mode. The separation applies the algorithm introduced in Section 7.3.3.

We call "random grouping" the most simple algorithm which consists in randomly dividing a set of UEs into several groups. For example, for distributed MIMO applying , we randomly divide the U_d UEs into $U_d/6$ groups, and each group is served during one subframe.

Figure 7.7 compares the performance of PP-SUGA with "random grouping" for UEs served in single RRH mode with different values of U_s. Note that the value of $\overline{C}$ is the

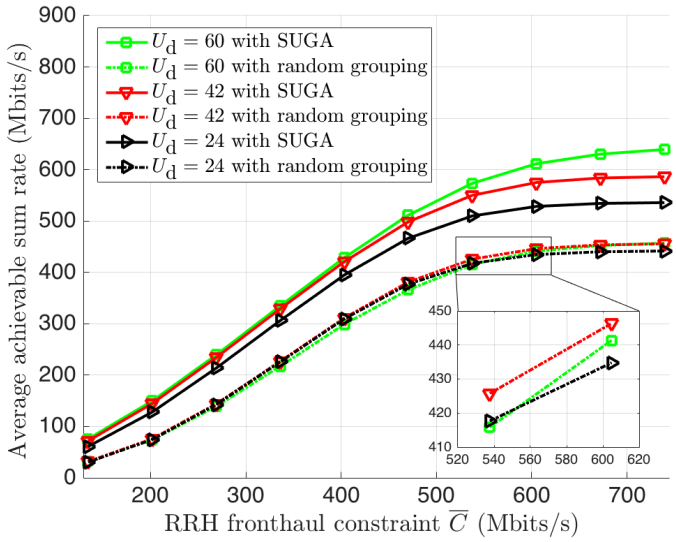

Fig. 7.8 Average achievable sum rate vs. fronthaul capacity constraint $\overline{C}$ for the UEs served in distributed MIMO mode.

capacity constraint for each fronthaul link and there are in total 3 fronthaul links in our simulation scenario.

We observe that, compared with "random grouping", applying PP-SUGA brings about 31% more average sum transmission rate when $U_d = 18$, around 25% when $U_d = 36$ and about 17% when $U_d = 60$. When the fronthaul capacity is larger than 300 Mbits/s, the average sum transmission rate barely increases with the increase of fronthaul constraint: the capacity is limited by the radio interface and not the fronthaul.

Figure 7.8 and Figure 7.9 compares the performance of SUGA with "random grouping" for UEs served in distributed MIMO mode with different values of U_d. Figure 7.8 shows the average sum transmission rates for the UEs served in distributed MIMO mode applying SUGA and "random grouping" with respect to different fronthaul capacity constraints. Figure 7.9 shows the corresponding ratios of increasing average achievable sum rates of applying SUGA compared with applying "random grouping". It is observed that SUGA brings significant performance improvement in low fronthaul capacity constraint regime while less significant in high fronthaul capacity constraint regime. SUGA can reduce the quantization noise for a fixed fronthaul capacity constraint, which in turn improves the average sum rate. When the fronthaul capacity constraint is large, the benefit is less significant.

Furthermore, the higher U_d is, the higher average sum rate can be achieved. This is because, with a larger number of UEs served in distributed MIMO mode, more diversity can be exploited by SUGA. By contrast, with single RRH mode, the average achievable sum rates continue to increase when the fronthaul capacity constraint is larger than 300 Mbits/s.

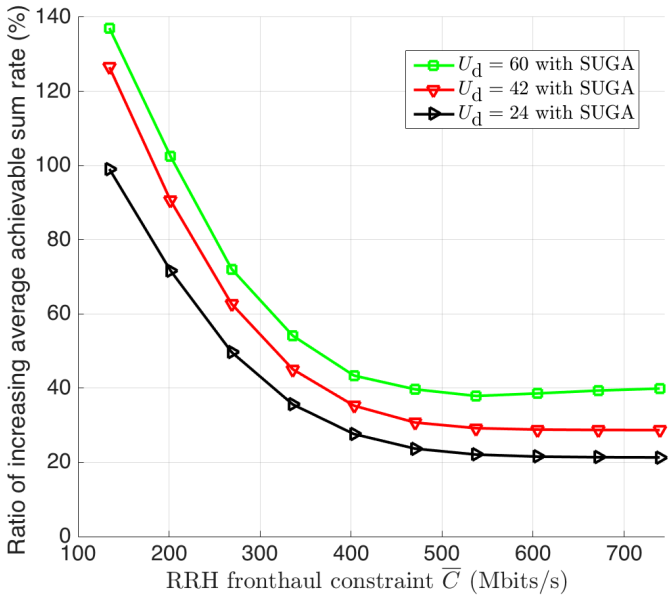

Fig. 7.9 Ratio of increasing average achievable sum rate compared with applying "Random user grouping" vs. fronthaul capacity constraint $\overline{C}$ for the UEs served in distributed MIMO mode.

This is the incentive for us to share part of the fronthaul capacity for UEs served in single RRH transmission mode to UEs served in distributed MIMO mode.

7.4.2 Performance Comparison of Different Transmission Strategies

In this subsection, we compare the performance of different transmission strategies with different values of U_s ($U_d = U - U_s$) from fully single RRH transmission ($U_s = U$) to fully distributed MIMO transmission ($U_s = 0$). The hybrid transmission with 24 UEs served in single RRH mode and 36 UEs served in distributed MIMO mode is denoted by "hybrid 24-36 transmission". Both S-FTS and B-FTS are considered for hybrid transmission.

The simulations results are illustrated in Figure 7.10 and Figure 7.11 is a zoom of the low fronthaul capacity constraint regime. We can observe that the performance of hybrid transmission is always better than fully single RRH transmission. Hybrid transmission has a better performance than fully distributed MIMO transmission in low fronthaul capacity regime while lower in high fronthaul capacity regime. Applying B-FTS (advanced fronthaul transmission for the UEs served in distributed MIMO mode) can largely increase the performance of hybrid transmission compared with applying S-FTS when the fronthaul capacity constraint is low.

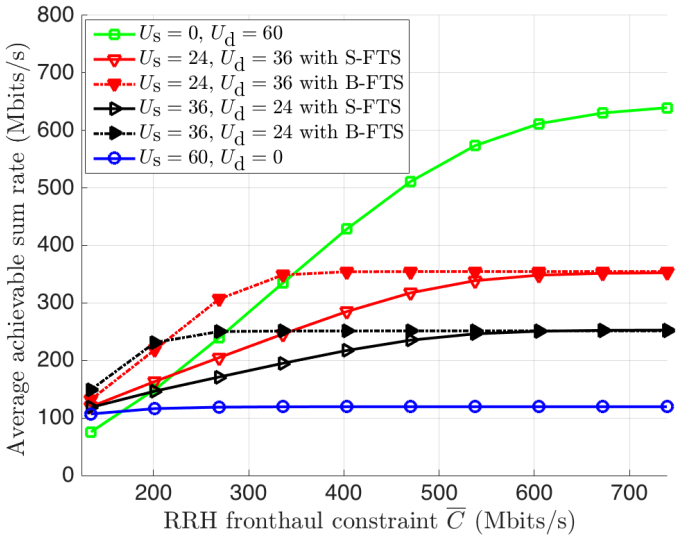

Fig. 7.10 Comparison of average achievable sum rate vs. fronthaul capacity constraint $\overline{C}$ for different transmission strategies (no zoom).

