Abstract

To satisfy the ever-increasing capacity demand and quality of service (QoS) requirements of users, massive MIMO (multiple-input multiple-output) has been attracted as a promising technique in next generation wireless networks. Employing large-scale antennas, however, at remote radio heads (RRHs) generates heavy traffic to be carried through fronthaul links in cloud radio access networks (C-RANs). In this dissertation, we investigate C-RAN architecture, fronthauling methods, and multiple-input multiple-output (MIMO) transmission strategies to overcome explosive fronthaul traffic while maintaining the potential of C-RAN and massive MIMO to the fullest.

Firstly, we proposed a partially-centralized C-RAN (PC-RAN) architecture where precoder, data symbol, and channel state information (CSI) are separately transported in fronthaul links. With the proposed PC-RAN, fronthaul traffic can be remarkably reduced with no or marginal performance degradation, compared with the conventional fully-centralized C-RAN (FC-RAN).

Secondly, we mathematically evaluated the performance of zero-

forcing based large-scale MIMO. We derive wireless performance and fronthaul traffic taking account of cooperative processing among RRHs in C-RAN environments. Through extensive simulations, we confirmed the accuracy of our analytical model and provided intuition on trade-off between wireless performance and fronthaul traffic volume.

Thirdly, we investigated a joint beamforming and resource allocation problem of a single RRH for a constrained fronthaul capacity. We provide a heuristic algorithm to decide beamforming configuration and bandwidth allocation for each beamforming technique. The simulation results show that the proposed algorithm further improves the wireless sum-rates and achieves near optimal performance in our proposed partially-centralized C-RANs.

Lastly, we investigated the performance of ZF and MRT with two fronthauling methods in fronthaul-constrained C-RANs. We provide an algorithm to decide the optimal fronthauling method and beamforming strategy to maximize the wireless sum-rate under a limited capacity of fronthaul link. Numerical results confirm that the sum-rate gain is greater when both fronthauling solutions are available.

Keywords: C-RAN, cloud radio access network, massive MIMO, large-scale antenna system, fronthaul.

Student Number: 2009-20803

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Background

To meet ever-increasing mobile traffic demand and QoS requirements contributed by new applications and services such as HDTV, VoIP, real-time gaming, and machine type communication, the fifth generation (5G) wireless communication technologies are expected to attain 1000 times higher mobile data volume per unit area, 10-100 times higher number of connecting devices and user data rate, 10 times longer battery life and five times reduced latency [1].

To enhance network capacity and energy efficiency, one prominent way is to spatially reuse the limited spectrum resource by either adding more cells. However, increasing cell density results in growing inter-cell interference as well as high CAPital EXpenditure (CAPEX) and OPerating EXpenditure (OPEX). More specifically, CAPEX in-

creases as base stations are the most expensive components of a mobile network infrastructure, while OPEX increases as base stations consumes a considerable amount of power to operate, e.g., China Mobile estimates 72% of total power consumption originates from the cell sites[2]. Mobile network operators need to cover the expenses for network construction, operation, maintenance and upgrade; meanwhile, the Average Revenue Per User (ARPU) stays flat.

C-RAN is a novel cellular architecture to answer the above mentioned challenges. In C-RANs, the traditional role of the base stations is decoupled into two parts: the remote radio heads (RRHs) distributed at cell-sites and the baseband units (BBUs) clustered as a BBU pool in a centralized cloud server. Due to the centralization of computing resources and radio signal generation functions, equipment at cell-sites can be cheaper and computing resources can be more efficiently utilized in a cloud system. As L2/L3 and most of PHY-layer functions are processed at the central BBU pool, it becomes easy to upgrade the RAT and support multi-radio access technology (RAT). For mobile network operators, these modifications result in reduced CAPEX and OPEX. Moreover, centralized radio signal processing of multiple RRHs makes it easy to optimize network-wide performance by exploiting, in real-time, inter-cell coordination techniques such as Coordinated Multi-Point (CoMP), joint spectrum resource allocation, and scheduling as well as multiple associations.

In parallel, large-scale antenna system (also known as massive

MIMO) is regarded as a key technology to boost up system capacity for next generation wireless systems. Due to the significant potential of improving the spectral efficiency and spatial multiplexing gain with tens or hundreds of antennas, massive MIMO has recently received tremendous attention from both academia and industry. With a large number of antennas, a base station can simultaneously serve tens or hundreds of users (but less than the number of antennas) at a channel use by utilizing proper multi-user beamforming techniques. Theoretically, as the number of antennas approaches infinity, the effect of uncorrelated interference and thermal noise can vanish, thanks to the law of large numbers [3] [4].

The challenge, however, comes when implementing large-scale antenna systems on C-RANs. The practical fronthaul is often capacity constrained or time-delay constrained, which has a significant impact on wireless performance of C-RANs[5]. The centralization concept that has seemed absolutely beneficial will, as the number of antennas increases at cell sites, precipitate explosive data volume into fronthaul links. A BBU and an RRH transport radio signals to each other over a fronthaul link, which is commonly a high-speed and constant bit-rate optical fiber or mmWave link. As a fronthaul communication protocol between a BBU and an RRH, Common Public Radio Interface (CPRI)[6] or Open Base Station Architecture Initiative (OBSAI)[7] can be used. Considering current CPRI link rate is approximately a 10 Gbps, which can support 8 RRH antennas in a 20-MHz LTE

system[2] in FC-RANs, it is essential to extensively increase fronthaul resources, i.e., fiber cables, optical devices, or spectrum (in wireless fronthaul) to support tens or hundreds of antennas at cell sites. Such a fronthaul cost burden may limit scaling up antennas of mobile network operators.

C-RAN architecture is classified as a fully-centralized solution or a partially-centralized solution according to the location of PHY-layer functionality [2]. In fully-centralized C-RAN (FC-RAN), either In-phase and Quadrature (IQ) data of baseband signal or RF signal is transported in fronthaul links between BBU pool and RRHs, resulting in inefficient use of limited fronthaul resources[8]. On the other hand, with partially centralized approach, PHY-layer functions are processed at RRHs while L2/L3 functions are still processed in the clustered BBU pool, thus reducing the fronthaul traffic burden on the fronthaul links. However, the original concept of partially-centralized C-RAN (PC-RAN)[2] where whole baseband signal processing is performed in RRHs, cannot fully utilize the potential of C-RAN due to limited L1/L2 cross-layer optimization and also limitation in applying advanced signal processing among RRHs. Furthermore, implementing large-scale antennas at RRHs and applying complex MIMO techniques to many users with large-scale antennas requires expensive RRH cost which deteriorates the original benefit of C-RAN.

1.2 Contributions and Outline

In this dissertation, we propose a novel partial centralization approach to enable large-scale antenna operation without extraordinary fronthaul resource supplement in C-RANs. We also analyze network-wide performance of large-scale MIMO operation in our proposed partially-centralized C-RAN and provide desirable operation strategy to efficiently utilized fronthaul resource while achieving the potential of large-scale antenna and cloud-based system.

This dissertation is organized as follows.

In Chapter 2, we provide an overview of the challenging issues of C-RAN in adopting massive MIMO and propose a partially-centralized C-RAN (PC-RAN) architecture which is designed for large-scale antenna operation. The proposed C-RAN is fronthaul-efficient and flexible in that per-stream data symbols are transported, instead of per-antenna IQ samples and precoder and CSI traffic which depends on the number of antennas are further reducible without significant performance degradation compared with conventional fully-centralized C-RAN (FC-RAN). We also discuss challenging issues of heterogeneous C-RAN environment where high power macro-RRH and low-power small-RRHs coexist hierarchically.

In Chapter 3, we evaluate the performance of large-scale antenna C-RANs taking account of cooperative MIMO processing scenario among RRHs. Stochastic geometric approach will be applied to ob-

tain statistical performances of large-scale multi-user MIMO as well as the amount of expected fronthaul traffic.

In Chapter 4, we investigate a joint beamforming and resource allocation problem of a single RRH for a constrained fronthaul capacity. We consider two representative MIMO techniques, zero-forcing and maximimum ratio transmission (MRT), and provide a heuristic algorithm to decide optimal beamforming configuration and bandwidth allocation for each beamforming technique.

In Chapter 5, we investigated the performance of ZF and MRT with two fronthauling methods in fronthaul-constrained C-RANs, which are transport of IQ sample data and transport of precoder, data symbol, and channel information. We provide an algorithm to decide the optimal fronthauling method and beamforming strategy to maximize the wireless sum-rate under a limited capacity of fronthaul link. Numerical results confirm that the sum-rate gain is greater when both fronthauling solutions are available.

Finally, we conclude the dissertation in Chapter 6.

Chapter 2

A Partially-centralized C-RAN Architecture for Massive MIMO

2.1 Introduction

From 2010 to 2015, mobile data traffic has grown, experts believe, more than 24 times; from 2010 to 2020, they expect it to grow more than 500 times [9]. To satisfy the explosive capacity demand, mobile operators need to increase network capacity by adding more cells, thereby creating a complicated structure of heterogeneous and small cell networks (HetSNets).

To implement HetSNets, mobile operators are closely looking at a cost effective and performance-optimizing radio access architecture

known as cloud radio access network (C-RAN)[2]. In conventional RANs, Layer-2 and PHY-layer functions are processed at antenna sites while in C-RANs, most base station functionalities are processed at central base band unit (BBU) pools, and the remaining, minimal radio frequency functionalities are processed at remote radio heads (RRHs).

Owing to the centralization of computing resources and radio signal generation functions, equipment at cell-sites can be cheaper and computing resources more efficiently utilized in a cloud system. As the baseband radio signals are manufactured at the central BBU pool, it becomes easy to upgrade the radio access technology (RAT) and support multi-RAT. For mobile operators, these modifications result in reduced capital expenditure (CAPEX) and operating expenditure (OPEX).

Moreover, clustering of cross-tier and co-tier cells makes it easy to optimize network-wide performance by exploiting, in real-time, inter-cell coordination techniques such as Coordinated Multi-Point (CoMP), joint spectrum resource allocation, and scheduling as well as multiple associations.

Due to strict requirements on capacity and delay that stimulate such centralized resource management and baseband processing, underlay transport networks among BBU pools and RRHs are commonly constructed via optical networks. Fig. 2.1 shows a HetSNet C-RAN (H-CRAN) and fronthaul transport network scenario where multiple

RRHs for macro and small cells are single- or multi-hop connected to central locations (i.e., BBU pools), sharing fiber links.

In boosting per-cell capacity, a key enabler is the large-scale antenna systems (a.k.a. massive MIMO), in parallel. A large array of antenna elements enable narrow beamforming for many simultaneous users with a large degree of freedom. If detailed channel state information (CSI) is available, a base station can simultaneously serve tens-or hundreds of terminals (but fewer than the number of antennas) by exploiting proper transmit precoding or receive combining with the large set of antennas.

The challenge arises when large-scale antenna systems must be implemented with H-CRAN. The centralization concept that seems absolutely beneficial precipitates, as the number of antennas increases at cell sites, explosive data volume into fronthaul links. Therefore, to support increasing fronthaul data, mobile operators are forced to add more fiber cables and optical devices.

In this article, we provide an overview of the challenging issues of H-CRAN with large-scale antennas and investigate C-RAN architectures in terms of large-scale antenna operation in H-CRANs. Afterwards, we provide a partially centralized approach that reduces, at remarkable levels, fronthaul overheads, offering a flexible and scalable solution in large-scale antennas C-RAN. The proposed approach can also be adopted in H-CRANs, and H-CRAN related issues are also discussed.

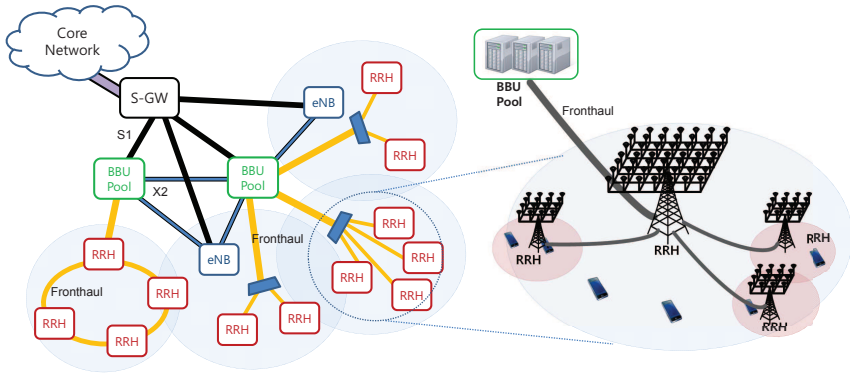


Figure 2.1: H-CRAN architecture and deployment scenario.

2.2 System Architecture and Challenging Issues

2.2.1 Heterogeneous C-RAN with Large-scale Antennas

In C-RANs, BBUs, which process radio signals, are separated from cell sites and concentrated at a data center (BBU pool), whereas in conventional RANs a BBU and a radio unit are located together. The cloud base station structure allows more advanced techniques for the management of radio resources, interference, computing resources, and energy consumption[10]. In addition, due to the centralized infrastructure, mobile operators are able to significantly reduce the installation and operation cost.

In H-CRANs, high-power macro cells with a large number of an-

tennas and low-power small cells with fewer antennas can be combined to enhance network capacity and access network connectivity. The performance of each user particularly in hot-spots or indoor buildings can be highly improved thanks to abundant reuse of limited spectrum resources and better link qualities. Meanwhile, mobile users can be adaptively or simultaneously served by small cells and macro cells to achieve high data rates and to maintain connectivity with reduced control-plane signaling for handover.

As large-scale antennas are deployed in H-CRANs, the potential of H-CRAN can further be maximized with spatial multiplexing and array gain. At hotspots, many users can be spatially multiplexed to further increase per-user throughput; even cell-edge users who require high data rates can be satisfied by large-scale antenna beamforming of macro cells or by CoMP between macro cells and near small cells. Abundant spatial degree of freedom provides a chance to handle cross-tier or co-tier interference in complicated H-CRANs, avoiding under-utilization of expensive spectrum resources.

In H-CRAN design, conventional or already deployed macro base stations, termed eNB in LTE, which perform the entire baseband processing and upper layer functions can be combined with newly deployed RRHs that work for small and macro cells as well. Such macro base stations can be connected to C-RANs through X2 interface for inter-cell coordination in handover and interference management. With standalone baseband processing and non-ideal backhaul, how-

ever, fully centralized coordination with large-scale antennas is a challenging issue, due to a fundamental difficulty in CSI gathering[11]. On the other hand, designing H-CRAN with only RRHs provides chances of flexible cell virtualization and computing resource pooling, as well as PHY-layer multi-cell coordination.

Suitable for H-CRAN deployment are wavelength-division multiplexing (WDM) and optical transport network (OTN) solutions, especially when fiber resources are limited and fronthaul transport network topology is complex[12]. The bandwidth of the fronthaul link can be improved greatly as 40 to 80 optical wavelengths can be multiplexed in a single optical fiber.

Considering that a tolerable round trip time of 400 usec for LTE-A (Long Term Evolution-Advanced) and a propagation delay of 5nsec/m in the optical fiber, a maximum distance of 40 km can be supported between a BBU pool and RRHs. This allows hundreds or even thousands of RRHs including macro-RRHs, to be connected to one BBU pool. In such a fronthaul transport network, numerous RRHs should share some fiber resources with other RRHs, limiting the end-to-end fronthaul resource for each RRH. The system architecture of H-CRAN is depicted in Fig. 2.1.

2.2.2 Challenging Issues in Massive MIMO on H-CRAN

Joint MIMO Processing and User Scheduling: High-power macro-RRH may cover tens/hundreds of low-power small-RRHs in

H-CRANs. Coverage-different and numerous interference-connected cells in H-CRANs make RRH clustering and joint precoding/decoding more complicated than C-RAN. When a macro-RRH is equipped with many more antennas, the radio signals of macro-RRH can be generated to eliminate or mitigate interferences to small-RRH users while serving macro-RRH users with abundant degree of freedom. However, if the degree of freedom of a macro-RRH is not relatively large compared with the number of users served by small-RRHs, there should be performed tightly-coupled precoding and scheduling over the macro-RRH and small-RRHs. The number of active antennas at macro- and small-RRHs can be adaptive to the number of small-RRH and macro- and small-RRH users.

CSI Gathering and Pilot Resource Management:

In H-CRANs, the channel matrices between users and macro/small-RRH are large and the influence of the channel estimation error is non-negligible[13]. Therefore, which channel parameters are measured and how the parameters are measure need to be studied. To obtain huge CSI with reasonable pilot overheads, a promising option is TDD-based large-scale antenna systems [4]. In TDD-based H-CRAN, the received pilot signals of macro- and small-RRHs can be jointly processed at the BBU pool. To avoid correlated interference by pilot contamination, which limits the performance of massive MIMO even with an infinite number of antennas, integrated pilot resource management and channel estimation technique needs to be investigated. The main consid-

erations are asymmetric coverage and the number of associated users of macro- and small-RRHs as well as the applied precoding schemes and resource sharing/partitioning among macro/small-RRHs.

Adaptive Cell Configuration: In an H-CRAN consisting of high power macro-RRHs and low power small-RRHs, the entire network capacity and per user performance can be enhanced by exploiting multi-RRH joint transmission/reception with flexible cell configuration as in [14, 15]. To perform cooperative MIMO processing and resource allocation, RRHs should be adaptively clustered and occupy the necessary computing resource in a BBU pool. However, the required computing resources of each cell dynamically varies in H-CRAN environments due to various patterns of cell coverage and traffic loading. The BBU pool should assign computing resources to RRHs considering their baseband processing complexity. For real-time baseband processing and efficient computing resource use, an accurate amount of required computing resources should be modeled for large-scale antenna H-CRANs.

Fronthaul Resource Management and Topology Optimization: In complicated heterogeneous networks, where many RRHs are multi-hop connected to the BBU pool, a significant problem that must be solved is fronthaul resource managements, especially for fronthaul resources shared by hot-spot RRHs. In addition, fronthaul transport network topology and fiber resource installation should be decided considering each cell's user traffic and fronthaul data volume. Re-

source and topology should be jointly optimized adaptive to heterogeneous number of antennas, coverage, and traffic or on/off state of each RRHs. Depending on the fronthaul resource medium, i.e., wireless channels or wavelengths in optical fiber, different approaches are needed for resource management and topology optimization.

Fronthauling and Architecture Design: Explosive fronthaul data volume is a very challenging issue, especially in H-CRANs, consisting of macro-RRHs, which are probably equipped with a large number of antennas, and small-RRHs, which are equipped with fewer antennas but much more densely deployed than macro-RRHs. The fronthaul data explosively increases as wireless networks are evolved with scaling the number of antennas, cell density, and spectrum expansion. What are needed to tackle such explosively increasing fronthaul overheads in H-CRANs, are efficient and flexible fronthauling and H-CRAN architecture.

2.3 Fully Centralized C-RAN and Fronthaul Transport Solutions

In this section, we describe existing fully centralized C-RAN (FC-RAN) solutions; one is radio over fiber (RoF), which carries radio signals directly over the optical fiber, and the other is digitized IQ data transport, which carries digitized IQ sample data over it.

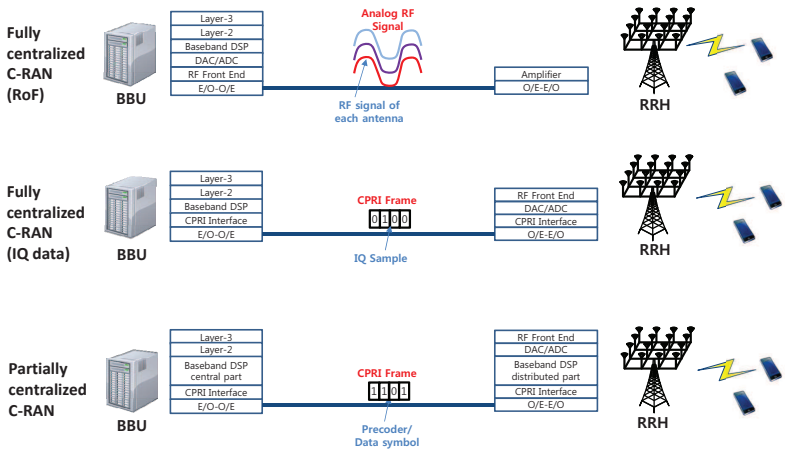


Figure 2.2: Fronthaul transport and C-RAN architecture solutions.

2.3.1 Radio over Fiber

A way of communication between the BBU pool and RRHs is to employ radio frequency signal transmission based on RoF techniques [14, 16, 17]. In RoF systems, radio signals are carried in optical form between the BBU pool and RRHs before being radiated over the air.

In the BBU pool, digital baseband signals generated by a digital signal processor (DSP) are converted to analog radio frequency signals via a digital-analog converter (DAC) function. Then, the RF signal is carried over an optical signal using electro-optic modulators (EOM) such as Mach-Zehnder modulators and electro-absorption modulators. As radio frequency signals are completely manipulated in the BBU pool, the design of RRHs can be further simplified. Due to its transparency with respect to radio access technologies, RRHs can transmit

(or receive) any radio signal (LTE, 3G, WiFi, etc) that can be transported over a fiber cable. By adopting WDM in fronthaul link, a set of radio signals for multiple RRHs[16] or multiple antennas can be multiplexed in a single fiber.

Due to high carrier frequency and large frequency gap between wavelength channels (12.5-100 GHz), one wavelength can carry an antenna's radio signal nearly regardless of its bandwidth. With regard to bandwidth expansion in future cellular systems, this is an attractive prospect. However, each wavelength carries the radio signal of only one antenna at a time. In implementing H-CRANs, carrying radio frequency signals of large-scale antenna macro-RRHs and numerous small-RRHs requires huge wavelengths and fibers in the shared fronthaul links between RRHs and a BBU pool. Therefore, adaptive wavelength resource management and limiting RF-chains with proper large-scale antenna techniques are essential for RoF based large-scale antenna H-CRAN deployment.

2.3.2 Digitized IQ Data Transport

In conventional FC-RAN, the BBU pool and RRHs exchange baseband signal in the form of digitized IQ samples. The IQ samples are encapsulated using a fronthaul transport interface such as CPRI. In contrast to the RoF transport solution, digitized IQ sample information can be transported over both wired and wireless fronthaul links, resulting in easier implementation of small cells in H-CRANs.