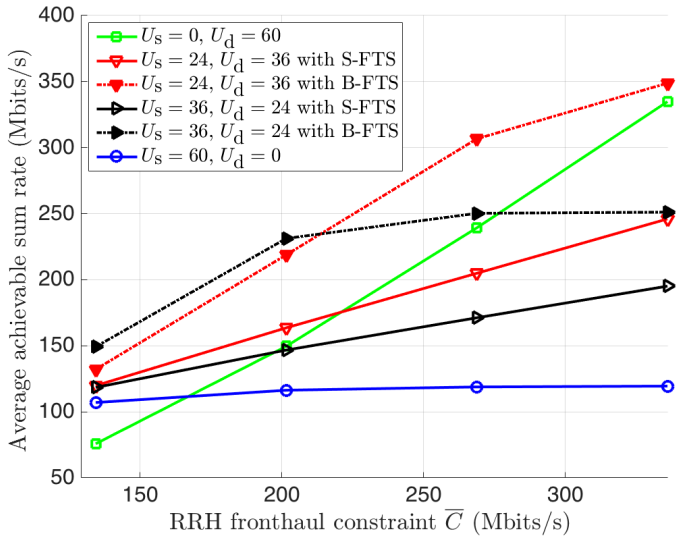

Fig. 7.11 Comparison of average achievable sum rate vs. fronthaul capacity constraint $\overline{C}$ for different transmission strategies (zoom).

Hybrid 24-36 transmission with B-FTS has a slightly lower performance than hybrid 36-24 transmission with B-FTS when the fronthaul capacity constraint is less than 200 Mbits/s. This is because with more UEs served in single RRH mode, more fronthaul capacity can be shared to each UE served in distributed MIMO mode. With the increasing of fronthaul capacity constraint, hybrid 24-36 transmission with B-FTS begins to surpass hybrid 36-24

transmission. This is because distributed MIMO mode has a better performance than single RRH mode when fronthaul capacity is high and there are more UEs served in distributed MIMO mode in hybrid 24-36 transmission than in hybrid 36-24 transmission. However, hybrid 24-36 transmission consumes more CSIs than hybrid 36-24 transmission.

7.5 Conclusion

In this chapter, we proposed a hybrid transmission strategy in a C-RAN downlink system with limited fronthaul capacity. In this downlink scheme, a cluster of RRHs serves a set of UEs according to two different modes: single RRH mode or distributed MIMO mode. The downlink mode for the UEs is decided in the BBU pool based on the statistical CSIs. Several low complexity user grouping algorithms are proposed to decide which UEs are served at the same time in order to improve the system throughput. Also, a new fronthaul transmission scheme is proposed to exploit the unused fronthaul resources by UEs served in single RRH mode to transmit data in advance for the benefit of UEs served in distributed MIMO mode.

Simulation results show that this strategy improves the system performance for hybrid transmissions when the fronthaul capacity is low. In particular, it outperforms both fully single RRH transmissions and fully distributed MIMO transmissions in low fronthaul capacity regime. With the increasing of fronthaul capacity, more UEs are to be served in distributed MIMO mode to improve system throughput.

Chapter 8

Conclusion

This chapter concludes the major contributions discussed in this thesis and gives possible directions of future work.

8.1 Major Contributions

C-RAN has the potential to increase the capacity of mobile networks while reducing operators' cost and energy consumption. Nevertheless, the large bit rate requirement in the fronthaul largely limits the future deployment of C-RAN. This study has been done in two aspects: 1) reduce C-RAN fronthaul throughput on the uplink; 2) design coordinated transmission in C-RAN downlink while considering user scheduling and fronthaul capacity allocation. The main contributions in the two aspects can be summarized separately as follows.

8.1.1 Reducing C-RAN fronthaul throughput on the uplink

This thesis investigates new functional split architectures between RRH and BBU, to reduce the transmission throughput between RRHs and BBUs. Two new architectures are proposed and modeled for the uplink. Part of physical layer functions of the BBU are moved to the RRH. For the proposed architectures, the transmission rate between RRHs and BBUs depends on the mobile network load, while that of current architecture is constant. Unlike most of relative works, we have done quantitative analyses on the impact of different functional splits. Simulation results illustrate that 30% to 40% bandwidth can be saved when all the radio channel capacity is used, and up to 70% bandwidth when half of the radio channel capacity is used.

8.1.2 Coordinated Transmission Design in C-RAN Downlink with Limited Fronthaul Capacity

CoMP can improve the system performance of the mobile network. Its implementation can be facilitated by C-RAN. However, the performance of CoMP is constrained by the high complexity and the limited fronthaul capacity of C-RAN. We design coordinated transmission scheme while considering fronthaul allocation. Low complexity user scheduling algorithms are also developed. This part of work is focused on the downlink. A cellular network model in 3GPP is adopted. A new RRH clustering scheme is proposed. It is a trade-off between disjoint clustering and user-centric clustering. We firstly compare different transmission strategies without and with cooperation between different cells while considering limited fronthaul capacity. Then we propose several user grouping algorithms to improve the performance of distributed MIMO. In distributed MIMO, several RRHs serve a number of UEs in parallel through Zero-Forcing. At last, a hybrid transmission strategy is proposed, in which part of UEs are served by only one RRH and the others served by several RRHs.

Comparison of Different Transmission Strategies

At first, we study the maximum throughput of different transmission strategies in a C-RAN cluster. Both RRH transmission power constraints and fronthaul capacity constraints are taken into account. The transmission strategies non-CoMP (e.g. single RRH mode) and CoMP (e.g. distributed MIMO) are considered. In single RRH mode, each UE is served by only one RRH. In distributed MIMO mode (D-MIMO), a cluster of RRHs serving a group of UEs applying Zero-Forcing (ZF) to do joint transmission. Simulation results show that distributed MIMO has a better performance than single RRH mode with high RRH power constraint, high fronthaul capacity constraint and when the UEs are located at cell edge area.

Improving Zero-Forcing Performance by User Grouping Algorithms

The UEs served by a cluster of RRHs usually have a total number of antennas larger than the RRHs. They are assigned to different subframes for serving. The performance D-MIMO applying ZF can be improved by appropriately choosing which UEs are served together in the same time frequency resource. With UE data and channel state information shared in BBU pool, C-RAN facilitates the centralized user scheduling. We propose several low complexity user grouping algorithms to maximize the average achievable sum rate.

Hybrid Transmission

The UEs in cell edge area benefit more from CoMP than UEs in cell center area. The extra fronthaul load and CSIs requirement in CoMP may counteract the throughput gain for UEs in cell center area [99]. We propose a hybrid transmission strategy to improve system performance with limited fronthaul capacity. The UEs in near RRH area are served in single RRH mode, and the UEs in cell edge area served in D-MIMO. This division is based on the UEs' statistical Channel State Informations (CSIs). A new fronthaul transmission scheme is also proposed. It allow the UEs served in D-MIMO exploit the unused fronthaul capacity for the UEs served in single RRH mode. This largely improves system performance when the fronthaul capacity is low. Simulation results show that the proposed hybrid mode and fronthaul strategy outperform both the case where all UEs are served in single RRH mode and the case where all UEs are served in D-MIMO when the fronthaul capacity is limited.

8.2 Future Work

This work can be extended in several possible directions for the future work. These directions are summarized as follows.

8.2.1 Coordinated Beamforming (CB)

In this thesis, we have not studied Coordinated Beamforming (CB). Distributed MIMO (D-MIMO) can achieve larger performance gain than CB. However, when we use data-sharing or compression-before-precoding (CBP) as the transmission scheme from BBU pool to RRHs, D-MIMO requires much higher fronthaul capacity than CB. Furthermore, the coordination between different cells for D-MIMO is in a signal level whereas that for CB is in a beamforming level. CB is easier to implement than D-MIMO. It would be interesting to compare single RRH mode, D-MIMO and CB with data-sharing or CBP.

8.2.2 Signaling Consumption

When we compare the performance of non-CoMP and CoMP transmission strategies, the channel state informations (CSIs) consumption has not been taken into account. In Frequency Division Duplex (FDD), extra downlink transmission resources are used for the CSIs measurements in the UEs' side. The feedback of CSIs from UEs to RRHs also costs significant uplink resources. Moreover, the transmission of CSIs from RRHs to the BBU

pool for processing generates large fronthaul throughput. In TDD, relying on reciprocity between the uplink and downlink channels, CSIs can be directly gotten in the RRHs' side. However, this still cannot avoid the CSIs transmission on the fronthaul for cooperative processing. Therefore, the evaluation of the performance of CoMP would be more realistic with considering the CSIs consumption.