The BBU generates a baseband signal for each antenna in the form of a digital IQ sample using a digital signal processing function after upper-layer protocol processing. As a result, the required fronthaul bit rate is given by

$$R_{\mathrm{IQ}} = \alpha M b_{\mathrm{IQ}} f_{\mathrm{s}} \tag{2.1}$$

where b_{IQ} is the number of bits to represent a pair of IQ samples and α is the redundancy in the fronthaul transport interface. In CPRI, $\alpha = \left(\frac{10}{8}\right)\left(\frac{16}{15}\right) = \frac{4}{3}$ considering on 8B/10B line encoding and the control channel portion of 1/16 [6]. The IQ sampling rate, f_s, is proportional to the wireless system bandwidth. This representation holds for both downlink and uplink. Assuming a 20-MHz LTE system ($f_s = 30.72$ MHz) with 64 RRH-antennas and 15 bits of IQ sample width ($b_{\mathrm{IQ}} = 30$), one end-to-end CPRI link should be able to support 78.64 Gbps of IQ data. The current maximum rate of a CPRI link is around 10 Gbps[6].

In H-CRANs where many macro-RRHs and small-RRHs are equipped with many antennas, network-wide fronthaul data drastically increases as cell density increases or antennas are scaled up. To this end, some options are available to reduce the fronthaul data volume for a given number of RRH-antennas and wireless bandwidth.

IQ data compression

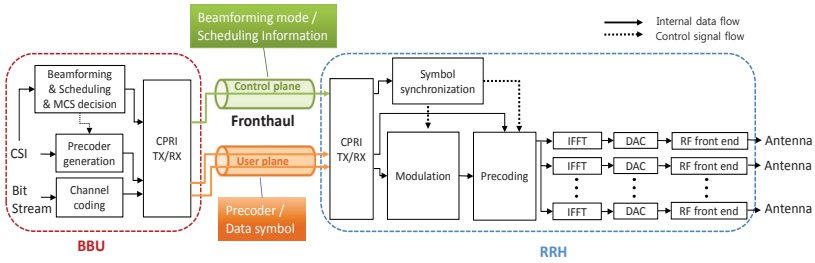
With compression/decompression at the BBU pool and RRHs, the IQ-sample data can be reduced at the cost of distortion in reconstruction of baseband signals and additional complexity both in the BBU pool and RRHs. In [18], the performance of existing IQ compression schemes are summarized. A joint compression/decompression scheme [19] exploiting correlation among the signals of neighbor RRHs might be extensible to H-CRANs.

Advanced MIMO technique

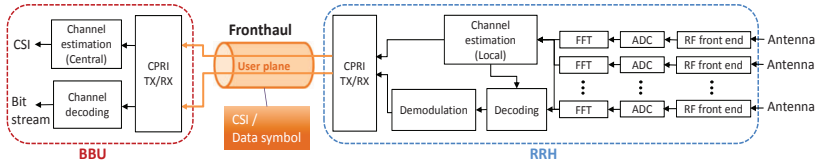
To reduce the number of active antennas, only the selected set of antennas can be used for radio signal transmission, as in [20]. Or limited RF-chains can be used with two-stage precoding techniques such as those in [21][22], with which long-term RF-precoding for antenna elements is performed at RRHs and short-term baseband-precoding for RF-chains is performed at the BBU pool. With such schemes, the benefits of large-scale antennas can be exploited with reasonable fronthaul overheads.

2.4 Partially Centralized C-RAN for Massive MIMO

The fundamental limitation of FC-RAN is that per-antenna baseband (or RF signal in RoF based C-RAN) information should be carried in



(a) Downlink structure



(b) Uplink structure

Figure 2.3: Example of the partially-centralized C-RAN (PC-RAN) structure for large-scale antenna operation.

fronthaul, resulting in huge requirements of fronthaul resources with order-of-antenna numbers when large-scale antenna systems are the case. As an alternative, partially centralized solution [2][18] can be adopted where L1 (PHY-layer) processing is co-located with RRHs to avoid fronthaul-inefficient IQ-sample transport. However, processing all L1 functions at RRHs as the original concept in [2] sacrifices the joint L1/L2 cooperation gain among multiple RRHs. One desirable solution for reducing fronthaul data but maintaining cooperation gain is dividing PHY-layer functionalities into the central part and the distributed part. Fig. 2.2 shows the key differences between the FC- and our proposed PC-RAN.

2.4.1 Basic Concept

Fig. 2.3 presents a PC-RAN architecture for TDD based large-scale antenna operations. One of the fundamental features of the proposed C-RAN is that the baseband processing part is divided into the BBU part and the RRH part. By doing so, precoder, data symbols, and channel information are separately transported instead of heavy IQ data.

The BBU pool may jointly decide the precoder and scheduled users with proper modulation and coding scheme (MCS) levels considering their channel conditions. The main products of the BBU processing are modulated data symbols and precoders (decoders for uplink). Precoding vectors and data symbols for the scheduled users are trans-

ported to corresponding RRHs which transmit radio signals through their antennas after precoding received data symbols.

Data symbol transport

In PC-RAN, by transporting constellation points (after deciding the modulation scheme), a data symbol can be represented with a smaller number of bits compared with a quantized IQ sample. Moreover, the data volume size can be reduced to the order of the number of users ($K \ll M$) from the number of antennas. Considering the pilot overhead for channel estimation in a slot, the required bit rate for data symbol delivery of K users in TDD systems is

$$R_{\mathrm{DS}} = \alpha \left(1 - \frac{\tau}{T}\right) K b_{\mathrm{DS}} f_{\mathrm{sym}}, \tag{2.2}$$

where f_{sym} is the symbol rate. In 20 MHz OFDM systems with 1200 subcarriers ($N_{sc} = 1200$), $f_{\mathrm{sym}} = \frac{N_{sc}}{T_s + T_g} = 16.8$ MHz. $\frac{\tau}{T}$ is the ratio of the number of symbols used for the uplink pilot, τ, to the number of total symbols, T, in a transmission slot. The number of bits to represent one data symbol, b_{DS}, depends on the representation method as described below.

Precoder transport

For consecutive symbols during a coherence time (or slot time during which the channel condition is assumed to remain constant), the same

precoder can be applied. Even though the length of the precoding vector increases with the number of RRH antennas, the precoding data volume is bearable due to its low transport frequency compared to data symbols. The required bit rate for a precoder update to support K users (or streams) with M antennas is

$$R_{\mathrm{Pre}} = \alpha M K b_{\mathrm{IQ}} f_{\mathrm{Pre}}. \tag{2.3}$$

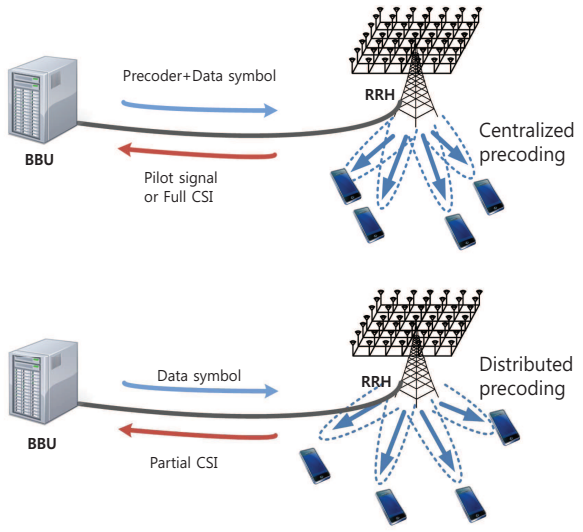
The frequency of precoder update, f_{Pre}, can be set as $(\frac{1}{T})f_{\mathrm{sym}}$ according to the transmission slot of the wireless physical channel.

2.4.2 Centralized and Distributed Precodings

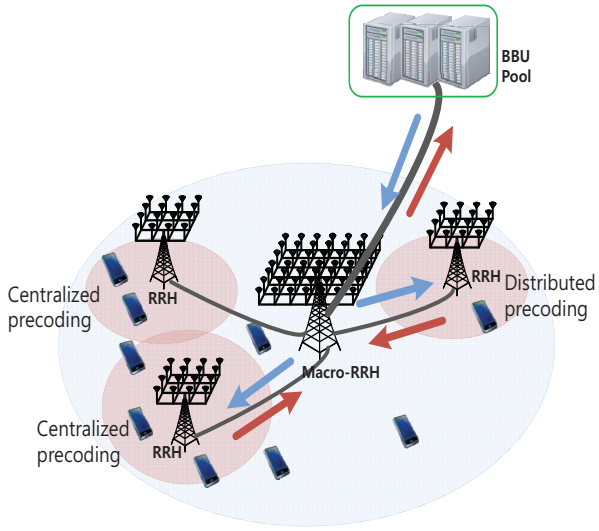
Multiple antenna precoding schemes have different characteristics with respect to the multiplexing order, user channel environments, and algorithm complexity. In the proposed PC-RAN, precoder design can, as shown in Fig. 2.4(a), be done in two ways; *centralized* and *distributed* precodings according to the need for centralized cooperative processing over multiple RRHs, the computational burden on the RRH, and available fronthaul bandwidth.

Centralized precoding

To overcome cross-tier and co-tier interference in densely deployed H-CRANs, the BBU pool may perform a cross-tier and/or co-tier cooperative processing over multiple RRHs, aiming at network-wide



(a) Centralized and distributed precodings for a single RRH



(b) Centralized/distributed precoding scenario in H-CRAN

Figure 2.4: Centralized and distributed precoding scenarios. In ultra-dense H-CRAN, distributed precoding can be exploited to reduce fronthaul burden according to RRH locations and user channel conditions.

performance optimization. For some users or RRHs that experience severe interference, it would be desirable to use zero-forcing based multiuser MIMO or network MIMO schemes. Large-scale zero-forcing based precoding for a macro-RRH and multiple small-RRHs should be performed in a BBU pool where the required computing resources can be flexibly assigned and CSI of entire RRHs are gathered.

To perform centralized precoding, the following information should be exchanged between a BBU pool and RRHs.

• Data symbol: modulation scheme and bit symbols for scheduled users.

• Precoding vector: precoding weight vectors to be precoded with data symbols of spatially multiplexed users.

• Estimated full CSI or pilot signal: information needed to decide the precoder in a BBU pool.

There are two ways to deliver the CSI of each user to the BBU: 1) The RRH transports the locally estimated CSIs (LE-CSIs) of selected users to the BBU, and 2) the RRH transports the received pilot signal during the channel training period in the form of an IQ sample, as in conventional C-RANs. The second option is needed when the BBU pool performs cooperative MIMO processing and channel estimation over multiple RRHs to eliminate inter-cell interference or pilot contamination[23].

Distributed precoding

For distributed precoding, RRHs generate precoding vectors for scheduled users using LE-CSIs during the channel training period. One representative application of distributed precoding is conjugate beamforming, which aims to maximize the desired signal power regardless of interference and requires low complexity to obtain the precoder [24]. Zero-forcing based precoding also can be used if the computational complexity is supportable with not so many antennas and spatially multiplexed users. Non-zero-forcing based limited coordinations among RRHs are also possible in distributed precoding. The following information should be exchanged between a BBU pool and RRHs.

• Data symbol stream: Modulation scheme and bit symbols for scheduled users.

• Precoding vector: Not necessary. The precoding vector can be locally calculated at each RRH using LE-CSIs if the precoding technique and scheduling information is informed by the BBU.

• Partial channel information: For distributed precoding operation, there is no need for the BBU to know instantaneously the full CSI of users because RRHs calculate the precoder for each user using LE-CSI.

The separation of precoder and data symbol in fronthaul is more beneficial in the distributed precoding scenario where precoding weights not transported over fronthaul and the information about CSI or

training signal can be further reduced. However, the BBU may need to know the channel condition to decide the appropriate precoding technique and to schedule each user. Especially, for the MCS decision, the BBU has to receive the expected SINR of considered users from RRHs or infer it from the effective channel gain of the users and long-term channel gains between the users and neighbor RRHs.

2.4.3 CSI Estimation and Report

The TDD operation of large-scale antenna systems does not need pilot resources proportional to the antenna array size because uplink and downlink channels are estimated by the channel reciprocity and pilot signals sent by users. In such TDD operation scenarios, the RRH can estimate the channel of each user if it knows the location of pilot resources assigned to each user. For such localized channel estimation, the BBU should inform the RRH of the pilot resource map of users served by the BBU. The required uplink fronthaul bit rate to report locally-estimated CSI (LE-CSI) is $R_{\mathrm{LE-CSI}} = \alpha M K b_{\mathrm{IQ}} f_{\mathrm{CSI}}$. The frequency of CSI report for each user, f_{CSI}, is given as $(\frac{1}{T})f_{\mathrm{sym}}$ (in this case, $R_{\mathrm{LE-CSI}} = R_{\mathrm{Pre}}$) and it can be adaptively chosen by balancing CSI accuracy and the fronthaul resource budget.

Another option is to transport IQ samples of received pilot signals to the BBU pool, especially when cooperative channel estimation against pilot contamination or for CoMP is needed. In such a case, the required uplink bit rate is $R_{\mathrm{pilot,IQ}} = \alpha M(\frac{\tau}{T})b_{\mathrm{IQ}} f_{\mathrm{sym}}$.

When only partial channel information, such as expected SINR or effective channel gain of users, is needed for distributed precoding, the required uplink bit rate to report the partial channel information obtained using LE-CSI is $R_{\mathrm{LE-CSI,partial}} = \alpha K b_{\mathrm{real}} f_{\mathrm{CSI}}$, where b_{real} is the number of bits needed to represent one real value.

2.4.4 Operation of Centralized/Distributed Precoding

In PC-RAN, the required fronthaul bandwidth for centralized precoding is much greater than that for distributed precoding because the required bit rates for the precoder and pilot signal IQ data outweigh that of data symbols and increase with the number of antennas. On the other hand, the fronthaul data volume of distributed precoding operations does not depend on the antenna array size because it is unnecessary to exchange precoder and full CSI (or pilot signal IQ data) between a BBU pool and an RRH. In this sense, the adaptive operation of centralized and distributed precoding provides chances to flexibly utilize fronthaul resources as well as to balance the processing load for precoder calculation. For example, a centralized precoding mode can be applied only for RRHs that may interfere significantly with other RRH users to reduce fronthaul traffic overhead and required computing resource for precoder calculation in the BBU pool.

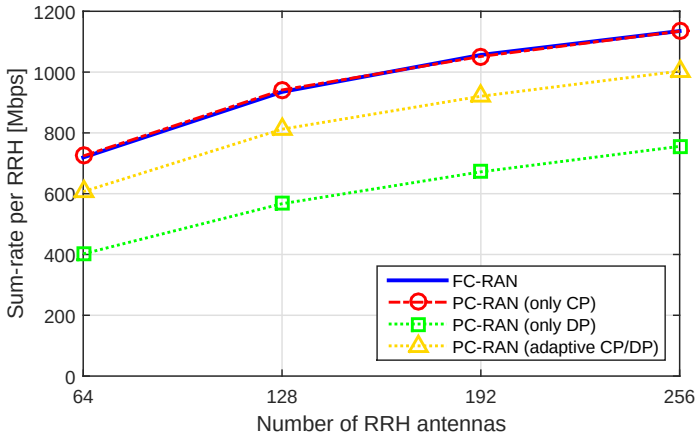


Figure 2.5: Average sum-rate per RRH.

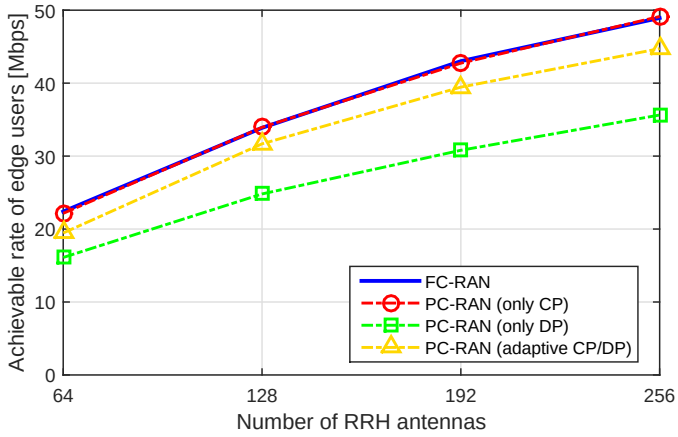


Figure 2.6: Average achievable rate of edge users.

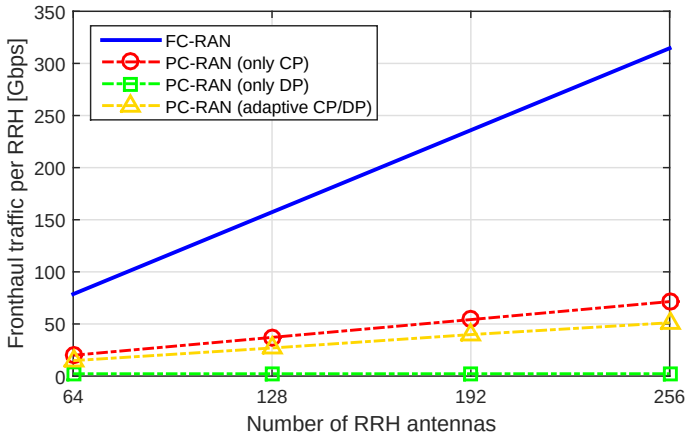


Figure 2.7: Average fronthaul traffic per RRH.

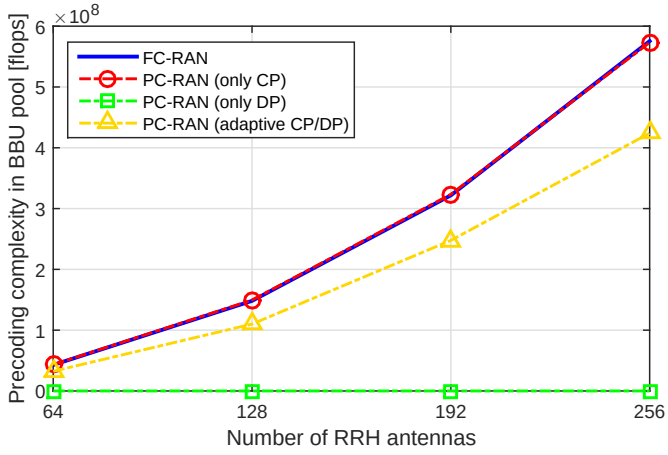


Figure 2.8: Average precoding complexity in BBU pool per RRH.

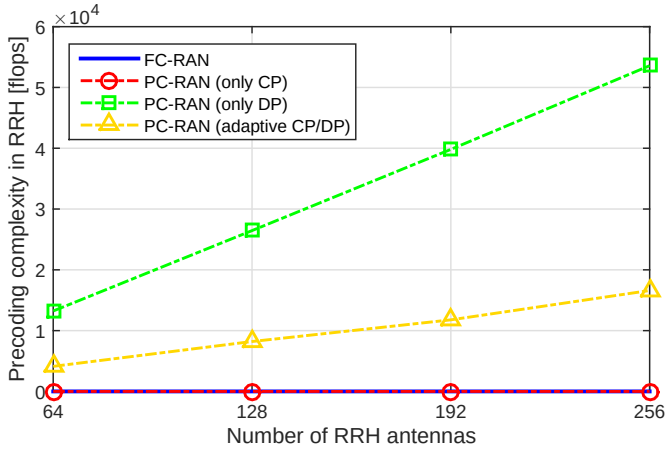


Figure 2.9: Average precoding complexity at each RRH.

2.4.5 Performance Evaluation & Discussion

Figs. 2.5, 2.6, and 2.7 show the average sum-rates of each RRH, average achievable rates of edge users, and average fronthaul traffic of each RRH for different number of RRH antennas. We can see that the fronthaul traffic of PC-RAN is remarkably less than FC-RAN, especially, as the number of antennas increases. More interestingly, with adaptive operation of centralized/distributed precoding, fronthaul traffic is further reduced. At the same time, the wireless sum-rate is somewhat degraded due to uncontrolled interference but is still much better than when only distributed precoding is applied. The main challenge is to find a good compromise between cell/network throughput and fronthaul overhead.

Figs. 2.8 and 2.9 show the precoding complexity [25] for precoder design of each RRH in BBU pool and in RRH, respectively. As can

be seen, as more antennas are deployed, the precoding complexity in BBU pool drastically increases in the case of FC-RAN and centralized precoding operation of PC-RAN. By adaptively applying distributed precoding to RRHs which have no or few interfered users, the required processing load in BBU pool can be remarkably reduced with some wireless performance degradation.

As investigated in the previous section, the partially centralized C-RAN approach provides chances to significantly reduce fronthaul traffic or flexibly balance between the wireless performance and fronthaul overhead. For such a flexible operation, however, RRHs should be more complicated than FC-RANs to perform local channel estimation and precoding while tightly synchronized with their BBU pool. The features of the addressed C-RAN architectures and corresponding fronthauling solutions are summarized in Table 2.1. Finally, we remark on the following related issues of our proposed PC-RAN.

Adaptive CSI Gathering: For large-scale antenna operation, CSI or pilot signal transport requires a large amount of fronthaul resources. In PC-RAN, the CSI traffic can be remarkably reduced according to the precoding scenario (i.e., centralized or distributed) and the target accuracy of CSI values thanks to local CSI estimation at RRHs. For example, in the case of centralized precoding, RRHs can report locally estimated CSIs less frequently than pilot signal transmissions at the expense of some precoder performance degradation. On the other hand, for distributed precoding, RRHs do not need to report full CSIs.

Table 2.1: Characteristics of C-RAN architecture - fronthaul transport solutions

	Signal form in fronthaul	Joint RRH processing	Spectrum expansion	Antenna scalability	RRH complexity	Possible medium
FC-RAN (RoF)	Analog RF	Possible	Almost unlimited	Bad	Very low	Wired (fiber)
FC-RAN (IQ data)	Digital baseband	Possible	Limited	Bad	Low	Wired /Wireless
PC-RAN (centralized precoding)	Digital baseband	Possible	Limited	Good	Normal	Wired /Wireless
PC-RAN (distributed precoding)	Digital baseband	Limited	Limited	Good	Normal	Wired /Wireless

However, it is possible that RRHs report part of locally estimated CSIs for more accurate interference estimation or for neighbor RRHs that may perform centralized precoding. Thus, CSI reporting of RRHs can be adaptively controlled by balancing between fronthaul overhead and performance gain of edge-users. This is also one of the key merits of applying PC-RAN, but it is quite challenging to consider the relation between precoding, CoMP communication, and channel estimation.

Fronthaul Resource Constrained Precoding: The required fronthaul bit rates depend on many factors including where the precoding is calculated, the number of active antennas (or RF-chains) at RRHs, user multiplexing order, uplink sounding period, precoder and scheduling granularity. Intuitively, by jointly deciding a proper precoding scheme and user scheduling with a consideration on the given fronthaul conditions and required wireless performances, limited fronthaul resources can be more efficiently utilized [26] [27]. In H-CRANs, as numerous macro- and small-RRHs are connected to a shared fronthaul link to a BBU pool, available fronthaul resources for each RRH can be limited when the network is heavily loaded. In such a case, the number of active antennas and multiplexed users can be optimized considering the number of associated users in each RRH and cross-tier interference.

Joint Processing in Uplink: In downlink, data symbols and precoders of coordinating RRHs are simultaneously decided in the BBU pool, and generated data symbols and precoders are transported to

corresponding RRHs. In uplink, however, signals from coordinating RRHs should be collected in the BBU pool and jointly detected to achieve joint processing gain to the fullest. Thus, it is difficult to exploit the separation of data symbol and precoder of PC-RAN when uplink multi-RRH cooperation is needed.

Frequency Division Duplexing Operation: The concept of precoder and data symbol separation holds for FDD systems. In FDD systems, however, downlink CSIs are not available at RRHs because CSI feedback from user terminals is transparent and indistinguishable to RRHs. Thus, in FDD systems, it is difficult to exploit distributed precoding and the CSI-related fronthaul burden due to the fact that large-scale antennas can not be relieved.

Advanced precoding techniques for large-scale antennas: If a two-stage baseband/RF precoding scheme [21][22] is adopted in FC-RAN, the antenna-scale dependent fronthaul data volume can be reduced proportional to the number of RF-chains, S. The two-stage can also be applied in the proposed PC-RAN. Then, the size of precoder and CSI information can also be reduced. When $f_{\mathrm{Pre,BB}}$ and $f_{\mathrm{CSI,BB}}$ are update frequency of baseband precoding and effective CSI for baseband precoding, and $f_{\mathrm{Pre,RF}}$ and $f_{\mathrm{CSI,RF}}$ are for RF precoding, the required fronthaul rates for precoder and CSI are $R_{\mathrm{Pre}} = \alpha(SKb_{\mathrm{IQ}}f_{\mathrm{Pre,BB}} + MKb_{\mathrm{IQ}}f_{\mathrm{Pre,RF}})$ and $R_{\mathrm{CSI}} = \alpha(SKb_{\mathrm{IQ}}f_{\mathrm{CSI,BB}} + MKb_{\mathrm{IQ}}f_{\mathrm{CSI,RF}})$, respectively, in a centralized precoding operation. For distributed precoding, the required fronthaul rates, while very

low, are the same regardless of adoption of two-stage precoding technique.

Fronthaul data compression: IQ data compression/decompression schemes[18] can reduce the fronthaul traffic of FC-RAN at the cost of some wireless performance degradation. Such IQ data compression schemes probably can also be applied to PC-RAN in a transporting precoder and CSIs. The impact of IQ data compression/decompression in PC-RAN still needs to be investigated. In PC-RAN, the number of bits for data symbol transport also can be reduced by grouping the data symbols based on applied modulation scheme instead of informing the constellation point of symbols.

2.5 Summary

A promising solution to the problem of enhancing network capacity in cellular networks is deploying more small cells and installing a great number of antennas. Large-scale antenna operation on H-CRANs requires not just extreme processing resources but huge fronthaul data delivery. In this article, we explored existing C-RANs and promoted a partially centralized approach in H-CRANs with large-scale antennas. The partially centeralized C-RAN approach provides chances for mobile operators to efficiently utilize fronthaul resources and to flexibly adopt proper precoding or CoMP schemes according to available fronthaul resources and required wireless performance in H-CRANs.

Chapter 3

Performance Analysis of Large-scale MIMO in C-RANs

3.1 Introduction

Multiple-input multiple-output (MIMO) is a key technique to enhance system capacity for the given spectral resources. Recently, to exploit more antenna array gain and spatial multiplexing gain, large-scale MIMO (also known as massive MIMO) have attracted significant attention for next generation cellular systems.

In parallel, cloud radio access network (C-RAN) is a promising key feature of cellular system evolution due to its low installation and operation cost and performance gain of centralized optimization.

In C-RANs, baseband signals of distributed RRHs are processed at the central BBU pool. Such a cloud-based baseband signal makes it easy to apply diverse coordinated multi-point transmission (CoMP) schemes to cancel out inter-cell interference and intensify desired signal power of cell-edge users[28, 29, 30].

Although the implementation of large-scale MIMO on C-RAN is a promising scenario in future cellular networks, understanding of statistical performance of such a system is necessary for efficient network design. In a random spatial model, the locations of base stations are modeled by a two-dimensional point process, the simplest being the Poisson Point Process (PPP)[31, 32, 33, 34]. Main advantages of this model is that powerful tools from stochastic geometry can be used to derive network-wide performance for general multi-tier wireless networks in closed form, which was not even possible for single-tier networks using deterministic grid model[33].

There are several works that used stochastic geometry approach to deal with interference from other base stations in the case of multiple antenna transmission or cooperative transmission among adjacent base stations. In [32], the authors investigated the performance of zero-forcing precoding with limited feedback in single-hop ad hoc networks. In [35], mathematical framework was proposed to analyze interference from co-tier and cross-tier considering a typical cell of fixed shape and size and PPP of other base stations' location for the fixed number of spatially multiplexed users in each cell. In [36],

the authors proposed a methodology to treat pairwise cooperation of single-antenna base stations.

In this paper, we analyze the downlink performance of zero-forcing precoding in C-RAN with/without cooperation among RRHs, leveraging recent applications of stochastic geometry. In our cooperative processing scenario, RRHs help (joint transmit and null interference to) edge victim users by applying ZF precoding including them. Rather than pair-wise cooperation between two RRHs such as [37][36], any set of RRHs can cooperate for edge users whenever cooperation is necessary. Also, our cooperation scenario does not need clustering based ZF precoding processing [38][39][40] which causes inter-cluster interference and needs much larger computational complexity in designing precoders. We derived the signal-to-interference ratio (SIR) performance of large-scale zero-forcing (ZF), considering randomness of the number of associated users and spatially-multiplexed users by each RRH. As a by-product, we can calculate the expected fronthaul traffic volume in the partially-centralized C-RAN which depends on the number of spatially multiplexed users. In our analysis, we leverage some results of [35] but model randomness of spatially multiplexed users without the assumption of fixed shape and size of a typical cell. We also consider joint transmission cases as well as interference nulling among cooperative RRHs for edge users. Furthermore, we derived the probability density function (PDF) and cumulative density function (CDF) of signal to interference ratio (SIR) of users at a target distance

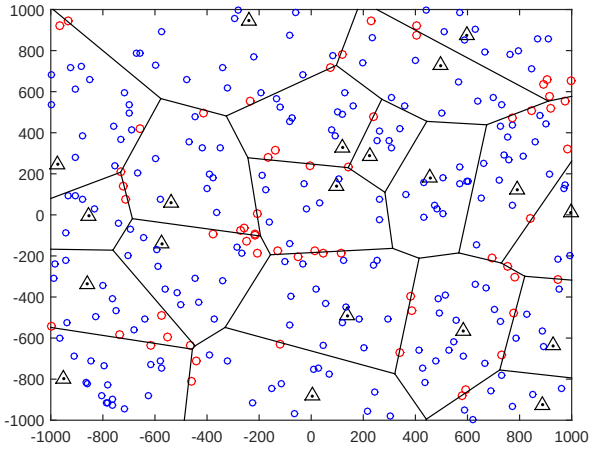


Figure 3.1: Stochastic geometric model where RRHs and users are randomly deployed according to Poisson Point Process. Interfered users are red-colored.

from its serving RRH.

Through extensive simulation, we confirm that our analytical model well matches the simulation results and analyze network-wide wireless/fronthaul performance of partially-centralized C-RAN with respect to various system parameters such as the number of active antennas, RRH and user density, and the number of maximum users to be spatially multiplexed.

3.2 System Model

3.2.1 Deployment and Operation Scenario

In the stochastic geometry model for cellular systems, as seen in Fig. 3.1, RRHs and users are located following PPP with density of λ_b and λ_u, respectively. We assume that transmit power P_{tx} is the same for all RRHs and users are associated with their nearest RRH. Thus, RRHs form the well-known Poisson Voronoi (PV) tesselation. Although users are associated with only one RRH, more than one interfering RRH can support a interfered user when the user is scheduled by its serving RRH. In our operation scenario, zero-forcing (ZF) precoding is applied for all RRHs and the RRHs collaborate to serve edge users while serving their own users at the same time. The precoders of each RRH are calculated at BBU pool for the selected user set. Specifically, for the given transmission slot, each RRH transmits to K_{m} users among K_{s} associated users. Random variable K_{s} is assumed to be independent and identically distributed (i.i.d) over the RRHs, thus K_{m} is also.

We consider relative thresholding-based cooperation among RRHs. If the ratio of the average received power from neighbor RRH b and that of serving RRH is large than T, that is,

$$\frac{E[P_{tx}L_b r_b^{-\alpha}]}{E[P_{tx}L_o r_o^{-\alpha}]} \geq T \Leftrightarrow r_b \geq r_o T^{-\frac{1}{\alpha}}, \tag{3.1}$$

the user is considered as one of the interfered users by RRH b. Among the interfered users, the set of users who are actually scheduled by their serving RRH become victim users. The number of interfered users and victim users of a typical RRH are denoted by K_{i} and K_{v}, respectively.

3.2.2 Interference and Desired Signal

To model the aggregate received power of a target user, let B_{h} be the set of RRHs helping the target user. Because the precoders of helping RRHs are calculated with zero-forcing precoding, their interference is nullified in to the target user. The received interference power is given by

$$\mathcal{I} = \sum_{b \in \Phi_{\mathrm{i}} \setminus B_{\mathrm{h}}} \frac{I_b}{\ell(r_b)}, \tag{3.2}$$

where I_b is the interference at unit distance and $\ell(r_b)$ is pathloss from RRH b at distance of r_b. The interference at unit distance is $I_b = P_{tx} G_b$, where G_b is instantaneous effective channel gain of interference from RRH b and can be modeled as Gamma distributed random variable with $\Gamma\left[U_b, 1/U_b\right]$.

In RRH cooperation scenario, the desired signal of a typical can be represented by the sum of the signal from the serving RRH and coherent desired signals from helping RRHs within cooperation region of the target user. Therefore, the aggregate desired signal of the target

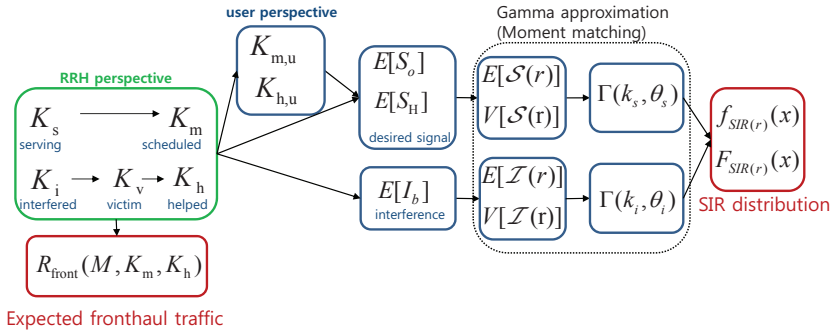


Figure 3.2: Logical flow of derivation.

user at r is given by

$$\mathcal{S}(r) = \frac{S_o}{\ell(r)} + \sum_{b \in B_{\mathrm{h}}} \frac{S_b}{\ell(r_b)}, \tag{3.3}$$

where $S_o = P_{tx} L_o$ and $S_{\mathrm{H}} = P_{tx} L_b$ are desired signal from serving RRH and neighbor RRH b, respectively. The desired signal channel gain L_o and L_b are Gamma distributed random variables with $\Gamma\left[M + U - 1, 1/U\right]$.

3.3 Analytical Model using Stochastic Geometry

In this section, we analyze the SIR performance of users and fronthaul traffic of RRHs while considering random user popularity of RRHs. To derive distributions on SIR of users, we approximate the aggregate desired signal and the aggregate interference as Gamma random

variable with moment matching technique [35]. Overall derivation procedure is represented in Fig. 3.2.

3.3.1 User Distribution

To consider randomness of the number of spatially multiplexed users in each RRH, we analyze the number of associated users, interfered users, and actual victim users. It is known that the cell area probability density function (PDF) of a PV tesselation is accurately predicted by Gamma distribution with Q = 3.575 [41][42] as follows.

$$f(A) = \frac{(Q\lambda_b)^Q A^{Q-1}}{\Gamma(Q)} \exp(-Q\lambda_b A) \quad ,A \geq 0, \tag{3.4}$$

where A denotes the cell area size and $\Gamma(Q) = \int_0^\infty x^{Q-1}e^{-x}dx$ is the Gamma function. Since the users are located according to PPP, the number of users associated with a typical RRH, K_s, follows Poisson distribution for given cell area A, i.e., $K_\mathrm{s} \sim Poiss(\lambda_b A)$. Thus, we get the probability mass function (PMF) of the number of serving users of a typical RRH as

$$\begin{aligned}
&p_{K_\mathrm{s}}(k) = \textstyle\int_0^\infty p_{K|A}(k) f(A) dA \\
&= \textstyle\int_0^\infty \frac{(\lambda_u A)^k \mathrm{e}^{-\lambda_u A}}{k!} \frac{(Q\lambda_b)^Q A^{Q-1}}{\Gamma(Q)} \exp(-Q\lambda_b A) dA \\
&= \textstyle\frac{(Q\lambda_b)^Q}{\Gamma(Q)} \frac{(\lambda_u)^k}{k!} \int_0^\infty A^{Q+k-1} \exp(-(Q\lambda_b + \lambda_u)A) dA \\
&= \textstyle\frac{\Gamma(Q+k)}{\Gamma(Q)k!} \left(\frac{Q}{Q+\lambda_u/\lambda_b}\right)^Q \left(\frac{\lambda_u/\lambda_b}{Q+\lambda_u/\lambda_b}\right)^k,
\end{aligned} \tag{3.5}$$

and its mean is $E[K_\mathrm{s}] = \frac{\lambda_u}{\lambda_b}$.

Since we know the *pmf* of the number of associated users, we can compute the *pmf* of the number of associated users who are spatially multiplexed as follows.

$$p_{K_{\mathrm{m}}}(k) = \begin{cases} p_{K_{\mathrm{s}}}(k) & ,k \leq K_{\mathrm{m,max}} \\ 1 - \sum_{k=0}^{K_{\mathrm{m,max}}-1} p_{K_{\mathrm{s}}}(k) & ,k = K_{\mathrm{m,max}} \\ 0 & ,\text{otherwise} \end{cases} \tag{3.6}$$

The maximum number of spatially multiplexed users, $K_{\mathrm{m,max}}$, can be determined by considering the number of supportable users in fronthaul link, or the number of CSI-valid users constrained by limited pilot resources. Then, the mean number of users who are associated with and spatially-multiplexed by a typical RRH can be obtained as

$$\begin{aligned} E[K_{\mathrm{m}}] &= \sum_{k=0}^{K_{\mathrm{m,max}}-1} k p_{K_{\mathrm{S}}}(k) + \sum_{k=K_{\mathrm{m,max}}}^{\infty} K_{\mathrm{m,max}} p_{K_{\mathrm{S}}}(k) \\ &= \sum_{k=0}^{K_{\mathrm{m,max}}-1} k p_{K_{\mathrm{S}}}(k) + K_{\mathrm{m,max}} \left(1 - \sum_{k'=0}^{K_{\mathrm{m,max}}-1} p_{K_{\mathrm{S}}}(k') \right) \\ &= K_{\mathrm{m,max}} - \sum_{k=0}^{K_{\mathrm{m,max}}-1} (K_{\mathrm{m,max}} - k) p_{K_{\mathrm{S}}}(k). \end{aligned} \tag{3.7}$$

We consider relative average received power based cooperation. To be specific, a typical user demands cooperation of RRH b when the relative received power from RRH b, $\frac{P_{rx,b}}{P_{rx,o}}$, is larger than threshold T where $P_{rx,o}$ is the average received power from its serving RRH.

Lemma 3.3.1. The *pmf* of the number of interfering RRHs of a typ-

ical user, N_{i}, follows geometric distribution as

$$p_{N_{\mathrm{i}}}(n) = T^{2/\alpha}(1 - T^{2/\alpha})^n. \tag{3.8}$$

Therefore, the average number of interfering RRHs of a typical users is

$$E[N_{\mathrm{i}}] = \frac{1 - T^{2/\alpha}}{T^{2/\alpha}}, \tag{3.9}$$

and the average number of interfered users from a typical RRH is

$$E[K_{\mathrm{i}}] = \frac{\lambda_u}{\lambda_b} E[N_{\mathrm{i}}] = \frac{\lambda_u}{\lambda_b} \left(\frac{1}{T^{2/\alpha}} - 1 \right) = \frac{\lambda_{u,i}}{\lambda_b}. \tag{3.10}$$

Then, we approximate the *pmf* of the number of interfered users by adopting (3.5) and substituting $\lambda_{u,i}$ to λ_u which are obtained from (3.10). Thus, we obtain

$$p_{K_{\mathrm{i}}}(k) = \frac{\Gamma(Q+k)}{\Gamma(Q)k!} \left(\frac{Q}{Q + \lambda_{u,i}/\lambda_b} \right)^Q \left(\frac{\lambda_{u,i}/\lambda_b}{Q + \lambda_{u,i}/\lambda_b} \right)^k, \tag{3.11}$$

Among K_{i} interfered users, K_{v} users who are scheduled by the serving RRH become victim users. Then, the *pmf* of the number of victim users is given by

$$p_{K_{\mathrm{v}}}(k) = \frac{\Gamma(Q+k)}{\Gamma(Q)k!} \left(\frac{Q}{Q + p_{\mathrm{S}}\lambda_{u,i}/\lambda_b} \right)^Q \left(\frac{p_{\mathrm{S}}\lambda_{u,i}/\lambda_b}{Q + p_{\mathrm{S}}\lambda_{u,i}/\lambda_b} \right)^k, \tag{3.12}$$

where p_{S} is the scheduling probability that an interfered user is scheduled by its serving RRH which can be obtained from $p_S = E[K_{\mathrm{m}}]/E[K_{\mathrm{s}}]$.

Similar to the number of scheduled users K_{m}, *pmf* of the number of helped users, K_{h}, among victim users interfered by a typical RRH can be represented as

$$p_{K_{\mathrm{h}}}(k) = \begin{cases} p_{K_{\mathrm{v}}}(k) & ,k \leq K_{\mathrm{h,max}} \\ 1 - \sum_{k=0}^{K_{\mathrm{h,max}}-1} p_{K_{\mathrm{v}}}(k) & ,k = K_{\mathrm{h,max}} \\ 0 & ,\text{otherwise} \end{cases} \tag{3.13}$$

The maximum number of users whom one RRH can help, $K_{\mathrm{h,max}}$, can be properly decided with $K_{\mathrm{m,max}}$ to balance the performance of edge users and normal users in a system.

In some cases, we need to know belonging probability that a typical user belongs to a set of users with a certain size. For example, the *pmf* of the number of users associated with an RRH seen by a typical RRH and that seen by a typical user are different.

Proposition 3.3.2. From a typical user's perspective, the *pmf* of the number of associated users with the same RRH is

$$p_{K_{\mathrm{s,u}}}(k) = \frac{k p_{K_{\mathrm{s}}}(k)}{E[K_{\mathrm{s}}]}. \tag{3.14}$$

Proof. Let B_k be the set of RRHs which serve k users and B be the set of all RRHs in the network. Then,

$$
\begin{aligned}
&p_{K_{\mathrm{s,u}}}(k) = \Pr[\text{a user is associated with any } b \in B_k] \\
&= \frac{k|B_k|}{\sum_{j=0}^{\infty} j|B_j|} = \frac{k|B_k|/|B|}{\sum_{j=0}^{\infty} j|B_j|/|B|} = \frac{kp_{K_{\mathrm{s}}}(k)}{\sum_{j=0}^{\infty} jp_{K_{\mathrm{s}}}(j)} \\
&= \frac{kp_{K_{\mathrm{s}}}(k)}{E[K_{\mathrm{s}}]}.
\end{aligned} \tag{3.15}
$$

3.3.2 Interference

In this section, we develop a model for interference of C-RAN taking account for cooperation of interfering RRHs. Assuming the unit-distance interference distributions are independent of RRH locations, the mean and variance of the aggregate interference can be obtained using the Campbell's theorem [43] which are given by

$$
\begin{aligned}
&E\left[\sum_{X\in\Phi} f(X)\right] = \lambda \textstyle\int_{R^2} f(x)dx, \\
&\text{and,} \\
&V\left[\sum_{X\in\Phi} f(X)\right] = \lambda \textstyle\int_{R^2} f^2(x)dx.
\end{aligned} \tag{3.16}
$$

Therefore, from from (3.2),

$$
\begin{aligned}
&E[\mathcal{I}(r))] = E[I_b]\left(\textstyle\int_r^{rT^{-\frac{1}{\alpha}}} \frac{\lambda_b(1-p_{\mathrm{H}})}{\ell(r')} 2\pi r' dr' + \int_{rT^{-\frac{1}{\alpha}}}^{\infty} \frac{\lambda_b}{\ell(r')} 2\pi r' dr'\right) \\
&= E[I_b]\tfrac{2\pi\lambda_b}{(\alpha-2)}\left((1-p_{\mathrm{H}}) + p_{\mathrm{H}}T^{\frac{\alpha-2}{\alpha}}\right) r^{2-\alpha}.
\end{aligned} \tag{3.17}
$$

The first term in the first line corresponds to the interfering RRHs who are in the range of cooperation region but not helping the target

user. p_{H} denotes the probability that a victim user is helped by the corresponding interfering RRH. Thus, our model considers practical cooperation scenario where some RRHs help victim users in probabilistic manner. The probability that a victim user is helped by the corresponding interfering RRH can be represented by

$$p_{\mathrm{H}} = \frac{E[K_{\mathrm{h}}]}{E[K_{\mathrm{v}}]}, \tag{3.18}$$

where $E[K_{\mathrm{v}}] = E[K_{\mathrm{i}}]p_{\mathrm{S}}$. The second term in the first line corresponds to the interference source from out of cooperation region. In the case of $p_{\mathrm{H}} = 0$ or $T = 1$, we can confirm that all RRHs located far than serving RRH contributes to the aggregate interference. Similarly, the variance of the aggregate interference is

$$\begin{aligned} V[\mathcal{I}(r))] &= E[I_b^2]\left(\int_r^{rT^{-\frac{1}{\alpha}}} \tfrac{\lambda(1-p_{\mathrm{H}})}{\ell(r')^2} 2\pi r' dr' + \int_{rT^{-\frac{1}{\alpha}}}^{\infty} \tfrac{\lambda}{\ell(r')^2} 2\pi r' dr'\right) \\ &= E[I_b^2]\tfrac{\pi\lambda_b}{(\alpha-1)}\left((1-p_{\mathrm{H}}) + p_{\mathrm{H}}T^{\frac{2(\alpha-1)}{\alpha}}\right) r^{2-2\alpha}. \end{aligned} \tag{3.19}$$

The n-th moment of I_b can be computed by

$$E[I_b^n] = \sum_{k}^{K_{\mathrm{m,max}}} \sum_{j}^{K_{\mathrm{h,max}}} E[I_b^n\,|k,j\,] p_{K_{\mathrm{m}}}(k) p_{K_{\mathrm{h}}}(j) \tag{3.20}$$

The first and second moments of interference at unit distance ($E[I_b^n|K_{\mathrm{m}}, K_{\mathrm{m}}]$) for given ($K_{\mathrm{m}}$, K_{m}) are represented in Table. 3.1.