8.2.3 Imperfect Channel State Information (CSI)

In this thesis, we assume that perfect channel state information (CSI) exists in the BBU pool. However, in practical, it is difficult to obtain perfect CSIs. The performance of CoMP is sensitive to imperfect and outdated CSIs. It is interesting to develop CoMP transmission schemes which are robust to the CSI errors. This can also reduce the need for CSI feedback, which in turn leads to less transmission costs both in radio interface and fronthaul.

8.2.4 Heterogeneous Cloud Radio Access Networks (H-CRAN)

The studies on Heterogeneous networks (HetNets) are hot in recent years. The deployment of low-power nodes (e.g. pico base station, small cell base station) in UEs dense areas is a promising solution to satisfy the high traffic demands. Nevertheless, the coexistence of macro sites and dense low-power nodes results in complex interference environment. Heterogeneous Cloud Radio Access Networks (H-CRAN) has been proposed to facilitate the management of the interference. H-CRAN is a mix of HetNets and C-RAN. The overall interference management in HetNets is much more challenging than in a macro cellular network. To develop low complexity CoMP strategies while considering UEs fairness, limited fronthaul capacity and energy consumption in H-CRAN would be very interesting.

References

[1] Cisco Visual Networking Index: Forecast and Methodology, 2016-2021.

[2] Stefania Sesia, Issam Toufik, and Matthew Baker. *LTE, The UMTS Long Term Evolution: From Theory to Practice, Second Edition*. Wiley Publishing, 2011.

[3] Dahlman Erik, Parkvall Stefan, and Skold Johan. *4G LTE/LTE Advanced for Mobile Broadband, Second Edition*. Elsevier, 2014.

[4] Sumitomo Electric Group Corporate.

[5] China Mobile Research Institute. The Road Towards Green RAN (White Paper).

[6] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M.A. Uusitalo, B. Timus, and M. Fallgren. Scenarios for 5G mobile and wireless communications: the vision of the METIS project. *IEEE Communications Magazine*, 52(5):26–35, May 2014.

[7] Insoo Hwang, Bongyong Song, and S.S. Soliman. A holistic view on hyper-dense heterogeneous and small cell networks. *IEEE Communications Magazine*, 51(6):20–27, June 2013.

[8] R. Q. Hu and Y. Qian. An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems. *IEEE Communications Magazine*, 52(5):94–101, May 2014.

[9] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, February 2014.

[10] E. Bjornson, E. G. Larsson, and T. L. Marzetta. Massive MIMO: ten myths and one critical question. *IEEE Communications Magazine*, 54(2):114–123, February 2016.

[11] ERICSSON Report. The 5G Business Potential - Industry digitalization and the untapped opportunities for operators. 2017.

[12] A. Capone, S. DElia, I. Filippini, A. E. C. Redondi, and M. Zangani. Modeling Energy Consumption of Mobile Radio Networks: An Operator Perspective. *IEEE Wireless Communications*, PP(99):2–8, 2017.

[13] S. Chen and J. Zhao. The requirements, challenges, and technologies for 5g of terrestrial mobile telecommunication. *IEEE Communications Magazine*, 52(5):36–43, May 2014.

[14] C. L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li. Recent Progress on C-RAN Centralization and Cloudification. *IEEE Access*, 2:1030–1039, 2014.

[15] U. Dropmann, X. Lagrange, and P. Godlewski. Architecture of a multi-cell centralized packet access system. In *PIMRC'95 : Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 27-29-September, Toronto, Canada*, pages 279–283, 1995.

[16] A. Checko, H.L. Christiansen, Ying Yan, L. Scolari, G. Kardaras, M.S. Berger, and L. Dittmann. Cloud RAN for Mobile Networks – A Technology Overview. *IEEE Communications Surveys Tutorials*, 17(1):405–426, 2015.

[17] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba. Network function virtualization: State-of-the-art and research challenges. *IEEE Communications Surveys Tutorials*, 18(1):236–262, Firstquarter 2016.

[18] M. Agiwal, A. Roy, and N. Saxena. Next Generation 5g Wireless Networks: A Comprehensive Survey. *IEEE Communications Surveys Tutorials*, 18(3):1617–1655, 2016.

[19] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang. Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues. *IEEE Communications Surveys Tutorials*, 18(3):2282–2308, 2016.

[20] V. Suryaprakash, P. Rost, and G. Fettweis. Are heterogeneous cloud-based radio access networks cost effective? *IEEE Journal on Selected Areas in Communications*, 33(10):2239–2251, Oct 2015.

[21] ZTE Shenzhen China. Green Technology Innovations white papers. *Tech. Rep.*, 2011.

[22] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis. Fronthaul and backhaul requirements of flexibly centralized radio access networks. *IEEE Wireless Communications*, 22(5):105–111, October 2015.

[23] M. Peng, C. Wang, V. Lau, and H. V. Poor. Fronthaul-constrained cloud radio access networks: insights and challenges. *IEEE Wireless Communications*, 22(2):152–160, April 2015.

[24] Common Public Radio Interface (CPRI); Interface Specification V6.0. August 2013.

[25] P. Chanclou, A. Pizzinat, F. Le Clech, T. L. Reedeker, Y. Lagadec, F. Saliou, B. Le Guyader, L. Guillo, Q. Deniel, S. Gosselin, S. D. Le, T. Diallo, R. Brenot, F. Lelarge, L. Marazzi, P. Parolari, M. Martinelli, S. O'Dull, S. A. Gebrewold, D. Hillerkuss, J. Leuthold, G. Gavioli, and P. Galli. Optical fiber solution for mobile fronthaul to achieve cloud radio access network. pages 1–11, July 2013.

[26] Anna Pizzinat, Philippe Chanclou, Fabienne Saliou, and Thierno Diallo. Things you should know about fronthaul. *J. Lightwave Technol.*, 33(5):1077–1083, Mar 2015.

[27] V. N. Ha, L. B. Le, and N. D. DAO. Coordinated Multipoint Transmission Design for Cloud-RANs With Limited Fronthaul Capacity Constraints. *IEEE Transactions on Vehicular Technology*, 65(9):7432–7447, September 2016.

[28] S. H. Park, O. Simeone, O. Sahin, and S. Shamai. Robust and Efficient Distributed Compression for Cloud Radio Access Networks. *IEEE Transactions on Vehicular Technology*, 62(2):692–703, February 2013.

[29] X. Rao and V. K. N. Lau. Distributed Fronthaul Compression and Joint Signal Recovery in Cloud-RAN. *IEEE Transactions on Signal Processing*, 63(4):1056–1065, February 2015.

[30] L. Liu and R. Zhang. Optimized Uplink Transmission in Multi-Antenna C-RAN With Spatial Compression and Forward. *IEEE Transactions on Signal Processing*, 63(19):5083–5095, October 2015.

[31] D. Wubben, P. Rost, J.S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis. Benefits and Impact of Cloud Computing on 5g Signal Processing: Flexible centralization through cloud-RAN. *IEEE Signal Processing Magazine*, 31(6):35–44, November 2014.

[32] G. Berardinelli, L. A. M. Ruiz de Temino, S. Frattasi, M. I. Rahman, and P. Mogensen. Ofdma vs. sc-fdma: performance comparison in local area imt-a scenarios. *IEEE Wireless Communications*, 15(5):64–72, October 2008.

[33] Aws Yonis, Mohammad Faiz Liew Abdullah, and M.F. Ghanim. Lte-fdd and lte-tdd for cellular communications. pages 1467–1471, 2012.

[34] Open Base Station Architecture Initiative.

[35] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra. An overview of the cpri specification and its application to c-ran-based lte scenarios. *IEEE Communications Magazine*, 54(2):152–159, February 2016.

[36] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy. FluidNet: A Flexible Cloud-Based Radio Access Network for Small Cells. *IEEE/ACM Transactions on Networking*, 24(2):915–928, April 2016.

[37] S. Namba, T. Warabino, and S. Kaneko. BBU-RRH switching schemes for centralized RAN. In *7th International Conference on Communications and Networking in China*, pages 762–766, August 2012.

[38] Bin Guo, Wei Cao, An Tao, and D. Samardzija. Cpri compression transport for lte and lte-a signal in c-ran. In *7th International Conference on Communications and Networking in China*, pages 843–849, Aug 2012.

[39] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela. Compressed transport of baseband signals in radio access networks. *IEEE Transactions on Wireless Communications*, 11(9):3216–3225, September 2012.

[40] RAN Evolution of LTE in Release 14.

[41] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, and V. Jungnickel. Coordinated multipoint: Concepts, performance, and field trial results. *IEEE Communications Magazine*, 49(2):102–111, February 2011.

[42] Daewon Lee, Hanbyul Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana. Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges. *IEEE Communications Magazine*, 50(2):148–155, February 2012.

[43] M. Feng, X. She, L. Chen, and Y. Kishiyama. Enhanced dynamic cell selection with muting scheme for dl comp in lte-a. In *2010 IEEE 71st Vehicular Technology Conference*, pages 1–5, May 2010.

[44] J. w Cho, J. Mo, and S. Chong. Joint network-wide opportunistic scheduling and power control in multi-cell networks. *IEEE Transactions on Wireless Communications*, 8(3):1520–1531, March 2009.

[45] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman. Pricing and power control in a multicell wireless data network. *IEEE Journal on Selected Areas in Communications*, 19(10):1883–1892, October 2001.

[46] H. Zhang, L. Venturino, N. Prasad, P. Li, S. Rangarajan, and X. Wang. Weighted Sum-Rate Maximization in Multi-Cell Networks via Coordinated Scheduling and Discrete Power Control. *IEEE Journal on Selected Areas in Communications*, 29(6):1214–1224, June 2011.

[47] T. Alpcan, T. Basar, and S. Dey. A power control game based on outage probabilities for multicell wireless data networks. *IEEE Transactions on Wireless Communications*, 5(4):890–899, April 2006.

[48] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu. Multi-Cell MIMO Cooperative Networks: A New Look at Interference. *IEEE Journal on Selected Areas in Communications*, 28(9):1380–1408, December 2010.

[49] W. Yu, T. Kwon, and C. Shin. Multicell coordination via joint scheduling, beamforming and power spectrum adaptation. In *2011 Proceedings IEEE INFOCOM*, pages 2570–2578, April 2011.

[50] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428)*, volume 1, pages 321–331 vol.1, March 2003.

[51] H. Dahrouj and W. Yu. Coordinated beamforming for the multicell multi-antenna wireless system. *IEEE Transactions on Wireless Communications*, 9(5):1748–1759, May 2010.

[52] L. Venturino, N. Prasad, and X. Wang. Coordinated linear beamforming in downlink multi-cell wireless networks. *IEEE Transactions on Wireless Communications*, 9(4):1451–1461, April 2010.

[53] S. He, Y. Huang, S. Jin, and L. Yang. Coordinated Beamforming for Energy Efficient Transmission in Multicell Multiuser Systems. *IEEE Transactions on Communications*, 61(12):4961–4971, December 2013.

[54] E. Pateromichelakis, M. Shariat, A. u Quddus, and R. Tafazolli. On the Evolution of Multi-Cell Scheduling in 3gpp LTE / LTE-A. *IEEE Communications Surveys Tutorials*, 15(2):701–717, 2013.

[55] Ralph Tanbourgi, Sarabjot Singh, Jeffrey G. Andrews, and Friedrich K. Jondral. Analysis of Non-Coherent Joint-Transmission Cooperation in Heterogeneous Cellular Networks. *arXiv:1402.2707 [cs, math]*, February 2014. arXiv: 1402.2707.

[56] T. R. Lakshmana, B. Makki, and T. Svensson. Frequency allocation in non-coherent joint transmission CoMP networks. In *2014 IEEE International Conference on Communications Workshops (ICC)*, pages 610–615, June 2014.

[57] Sun Yunfeng, Jiang Jing, and Liujun Hu. Joint transmission method in Coordinated Multi-Point transmission and reception. *ZTE communications*, No. 1, 2010.

[58] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela. Network coordination for spectrally efficient communications in cellular systems. *IEEE Wireless Communications*, 13(4):56–61, August 2006.

[59] C. F. Ball, R. Müllner, J. Lienhart, and H. Winkler. Performance analysis of closed and open loop mimo in lte. In *2009 European Wireless Conference*, pages 260–265, May 2009.

[60] Symmetricom. Timing and synchronization for LTE TDD and LTE Advanced mobile networks (White Paper).

[61] Cloud Radio Access Networks edited by Tony Q. S. Quek.

[62] J. Kang, O. Simeone, J. Kang, and S. Shamai. Fronthaul Compression and Precoding Design for C-RANs over Ergodic Fading Channels. *IEEE Transactions on Vehicular Technology*, PP(99):1–1, 2015.

[63] Sangkyu Park, Chan-Byoung Chae, and Saewoong Bahk. Before/after precoding massive MIMO systems for cloud radio access networks. *Journal of Communications and Networks*, 15(4):398–406, August 2013.

[64] J. Kang, O. Simeone, J. Kang, and S. Shamai. Layered Downlink Precoding for C-RAN Systems With Full Dimensional MIMO. *IEEE Transactions on Vehicular Technology*, 66(3):2170–2182, March 2017.

[65] Osvaldo Simeone, Andreas Maeder, Mugen Peng, Onur Sahin, and Wei Yu. Cloud Radio Access Network: Virtualizing Wireless Access for Dense Heterogeneous Systems. *arXiv:1512.07743 [cs, math]*, December 2015. arXiv: 1512.07743.

[66] J. Lorca and L. Cucala. Lossless compression technique for the fronthaul of lte/lte-advanced cloud-ran architectures. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–9, June 2013.

[67] U. Dötsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier. Quantitative analysis of split base station processing and determination of advantageous architectures for lte. *Bell Labs Technical Journal*, 18(1):105–128, June 2013.

[68] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen. Evaluating c-ran fronthaul functional splits in terms of network level energy and cost savings. *Journal of Communications and Networks*, 18(2):162–172, April 2016.

[69] H.G. Myung, Junsung Lim, and D. Goodman. Single carrier FDMA for uplink wireless transmission. *IEEE Vehicular Technology Magazine*, 1(3):30–38, September 2006.

[70] China Mobile Research Institute. Next generation fronthaul interface (White Paper).

[71] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA) Radio Frequency (RF) system scenarios version 13.0.0 Release 13. TR 36.942, 3rd Generation Partnership Project (3GPP), January 2016.

[72] J. Gong, S. Zhou, Z. Niu, L. Geng, and M. Zheng. Joint Scheduling and Dynamic Clustering in Downlink Cellular Networks. In *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, pages 1–5, December 2011.

[73] D. Liu, S. Han, C. Yang, and Q. Zhang. Semi-dynamic User-Specific Clustering for Downlink Cloud Radio Access Network. *IEEE Transactions on Vehicular Technology*, 65(4):2063–2077, April 2016.

[74] B. Dai and W. Yu. Sparse Beamforming and User-Centric Clustering for Downlink Cloud Radio Access Network. *IEEE Access*, 2:1326–1339, 2014.

[75] A. Papadogiannis, D. Gesbert, and E. Hardouin. A Dynamic Clustering Approach in Wireless Networks with Multi-Cell Cooperative Processing. In *2008 IEEE International Conference on Communications*, pages 4033–4037, May 2008.