3.3.3 Desired Signal

Similar to interference, the mean and variance of the aggregate desired signal of a target user at the distance of r from its serving RRH can be computed as follows.

$$\begin{aligned} E[\mathcal{S}(r))] &= \tfrac{E[S_o]}{\ell(r)} + E[S_{\mathrm{H}}] \int_r^{rT^{-\frac{1}{\alpha}}} \tfrac{\lambda_b p_{\mathrm{H}}}{\ell(r')} 2\pi r' dr' \\ &= E[S_o] r^{-\alpha} + E[S_{\mathrm{H}}] \tfrac{2\pi\lambda_b p_{\mathrm{H}} (1 - T^{(1-\frac{2}{\alpha})})}{(\alpha - 2)} r^{2-\alpha} \end{aligned} \tag{3.21}$$

$$\begin{aligned} V[\mathcal{S}(r))] &= \tfrac{V[S_o]}{\ell(r)^2} + E[S_{\mathrm{H}}^2] \lambda_b p_{\mathrm{H}} \int_r^{rT^{-\frac{1}{\alpha}}} \tfrac{1}{\ell(r')^2} 2\pi r' dr' \\ &= V[S_o] r^{-2\alpha} + E[S_{\mathrm{H}}^2] \tfrac{\pi\lambda_b p_{\mathrm{H}} (1 - T^{(2-\frac{2}{\alpha})})}{(\alpha - 1)} r^{2(1-\alpha)} \end{aligned} \tag{3.22}$$

To obtain expectation and variance of aggregate desired signal, the n-th moment of S_o and S_{H} are calculated by

$$E[S_o^n] = \sum_{k}^{K_{\mathrm{m,max}}} \sum_{j}^{K_{\mathrm{h,max}}} E[S_o^n \,|k, j] p_{K_{\mathrm{m,u}}}(k) p_{K_{\mathrm{h}}}(j), \tag{3.23}$$

and

$$E[S_{\mathrm{H}}^n] = \sum_{k}^{K_{\mathrm{m,max}}} \sum_{j}^{K_{\mathrm{h,max}}} E[S_{\mathrm{H}}^n \,|k, j] p_{K_{\mathrm{m}}}(k) p_{K_{\mathrm{h,u}}}(j). \tag{3.24}$$

Similar to (3.14), the n-th moment of S_o is calculated considering the *pmf* of the number of scheduled users by the serving RRH from user perspective,

$$p_{K_{\mathrm{m,u}}}(k) = \frac{k p_{K_{\mathrm{m}}}(k)}{E[K_{\mathrm{m}}]}. \tag{3.25}$$

Table 3.1: Moments of interference and desired signal transmitted from a certain RRH b with (K_{m}, K_{h})

	Interference ($X = I_b$)	Desired signal ($X = S_b$)
$E[X\,\vert K_{\mathrm{m}}, K_{\mathrm{h}}]$	P_{tx}	$P_{tx}\frac{M-(K_{\mathrm{m}}+K_{\mathrm{h}})+1}{K_{\mathrm{m}}+K_{\mathrm{h}}}$
$E[X^2\,\vert K_{\mathrm{m}}, K_{\mathrm{h}}]$	$P_{tx}^2\frac{(K_{\mathrm{m}}+K_{\mathrm{h}}+1)}{K_{\mathrm{m}}+K_{\mathrm{h}}}$	$P_{tx}^2\frac{(M-(K_{\mathrm{m}}+K_{\mathrm{h}})+1)(M-(K_{\mathrm{m}}+K_{\mathrm{h}})+2)}{(K_{\mathrm{m}}+K_{\mathrm{h}})^2}$
$V[X\,\vert K_{\mathrm{m}}, K_{\mathrm{h}}]$	$P_{tx}^2\frac{1}{K_{\mathrm{m}}+K_{\mathrm{h}}}$	$P_{tx}^2\frac{M-(K_{\mathrm{m}}+K_{\mathrm{h}})+1}{(K_{\mathrm{m}}+K_{\mathrm{h}})^2}$

The first and second moments of desired signal at unit distance ($E[S_b^n|K_{\mathrm{m}}, K_{\mathrm{m}}]$) for given ($K_{\mathrm{m}}$, K_{h}) are represented in Table. 3.1.

3.3.4 Signal to Interference Ratio

Using the Gamma approximated aggregate interference and aggregate desired signal, we can obtain distribution on SIR.

Proposition 3.3.3. Suppose that the aggregate desired signal and the aggregate interference of a target user at the distance of r from their serving RRHs are $\Gamma[k_{\mathrm{s}}, \theta_{\mathrm{s}}]$ distributed and $\Gamma[k_{\mathrm{i}}, \theta_{\mathrm{i}}]$, respectively. Then the probability distribution function (PDF) of SIR of the target user is

$$f_{\mathrm{SIR}(r)}(x) = \frac{\Gamma(k_s + k_i)}{\Gamma(k_s)\Gamma(k_i)} \left(\frac{\theta_s}{\theta_i}\right)^{k_i} \left(x + \frac{\theta_s}{\theta_i}\right)^{-(k_s+k_i)} x^{k_s-1}. \qquad (3.26)$$

Proof.

$$
\begin{aligned}
&f_{\mathrm{SIR}(r)}(x) = \tfrac{dF_{\mathrm{SIR}(r)}(x)}{dx} \\
&= \tfrac{d}{dx}\int_0^\infty \Pr[S \le xI \,|I = v] f_{I(r)}(v)dv \\
&= \tfrac{d}{dx}\int_0^\infty \Pr[S \le xv] f_{I(r)}(v)dv \\
&= \tfrac{d}{dx}\int_0^\infty \int_0^{\frac{xv}{\theta_\mathrm{s}}} \tfrac{t^{k_s-1}e^{-t}}{\Gamma(k_s)}dt \tfrac{v^{k_i-1}e^{-v}}{\theta_i^{k_i}\Gamma(k_\mathrm{i})}dv \\
&= \tfrac{1}{\Gamma(k_s)\Gamma(k_i)\theta_i^{k_i}}\int_0^\infty (xv/\theta_s)^{k_s-1}e^{-\frac{xv}{\theta_\mathrm{s}}}(v/\theta_s)v^{k_i-1}e^{-v}dv \\
&= \tfrac{x^{k_s-1}}{\Gamma(k_s)\Gamma(k_i)\theta_s^{k_s}\theta_i^{k_i}}\int_0^\infty v^{k_s+k_i-1}e^{-(\frac{x}{\theta_s}+\frac{1}{\theta_i})v}dv \\
&= \tfrac{x^{k_s-1}}{\Gamma(k_s)\Gamma(k_i)\theta_s^{k_s}\theta_i^{k_i}} \cdot \left(\tfrac{\Gamma(k_s+k_i)}{(x/\theta_s+1/\theta_i)^{k_s+k_i}}\right) \\
&= \tfrac{\Gamma(k_s+k_i)}{\Gamma(k_s)\Gamma(k_i)}\left(\tfrac{\theta_s}{\theta_i}\right)^{k_i}\left(x+\tfrac{\theta_s}{\theta_i}\right)^{-(k_s+k_i)}x^{k_s-1}
\end{aligned}
\tag{3.27}
$$

$k_s = \frac{V[\mathcal{S}]+E[\mathcal{S}]^2}{V[\mathcal{S}]}$ and $\theta_s = \frac{V[\mathcal{S}]}{E[\mathcal{S}]}$

Using the *pdf* of SIR, we can approximate cumulative distribution function (CDF) of SIR as follows.

Proposition 3.3.4.

$$
F_{SIR(r)}(x) = 1 - \left(\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}} \sum_{m=1}^{k_\mathrm{s}} \frac{\Gamma(k_\mathrm{s}+k_\mathrm{i}-m)}{\Gamma(k_\mathrm{i})(k_s-m)!} x^{k_\mathrm{s}-m}\left(x+\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{m-k_\mathrm{s}-k_\mathrm{i}}
\tag{3.28}
$$

Proof.

$$
\begin{aligned}
&\Pr[SIR(r) \le x] = \textstyle\int_0^x f_{SIR(r)}(y)dy \\
&= \tfrac{\Gamma(k_\mathrm{s}+k_\mathrm{i})}{\Gamma(k_\mathrm{s})\Gamma(k_\mathrm{i})} \left(\tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}} \textstyle\int_0^x \left(y + \tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{-(k_\mathrm{s}+k_\mathrm{i})} y^{k_\mathrm{s}-1} dy \\
&= \tfrac{\Gamma(k_\mathrm{s}+k_\mathrm{i})}{\Gamma(k_\mathrm{s})\Gamma(k_\mathrm{i})} \left(\tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}} \left\{ \left[\tfrac{1}{1-k_s-k_i} \tfrac{y^a}{(y+\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}})} \right]_0^x \right. \\
&\quad \left. - \tfrac{1}{1-k_s-k_i} \textstyle\int_0^x \left(y + \tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{-(k_\mathrm{s}+k_\mathrm{i}-1)} (k_s-1) y^{k_\mathrm{s}-2} dy \right\} \\
&\underset{(a)}{=} \tfrac{\Gamma(k_\mathrm{s}+k_\mathrm{i})}{\Gamma(k_\mathrm{s})\Gamma(k_\mathrm{i})} \left(\tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}} \left\{ - \sum_{m=1}^{k_\mathrm{s}} \left(\prod_{\ell=1}^{m} \tfrac{1}{k_\mathrm{s}+k_\mathrm{i}-\ell} \right) \tfrac{{}^{k_\mathrm{s}-1}\mathrm{P}_{m-1} x^{k_\mathrm{s}-m}}{\left(x+\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{s}+k_\mathrm{i}-m}} \right. \\
&\quad \left. + \left(\prod_{\ell=1}^{k_\mathrm{s}} \tfrac{1}{k_\mathrm{s}+k_\mathrm{i}-\ell} \right) \tfrac{(k_\mathrm{s}-1)!}{\left(\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}}} \right\} \\
&= \tfrac{\prod_{\ell=1}^{k_\mathrm{s}} (k_\mathrm{s}+k_\mathrm{i}-\ell)}{(k_s-1)!} \left(\tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}} \left\{ - \sum_{m=1}^{k_\mathrm{s}} \left(\prod_{\ell=1}^{m} \tfrac{1}{k_\mathrm{s}+k_\mathrm{i}-\ell} \right) \tfrac{(k_s-1)! x^{k_\mathrm{s}-m}}{(k_s-m)! \left(x+\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{s}+k_\mathrm{i}-m}} \right. \\
&\quad \left. + \left(\prod_{\ell=1}^{k_\mathrm{s}} \tfrac{1}{k_\mathrm{s}+k_\mathrm{i}-\ell} \right) \tfrac{(k_\mathrm{s}-1)!}{\left(\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}}} \right\} \\
&= 1 - \left(\tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}} \sum_{m=1}^{k_\mathrm{s}} \left(\prod_{\ell=m+1}^{k_\mathrm{s}} k_\mathrm{s}+k_\mathrm{i}-\ell \right) \tfrac{x^{k_\mathrm{s}-m}}{(k_s-m)! \left(x+\frac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{s}+k_\mathrm{i}-m}} \\
&= 1 - \left(\tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{k_\mathrm{i}} \sum_{m=1}^{k_\mathrm{s}} \tfrac{\Gamma(k_\mathrm{s}+k_\mathrm{i}-m)}{\Gamma(k_\mathrm{i})(k_s-m)!} x^{k_\mathrm{s}-m} \left(x + \tfrac{\theta_\mathrm{s}}{\theta_\mathrm{i}}\right)^{m-k_\mathrm{s}-k_\mathrm{i}}
\end{aligned}
\tag{3.29}
$$

where (a) holds when k_s is an integer.

3.3.5 Fronthaul Traffic Analysis

From [8], the fronthaul data traffic of partially-centralized C-RAN given M, K_m and K_h can be represented by

$$
R_\mathrm{front} = \alpha_\mathrm{front} \left(MKb_\mathrm{IQ} f_\mathrm{Pre} + (1 - \frac{\tau}{T_\mathrm{slot}}) Kb_\mathrm{DS} f_\mathrm{sym} + M(\frac{\tau}{T}) b_\mathrm{IQ} f_\mathrm{sym} \right), \tag{3.30}
$$

where α_{front} is fronthaul interface efficiency, f_{Pre} is precoder update frequency, f_{sym} is data symbol update, T_{slot} is the number of symbols during a transmission slot, τ is the number of symbols for pilot signal within a transmission slot, b_{IQ} is the number of bits for one IQ sample, and b_{DS} is the number of bits for one modulated data symbol.

Then, the average fronthaul traffic of an RRH can be computed using (3.30) as

$$E[R_{\text{front}}] = \sum_{k=0}^{K_{\text{m,max}}} \sum_{j=0}^{K_{\text{h,max}}} R_{\text{front}}(M, k+j) p_{K_{\text{m}}}(k) p_{K_{\text{h}}}(j). \tag{3.31}$$

In contrast, the fronthaul traffic of an RRH in fully-centralized C-RAN is given by

$$R_{\text{IQ}} = \alpha M b_{\text{IQ}} f_{\text{s}}, \tag{3.32}$$

where f_{s} is sampling frequency of IQ baseband signal. Note that in fully-centralized C-RAN, the fronthaul traffic does not depend on the number of users.

3.4 Simulations

In this section, we consider a system model of the form in Fig. 3.1 with 64-antenna RRHs and a single-antenna users. The locations of RRHs and users follows PPP, and ,therefore, the performance can be interpreted as the worst case performance of practical scenario.

To evaluate the performance of edge users, we compared the per-

Table 3.2: Simulation parameters

P_{tx}	2 W
a	4/3
f_s	30.72 MHz
T_{slot}	168
f_{sym}	16.8 MHz
f_{Pre}	$f_{\text{sym}}/T_{\text{slot}}$
b_{IQ}	30 bits
b_{DS}	9 bits

formance at the distance of $r_e = \sqrt{\frac{-\ln(1-\beta)}{\lambda_b \pi}}$, which satisfy that the expected portion of users whose distance to their own serving RRH is far than r_e is β. By setting $\beta = 0.95$, we can see the statistical lower 5-percentile user performance. Used parameters are summarized in Table 3.4.

Fig. 3.3 shows the relative frequency of SIR of target users and *pdf* obtained by 3.3.3. It can be seen that the derived *pdf* of SIR quite well fits the simulation results for different number of antennas and cooperation scenario. Fig. 3.4 plots the cumulative distribution function of SIR which can be interpreted as coverage probability. As can be seen in the plot, coverage probability is improved as cooperation (with thresholding parameter $T = 0.6$) is applied. Also, as the number of maximum multiplexed is reduced, SIR performance is more improved.

In Fig. 3.5, the expected fronthaul traffic and the generated fronthaul traffic in simulation are plotted. The fronthaul traffic is slightly increased by applying cooperation for victim users because the number of multiplexed users by each RRH increases in PC-RANs. In

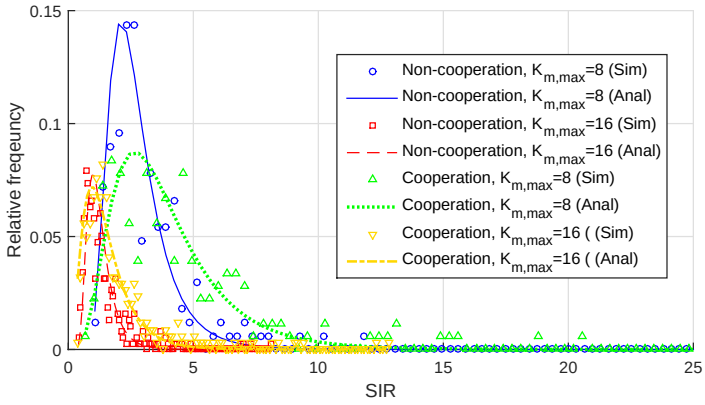


Figure 3.3: Probability distribution function of SIR of statistical 5-percentile users.

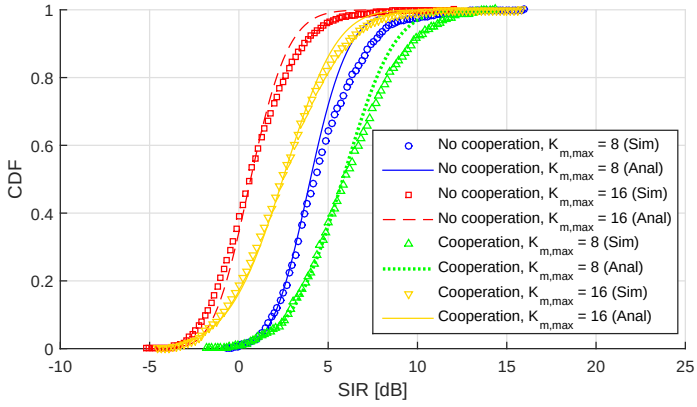


Figure 3.4: Coverage probability (*cdf* of SIR).

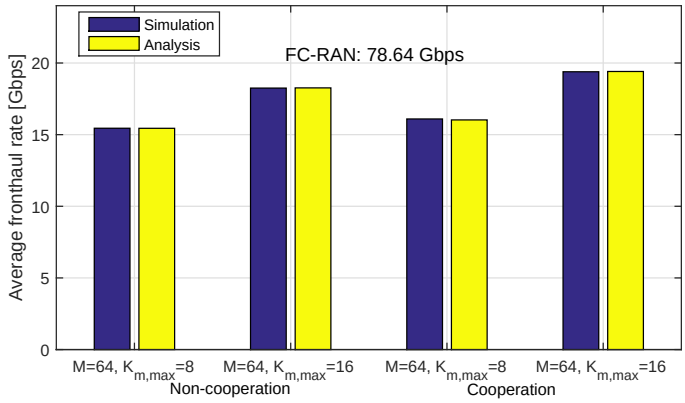


Figure 3.5: Fronthaul traffic.

PC-RANs, the data symbols and precoder of helped users are also transported to helping RRHs. Therefore, it is desirable to control the degree of cooperation regarding edge user performance as well as fronthaul traffic burden.

3.5 Summary

Large-scale antenna operation in C-RAN is a promising scenario for next generation cellular systems. In this work, we propose an analytical model to evaluate edge user performance in large-scale MIMO with zero-forcing precoding considering thresholding based cooperation of neighbor RRHs. The sum of desired signal and interference from multiple RRHs are approximated as Gamma random variables. Through the extensive simulations, it was shown that our mathematical model well approximates the results obtained from simulation.

Future work is needed to analyze considering both ZF and MRT precoding schemes.

Chapter 4

Large-scale ZF and MRT Beamforming in Partially-centralized C-RANs with Limited Fronthaul Capacity

4.1 Introduction

For future wireless communication systems, one technology that has attracted significant attention as being critical is the multiple-input multiple-output (MIMO) system with a large number antennas, the

so called massive MIMO (a.k.a, large-scale antenna system), has attracted significant attention as a key technology in future wireless communication systems [44] [45]. With a large number of antennas, a base station can simultaneously serve tens or hundreds of terminals (but fewer than the number of antennas) per channel use by utilizing proper multi-user beamforming techniques. Theoretically, as the number of antennas approaches infinity, the issues of fast fading channels, interference, and the effect of thermal noise effect can vanish, thanks to the law of large numbers, vanish [3] [4].

There have been numerous studies that address multiuser MIMO techniques in massive MIMO systems. In [3], the performances of maximal ratio transmission (MRT) in downlink and maximal ratio combining (MRC) in uplink were analyzed assuming an infinite number of base station antennas. It was shown that the system capacity of an infinite-scale antenna system is only limited by pilot overhead for channel state information (CSI) estimation of multiplexed users and pilot contamination due to pilot resource reuse in neighbor cells. In [46] and [47], the performance of zero-forcing beamforming (ZF) and MRT beamforming with a large but finite number of antennas was investigated under perfect CSI and imperfect CSI scenarios, respectively. One common result is that, ZF beamforming generally outperforms MRT beamforming for the same number antennas or users but, because of matrix inversion, requires a much larger processing load. There have also been architectural investigations into

radio access network with large-scale antennas [24, 48, 49]. Argos [24] is a large-scale antenna system that employs conjugate beamforming and it was implemented with 64 antennas. The authors proposed a localized conjugate beamforming scheme with a calibration method to maintain channel reciprocity of time division duplexing (TDD) systems. In [49], the authors presented a scalable architecture that enables real-time signal processing (for ZF beamforming) in large-scale MIMO systems was presented.

Another feature of future cellular systems is cloud-based base station architecture, also called cloud radio access network (C-RAN) [2]. C-RAN has been proposed as a cost-efficient and performance-optimizing radio access architecture solution. In conventional RANs, the baseband units (BBUs) and the remote radio heads (RRHs) are located together. In C-RAN, however, BBUs, which process baseband signals, are separate from cell sites and concentrated at a data center (BBU pool). A BBU and an RRU transport between each other a baseband signal through the fronthaul link, which might have a low latency and a high bandwidth wired or wireless link such as optical fiber or mmWave, to each other. The cloud base station structure allows more advanced techniques for the management of radio resources, interference, computing resources, and energy consumption. In addition, due to the centralized infrastructure, service providers might reduce the installation and operation cost. The authors in [50] investigated a processing resource assignment scheme according to each

cell-site's traffic load and resource pooling gain of cloud-based architectures. In [14], flexible cloud radio access network was proposed to efficiently utilize the baseband processing resource according to mobile environments. The processing resources of in a BBU pool are dynamically assigned to cell-sites based on whether the corresponding RRH operates as fractional frequency reuse (FFR) or distributed antenna system (DAS) mode.

In C-RAN, employing large-scale antennas generates a huge amount of digital signal samples in real time. Because the amount of baseband signal is proportional to the number of antennas in general[6], the fronthaul bandwidth of C-RAN systems with many antennas should be much higher than the backhaul bandwidth of the conventional systems (no C-RAN and no large-scale antenna system), imposing huge extra cost for fronthaul bandwidth expansion to operators. As a fronthaul communication protocol between a BBU and an RRH, we consider Common Public Radio Interface (CPRI)[6]. The current CPRI standards specify approximately a 10 Gbps line bit rate per one CPRI link, which can support 8 RRH antennas in a 20MHz LTE system[8]. Despite there being some studies that have adapted an architectural perspective, such as [24] [49] [50] [14], the main focus of these was on either scalability against computational complexity of large scale antenna systems or efficient usage of processing resources in centralized multicell operations.