[76] Yuanming Shi, Jun Zhang, and K.B. Letaief. Group Sparse Beamforming for Green Cloud-RAN. *IEEE Transactions on Wireless Communications*, 13(5):2809–2823, May 2014.

[77] Ruxuan Jiao, Xiangming Wen, Zhaoming Lu, Yawen Chen, Hua Shao, and Wenpeng Jing. Dynamic user-centric clustering algorithm based on energy efficiency in Cloud-RAN. In *2017 24th International Conference on Telecommunications (ICT)*, pages 1–7, May 2017.

[78] M. Hong, R. Sun, H. Baligh, and Z. Q. Luo. Joint Base Station Clustering and Beamformer Design for Partial Coordinated Transmission in Heterogeneous Networks. *IEEE Journal on Selected Areas in Communications*, 31(2):226–240, February 2013.

[79] Vincent W. S. Wong, Robert Schober, Derrick Wing Kwan Ng, and Li-Chun Wang. *Key Technologies for 5G Wireless Systems*. Cambridge University Press, Cambridge, United Kingdom, 2017.

[80] Ekram Hossain, Dong In Kim, and Vijay K. Bhargava. *Cooperative Cellular Wireless Networks*. Cambridge University Press, New York, NY, USA, 2011.

[81] Q. Zhang, C. Yang, and A. F. Molisch. Downlink Base Station Cooperative Transmission Under Limited-Capacity Backhaul. *IEEE Transactions on Wireless Communications*, 12(8):3746–3759, August 2013.

[82] R. Zakhour and D. Gesbert. Optimized data sharing in multicell mimo with finite backhaul capacity. *IEEE Transactions on Signal Processing*, 59(12):6102–6111, Dec 2011.

[83] C. T. K. Ng and H. Huang. Linear Precoding in Cooperative MIMO Cellular Networks with Limited Coordination Clusters. *IEEE Journal on Selected Areas in Communications*, 28(9):1446–1454, December 2010.

[84] Vu Nguyen Ha, Long Bao Le, and Ngoc-Dung Dao. Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints. In *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1862–1867, April 2014.

[85] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai (Shitz). Downlink multicell processing with limited-backhaul capacity. *EURASIP Journal on Advances in Signal Processing*, 2009(1):840814, Jun 2009.

[86] O. Somekh, O. Simeone, Y. Bar-Ness, A. M. Haimovich, and S. Shamai. Cooperative Multicell Zero-Forcing Beamforming in Cellular Downlink Channels. *IEEE Transactions on Information Theory*, 55(7):3206–3219, July 2009.

[87] Q.H. Spencer, A.L. Swindlehurst, and M. Haardt. Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels. *IEEE Transactions on Signal Processing*, 52(2):461–471, February 2004.

[88] S. H. Park, O. Simeone, O. Sahin, and S. Shamai. Joint Precoding and Multivariate Backhaul Compression for the Downlink of Cloud Radio Access Networks. *IEEE Transactions on Signal Processing*, 61(22):5646–5658, November 2013.

[89] S. Huang, H. Yin, J. Wu, and V. C. M. Leung. User Selection for Multiuser MIMO Downlink With Zero-Forcing Beamforming. *IEEE Transactions on Vehicular Technology*, 62(7):3084–3097, September 2013.

[90] J. Wang, D. J. Love, and M. D. Zoltowski. User selection with zero-forcing beamforming achieves the asymptotically optimal sum rate. *IEEE Transactions on Signal Processing*, 56(8):3713–3726, Aug 2008.

[91] J. C. Mundarath, P. Ramanathan, and B. D. Van Veen. A distributed downlink scheduling method for multi-user communication with zero-forcing beamforming. *IEEE Transactions on Wireless Communications*, 7(11):4508–4521, November 2008.

[92] G. Dimic and N. D. Sidiropoulos. On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm. *IEEE Transactions on Signal Processing*, 53(10):3857–3868, October 2005.

[93] S. Akhlaghi, A. K. Khandani, and A. Falahati. User Selection and Signaling Over Time-Varying MIMO Broadcast Channels. In *23rd Biennial Symposium on Communications, 2006*, pages 31–34, 2006.

[94] Taesang Yoo and A. Goldsmith. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming. *IEEE Journal on Selected Areas in Communications*, 24(3):528–541, March 2006.

[95] L. Liu, S. Bi, and R. Zhang. Joint Power Control and Fronthaul Rate Allocation for Throughput Maximization in OFDMA-Based Cloud Radio Access Network. *IEEE Transactions on Communications*, 63(11):4097–4110, November 2015.

[96] Jinkyu Kang, Osvaldo Simeone, Joonhyuk Kang, and Shlomo Shamai. Layered Downlink Precoding for C-RAN Systems with Full Dimensional MIMO. *arXiv:1511.08084 [cs, math]*, November 2015. arXiv: 1511.08084.

[97] Liu Liang, Suzhi Bi, and Rui Zhang. Joint Power Control and Fronthaul Rate Allocation for Throughput Maximization in OFDMA-based Cloud Radio Access Network. *arXiv:1407.3855 [cs, math]*, July 2014. arXiv: 1407.3855.

[98] S. K. Mohammed and E. G. Larsson. Improving the Performance of the Zero-Forcing Multiuser MISO Downlink Precoder Through User Grouping. *IEEE Transactions on Wireless Communications*, 15(2):811–826, February 2016.

[99] Q. Zhang and C. Yang. Transmission Mode Selection for Downlink Coordinated Multipoint Systems. *IEEE Transactions on Vehicular Technology*, 62(1):465–471, January 2013.

[100] N. Seifi, M. Viberg, R. W. Heath, J. Zhang, and M. Coldrey. Coordinated single-cell vs multi-cell transmission with limited-capacity backhaul. In *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, pages 1217–1221, November 2010.

[101] X. Yi and N. Liu. An achievability scheme for downlink multicell processing with finite backhaul capacity: The general case. In *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, pages 1–5, October 2015.

Publication list

[DLG16a] Jialong Duan, Xavier Lagrange, and Frédéric. Guilloud. Maximum throughput in a C-RAN cluster with limited fronthaul capacity. In *2016 19th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, November 2016.

[DLG16b] Jialong Duan, Xavier Lagrange, and Frédéric. Guilloud. Performance Analysis of Several Functional Splits in C-RAN. In *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pages 1–5, May 2016.

[DLG17] Jialong Duan, Xavier Lagrange, and Frédéric. Guilloud. Analysis of different user grouping algorithms in a c-ran downlink system. In *2017 IEEE International Conference on Communications (ICC)*, May 2017.

Appendix A

Proof of Theorem 1

Remind that $\tilde{\Omega} = \text{diag}([\tilde{\sigma}_{1,t}^2 \mathbf{I}_{N_{\text{TA}}},\ ...,\ \tilde{\sigma}_{M,t}^2 \mathbf{I}_{N_{\text{TA}}}])$ as mentioned in Section 6.2.2.

Let $\Sigma = \text{diag}([\tilde{\sigma}_{1,t} \mathbf{I}_{N_{\text{TA}}},\ ...,\ \tilde{\sigma}_{M,t} \mathbf{I}_{N_{\text{TA}}}])$, then

$$\mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\text{H}} = \mathbf{H}_u \Sigma \Sigma \mathbf{H}_u^{\text{H}} = (\mathbf{H}_u \Sigma)(\mathbf{H}_u \Sigma)^{\text{H}} \tag{A.1}$$

For any matrix U, the matrix UU^{H} is positive semidefinite. Therefore, each eigenvalue of $\mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\text{H}}$ is non-negative: $\lambda_{u,i} \geq 0,\ i = 1,\ 2, \ldots\ N_{\text{UA}}$, where $\lambda_{u,1}, \ldots\ \lambda_{u,N_{\text{UE}}}$ are the N_{UA} eigenvalues of $\mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\text{H}}$.