In this paper, we propose a C-RAN structure and operation strat-

egy suited for massive MIMO operation. The main contributions of this work are as follows:

• We propose massive MIMO partially-centralized C-RAN (*MPC-RAN*) that can be classified as a partially-centralized C-RAN in which not heavy IQ (In-phase and Quadratic-phase) sample data but data symbols are transported over the fronthaul.

In our proposed *MPC-RAN*, precoder design can be performed at both the BBU pool or RRHs according to beamforming computational complexity and fronthaul traffic load. The fundamental key feature is that data symbols, precoders and CSI are separately managed and transported in C-RAN. Such separation of internal information enables flexible operation of C-RAN according to fronthaul data traffic and computational complexity of massive MIMO.

• We provide an operational insight of the proposed *MPC-RAN*. Fronthaul bandwidth requirements for each according to beamforming modes (MRT/ZF) are modeled. We formulate a wireless sum rate maximization problem for a given fronthaul link capacity and provide an algorithm that finds suboptimal beamforming configurations including beamforming mode, the number of active antennas, the number of multiplexed users and subchannel allocation for each beamforming configuration.

The rest of the paper is organized as follows. In Section 4.2, we describe our motivation and proposed C-RAN architecture. In Section 4.3, we present our system model. In Section 4.4, we formulate

a sum rate maximization problem and propose a complexity-reduced algorithm to find the optimal beamforming configuration and sub-channel allocation. In Section 4.5, we discuss our simulation results, and we conclude the paper in Section 5.6.

4.2 Motivation and Proposed Architecture

4.2.1 Cloud Radio Access Networks

In C-RAN, most functionalities and processing resource of conventional base stations are centralized at a central BBU pool and only radio frequency functions are performed at distributed RRHs. Baseband signals for RRH-antennas are entirely processed in BBUs and transported to distributed RRHs through fronthaul links in the form of IQ sample. Then, the RRHs then radiate the transmit signal after converting the IQ samples to waveform. Generally, such an architecture and operation is called *fully-centralized C-RAN* [2] and throughout this paper we will refer to it as *FC-RAN* throughout the paper.

In *FC-RAN*, the required fronthaul link rate is scaled with the number of active antennas. For example, to convey IQ samples of 8 antennas for a 20MHz LTE (Long Term Evoluation) system, 9.8304 Gbps of transmission rate is required in the fronthaul link [6]. If the radio access system is scaled to hundreds of antennas or more, fronthaul link capacity also should also be enhanced up to the order of hundreds of Gbps. Furthermore, spectrum expansion in future

cellular systems with carrier aggregation or wider spectrum operation in tens of GHz carrier frequency (a.k.a, mmWave) also requires scaled fronthaul capacity in C-RAN. Operating 256 antennas for a 100MHZ bandwidth requires a 1.573 Tbps link rate for an RRH. Thus, scaling in C-RAN arouses huge link cost albeit C-RAN has been adapted by mobile operators due to its merits in terms of deployment and operation cost.

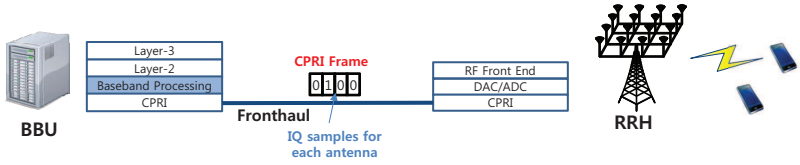
Such an inefficiency of fronthaul link and cost burden to the operator with an *FC-RAN* operation is our motivation for devising a novel C-RAN architecture and operational mechanism which is better suited for large-scale antenna operation in future cellular systems.

4.2.2 Partially-centralized C-RAN for Massive MIMO

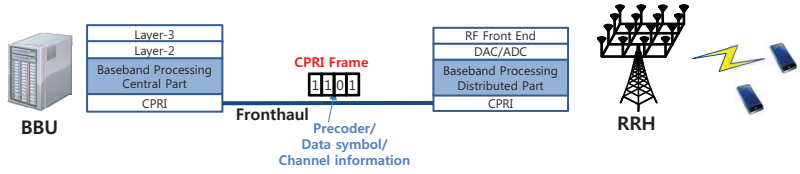
In contrast to *FC-RAN*, the C-RAN structure where baseband functions are located at RRHs is referred to as *partial centralization*. As a C-RAN solution suited for massive MIMO operation, we propose a type of *partial C-RAN*, which can utilize many antennas at the RRH by reducing the amount of data to be carried through fronthaul. In the proposed *MPC-RAN* structure, a part of the baseband processing functions is performed at BBUs and the other part is performed at RRHs for massive MIMO operation. Fig. 4.1 shows the architecture and fronthaul data of *FC-RAN* and our proposed *MPC-RAN*.

Our proposed *MPC-RAN* structure has the following features:

• In the fronthaul link between the BBU pool and the RRH, beam-



(a)



(b)

Figure 4.1: C-RAN architecture comparison: (a) Fully-centralized C-RAN (FC-RAN) (b) Massive MIMO Partially-centralized C-RAN (MPC-RAN)

forming weights for the precoder and data symbols are separately transported.

• An RRH precodes each user's data symbols using the precoding vector for the user and generates baseband signals. Such a baseband processing division allows for separate treatment of the precoding part and data stream part within C-RAN and fronthaul.

• An RRH estimates CSI (Channel State Information) between its antennas and users, $\boldsymbol{H}$, with received uplink pilot signals from users.

• For MRT beamforming, RRHs generate a precoder using the estimated CSI as MRT beamforming is simply conjugating estimated channel vectors for each user k, i.e., $\boldsymbol{f}_k = c_{\mathrm{MRT}}\boldsymbol{h}_k^*$. As a consequence, the RRH only needs to report the expected effective channel gains of users with the adopted precoders to its BBU instead of full CSI.

• For ZF beamforming, a precoder is calculated in the BBU pool to avoid a heavy precoding matrix calculation computation burden on RRHs. In addition, in the case of joint processing to handle inter-cell interference among multiple RRH-sites, precoders should be calculated at the BBU pool. To compute the precoding matrix, the BBU should know the channel gain vectors of each user.

• Other baseband processing-related functions such as modulation and coding scheme (MCS) selection, channel coding, user scheduling, and resource allocation are performed at the BBU pool.

One might argue that decoupling some baseband functions from BBUs and burdening to RRHs attenuates the resource pooling gain of

the centralized base station architecture. Nevertheless, the goal of this paper is to understand the impact of architecture, baseband signal transport method, and beamforming techniques on wireless performance and the potential of *partial centralization* in a massive MIMO scenario.

4.3 System Model

4.3.1 Massive MIMO Scenario

Throughout the paper, we consider massive MIMO TDD operation of a single RRH. The RRH is equipped with M_{tot} antennas and serves K_{tot} users. Among K_{tot} users, $\min\{\tau, K_{\text{tot}}\}$ users transmit reverse link pilot signal on the first τ symbols in each subchannel (consisting of T symbols) on which the channel gains are assumed to be constant.

The RRH serves K users among a set of users, $\mathbb{K}$, whose channel information is valid, using M active antennas. During the remaining $T - \tau$ symbols, the RRH transmits data to the K users using the precoder $\boldsymbol{F}$. Then, the RRH transmits $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{q}$ through M antennas. In this paper, MRT and ZF beamforming are considered as precoder $\boldsymbol{F}$.

$$\begin{aligned} &\text{MRT:} \quad \boldsymbol{F} = c_{\text{MRT}}\boldsymbol{H}^* \\ &\text{ZF:} \quad \boldsymbol{F} = c_{\text{ZF}}\boldsymbol{H}^*(\boldsymbol{H}\boldsymbol{H}^*)^{-1}, \end{aligned} \tag{4.1}$$

where c_{MRT} and c_{ZF} are constants for power normalization. Even

Table 4.1: Notations

b_{IQ}	the number of bits to represent an IQ sample
b_{DS}	the number of bits to represent a data symbol
T	the number of symbols in a subchannel
f_{sym}	symbol frequency
C_{front}	fronthaul link capacity
M	the number of active antennas
K	the number of multiplexed users
M_{tot}	the number of total antennas of the RRH
K_{tot}	the number of total users served by the RRH
ρ	normalized forward link received power
τ	the length of a reverse link pilot sequence

though we address only two representative precoders, MRT and ZF, any precoders can be applied by considering their features. The main notations referred to in the paper are summarized in Table 4.1.

If the entries of channel matrix follow complex normal distribution ($\mathcal{CN}(0,1)$), from [46][51], the asymptotic capacity of ZF and MRT can be represented by

$$C_{\mathrm{ZF}} = K_{\mathrm{ZF}} \log_2 \left(1 + \frac{\rho(M_{\mathrm{ZF}} - K_{\mathrm{ZF}} + 1)}{K_{\mathrm{ZF}}}\right), \tag{4.2}$$

$$C_{\mathrm{MRT}} = K_{\mathrm{MRT}} \log_2 \left(1 + \frac{\rho(M_{\mathrm{MRT}} + 1)}{\rho(K_{\mathrm{MRT}} - 1) + K_{\mathrm{MRT}}}\right). \tag{4.3}$$

Taking account of pilot signal overhead, the achievable sum rate is represented by $\gamma = (1 - \frac{\tau}{T})C$.

The number of users K can be chosen to maximize the achievable sum rate. In C-RAN, however, the achievable sum rate is restricted due to limited fronthaul link capacity even if a great number of an-

tennas are deployed at an RRH.

4.3.2 Required Transmission Rate for MIMO Operation

FC-RAN

In conventional *FC-RAN*, the BBU generates a baseband signal for each antenna in the form of a digital IQ sample using a digital signal processing function after upper layer protocol processing. As a result, the required fronthaul bit rate is given by

$$R_{\mathrm{IQ}} = \alpha M b_{\mathrm{IQ}} f_{\mathrm{s}} \tag{4.4}$$

where b_{IQ} is the number of bits to represent a pair of IQ samples and α is the redundancy in the fronthaul transport interface. [1] The IQ sampling rate, f_s, is proportional to the wireless system bandwidth. For simplicity, we omit α for the rest of the paper.

MPC-RAN

In *MPC-RAN*, data symbols are precoded at the RRH. In *MPC-RAN*, the following information should be exchanged between a BBU pool and RRHs - 1) Data symbol: data symbols of each user to be precoded at the RRH before radio transmission; 2) Precoder: precoding vectors for each user; 3) channel information: CSI or effective channel gains

[1] In CPRI, $\alpha = \left(\frac{10}{8}\right)\left(\frac{16}{15}\right) = \frac{4}{3}$ considering on 8B/10B line encoding and the control channel portion of 1/16 [6].

of users, which are used for modulation and coding scheme (MCS) and/or beamforming decision. The transport rate for data symbols of K users is

$$\phi_{\mathrm{DS}}(K) = \left(1 - \frac{\tau}{T}\right) K b_{\mathrm{DS}} f_{\mathrm{sym}}, \tag{4.5}$$

where b_{DS} denotes the number of bits to represent the constellation points of data symbols, and f_{sym} denotes symbol frequency (i.e., in 20 MHz LTE system, $f_{\mathrm{sym}} = \frac{N_{\mathrm{sc}}}{T_{\mathrm{s}}+T_{\mathrm{g}}} = 16.8MHz$, assuming 1200 useful subcarriers, N_{sc}, T_{s} is a useful symbol duration and T_{g} is a guard time interval). Note that the amount of traffic for data symbol transport does not depend on the number of antennas. To precode K users' data symbols, K precoding vectors, each of which has size of $M \times 1$, are transported every T symbols. Then, the required bit-rate for transport K users' precoding vector is represented by

$$\phi_{\mathrm{Pre}}(M, K) = \frac{1}{T} M K b_{\mathrm{IQ}} f_{\mathrm{sym}}. \tag{4.6}$$

The fronthaul data rate for channel information depends on the necessary channel information at the BBU. If the BBU has to know the full CSI between the M RRH antennas and K users, the required fronthaul rate is

$$\phi_{\text{Full-CSI}}(M, K) = \frac{1}{T} M K b_{\mathrm{IQ}} f_{\mathrm{sym}}. \tag{4.7}$$

For ZF beamforming transmission, the BBU transports data sym-

bols and corresponding precoder which is calculated at the BBU to the RRH for receiving users. Also, because precoder is calculated at the BBU, the RRH transports full CSI($\{\boldsymbol{h}_1, ..., \boldsymbol{h}_{K_{\mathrm{ZF}}}\}$). Therefore, the total required fronthaul transport rate for ZF beamforming operation can be represented by

$$\phi_{\mathrm{ZF}}(M_{\mathrm{ZF}}, K_{\mathrm{ZF}}) = \phi_{\text{Full-CSI}} + \phi_{\mathrm{DS}} + \phi_{\mathrm{Pre}}. \tag{4.8}$$

For MRT operation, an RRH generates precoding vectors for users using the CSI estimated at the RRH. Thus, the BBU needs to know only the effective channel gain of the receiving users ($\{\|\boldsymbol{h}_1\|^2, ..., \|\boldsymbol{h}_{K_{\mathrm{MRT}}}\|^2\}$) for proper MCS decision, not all the channel coefficient vectors between the antennas and users ($\{\boldsymbol{h}_1, ..., \boldsymbol{h}_{K_{\mathrm{MRT}}}\}$). In this case, the required fronthaul rate for channel information is

$$\phi_{\text{Partial-CSI}}(K) = \frac{1}{T} K b_{\mathrm{real}} f_{\mathrm{sym}}. \tag{4.9}$$

Then, the total required fronthaul bit-rate for MRT transmission is

$$\phi_{\mathrm{MRT}}(K_{\mathrm{MRT}}) = \phi_{\text{Partial-CSI}} + \phi_{\mathrm{DS}}. \tag{4.10}$$

Remark 1. For MRT beamforming, the required fronthaul bit rate is independent of the number of antennas at the RRH. This is one of the key merits of *MPC-RAN* which makes the radio access network scalable in the number of antennas.

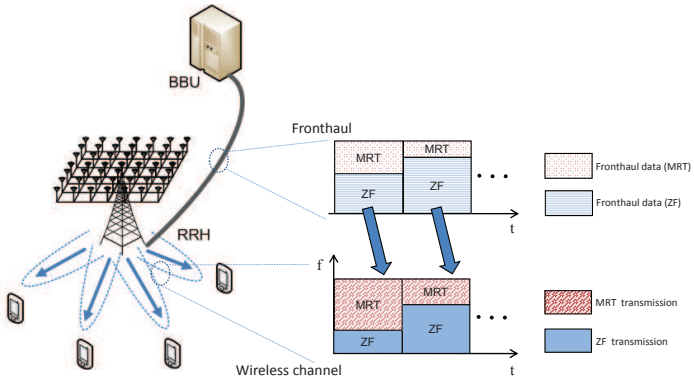


Figure 4.2: Hybrid MRT/ZF operation in *MPC-RAN*

Remark 2. The required transmission rate of ZF beamforming is always higher than that of MRT beamforming for the same beamforming configuration, i.e., the number of receiving users, active antennas, and symbols for reverse link pilot. This is due to the CSI and precoding vector information which increases in proportion to the number of active antennas.

4.4 Operation Strategy

In this section, we present an operation strategy to efficiently utilize *MPC-RAN* with capacity-limited fronthaul link. In conventional *FC-RAN*, where the available number of active antennas is predetermined and fixed by fronthaul capacity. From (4.2) and (4.3), we know that, without fronthaul constraints, the sum rate maximization for each beamforming can be achieved by simply deciding the number of mul-

tiplexed users. Thus, the beamforming precoder and the number of users (or even the scheduled user set) are decided (as in [47]) without consideration of fronthaul capacity. As shown in 4.3.2, however, different fronthaul link rates are required for baseband signal representation according to beamforming configuration in *MPC-RAN*. Moreover, the required fronthaul rate depends on the number of multiplexed users and active antennas (ZF beamforming). For this reason, the beamforming strategy should be decided taking the required fronthaul rates into account as well as the wireless performance.

4.4.1 Problem Formulation

Our objective is to maximize the achievable sum rate of the cell under the given fronthaul link capacity. To meet the fronthaul link capacity constraint, the number of users or the number of active antennas over entire subchannels should be jointly decided.

We consider asymptotically achievable sum rate as in [51] [46] which is averaged over a frequency selective channel, instead of instantaneous achievable sum rate. This approach can be justified in the way that effective channel gain becomes harder as the number of antennas increases. For these reasons, instead of solving a joint beamforming and scheduling problem, we solve a beamforming configuration and subchannel allocation problem that is enough to provide insight on the desirable operation strategy on the proposed *MPC-RAN* and computationally tractable.

Fig. 4.2 shows the operation scenario conceptually. Considering the influence of beamforming mode, the number of active antennas, and the multiplexed users on wireless sum rate and the given fronthaul capacity, we can write a beamforming configuration problem as follows.

P1-beamforming configuration:

$$\begin{aligned} &\underset{\{B_s,M_s,K_s\}}{\text{maximize}} \quad \sum_{s\in\mathbb{S}} \gamma_s(B_s,M_s,K_s) \\ &\text{subject to} \\ &\qquad \sum_{s\in S} \phi_s(B_s,M_s,K_s) \le C_{\text{front}}, \\ &\qquad K_s \le M_s, \qquad\qquad \forall s \in \mathbb{S}. \end{aligned} \tag{4.11}$$

The achievable sum rate on subchannel $s \in \mathbb{S}$, γ_s, depends on the beamforming scheme B_s, the number of active antennas M_s, and the number of multiplexed users K_s on subchannel s. The beamforming configuration problem over multiple subchannels is an integer programming problem and therefore NP-hard. The complexity of an exhaustive search to find the optimal is $O\left((|\mathbb{B}|M_{\text{tot}}|\mathbb{K}|)^{|\mathbb{S}|}\right)$. Under the assumption that the average sum rate on each subchannel depends only on the beamforming configuration, the problem can be modified to the following problem that decides $\mathbf{y}$, which indicates the number of subchannels for each possible beamforming configuration $v \in \mathbb{V}$

where $\mathbb{V}$ is set of all possible beamforming configurations.

P2-beamforming configuration:

$$\begin{aligned} &\underset{\mathbf{y}}{\text{maximize}} \quad \sum_{v\in\mathbb{V}} \gamma_v y_v \\ &\text{subject to} \\ &\qquad \sum_{v\in\mathbb{V}} \phi_v y_v \le C_{\text{front}}, \\ &\qquad \sum_{v\in\mathbb{V}} y_v \le N_{\text{sub}}. \end{aligned} \tag{4.12}$$

However, the problem **P2-beamforming configuration** is also integer programming, and the complexity of exhaustive search is $O(|\mathbb{S}|^{|\mathbb{V}|})$. To achieve a suboptimal solution with tractable complexity, we design a heuristic algorithm.

4.4.2 Heuristic Algorithm : Beamforming Configuration and Subchannel Allocation

Now we describe our proposed algorithm which decides the beamforming configuration (including beamforming mode, active antenna numbers, and the number of spatially multiplexed users) and the number of subchannels for each beamforming configuration. To design a low complexity algorithm that finds suboptimal close to the optimal, we select only two representative beamforming configurations at most, one from ZF and another one from MRT.

To find the optimal combination of the number of receiving users, active antennas, and subchannels for each beamforming, the algorithm

decides the optimal bandwidth allocation for each beamforming configuration. The bandwidth allocation part decides the optimal bandwidth satisfying the given fronthaul capacity for each beamforming configuration.

Algorithm 1 Beamforming Configuration Algorithm

Initialization:
$M^{\star}_{\mathrm{ZF}}, K^{\star}_{\mathrm{ZF}}, K^{\star}_{\mathrm{MRT}}, N^{\star}_{\mathrm{ZF}}, N^{\star}_{\mathrm{MRT}}, R^{\star} \leftarrow 0$
$M_{\mathrm{MRT}} = M^{\star}_{\mathrm{MRT}} \leftarrow M_{\mathrm{tot}}$
for each $(M_{\mathrm{ZF}}, K_{\mathrm{ZF}}, K_{\mathrm{MRT}})$ **do**
    $\gamma_{\mathrm{ZF}} \leftarrow \gamma_{\mathrm{ZF}}(M_{\mathrm{ZF}}, K_{\mathrm{ZF}})$
    $\gamma_{\mathrm{MRT}} \leftarrow \gamma_{\mathrm{MRT}}(M_{\mathrm{MRT}}, K_{\mathrm{MRT}})$
    $\phi_{\mathrm{ZF}} \leftarrow \phi_{\mathrm{ZF}}(K_{\mathrm{ZF}})$
    $\phi_{\mathrm{MRT}} \leftarrow \phi_{\mathrm{MRT}}(M_{\mathrm{MRT}}, K_{\mathrm{MRT}})$
    Obtain $(\tilde{W}_{\mathrm{ZF}}, \tilde{W}_{\mathrm{MRT}})$ from Table 4.3 or 4.4.
    if $\phi_{\mathrm{MRT}} < \phi_{\mathrm{ZF}}$ **then**
        $N_{\mathrm{ZF}} = \left\lfloor \tilde{W}_{\mathrm{ZF}} N_{\mathrm{sub}} \right\rfloor$, $N_{\mathrm{MRT}} = \left\lceil \tilde{W}_{\mathrm{MRT}} N_{\mathrm{sub}} \right\rceil$
    else
        $N_{\mathrm{ZF}} = \left\lceil \tilde{W}_{\mathrm{ZF}} N_{\mathrm{sub}} \right\rceil$, $N_{\mathrm{MRT}} = \left\lfloor \tilde{W}_{\mathrm{MRT}} N_{\mathrm{sub}} \right\rfloor$
    end if
    if $\gamma_{\mathrm{ZF}} N_{\mathrm{ZF}} + \gamma_{\mathrm{MRT}} N_{\mathrm{MRT}} > R^{\star}$ **then**
        $R^{\star} \leftarrow \gamma_{\mathrm{ZF}} N_{\mathrm{ZF}} + \gamma_{\mathrm{MRT}} N_{\mathrm{MRT}}$
        $M^{\star}_{\mathrm{ZF}} \leftarrow M_{\mathrm{ZF}}$
        $K^{\star}_{\mathrm{ZF}} \leftarrow K_{\mathrm{ZF}}$
        $K^{\star}_{\mathrm{MRT}} \leftarrow K_{\mathrm{MRT}}$
        $N^{\star}_{\mathrm{ZF}} \leftarrow N_{\mathrm{ZF}}$
        $N^{\star}_{\mathrm{MRT}} \leftarrow N_{\mathrm{MRT}}$
    end if
end for

For each beamforming configuration, the optimal bandwidth is chosen from Table 4.3 and 4.4. [2] After that, the continuous band-

[2] According to given beamforming configuration, the second case ($\phi_{\mathrm{MRT}} \geq \phi_{\mathrm{ZF}}$) can also happen.

width is In deciding the number of active antennas for MRT beamforming, we use the following proposition.