As $\mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\text{H}}$ in (6.24) is a Hermitian positive semidefinite matrix, it is diagonalizable. Thus there exists an invertible matrix $\mathbf{P}$ which lets

$$\mathbf{H}_u \tilde{\Omega} \mathbf{H}_u^{\text{H}} = \mathbf{P}^{-1} \Lambda_u \mathbf{P} \tag{A.2}$$

where $\Lambda_u = \text{diag}\,([\lambda_{u,1}, \ldots\ \lambda_{u,N_{\text{UE}}}])$.

Using (A.2) and because $\det(\mathbf{P}^{-1})\det(\mathbf{P}) = 1$, therefore the achievable transmission rate for

$$\begin{aligned} R_u &= \log \frac{\det(\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\mathbf{H}_u\tilde{\Omega}\mathbf{H}_u^{\mathrm{H}})}{\det(\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\mathbf{H}_u\tilde{\Omega}\mathbf{H}_u^{\mathrm{H}})} \\ &= \log \frac{\det(\mathbf{P})\det(\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\mathbf{P}^{-1}\Lambda_u\mathbf{P})\det(\mathbf{P}^{-1})}{\det(\mathbf{P})\det(\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\mathbf{P}^{-1}\Lambda_u\mathbf{P})\det(\mathbf{P}^{-1})} \\ &= \log \frac{\det(\mathbf{P}(1+\gamma_t^2)\mathbf{I}_{N_{\mathrm{UA}}}\mathbf{P}^{-1}+\gamma_t^2\mathbf{P}\mathbf{P}^{-1}\Lambda_u\mathbf{P}\mathbf{P}^{-1})}{\det(\mathbf{P}\mathbf{I}_{N_{\mathrm{UA}}}\mathbf{P}^{-1}+\mathbf{P}\mathbf{P}^{-1}\gamma_t^2\Lambda_u\mathbf{P}\mathbf{P}^{-1})} \\ &= \log \frac{\det((1+\gamma_t^2)\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\Lambda_u)}{\det(\mathbf{I}_{N_{\mathrm{UA}}}+\gamma_t^2\Lambda_u)} \\ &= \log \frac{\prod_{i=1}^{N_{\mathrm{UA}}}(1+\gamma_t^2+\gamma_t^2\lambda_{u,i})}{\prod_{i=1}^{N_{\mathrm{UA}}}(1+\gamma_t^2\lambda_{u,i})} \\ &= \sum_{i=1}^{N_{\mathrm{UA}}} \log \frac{1+\gamma_t^2+\gamma_t^2\lambda_{u,i}}{1+\gamma_t^2\lambda_{u,i}} \end{aligned} \tag{A.3}$$

If $\gamma_t^2 = 0$, then $R_u = 0$.

If $\gamma_t^2 > 0$, (A.3) can be changed to

$$R_u = \sum_{i=1}^{N_{\mathrm{UA}}} \log(1 + \frac{1}{\frac{1}{\gamma_t^2} + \lambda_{u,i}}) \tag{A.4}$$

As $\lambda_{u,i} \geq 0 \; \forall u \in \mathcal{N}_{\mathrm{UE},t} \; i = 1, \ldots N_{\mathrm{UA}}$, it can easily be proved that the maximum value of γ_t^2:

$$\gamma_t^2 = \min\{\frac{\overline{P}}{\mathrm{tr}\big(\tilde{\mathbf{V}}_{t,1}^r(\tilde{\mathbf{V}}_{t,1}^r)^{\mathrm{H}} + \tilde{\sigma}_{1,t}^2\mathbf{I}_{N_{\mathrm{TA}}}\big)}, \ldots \frac{\overline{P}}{\mathrm{tr}\big(\tilde{\mathbf{V}}_{t,M}^r(\tilde{\mathbf{V}}_{t,M}^r)^{\mathrm{H}} + \tilde{\sigma}_{M,t}^2\mathbf{I}_{N_{\mathrm{TA}}}\big)}\} \tag{A.5}$$

maximizes (A.4). The maximized value is obviously larger than 0.

Therefore, problem (6.24) can be solved by (A.5). The proof is completed.

Appendix B

Coordination dans les réseaux d'accès radio centralisés avec liaisons de transport à débit limité

B.1 Contexte de la thèse

Afin de satisfaire l'explosion du trafic de données, les opérateurs doivent investir largement pour augmenter la capacité du réseau mobile. Le réseau d'accès radio centralisé (*Centralized/Cloud Radio Access Network*, C-RAN) peut fortement augmenter la capacité des réseaux mobiles tout en réduisant le coût et la consommation d'énergie des opérateurs.

Dans le LTE (Long Term Evolution), la plupart des eNodeBs (eNB) contiennent deux parties principales: les têtes radios distantes (*remote radio head*, RRH) et les unités de bande de base (*baseband unit*, BBU). C-RAN éloigne les BBUs des RRHs et les place dans une grappe de BBU pour un traitement centralisé. La centralisation des BBUs facilite la coopération entre différentes cellules. Il permet de nombreux algorithmes de cooperation avancés, tels que la transmission coordonnée Multi-point (CoMP). CoMP peut améliorer l'efficacité spectrale.

Cependant, le déploiement commercial de C-RAN fait face à de nombreux défis. Une limitation principale de la faisabilité de C-RAN est le flux considérable engendré sur les liaisons de transport, appelées également fronthaul. Du point de vue du système, une architecture C-RAN est satisfaisante si elle fournit un débit élevé avec un coût limité, en

d'autres termes, une charge réduite sur le fronthaul. L'objectif de cette thèse est de maximiser le débit de l'utilisateur tout en tenant compte des contraintes de fronthaul. De plus, dans C-RAN, il est complexe de réaliser l'ordonnancement des utilisateurs, l'allocation des ressources, les algorithmes coordonnés avancés avec des ressources de fronthaul limitées. Nous essayons également de résoudre ce problème dans cette thèse.

B.2 Les contributions

Cette thèse est divisée en deux parties. Dans la première partie, nous proposons de nouveaux schémas de division fonctionnelle qui déplacent une partie des fonctions dans BBU vers RRH afin de réduire le débit sur fronthaul. Comme les découpages fonctionnels sont plus complexes sur le lien montant que sur le lien descendant, nous nous concentrons sur le lien montant dans cette étude. Dans la deuxième partie, nous étudions l'application de CoMP sur le lien descendant dans C-RAN pour améliorer la capacité du réseau. Nous approfondissons l'étude de l'ordonnancement des utilisateurs et de l'allocation des ressources de fronthaul. Nous analysons le compromis entre la charge sur le fronthaul et le débit sur la chaîne radio. Les principales contributions de cette thèse sont résumées ci-dessous.

B.2.1 Réduction du débit de fronthaul sur la liaison montante

Cette thèse étudie les nouvelles architectures de RRH et BBU pout réduire le débit de transmission entre RRH et BBU. Deux nouvelles architectures sont proposées et modélisées sur la liaison montante. Une partie des fonctions dans le couche physique de BBU sont déplacées vers RRH. Pour les architectures proposées, le débit entre RRH et BBU dépend la charge du réseau mobile, tandis qu'il est constant dans l'architecture actuelle. Contrairement à la plupart des travaux précédents, nous avons effectué des analyses quantitatives sur l'impact des différentes divisions fonctionnelles. Les résultats de la simulation illustrent que 30% à 40% de la bande passante peut être économisée lorsque toute la capacité du canal radio est utilisée et jusqu'à 70% lorsque la moitié de la capacité du canal radio est utilisée. Cette partie du travail a été publiée dans la conférence internationale VTC au printemps 2016 [DLG16b].

B.2.2 Conception de transmission coordonnée dans C-RAN avec capacité limitée de fronthaul

CoMP peut améliorer les performances du réseau mobile. Cependant, la performance de CoMP est limitée par la complexité et la capacité bornée de fronthaul. Nous concevons un schéma de transmission coordonné tout en considérant l'allocation des resources sur le fronthaul. Cette partie du travail se concentre sur le lien descendant. Un nouveau schéma de constitution de grappes des RRHs est proposé. Nous comparons d'abord différentes stratégies de transmission sans et avec la coopération entre différentes cellules tout en considérant la capacité de fronthaul limitée. Ensuite, nous proposons plusieurs algorithmes de groupement d'utilisateurs pour améliorer les performances du MIMO (*Multiple-Input Multiple-Output*) distribués, appelés D-MIMO. Dans D-MIMO, plusieurs RRHs servent un certain nombre d'UE en parallèle par Zero-Forcing (ZF). Enfin, une stratégie de transmission hybride est proposée, dans laquelle une partie des utilisateurs sont servies par un seul RRH et les autres servis par plusieurs RRHs.

Comparaison des différentes stratégies de transmission

Au début, nous étudions le débit maximal de différentes stratégies de transmission dans une grappe de RRHs de C-RAN. Les contraintes de puissance de RRH et de capacité de fronthaul sont prises en compte. Les stratégies de transmission non-CoMP (par exemple en mode RRH unique) et CoMP (par exemple D-MIMO) sont considérées. En RRH unique mode, chaque utilisateur est servi par un seul RRH. En D-MIMO mode, une grappe de RRHs servent une groupe d'utilisateurs appliquant ZF pour effectuer une transmission conjointe. Les résultats de simulation montrent que D-MIMO a une meilleure performance que le mode RRH unique avec une puissance de RRH élevée, une capacité de fronthaul élevée et lorsque les utilisateurs sont situés en bordure de cellule. Cette partie du travail a été presentée dans la conférence internationale WPMC 2016 [DLG16a].

Amélioration zero-forcing par algorithme de regroupement d'utilisateurs

Les utilisateurs servis par une grappe de RRHs ont souvent un nombre total d'antennes plus grandes que les RRHs. Ils sont servis dans différentes sous-trames temporelles. La performance de D-MIMO appliquant ZF peut être améliorée en choisissant de manière appropriée quels utilisateurs sont servis ensemble dans la même ressource de fréquence et de temps. Avec les données des utilisateurs et les informations d'état du canal (*channel*

state information, CSI) partagées dans la grappe de BBU, C-RAN facilite l'ordonnancement centralisé des utilisateurs. Nous proposons plusieurs algorithmes de groupement d'utilisateurs de faible complexité pour maximiser le débit total.

Nous étudions d'abord le scénario avec une capacité de fronthaul illimitée et avec une seule antenne sur chacune utilisateur et RRH. Cette partie du travail a été presentée dans la conférence internationale ICC 2017 [DLG17].

Ensuite, nous étendons le travail au scénario avec une capacité de frontal limitée et avec plusieurs antennes sur les utilisateurs et sur les RRHs.

Transmission hybride

Les utilisateurs en bordure de cellule bénéficient plus de CoMP que les utilisateurs dans les zones centrales. La charge supplémentaire de fronthaul et l'exigence de CSI dans CoMP peuvent amoindrir le gain de débit pour les utilisateurs dans la zone centrale de cellule. Nous proposons une stratégie de transmission hybride pour améliorer les performances du système avec une capacité de fronthaul limitée. Les utilisateurs dans la région proche de RRH sont servis en mode RRH unique, et les utilisateurs en bordure de la cellule servis en mode D-MIMO. Cette division est basée sur les CSIs statistiques.

Dans un schéma typique de transmission sur fronthaul, les données transmises du pool de BBU aux RRHs pendant la sous-trame t sont transmises des RRHs aux utilisateurs sur l'interface radio pendant la sous-trame $t+1$. Ce schéma est appelé S-FTS (Synchronous Fronthaul Transmission Scheme). Un nouveau schéma de transmission de fronthaul est également proposé. Il permet aux utilisateurs servis en mode D-MIMO d'exploiter la capacité de fronthaul inutilisée pour les utilisateurs servis en mode RRH unique. On l'appelle B-FTS (Buffered Fronthaul Transmission Scheme).

On étudie un scénario où un nombre fixe d'utilisateurs est servi par une grappe de RRHs. Le nombre des utilisateurs servis en mode RRH unique est désigné par U_{s} et celui servis en D-MIMO mode par U_{d}. Les résultats de la simulation sont montrés sur la Figure B.1. Elle montre que, comparé avec S-FTS, B-FTS améliore grandement les performances du système lorsque la capacité de fronthaul est faible. En même temps, on peut observer que les performances du mode hybride proposé avec B-FTS dépassent le cas où tous les utilisateurs sont servis en mode RRH unique et le cas où tous les utilisateurs sont servis en mode D-MIMO quand la capacité de fronthaul est limitée.

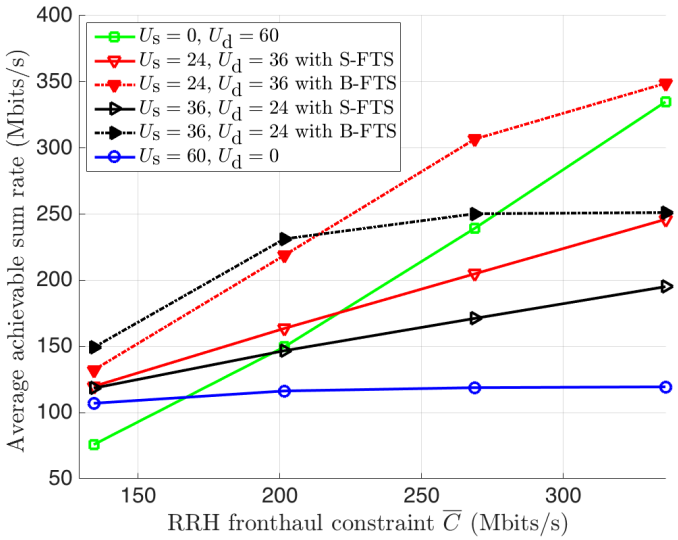

Fig. B.1 Comparaison des débits totaux vs. capacité de fronthaul $\overline{C}$ pour différentes stratégies de transmission.

B.3 Perspectives

Dans l'avenir, nous prendrons en compte la consommation des CSIs pour différents schémas de transmission. En outre, nous supposons que tous les CSIs disponible dans les BBUs sont parfaits. Mais, c'est difficile d'obtenir les CSIs parfaits dans la pratique. Nous allons aussi considérer les erreurs sur les CSIs à cause des délais et de la quantification.

Nomenclature

Roman Symbols

$\mathcal{C}$	a cluster of RRHs
$\overline{C}$	fronthaul capacity constraint
$d_{u,v}$	distance between RRH v and UE u
E_s	average symbol energy
f	index of subcarrier
F_{os}	oversampling factor
f_{s}	sampling frequency
$\mathbf{H}$	channel matrix
M	number of RRHs in a cluster
N_c	number of collections
$N_{g,\mathrm{d}}$	number of UEs served in distributed MIMO mode during one subframe
$N_{g,\mathrm{s}}$	number of UEs served in single RRH mode during one subframe
$N_{g,t}$	number of UEs served during subframe t
N_{Sc}	number of active subcarriers
N_T	number of subframes
N_{TA}^{v}	number of antennas for RRH v
$N_{T,\mathrm{d}}$	length of $\mathcal{T}_{\mathrm{d}}$
$\mathcal{N}_{\mathrm{TP}}$	set of RRHs in the cluster $\mathcal{C}$
$N_{T,\mathrm{s}}$	length of $\mathcal{T}_{\mathrm{s}}$
$\mathcal{N}_{\mathrm{UE,d}}$	set of UEs served in distributed MIMO mode
$\mathcal{N}_{\mathrm{UE,s}}$	set of UEs served in single RRH mode
$\mathcal{N}_{\mathrm{UE},t}$	set of UEs served during subframe t
N_{UA}^{u}	number of antennas for UE u
P_{dBm}	RRH power constraint in dBm
$\overline{P}$	RRH power constraint
$\mathbf{Q}$	quantization noise matrix

q_n	n-th quantization level
Q_q	length of quantization bits
Q_r	number of bits for reference of PRB
Q_s	number of bits for scaling factor
r_c	code rate of FEC applied
R_u	achievable transmission rate for UE u
t	index of subframe
$\mathcal{T}$	set of subframes
$\mathcal{T}_\mathrm{d}$	set of subframes when the UEs are served in distributed MIMO mode
$\mathcal{T}_\mathrm{s}$	set of subframes when the UEs are served in single RRH mode
U	number of UEs served by a cluster of RRHs
u	index of UE
U_d	number of UEs served in distributed MIMO mode
U_s	number of UEs served in single RRH mode
$\mathbf{V}$	precoding matrix
v	index of RRH
$\mathbf{X}$	signal transmitted by RRHs
$\mathbf{Y}$	signal received by UEs