Proposition 4.4.1. In *MPC-RAN*, using all the antennas always maximize the spectral efficiency for MRT beamforming, i.e., $M^{\star}_{\mathrm{MRT}} = M_{\mathrm{tot}}$.

Proof. This is quite simple in that the required fronthaul link rate for MRT in *MPC-RAN* does not depend on the number of active antennas M as in (4.10), while the achievable sum rate of MRT is non-decreasing with M.

4.5 Performance Evaluation and Discussion

4.5.1 Scenario

To evaluate the performance of the proposed *MPC-RAN* and operation strategy, we consider a single cell-site covered by an RRH. The parameters used are set based on LTE and CPRI standards [6] and summarized in Table 5.2. Assuming that BPSK, QPSK, 16-QAM, 64-QAM, and 256-QAM are used as modulation schemes, b_{DS} is set to be 9 bits because all the modulation schemes and constellation points (2+4+16+64+256 = 342 candidates) can be represented by 9 bits. We assume that other OFDM parameters are identical to those of LTE with 20 MHz system bandwidth. According to the fronthaul capacity and selected beamforming configuration, a part of antennas

Table 4.2: Simulation parameters

f_s	30.72 MHz
N_{sc}	1200
N_{sub}	100
N_{smooth}	14
T_{sl}	1 msec
f_{sym}	15 kHz
b_{IQ}	30 bits
b_{N}	15 bits
b_{DS}	9 bits
$K_{\text{tot}} = \tau$	48

may be activated among M_{tot} deployed antennas to meet the fronthaul capacity.

4.5.2 Optimality and Performance Ratio

We randomly picked 100 different scenarios varying the number of RRH antennas, total users, and average user SNR. For each scenario the solution of **problem P2** was obtained by branch-and-bound technique with linear programming relaxation [52], and we evaluated the proposed heuristic algorithm. To evaluate the heuristic algorithm, we measured the performance ratio, which is defined as the ratio between the sum-rate value obtained from our heuristic algorithm and that from the branch-and-bound technique. As shown in Fig. 4.3, the performance ratios exceeds 99% for all scenarios. Although the branch-and-bound technique can find better solution guaranteeing relative gap of 10^{-9} to the optimal, its complexity is still intractable for large scale beamforming scenario. On the other hand, the complexity

Table 4.3: Bandwidth Solution Table ($\phi_{\mathrm{MRT}} \leq \phi_{\mathrm{ZF}}$)

	$C_{\mathrm{front}} < \phi_{\mathrm{MRT}} N_{\mathrm{sub}}$ (Bottlenecked)	$\phi_{\mathrm{MRT}} N_{\mathrm{sub}} \leq C_{\mathrm{front}} < \phi_{\mathrm{ZF}} N_{\mathrm{sub}}$ (Semi-bottlenecked)	$C_{\mathrm{front}} \geq \phi_{\mathrm{ZF}} N_{\mathrm{sub}}$ (Non-bottlenecked)
$\frac{\gamma_{\mathrm{MRT}}}{\phi_{\mathrm{MRT}}} < \frac{\gamma_{\mathrm{ZF}}}{\phi_{\mathrm{ZF}}}$	$\tilde{W}_{\mathrm{MRT}} = 0$ $W^{\star}_{\mathrm{ZF}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{ZF}} N_{\mathrm{sub}}}$	$\tilde{W}_{\mathrm{MRT}} = 0$ $\tilde{W}_{\mathrm{ZF}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{ZF}} N_{\mathrm{sub}}}$	$\tilde{W}_{\mathrm{MRT}} = 0$ $\tilde{W}_{\mathrm{ZF}} = 1$
$\frac{\phi_{\mathrm{MRT}}}{\phi_{\mathrm{ZF}}} \leq \frac{\gamma_{\mathrm{MRT}}}{\gamma_{\mathrm{ZF}}} \leq 1$	$\tilde{W}_{\mathrm{MRT}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{MRT}} N_{\mathrm{sub}}}$ $\tilde{W}_{\mathrm{ZF}} = 0$	$\tilde{W}_{\mathrm{MRT}} = \frac{\phi_{\mathrm{ZF}} N_{\mathrm{sub}} - C_{\mathrm{front}}}{\phi_{\mathrm{ZF}} N_{\mathrm{sub}} - \phi_{\mathrm{MRT}} N_{\mathrm{sub}}}$ $\tilde{W}_{\mathrm{ZF}} = \frac{C_{\mathrm{front}} - \phi_{\mathrm{MRT}} N_{\mathrm{sub}}}{\phi_{\mathrm{ZF}} N_{\mathrm{sub}} - \phi_{\mathrm{MRT}} N_{\mathrm{sub}}}$	$\tilde{W}_{\mathrm{MRT}} = 0$ $\tilde{W}_{\mathrm{ZF}} = 1$
$\gamma_{\mathrm{ZF}} < \gamma_{\mathrm{MRT}}$	$\tilde{W}_{\mathrm{MRT}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{MRT}} N_{\mathrm{sub}}}$ $\tilde{W}_{\mathrm{ZF}} = 0$	$\tilde{W}_{\mathrm{MRT}} = 1$ $\tilde{W}_{\mathrm{ZF}} = 0$	$\tilde{W}_{\mathrm{MRT}} = 1$ $\tilde{W}_{\mathrm{ZF}} = 0$

Table 4.4: Bandwidth Solution Table ($\phi_{\mathrm{MRT}} > \phi_{\mathrm{ZF}}$)

	$C_{\mathrm{front}} < \phi_{\mathrm{ZF}} N_{\mathrm{sub}}$ (Bottlenecked)	$\phi_{\mathrm{ZF}} N_{\mathrm{sub}} \leq C_{\mathrm{front}} < \phi_{\mathrm{MRT}} N_{\mathrm{sub}}$ (Semi-bottlenecked)	$C_{\mathrm{front}} \geq \phi_{\mathrm{MRT}} N_{\mathrm{sub}}$ (Non-bottlenecked)
$\frac{\gamma_{\mathrm{ZF}}}{\phi_{\mathrm{ZF}}} < \frac{\gamma_{\mathrm{MRT}}}{\phi_{\mathrm{MRT}}}$	$W^*_{\mathrm{MRT}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{MRT}} N_{\mathrm{sub}}}$ $W^*_{\mathrm{ZF}} = 0$	$\tilde{W}_{\mathrm{MRT}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{MRT}} N_{\mathrm{sub}}}$ $\tilde{W}_{\mathrm{ZF}} = 0$	$\tilde{W}_{\mathrm{MRT}} = 1$ $\tilde{W}_{\mathrm{ZF}} = 0$
$\frac{\phi_{\mathrm{ZF}}}{\phi_{\mathrm{MRT}}} \leq \frac{\gamma_{\mathrm{ZF}}}{\gamma_{\mathrm{MRT}}} \leq 1$	$\tilde{W}_{\mathrm{MRT}} = 0$ $\tilde{W}_{\mathrm{ZF}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{ZF}} N_{\mathrm{sub}}}$	$\tilde{W}_{\mathrm{MRT}} = \frac{C_{\mathrm{front}} - \phi_{\mathrm{ZF}} N_{\mathrm{sub}}}{\phi_{\mathrm{MRT}} N_{\mathrm{sub}} - \phi_{\mathrm{ZF}} N_{\mathrm{sub}}}$ $\tilde{W}_{\mathrm{ZF}} = \frac{\phi_{\mathrm{MRT}} N_{\mathrm{sub}} - C_{\mathrm{front}}}{\phi_{\mathrm{MRT}} N_{\mathrm{sub}} - \phi_{\mathrm{ZF}} N_{\mathrm{sub}}}$	$\tilde{W}_{\mathrm{MRT}} = 1$ $\tilde{W}_{\mathrm{ZF}} = 0$
$\gamma_{\mathrm{MRT}} < \gamma_{\mathrm{ZF}}$	$\tilde{W}_{\mathrm{MRT}} = 0$ $\tilde{W}_{\mathrm{ZF}} = \frac{C_{\mathrm{front}}}{\phi_{\mathrm{ZF}} N_{\mathrm{sub}}}$	$\tilde{W}_{\mathrm{MRT}} = 0$ $\tilde{W}_{\mathrm{ZF}} = 1$	$\tilde{W}_{\mathrm{MRT}} = 0$ $\tilde{W}_{\mathrm{ZF}} = 1$

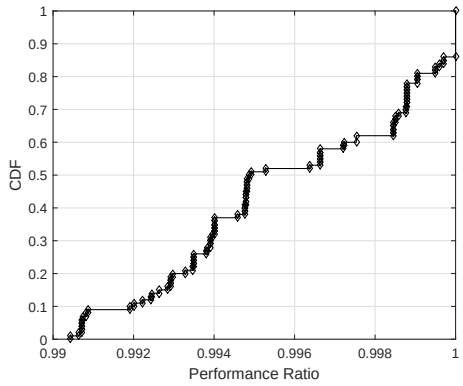


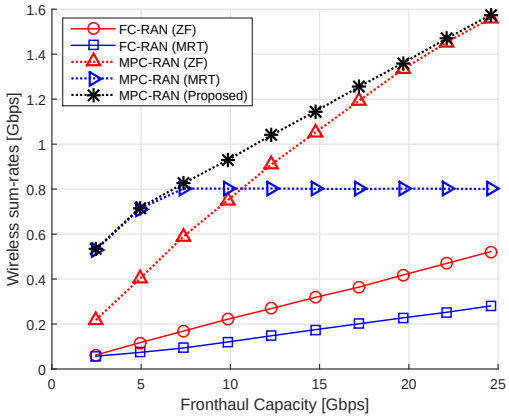
Figure 4.3: Performance ratio between the sum-rate obtained from branch-and-bound with LP relaxation and that obtained from the heuristic algorithm.

of the heuristic algorithm is $O\left(M_{\text{tot}}|\mathbb{K}|^2\right)$. Thus we can conclude that the proposed heuristic algorithm is efficient and easy to implement, achieving good approximation.

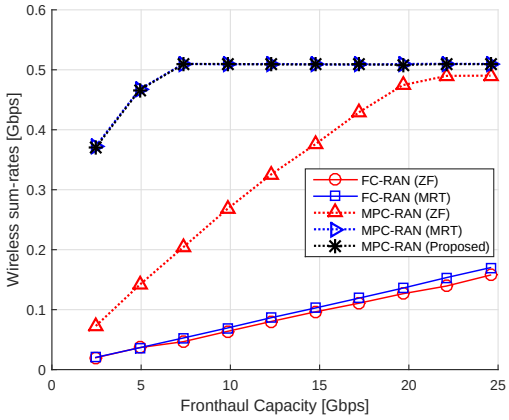
4.5.3 Simulation Results

Figs. 4.4 and 4.5 show the maximized wireless sum rates of *FC-RAN* and the proposed *MPC-RAN* with 64 RRH antennas and 256 RRH antennas. The maximized sum rates of *FC-RAN* increases with more available antennas as the fronthaul capacity increases. As more antennas can be activated, more users can be served simultaneously. As can be seen, for the given fronthaul capacity, *MPC-RAN* achieves higher sum rates than *FC-RAN*. The fundamental reason for this is that more antennas can be activated in *MPC-RAN* than in *FC-RAN*.

In *MPC-RAN*, MRT beamforming which utilizes the whole de-

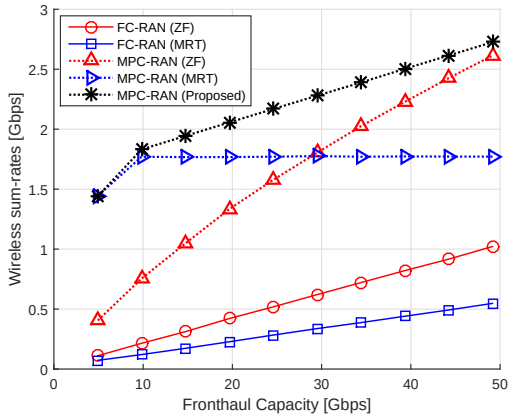


(a) $M_{\text{tot}} = 64$, $K_{\text{tot}} = \tau = 48$, $\rho = 10$ dB

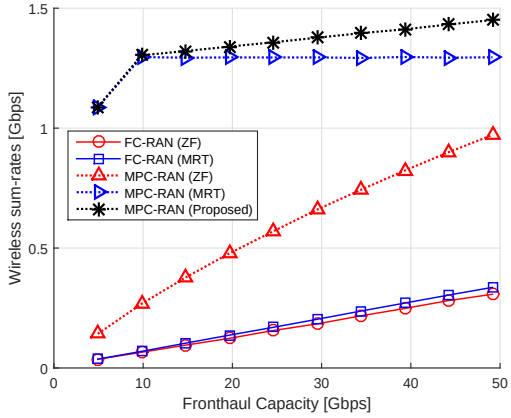


(b) $M_{\text{tot}} = 64$, $K_{\text{tot}} = \tau = 48$, $\rho = 0$ dB

Figure 4.4: Sum rate comparison over fronthaul capacity with $M = 64$.



(a) $M_{\text{tot}} = 256$, $K_{\text{tot}} = \tau = 48$, $\rho = 10$ dB



(b) $M_{\text{tot}} = 256$, $K_{\text{tot}} = \tau = 48$, $\rho = 0$ dB

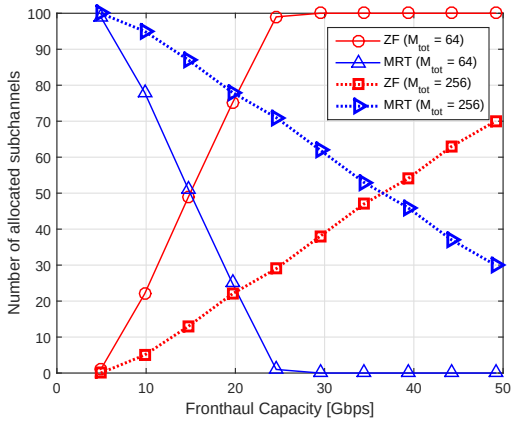
Figure 4.5: Sum rate comparison over fronthaul capacity $M = 256$.

ployed antennas outperforms ZF beamforming when fronthaul capacity is relatively low. As the fronthaul capacity increases, the maximized sum rate of ZF steadily increases and exceeds that of MRT whereas MRT saturates at much lower fronthaul capacity.

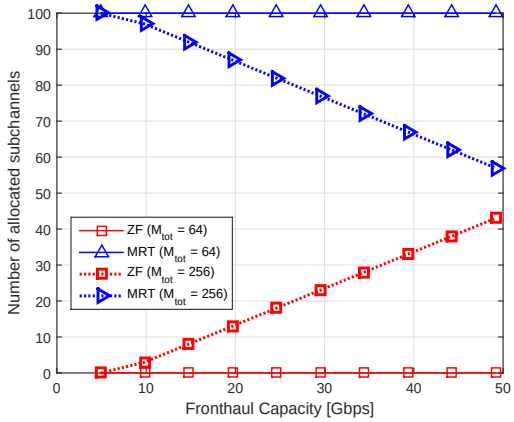
We can also confirm that our proposed beamforming configuration and subchannel allocation algorithm improves the sum-rates further depending on the fronthaul capacity. The number of subchannels among 100 allocated to each beamforming corresponding to Figs. 4.4 and 4.5 are shown in Fig. 4.6 for two average SNR values of 10 dB and 0 dB. As the fronthaul capacity increases, in most cases, more subchannels are allocated to ZF beamforming which achieves higher sum rates but requires much more fronthaul bandwidth than MRT beamforming. In 4.6(b), when average SNR is 0 dB and $M_{\text{tot}} = 64$, only MRT beamforming is used regardless of fronthaul capacity. This is because $\gamma(M^{\star}_{\text{MRT}}, K^{\star}_{\text{MRT}})$ is always larger than $\gamma(M^{\star}_{\text{ZF}}, K^{\star}_{\text{ZF}})$ even though fronthaul is not a bottleneck.

It can be interpreted that 1) to achieve the optimal performance MRT needs much less fronthaul bandwidth, and 2) if the same number of antennas are deployed and the fronthaul capacity is large enough, the performance of ZF is superior to MRT.

We can also see that in the lower SNR regime, the crossing points of MRT and ZF in both *FC-RAN* and *MPC-RAN* are shifted to left. This means MRT beamforming becomes preferred to ZF beamforming in a low SNR regime, which coincides with the general wireless



(a) Average SNR = 10 dB



(b) Average SNR = 0 dB

Figure 4.6: Subchannel allocation results of our proposed algorithm.

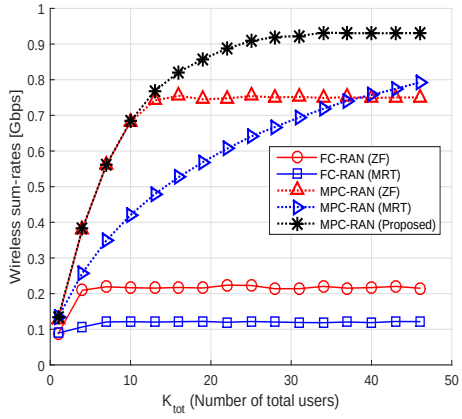
performance trend of MRT and ZF.

The required fronthaul link rate for MRT beamforming does not depend on the number of antennas. Thus, the maximum required capacity of fronthaul link is $\phi_{\mathrm{MRT,\,max}} = \phi_{\mathrm{MRT}}(|\mathbb{K}|)$. Thus, if $C_{\mathrm{front}} \geq \phi_{\mathrm{MRT}}(|\mathbb{K}|)$, the fronthaul link capacity is sufficient to operate *MPC-RAN* for MRT beamforming without any constraints on the number of users or the number of antennas. In contrast, the required transmission rates for CSI and the precoder of the ZF beamforming operation increase in proportion to both M and K. Therefore, the required fronthaul link rate increases without bound as the number of antennas M increases even though the number of users K is limited by $|\mathbb{K}|$.

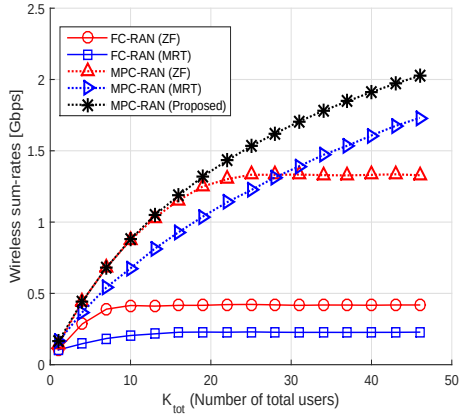
Fig. 4.7 shows the maximized sum rates for different number of cell users. Note that when the number of total users is small (i.e., small cell), ZF beamforming is preferred. This is because to achieve high sum rates, MRT should multiplex more users than ZF. Also we can confirm that proper mixing of ZF and MRT beamforming further improves the wireless sum rates.

4.5.4 Discussion

From the simulation results, the desirable beamforming strategy can be summarized in Table 5.4. We see that when fronthaul capacity is large, the performance of ZF is better than MRT. However, it is difficult to say that operation in such a region is the best network design from the viewpoint of operators. It may depend on the actual



(a) $M_{\text{tot}} = 64$, $\tau = 48$, $\rho = 10$ dB, $C_{\text{front}} = 9.8304$ Gbps



(b) $M_{\text{tot}} = 256$, $\tau = 48$, $\rho = 10$ dB, $C_{\text{front}} = 2{\times}9.8304$ Gbps

Figure 4.7: Sum rate comparison over the number of total users.

Table 4.5: Desired beamforming technique in MPC-RAN

	Low	High
Fronthaul capacity	MRT	ZF
Number of deployed antennas	ZF	MRT
Number of users in a cell	ZF	MRT
Average SNR	MRT	ZF

cost for fronthaul capacity enhancement and the traffic load statistics. Moreover, ZF beamforming takes up large processing resources in the BBU pool, leading to higher processing resource requirements from the BBU pool as well as additional energy consumption in BBU pool as well. In addition, because the performance of MRT is less sensitive to fronthaul capacity but increases more radically with the number of antennas than ZF, MRT may become gradually preferred or primary as scaling up RRH antennas. Although a long-term joint beamforming configuration and bandwidth partitioning problem has been investigated in this work, a short-term scheduling algorithm shall be studied considering instantaneous channel condition and combination of multiplexed users especially for a small number of deployed antennas (or active antennas probably due to insufficient fronthaul capacity) environments.

4.6 Summary

For performance and cost efficiency enhancement, next-generation cellular system should utilize much larger numbers of antennas on the

centralized base station structure. In this paper, we have proposed a new C-RAN architecture, *MPC-RAN*, and operation methodology for efficient usage of limited fronthaul link and scalability in the number of antennas. We formulated a sum rate maximization problem in *MPC-RAN*, which can be generalized to any type of C-RAN structure where baseband signal generation is completed at RRHs and not the at BBU pool. As a solution, we designed a heuristic beamforming configuration and subchannel allocation algorithm utilizing some properties of *MPC-RAN* to minimize algorithm complexity. Simulation results show that the proposed *MPC-RAN* operation achieved a much higher sum rate than did the *FC-RAN* operation in fronthaul-bottlenecked massive MIMO scenarios and that proper mixing of different beamforming configurations additionally improves the sum rate in the proposed *MPC-RAN*.

Chapter 5

Before/After Precoded Massive MIMO in C-RANs with Fronthaul Capacity Limitation

5.1 Introduction

Multiple-input multiple-output (MIMO) is a key technique to improve the throughput of future wireless communication systems. For a point-to-point link with multiple antennas, it has been shown that the capacity grows linearly with the minimum number of transmit and receive antennas [53] [54]. Practically, however, it is difficult to

increase the number of antennas at mobile terminals due to the limitation of space, which bounds the capacity gain of MIMO. Multi-user MIMO (MU-MIMO), where a base station communicates with multiple mobile users simultaneously in a spatial domain, provides a chance to increase the sum capacity through joint precoding (downlink) and joint decoding (uplink) at the base station even when each user has a single antenna [55]. Therefore, the evolution to cellular systems with a larger number of antennas at the base station is essential to enhance the capacity without extra spectral resources.

Numerous papers have researched on transmission mechanisms of multi-user MIMO. Simple zero-forcing (ZF) based linear algorithms were proposed in [56] and [57] for MU-MIMO where the transmitter and receivers are equipped with multiple antennas. A recent proposal to further maximize network capacity and to conserve energy is massive MIMO [58, 3, 59, 24]. In the literature, it was shown that, in a single-cell system, it is always advantageous to have an unlimited number of antennas at the transmitter [3] and also at the receiver [60]. In [3], the author proposed massive MIMO systems using a simple linear algorithm such as maximal ratio transmission (MRT) in downlink and maximal ratio combining (MRC) in uplink. It was shown that the system capacity with an unlimited number of antennas is only degraded by pilot contamination in a multi-cell network. In [4], the downlink performance of MRT and ZF beamforming for massive MIMO systems were investigated. In [61], the authors showed theo-

retically and numerically the effect of pilot contamination. With MRT precoding, inter-user interference is eliminated when the transmitter has unlimited number of antennas. The assumption of an infinite number of antennas, however, is not feasible in practice. In [46] and [47], the authors investigated the capacity of the numerous but finite antenna systems for perfect CSI and imperfect CSI scenarios, respectively. The authors in [24] implemented a many-antenna infrastructure system with 64 transmit antennas. From the experimental results of [24] and theoretical results of [46], it was shown that there is a large performance gap between MRT and ZF with respect to the number of served users when the number of transmit antennas is large but limited.