Greek Symbols

$\alpha_{u,v}$	path loss coefficient between RRH v and UE u
η	assumed PRB utilization ratio
γ	normalization factor
$\rho_{u,v}$	shadow fading coefficient between RRH v and UE u
$\theta_{u,v}$	angle between the line (RRH v, UE u) and the antenna orientation of RRH v

Other Symbols

$(\cdot)^\mathrm{H}$	hermitian matrix transpose
$(\cdot)^\star$	complex conjugate
$(\cdot)^\mathrm{T}$	matrix transpose
$\mathbb{C}$	complex field
$\Theta(\cdot)$	complexity of an algorithm
$\mathbf{I}_n$	identity matrix of size n
$\mathcal{N}[i]$	the i-th element in set $\mathcal{N}$
$\lvert\mathcal{N}\rvert$	the cardinal of $\mathcal{N}$
$I(\mathbf{X};\mathbf{Y})$	the mutual information between random variables $\mathbf{X}$ and $\mathbf{Y}$
$\mathbb{1}$	two-element set $\{0,1\}$

Acronyms / Abbreviations

BBU	Baseband Unit
BS	Base Station
CAP	Compression After Precoding
CBP	Compression Before Precoding
CDD	Cyclic Delay Diversity
CoMP	Coordinated Multiple-point
CP	Cyclic Prefix
CPRI	Common Public Radio Interface
C-RAN	Cloud/Centralized Radio Access Network
CS/CB	Coordinated Scheduling/Coordinated Beamforming
CSI	Channel State Information
DFT	Discrete Fourier Transform
DFTS-OFDM	Discrete Fourier Transform Spread Orthogonal Frequency-Division Multiplexing
D-MIMO	Distributed MIMO
DPC	Dirty Paper Coding
DPS	Dynamic Point Selection
E-UTRA	Evolved Universal Terrestrial Radio Access
EVM	Error Vector Magnitude
FDD	Frequency Division Duplexing
FFT	Fast Fourier Transform
GUGA	Greedy User Grouping Algorithm
ICIC	Inter-cell Interference Coordination
ICI	Inter-Cell Interference
JT	Joint Transmission
LTE	Long Term Evolution
MIMO	Multiple Input Multiple Output
MMSE	Minimum Mean Square Error
MU-MIMO	Multi-user Multiple Input Multiple Output
NFV	Network Function Virtualization
NGFI	Next Generation Fronthaul Interface
OFDMA	Orthogonal Frequency-Division Multiplexing Access
PAPR	Peak to Average Power Ratio
PP-GUGA	Pre Partitioned - Greedy User Grouping Algorithm
PP-RUGA	Pre Partitioned - Random User Grouping Algorithm

PP-SUGA	Pre Partitioned - Semi-orthogonal User Grouping Algorithm
PRB	Physical Resource Block
QAM	Quadrature Amplitude Modulation
RE	Resource Element
RRH	Remote Radio Head
RR	Round Robin Selection
SUGA	Semi-orthogonal User Grouping Algorithm
SU-JT	Single User Joint Transmission
TDD	Time Division Duplexing
UDA	User Division Algorithm
UE	User Equipment
VM	Virtual Machine
ZF	Zero Forcing

Résumé

Le réseau d'accès radio centralisé (C-RAN) peut fortement augmenter la capacité des réseaux mobiles. Cependant, la faisabilité de C-RAN est limitée par le débit considérable engendré sur les liaisons de transport, appelées également fronthaul. L'objectif de cette thèse est d'améliorer les performances de C-RAN tout en considérant les limitations du débit sur le frontaul, l'allocation de ressources et l'ordonnancement des utilisateurs.

Nous étudions d'abord les séparations fonctionnelles possibles entre les têtes radios distantes (RRH) et les unités de traitement en bande de base (BBU) sur la liaison montante pour réduire le débit de transmission sur le fronthaul : certaines fonctions de couche basse sont déplacées du BBU vers les RRH. Nous fournissons une analyse quantitative des améliorations de performances ainsi obtenues.

Nous nous concentrons ensuite sur la transmission coordonnée Multi-point (CoMP) sur le lien descendant. CoMP peut améliorer l'efficacité spectrale mais nécessite une coordination inter-cellule, ce qui est possible uniquement si une capacité fronthaul élevée est disponible. Nous comparons des stratégies de transmission avec et sans coordination inter-cellule. Les résultats de simulation montrent que CoMP doit être préféré pour les utilisateurs situés en bordure de cellule et lorsque la capacité du fronthaul est élevée. Nous en déduisons une stratégie hybride pour laquelle Les utilisateurs sont divisés en deux sous-ensembles en fonction de la puissance du signal. Les utilisateurs situés dans les zones centrales sont servis par un seul RRH avec une coordination simple et ceux en bordure de cellule sont servis en mode CoMP. Cette stratégie hybride constitue un bon compromis entre les débits offerts aux utilisateurs et les débits sur le fronthaul.

Mots clefs : C-RAN, RRH, Quantification, BBU, Fronthaul, Regroupement d'utilisateurs, Précodage

Abstract

Centralized/Cloud Radio Access Network (C-RAN) is a promising mobile network architecture, which can potentially increase the capacity of mobile networks while reducing operators' cost and energy consumption. However, the feasibility of C-RAN is limited by the large bit rate requirement in the fronthaul. The objective of this thesis is to improve C-RAN performance while considering fronthaul throughput reduction, fronthaul capacity allocation and users scheduling.

We first investigate new functional split architectures between Remote Radio Heads (RRHs) and Baseband Units (BBU) on the uplink to reduce the transmission throughput in fronthaul. Some low layer functions are moved from the BBU to RRHs and a quantitative analysis is provided to illustrate the performance gains.

We then focus on Coordinated Multi-point (CoMP) transmissions on the downlink. CoMP can improve spectral efficiency but needs tight coordination between different cells, which is facilitated by C-RAN only if high fronthaul capacity is available. We compare different transmission strategies without and with multi-cell coordination. Simulation results show that CoMP should be preferred for users located in cell edge areas and when fronthaul capacity is high. We propose a hybrid transmission strategy where users are divided into two parts based on statistical Channel State Informations (CSIs). The users located in cell center areas are served by one transmission point with simple coordinated scheduling and those located in cell edge areas are served with CoMP joint transmission. This proposed hybrid transmission strategy offers a good trade-off between users' transmission rates and fronthaul capacity cost.

Keywords: C-RAN, RRH, Quantization, BBU, Fronthaul, User grouping, Precoding

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

N° d'ordre : 2017IMTA0052

IMT Atlantique Bretagne-Pays de la Loire - www.imt-atlantique.fr

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 03
T +33 (0)2 29 00 11 11
F +33 (0)2 29 00 10 00

Campus de Nantes
4, rue Alfred Kastler - La Chantrerie
CS 20722
44307 Nantes Cedex 3
T +33 (0)2 51 85 81 00
F +33 (0)2 51 85 81 99

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
T +33 (0)2 99 12 70 00
F +33 (0)2 99 12 70 08