Another approach to enhance system performance of wireless service providers has been considered from an architecture, that is, cloud radio access networks (C-RAN a.k.a. cloud base station). In these networks, the baseband unit (BBU) processes the most PHY/MAC functions and generates (or decodes in uplink) IQ (In-phase and Quadrature-phase) data for transmission at the remote radio head (RRH) (or reception in uplink). The IQ-data is carried through a high speed wired-link such as optical fiber cables [2]. With such centralized BBUs and distributed RRHs structures, network-wide performance can be more enhanced by exploiting real-time joint scheduling or network MIMO as in [46], especially, in managing inter-cell interference. In addition, due to the centralized computational resources at the BBU, wireless

service providers can reduce the installation and operation cost. In [50], the resource management and resource pooling gain of IQ-based cloud base station were studied. None of previous work, however, considered IQ-data transfer issue in massive MIMO systems.

In a multi-antenna system, generally, the amount of IQ-data is proportional to the number of antennas per RRH [6]. For this reason, a large-scale antenna cellular system incurs a significant cost for the fronthaul between BBU and RRH. As seen in latter, the required transmission rate of IQ-data in a fronthaul link is significantly increased with RRH antennas, and the current fronthaul link rate is not sufficient to support large-scale antenna systems.

Such a challenge for massive MIMO systems in C-RAN motivates us to investigate efficient IQ-data transfer methods and MU-MIMO beamforming strategies to improve cellular system capacity under a given wired-link capacity and mobile environments. The contributions of this work are as follows:

• We investigate two types of IQ-data transfer methods in the fronthaul link: 1) *"after-precoding"* with which a BBU transfers IQ-data after precoding data symbols with beamforming matrix (or vector), and 2) *"before-precoding"* with which a BBU transfers beamforming weights for each data stream and data symbols separately before data symbols are precoded. The required bit-rate for "after-precoding" IQ-data transfer method only depends on the number antennas used for transmission/reception at the RRH. With "after-precoding", all the

information for IQ-data should be exchanged for each symbol between the BBU and the RRH. In contrast, with "before-precoding" IQ-data transfer method, data symbols for each user are exchanged for each symbol duration, but beamforming weights for each data stream are exchanged less frequently according to the channel coherence time.

• We provide an insight in operation of massive MIMO systems where the fronthaul link is a bottleneck due to a large number of antennas or users served simultaneously. We define a wireless capacity maximization problem under a given fronthaul link rate constraint. Our objective is to choose the number of active antennas, the number of users, beamforming method, and IQ-data transfer method to maximize the wireless sum-rate.

Throughout this paper, we focus on the downlink transmission in a single cell. A similar approach may be performed for uplink. To the best of our knowledge, this work is the first trial that addresses the IQ-data exchange issue in cloud MIMO systems, which will be critical to the design of future massive MIMO cellular systems.

The rest of the paper is organized as follows. Section 5.3 describe a considered system model. Section 5.4 presents operation algorithms for cloud massive MIMO system. Section 5.5 shows numerical results and discussions, and we conclude the paper in Section 5.6.

5.2 Notation

In this section, we list all the mathematical notation and symbols in this section.

$\boldsymbol{A}^T$: transpose of matrix $\boldsymbol{A}$,

$\boldsymbol{A}^*$: conjugate transpose of matrix $\boldsymbol{A}$,

$\boldsymbol{A}^{-1}$: inverse of matrix $\boldsymbol{A}$,

$\star$: this superscript denotes the optimal value,

N_{Sub} : the number of subcarriers in an orthogonal frequency division multiplexing (OFDM) system,

b_{IQ} : the number of bits to represent an IQ sample,

b_{DS} : the number of bits to represent a demodulated signal,

T_{Sym} : symbol duration, during which, N_{Sub} symbols are transmitted on wireless link in the OFDM system,

T_w : precoding matrix update interval,

f_{Sym} : symbol frequency ($= 1/T_{\text{Sym}}$),

f_w : precoding matrix update frequency ($= 1/T_w$),

ϕ : maximum fronthaul link rate,

m : the number of active antennas used to transmit,

k : the number of users served simultaneously (number of spatially multiplexed users),

s^D : IQ-data transfer method in fronthaul link, i.e., $s^D = \{$"after", "before"$\}$,

s^B : beamforming method of RRH, i.e., $s^B = \{$"ZF", "MRT"$\}$,

M : the number of total antennas of the RRH,

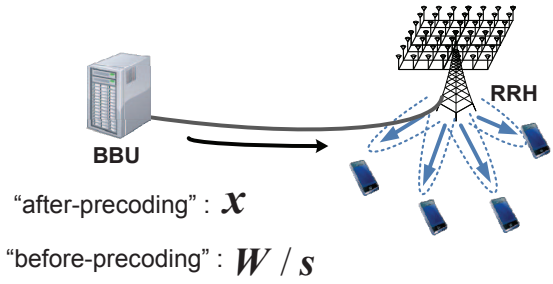


Figure 5.1: Two IQ-data Transfer Methods.

K : the number of total users served by (associated with) the RRH,

$k_{\max}$: maximum feasible k which satisfies constraint conditions,

ρ_f : normalized forward link power (the radiated power of the RRH divided by the variance of the noise),

ρ_r : normalized reverse link power (the radiated power of the mobile terminal divided by the variance of the noise),

τ_r : the number of symbols in a reverse link pilot sequence,

$\boldsymbol{H}$: $k \times m$ channel response matrix with independent $\mathcal{CN}(0,1)$ entries,

$\boldsymbol{Z}$: $k \times m$ channel estimate noise matrix with independent $\mathcal{CN}(0,1)$ entries,

$\boldsymbol{n}$: $k \times 1$ channel noise vector with independent $\mathcal{CN}(0,1)$ entries.

5.3 System model

5.3.1 Channel Estimation and Multi-user Beamforming

The independent Rayleigh fading channels between the m RRH active antennas and the k users are described by a $k \times m$ matrix, $\boldsymbol{H}$,[1] where the entries of $\boldsymbol{H}$ are independent $\mathcal{CN}(0,1)$ random variables. k users transmit orthogonal pilot sequences, consisting of τ_r symbols, to inform the BBU of $\boldsymbol{H}$. The processed signal after correlation of the received pilot signals with the conjugates of the respective pilot sequences is given by

$$\boldsymbol{V} = \sqrt{\tau_r \rho_r}\boldsymbol{H} + \boldsymbol{Z}. \tag{5.1}$$

The minimum mean-square error (MMSE) estmate for $\boldsymbol{H}$ is

$$\hat{\boldsymbol{H}} = \frac{\sqrt{\tau_r \rho_r}}{1 + \tau_r \rho_r}\boldsymbol{V}. \tag{5.2}$$

The signals that an RRH transmits to each antenna is represented by an $m \times 1$ vector, $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{d}$, where $\boldsymbol{W}$ and $\boldsymbol{d}$ denote an $m \times k$ precoding matrix and a $k \times 1$ data symbol vector, respectively. The received signal vector of receiving users is given by $\boldsymbol{y} = \sqrt{\rho_f}\boldsymbol{H}\boldsymbol{W}\boldsymbol{d} + \boldsymbol{n}$.

[1] Throughout this paper, we assume $k \leq m$. One might argue that k could be larger than m but for this, non-linear precoding is required. So we leave this issue for future work.

Then, the received signal at the j-th user is expressed as

$$y_j = \sqrt{\rho_f}\boldsymbol{h}_j^T \boldsymbol{w}_j d_j + \sum_{\ell=1,\ell\neq j}^{k} \sqrt{\rho_f}\boldsymbol{h}_j^T \boldsymbol{w}_\ell d_l + n_j, \tag{5.3}$$

where $\boldsymbol{h}_j^T$ is the column vector of the j-th user's channel.

In this paper, we focus on a single cell scenario consisting of a single BBU and a single RRH to investigate the effect of fronthaul link limitation on a massive MIMO system. Interference between RRHs can be eliminated or mitigated through clustered coordination [46], and our results hold without loss of generality because clustered RRHs and the fronthaul links to them are seen as a single RRH and a single link to the BBU.

We use the following precoding matrix for MRT and ZF beamforming methods.

$$\begin{aligned}
&\text{MRT}: \quad \boldsymbol{W} = c_1 \hat{\boldsymbol{H}}^* = [\boldsymbol{w}_1\ \boldsymbol{w}_2 \cdots \boldsymbol{w}_j \cdots \boldsymbol{w}_k], \\
&\text{ZF}: \quad \boldsymbol{W} = c_2 \hat{\boldsymbol{H}}^* (\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^*)^{-1} = [\boldsymbol{w}_1\ \boldsymbol{w}_2 \cdots \boldsymbol{w}_j \cdots \boldsymbol{w}_k],
\end{aligned}$$

where c_1 and c_2 are constants for power normalization. In a single cell, the ergodic capacity lower bounds of MRT beamforming and ZF beamforming are given by [47]

$$C(s^B, m, k) = \begin{cases} k \log\left(1 + \frac{m}{k} \frac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)}\right), s^B = \text{MRT} \\ k \log\left(1 + \frac{m-k}{k} \frac{\rho_f \tau_r \rho_r}{\rho_f+1+\tau_r \rho_r}\right), s^B = \text{ZF}. \end{cases} \tag{5.4}$$

For simplicity, we assume identical user environments, i.e., the considered users experience the same average SNR and have the same precoding matrix update interval.

5.3.2 IQ-data Transfer Methods

In C-RAN, a BBU generates the transmit signal for each antenna of an RRH and transfers it to the RRH via a fronthaul link. The transmit signals are the mix of each stream's signal for each antenna, which is, determined by the beamforming technique chosen by the BBU, and the precoded data symbol with a weight vector.

For MIMO transmission, the simplest way for a BBU to inform an RRH of transmit signals is to transfer all the IQ-data samples. That is, the BBU, after precoding the data symbols, transfers $\boldsymbol{x}$ to the RRH in the form of IQ-data. For the rest of the paper, it will be referred to as "after-precoding". With the "after-precoding" IQ-data transfer method, the required bit-rates of the wired-link between the BBU and the RRH is

$$r_a = N_{\mathrm{Sub}} f_{\mathrm{Sym}} b_{\mathrm{IQ}} m, \tag{5.5}$$

where b_{IQ} is the number of bits representing one IQ-data sample.[2] Note that the required bit-rate for IQ-data transfer linearly increases with the number of transmitting antennas; it is independent of the

[2]It is specified as 8−20 bits in the CPRI specification [6].

number of users.

The "after-precoding" IQ-data transfer method carries the complete transmit signal vector. Another way to exchange IQ-data between the BBU and the RRH is "before-precoding." With the "before-precoding" IQ-data transfer method, the data symbol vector $\boldsymbol{d}$ and the precoding matrix $\boldsymbol{W}$ can be carried separately before the data symbols are ever precoded. In this case, the length of the data symbol vector equals the number of users, k, and then the precoding matrix size is $k \times m$. Thus, the bit-rates required in the wired-link is

$$r_b = N_{\mathrm{Sub}} f_w b_{\mathrm{IQ}} mk + N_{\mathrm{Sub}} f_{\mathrm{Sym}} b_{\mathrm{DS}} k. \tag{5.6}$$

The number of bits to represent demodulated signal, b_{DS}, depends on the possible number of modulation symbols. The data symbols are updated every symbol time as the transmit signals in "after-precoding", but the data symbols for each user are commonly used in generating the transmit signal of all the transmitting antennas. The precoding vectors for each steam are updated according to coherence time. The coherence time corresponding velocity is shown in Fig. 5.2. Table.5.1 shows the rate requirements in a fronthaul link to maximize wireless sum-rates for 20 MHz bandwidth and different precoding update periods. In a low speed mobile scenario, "before-precoding" is more efficient because the information for precoding is less frequently transmitted. Thus, by switching the IQ-data transfer

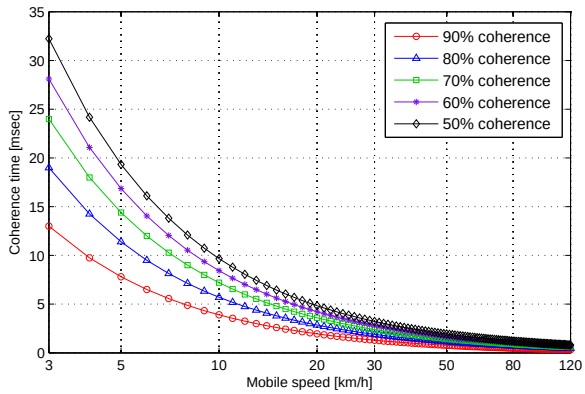


Figure 5.2: Coherence time in mobile environment.

mode adaptively, the wireless capacity can be increased and fronthaul link can be utilized more effectively. Even if only "after-precoding" is used, 78.64 Gbps is required for the 64-antenna system to have optimal performance,[3] which is about 800% of the highest link rate option (9.82 Gbps) in the current CPRI specification[6]. The "before-precoding" IQ-data transfer method requires an additional capability for an RRH to precode the unprecoded demodulated signals. This is not a heavy burden to the RRH because the calculation for precoding matrix (or vectors) calculation is still performed at the BBU. In this paper, we consider only the required bit-rate for IQ-data (user-plane data in CPRI specification) and the overhead for other related information to control the RRH (i.e., synchronization, control and management) is not considered.

[3]For optimal performance in a high mobility case, "before-precoding" seems too inefficient. For a given wired-link rate, however, "after-precoding" does not always outperform "before-precoding."

Table 5.1: Required fronthaul link rates for optimal sum-rates

$m = 64$, $T_w = 10msec$		
	after-precoding	before-precoding
MRT	78.64 Gbps ($k^\star = 64$)	47.31 Gbps ($k^\star = 64$)
ZF	78.64 Gbps ($k^\star = 36$)	26.62 Gbps ($k^\star = 36$)

$m = 64$, $T_w = 1msec$		
	after-precoding	before-precoding
MRT	78.64 Gbps ($k^\star = 64$)	349.31 Gbps ($k^\star = 64$)
ZF	78.64 Gbps ($k^\star = 36$)	196.49 Gbps ($k^\star = 36$)

5.4 Operation Algorithms for Cloud Massive MIMO System

5.4.1 The Number of Antennas with a Limited Fronthaul Capacity

It can be noticed, from (5.4), that the sum-rate C monotonically increases with the number of active antennas m. This means that if other variables are constant, using a larger m is always better. In this sense, with the "after-precoding" method where the IQ-data transfer bit-rate depends only on m but not on s^B and k, choosing a largest possible m always maximizes the wireless sum-rate, under a given fronthaul link capacity ϕ.

Thus, the optimal $m_a^\star$ in the "after-precoding" case can be obtained from (5.5):

$$m_a^\star = \min\{\lfloor \theta_1 \rfloor, M\}, \tag{5.7}$$

where $\theta_1 = \frac{\phi}{N_{\mathrm{Sub}} f_{\mathrm{Sym}} b_{\mathrm{IQ}}}$ and $\lfloor q \rfloor$ is the largest integer value which less

than or equal to q .

As in the "after-precoding" case, the largest possible m for the given k and ϕ leads to the maximum wireless sum-rate. In the "before-precoding" case, however, m and k have a tradeoff relation with the given fronthaul link capacity. Thus, the optimal $m_b^\star$ is obtained from (5.6):

$$m_b^\star = \min\left\{ \left\lfloor \frac{\theta_2}{k} - \theta_3 \right\rfloor, M \right\}, \tag{5.8}$$

where $\theta_2 = \frac{\phi}{N_{\mathrm{Sub}} f_w b_{\mathrm{IQ}}}$ and $\theta_3 = \frac{f_{\mathrm{Sym}} b_{\mathrm{DS}}}{f_w b_{\mathrm{IQ}}}$.

5.4.2 Beamforming Mode Selection

In this section, we propose a beamforming mode selection (MRT/ZF) algorithm for a large-scale antenna system with a limited fronthaul link capacity for each IQ-data transfer mode. We assume that the transmit power of the RRH and the number of users are fixed.[4]

Lemma 4.1: In "after-precoding" IQ-data transfer, if the number of total antennas is large so that $\theta_1 < M$ and the number of active antennas is chosen to maximize wireless capacity for the given fronthaul link capacity constraint, the sum rate of MRT is larger than that of ZF when $k > k^a_{\mathrm{cross}}$,

$$k^a_{\mathrm{cross}} = \frac{\rho_f \tau_r \rho_r}{(\rho_f + 1)(\tau_r \rho_r + 1)} \theta_1. \tag{5.9}$$

[4]Later, we will also investigate the case when the number of users is not fixed and controllable.

Proof:

$$
\begin{aligned}
& C(MRT, m, k) \geq C(ZF, m, k) \\
\Leftrightarrow & \frac{m}{k} \frac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)} \geq \frac{m-k}{k} \frac{\rho_f \tau_r \rho_r}{\rho_f+1+\tau_r \rho_r} \\
\Leftrightarrow & k \geq \left(1 - \frac{\rho_f+1+\tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)}\right) m \\
\Leftrightarrow & k \geq \frac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)} m.
\end{aligned}
$$

Because $\theta_1 < M$, $m^\star = \theta_1$. Thus, $k^a_{\text{cross}} = \frac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)} \theta_1$.

For the "before-precoding" case, the number of antennas is related with the wired capacity as well as the number of served users from (5.8).

Lemma 4.2: If the number of total antennas is large so that $\frac{\theta_2}{K} - \theta_3 < M$ and the number of active antennas is chosen to maximize wireless capacity for the given fronthaul link capacity constraint, the sum rate of MRT is larger than that of ZF when $k > k^b_{\text{cross}}$,

$$
\begin{aligned}
k^b_{\text{cross}} = & \frac{\rho_f \tau_r \rho_r}{2(\rho_f+1)(\tau_r \rho_r+1)} \\
& \cdot \left(-\theta_3 + \sqrt{\theta_3^2 + 4 \frac{(\rho_f+1)(\tau_r \rho_r+1)}{\rho_f \tau_r \rho_r} \theta_2}\right).
\end{aligned}
\tag{5.10}
$$

Proof:

$$
\begin{aligned}
& C(MRT, m, k) \geq C(ZF, m, k) \\
\Leftrightarrow\ & \tfrac{m}{k} \tfrac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)} \geq \tfrac{m-k}{k} \tfrac{\rho_f \tau_r \rho_r}{\rho_f+1+\tau_r \rho_r} \\
\Leftrightarrow\ & k \geq \left(1 - \tfrac{\rho_f+1+\tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)}\right) m \\
\Leftrightarrow\ & k \geq \tfrac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)} \left(\tfrac{\theta_2}{k} - \theta_3\right) \\
\Leftrightarrow\ & k^2 + \tfrac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)} \theta_3 k - \tfrac{\rho_f \tau_r \rho_r}{(\rho_f+1)(\tau_r \rho_r+1)} \theta_2 \geq 0
\end{aligned}
$$

and since M is large enough to satisfy $\frac{\theta_2}{K} - \theta_3 < M$ and $k \geq 0$, $k^b_{\text{cross}} = \frac{\rho_f \tau_r \rho_r}{2(\rho_f+1)(\tau_r \rho_r+1)} \left(-\theta_3 + \sqrt{\theta_3^2 + 4\frac{(\rho_f+1)(\tau_r \rho_r+1)}{\rho_f \tau_r \rho_r} \theta_2}\right)$.

5.4.3 Joint Optimization of Beamforming and IQ-data Transfer Mode

In this section, we address the joint beamforming and IQ-data transfer strategy in a cloud MIMO system. Our objective is to maximize the wireless sum-rate under the limitation of IQ-data transfer rates imposed by the capacity of the fronthaul link. The joint beamforming and IQ-data transfer strategy determines the optimal number of users, antennas, beamforming methods (ZF/MRT), and IQ-data transfer method.

From (5.5) and (5.6), we know that the numbers of active antennas and served users can be restricted by the fronthaul link rate and the IQ-data transfer method. Another constraint for the numbers of active antennas and served users is the degree of freedom in spatial domain (i.e., $k \leq m$). By considering these constraints, the sum-rate

maximization problem can be formulated as:

$$
\begin{aligned}
&\max_{\{s^D, s^B, k, m\}} C(s^B, m, k) \\
&\text{subject to} \\
&\quad 1 \le m \le M \\
&\quad 1 \le k \le K \\
&\quad k \le \tau_r \\
&\quad k \le m \\
&\quad r(s^D) \le \phi \\
&\quad r(s^D) = \begin{cases} N_{\mathrm{Sub}} f_{\mathrm{Sym}} b_{\mathrm{IQ}} m, & s^D = \text{“after”} \\ N_{\mathrm{Sub}} \left(f_w b_{\mathrm{IQ}} m + f_{\mathrm{Sym}} b_{\mathrm{DS}} \right) k, & s^D = \text{“before”} \end{cases}
\end{aligned}
\tag{5.11}
$$

The bit-rate of the fronthaul link is constrained by the fronthaul link capacity ϕ. Remark that even though $C(s^B, m, k)$ seems to be independent of the IQ-data transfer method, it is restricted by the IQ-data transfer method because the possible numbers of active antennas and users are determined by the IQ-data transfer method. To find the optimal s^D, s^B, and the integer variables k and m with exhaustive search, the computational complexity is $O(MK)$ (if $M > K$) or $O(M^2)$ (if $M \le K$).

The algorithm can be modified to reduce the complexity by replacing the number of antennas with (5.7) and (5.8) for “after-precoding” and “before-precoding”, respectively. Then, we can obtain the feasi-

bility conditions satisfying the constraints, $1 \leq m \leq M$, $1 \leq k \leq K$, $k \leq \tau_r$, and $k \leq m$. That is,[5]

$$k \leq \begin{cases} \min\{\lfloor \theta_1 \rfloor, M, K, \tau_r\}, & s^D = \text{“after”} \\ \min\left\{ \left\lfloor \frac{-\theta_3 + \sqrt{\theta_3^2 + 4\theta_2}}{2} \right\rfloor, M, K, \tau_r \right\}, & s^D = \text{“before”} \end{cases} \tag{5.12}$$

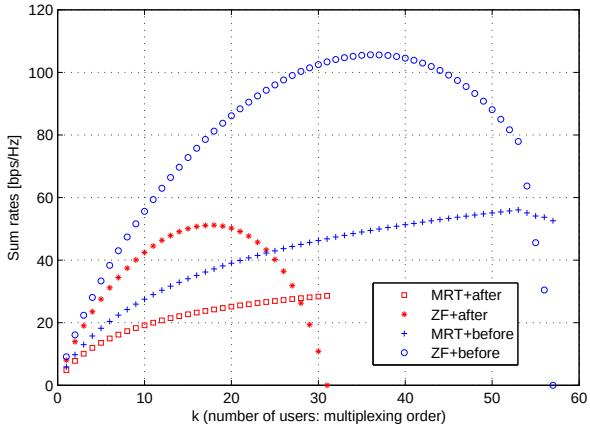
Now the problem can be modified to remove the number of antennas m by substituting (5.7), (5.8), and (5.12) into (5.11).

$$\begin{aligned} &\max_{\{s^D, s^B, k\}} C(s^B, k) \\ &\text{subject to} \\ &m = \begin{cases} \min\left\{\lfloor \theta_1 \rfloor, M\right\}, s^D = \text{“after”} \\ \min\left\{ \left\lfloor \frac{\theta_2}{k} - \theta_3 \right\rfloor, M \right\}, s^D = \text{“before”} \end{cases} \\ &k \leq \begin{cases} \min\left\{\lfloor \theta_1 \rfloor, M, K, \tau_r\right\}, s^D = \text{“after”} \\ \min\left\{ \left\lfloor \frac{-\theta_3 + \sqrt{\theta_3^2 + 4\theta_2}}{2} \right\rfloor, M, K, \tau_r \right\}, s^D = \text{“before”} \end{cases} \end{aligned} \tag{5.13}$$

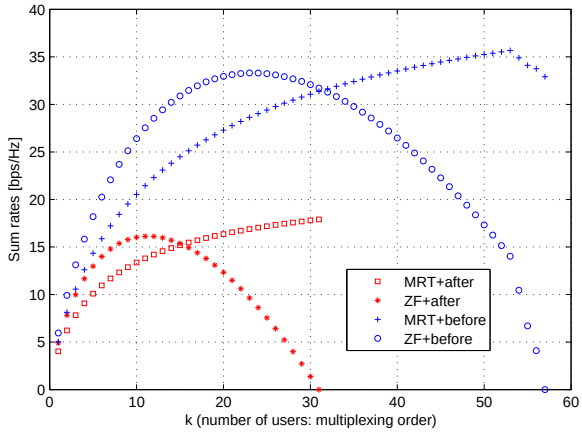
The remaining procedure is to find $k^\star = \arg \max_{k \leq k_{\max}} C(k)$ for each combination of s^D and s^B, where $k_{\max}$ denotes the maximum feasible k obtained from (5.12).

Figs. 5.3 and 5.4 show the wireless sum-rate given by (5.13) over the number of users k having different average SNRs and precoding matrix update intervals (T_w). Other parameters are summarized in

[5]It can be obtained using the relation with k and m.

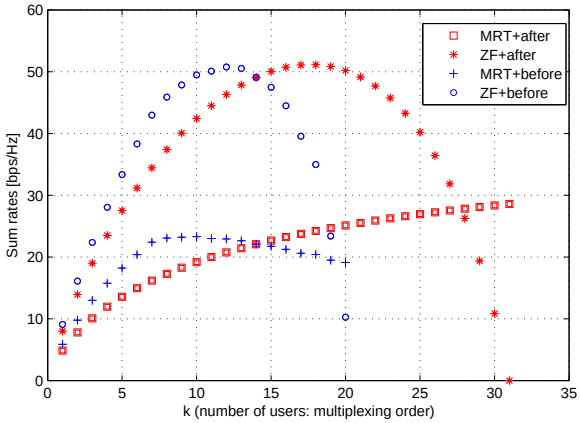


(a) SNR = 10 dB, Tw = 10 msec

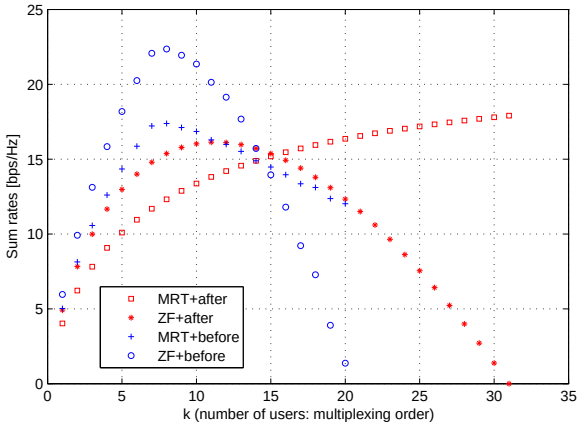


(b) SNR = 0 dB, Tw = 10 msec

Figure 5.3: Sum-rates with the number of users for different SNR values with precoding update interval of 10 ms (low mobility).



(a) SNR = 10 dB, Tw = 1 msec



(b) SNR = 0 dB, Tw = 1 msec

Figure 5.4: Sum-rates with the number of users for different SNR values with precoding update interval of 1 ms (high mobility).

Table 5.2. The sum-rate curve is plotted with feasible k. We can see that with the "after-precoding" method, the sum-rate curve is independent of the precoding matrix update interval and has the same tendency as was found in [46]. This is because the required IQ-data transfer rate is independent of the precoding matrix update interval. On the other hand, the sum rates of the "before-precoding" method in Figs. 5.3(a) and 5.3(b) are larger than those in Figs. 5.4(a) and 5.4(b). This is because as f_w decreases, $m^\star$ increases, and the sum-rate monotonically increases with the number of antennas used for transmission.

We see that with the "after-precoding" method, the maximum sum-rate of MRT is higher than that of ZF in the low SNR scenario, but, in the high SNR scenario, ZF has higher maximum sum-rates than MRT–a trend already shown in [46][47]. However, with the "before-precoding" method, ZF shows, in both high and low SNR scenarios, a higher maximum sum-rate than MRT. This is due to the flexibility of "before-precoding" in choosing m and k, whereas with the "after-precoding" method, m is fixed. Table 5.3 represents the values of $m^\star$, $k^\star$, $k_{\max}$ in Fig. 5.4(b). It is remarkable that with the "before-precoding" method, lower $k^\star$ and larger $m^\star$ are chosen to maximize the sum-rate.

Table 5.2: Parameters

N_{Sub}	2048
T_{Sym}	66.7 μsec
b_{IQ}	40 bits
b_{DS}	7 bits
ϕ	39.28 (4x9.82) Gbps
M	64
K	64

Table 5.3: Selected variables (SNR = 0 dB, $T_w = 1$ $msec$)

	$k_{\max}$	$k^\star$	$m^\star$
MRT, after	31	31	31
ZF, after	31	11	31
MRT, before	20	8	57
ZF, before	20	8	57

5.4.4 User Grouping and Resource Allocation

So far, we have assumed that users experience the same mobile environment. In practice, users may have different SNR values and weight vector update intervals according to their heterogeneous location and mobility. To apply the joint beamforming and IQ-data transfer strategy, users are grouped according to their SNR value and weight vector update interval. The expected throughput of each user is determined by these: the beamforming strategy, the number of users in the same group, and the allocated resource to the group.[6] If users in the same group share the resource fairly, the expected throughput of each user

[6]The number of groups and criteria affect the sum-rate performance. Detailed user grouping will be studied in future work.

in the group can be represented by

$$\overline{R}_i = v_i \frac{k_i^\star}{K_i} r_i(k_i^\star), \tag{5.14}$$

where v_i, K_i, $k_i^\star$, and $r_i(k_i^\star)$ denote the normalized resource fraction allocated to group-i, the total number of users in group-i, the number of users served simultaneously in group-i, and the achievable rate of a user in group-i. We analyze two main resource allocation policies: throughput fair and proportional fair resource allocations. In throughput fair resource allocation, the resource for each group is assigned such that the expected throughput of all users is equivalent. By taking

$$v_i = \frac{\frac{K_i}{k_i^\star r_i(k_i^\star)}}{\sum_{j=1}^{G} \frac{K_j}{k_j^\star r_j(k_j^\star)}} \tag{5.15}$$

users in different groups achieve the same expected throughput where G denotes the number of groups. Then, the average sum-rate of total users is

$$\overline{R}_{\text{tot}} = \frac{\sum_{j=1}^{G} K_j}{\sum_{j=1}^{G} \frac{K_j}{k_j^\star r_j(k_j^\star)}}. \tag{5.16}$$

In proportional fair resource allocation, the resource is allocated to each group proportional to the number of users in the group. That is,

$$v_i = \frac{K_i}{\sum_{j=1}^{G} K_j}. \tag{5.17}$$

Then the average sum-rate of total users is

$$\overline{R}_{\text{tot}} = \frac{\sum_{j=1}^{G} K_j k_j^{\star} r_j(k_j^{\star})}{\sum_{j=1}^{G} K_j}. \tag{5.18}$$

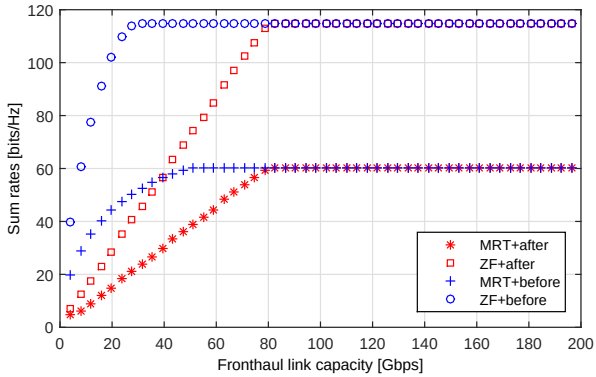
In a strict sense, proportional fair resource allocation can be dealt with in different ways. For example, we can reflect the access opportunities of users in the group considering the multiplexing order of the group instead of just considering the number of users in the group.

5.5 Numerical results and discussions

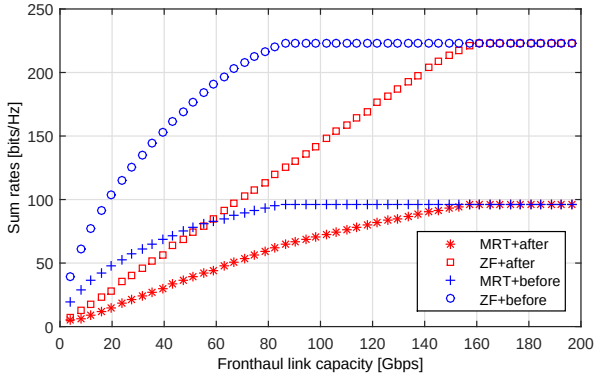
5.5.1 Fronthaul Link Capacity Effect

Figs. 5.5(a) and 5.5(b) show the maximized sum-rate over fronthaul link capacity for $M = 64$ and $M = 128$. The numbers of active antennas (m) and users (k) are chosen with the proposed joint beamforming and IQ-data transfer mode selection algorithm. In both figures, the wireless capacity increases as the fronthaul link capacity increases. Since the numbers of antennas and users are limited, however, the wireless capacity is saturated. These results provide a guideline in installation of the fronthaul link fiber cable for a given number of RRH antennas or deployment of RRH antennas for a given fronthaul link rate.

Figs. 5.6 and 5.7 illustrate the maximum sum-rates as the number of total antennas, M, goes to infinity in high and low SNR regime,

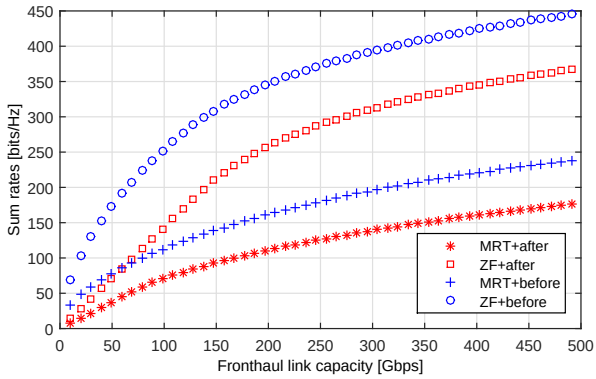


(a) $M = 64$, $K = 64$, SNR $= 10$ dB, $T_w = 10$ msec.

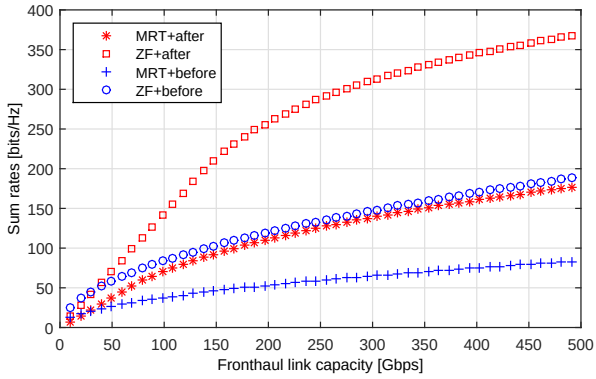


(b) $M = 128$, $K = 64$, SNR $= 10$ dB, $T_w = 10$ msec.

Figure 5.5: Effect of the limited fronthaul link capacity with limited numbers of antennas. Sum rates are saturated due to the fixed number of antennas.

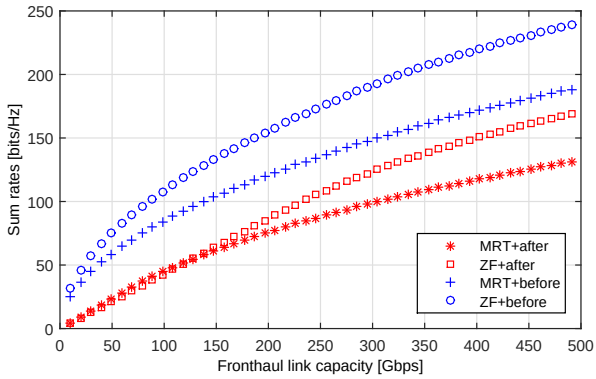


(a) $M = \infty$, $K = 64$, SNR $= 10$ dB, $T_w = 10$ msec.

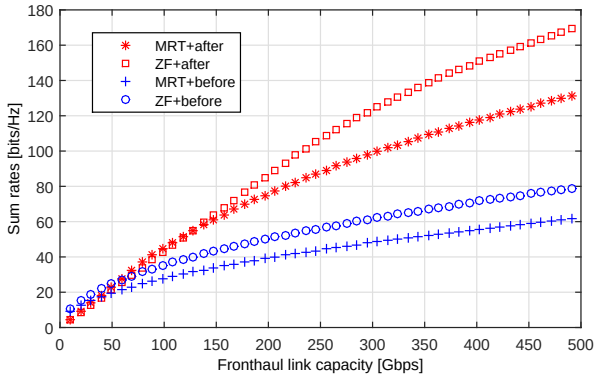


(b) $M = \infty$, $K = 64$, SNR $= 10$ dB, $T_w = 1$ msec.

Figure 5.6: Effect of the limited fronthaul link capacity with an infinite number of antennas with SNR $= 10$ dB.



(a) $M = \infty$, $K = 64$, SNR $= 0$ dB, $T_w = 10$ msec.



(b) $M = \infty$, $K = 64$, SNR $= 0$ dB, $T_w = 1$ msec.

Figure 5.7: Effect of the limited fronthaul link capacity with an infinite number of antennas with SNR $= 0$ dB.

Table 5.4: Desired beamforming technique for Massive-MIMO systems as $M \to \infty$

	"after-precoding"	"before-precoding"
Cell-center	ZF	ZF
Cell-boundary	MRT	MRT or ZF

respectively. Note that although M can be approximated to an infinite number, the number of total users (K) is limited in a practical scenario. Then, the number of antennas is chosen from (5.7) and (5.8), i.e, $m^\star = \min\{\lfloor\theta_1\rfloor, \infty\} = \lfloor\theta_1\rfloor$ for "after-precoding" and $m^\star = \min\left\{\left\lfloor\frac{\theta_2}{k^\star} - \theta_3\right\rfloor, \infty\right\} = \left\lfloor\frac{\theta_2}{k^\star} - \theta_3\right\rfloor$ for "before-precoding". As the fronthaul link capacity ϕ increases (also θ_2 increases), $m^\star$ increases whereas $k^\star$ is bounded to K. Because for a fixed number of users, the wireless capacities of ZF and MRT monotonically increase with the number of transmit antennas, the wireless capacity monotonically increases with the fronthaul link capacity when $M \to \infty$.

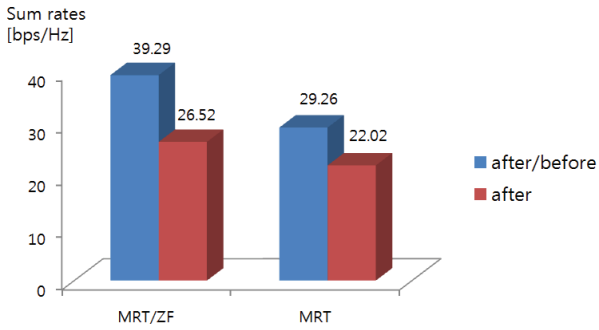
If the fronthaul link capacity is very large, then the maximum sum-rate is only determined by the beamforming strategy. The desired beamforming modes for a large number of antennas are summarized in Table 5.4.

5.5.2 Sum-rate with Resource Allocation Schemes

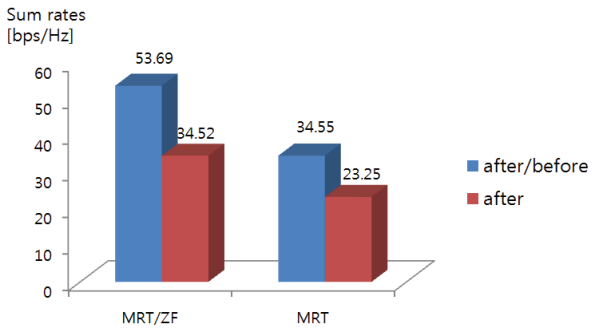
In this section, the performance of the proposed joint beamforming and IQ-data transfer method described in Section 5.4 with two types of resource allocation is verified through numerical simulations. The objective of simulations is to compare the adaptive "after/before

precoding" IQ-data transfer method and the conventional IQ-data transfer method using only "after-precoding." For fair comparison, when only "after-precoding" is available, other variables (s^B, m, and k) aside from s^D are decided in the same way as the adaptive "after/before precoding" case. In addition, in massive MIMO networks, ZF beamforming requires very high computational complexity as the precoding matrix size increases. For this consideration, we compare the case when only MRT is available as a beamforming technique. In simulations, four types of users, {Cell-center, high speed}, {Cell-center, low speed}, {Cell-boundary, high speed}, {Cell-boundary, low speed} and 50 users for each type are assumed. Users within each type are grouped and share the allocated resource of the group. The average SNR and the precoding vector update interval for each group are given in Table 5.5.

Fig. 5.8 shows the sum-rate of each resource allocation scheme. From the results, the adaptive "after/before precoding" to "after-precoding" gains 48.2% in throughput for throughput fair allocation and 55.5% for proportional fair allocation. The gain is due to the fact that some user groups have higher sum-rate with the "before-precoding" IQ-data transfer method. When only MRT is available for beamforming, the gain is somewhat reduced compared to when both ZF and MRT are available, though it is still larger than 34%.



(a) Throughput fair group resource allocation.



(b) Proportional fair group resource allocation.

Figure 5.8: Average sum-rates with resource allocation

Table 5.5: User type settings

	SNR	T_w
Cell-center, high speed	10 dB	10 msec
Cell-center, low speed	10 dB	1 msec
Cell-boundary, high speed	0 dB	10 msec
Cell-boundary, low speed	0 dB	1 msec

Table 5.6: Desired IQ-data transfer method for Massive-MIMO systems

Short coherence time	Long coherence time
"after-precoding"	"before-precoding"

5.5.3 Discussions

For the required bit-rate in the fronthaul link, we consider only IQ-data for transmit signal at the antennas without concerning about the overhead in channel state information (CSI) gathering. This is because if the downlink channel is estimated at mobiles and the CSI is fed back to the RRH as in current systems, the downlink CSI at the RRH will be carried as IQ-data to the BBU. In an uplink scenario, however, the required bit-rate will depend on uplink CSI gathering, such as the amount of pilot signals to measure the uplink channel, whether uplink channel is estimated at the RRH or the BBU, and how to represent the CSI estimated at the RRH and carry to the BBU. In our work, the number of antennas are selected for ergodic capacity maximization. Selecting each antennas (binary) and performing IQ-data transfer will further improve the capacity for given fronthaul link rate constraints. We consider two types of IQ-data transfer methods, but there can be other ways to deliver IQ-data efficiently such as IQ-data compression. Therefore, comprehensive work, which capture the CSI overhead in fronthaul links and more generalized IQ-data transfer, are left for future work.

5.6 Summary

In this paper, we investigated the effect of a limited number of antennas or users due to a fronthaul link constraint on the wireless

sum-rate, which also depends on the beamforming strategy and the IQ-data transfer mode. We proposed beamforming mode selection algorithms for each IQ-data transfer mode under an assumption of a large number of RRH antennas and a limited capacity of the fronthaul link. To further maximize the wireless capacity, we considered massive MIMO systems that support the adaptive IQ-data transfer mode in the fronthaul link. For the given fronthaul link capacity and user environments, the beamforming and the IQ-data transfer methods are jointly decided to maximize the sum-rate. Through analytical and numerical results, we confirmed that the wireless capacity of massive MIMO systems with a limited capacity of the wired link could be highly improved by the proposed adaptive IQ-data transfer method.

Chapter 6

Conclusion

6.1 Research Contributions

In this dissertation, we addressed a cloud radio access network for large-scale antenna operation. To tackle heavy fronthaul traffic of C-RAN with many RRH antennas, we proposed a novel partially-centralized C-RAN designed for flexible and fronthaul-efficient operation in large-scale MIMO.

Firstly, we provided an overview of the challenging issues of C-RAN for massive MIMO and proposed a partially-centralized C-RAN (PC-RAN) architecture for fronthaul efficiency and flexible processing loads distribution. With the proposed C-RAN, fronthaul traffic is significantly reduced even for the same wireless performance of fully-centralized C-RAN (FC-RAN) and further reducible without significant performance degradation.

Secondly, we evaluated the performance of large-scale antenna C-RANs taking account of cooperative MIMO processing scenario among RRHs. We exploited stochastic geometric approach to obtain distribution of SIR (signal-to-interference ratio) and related fronthaul traffic volume according to system parameters. Through extensive simulations, we confirmed the accuracy of our analytical model and provided intuition on the performance of large-scale antenna C-RANs considering tradeoff between wireless performance and fronthaul traffic volume.

Thirdly, we investigated joint beamforming and resource allocation problem of a single RRH for a constrained fronthaul capacity. We consider two representative MIMO techniques, ZF and MRT, and provide a heuristic algorithm to decide optimal beamforming configuration and bandwidth allocation for each beamforming technique. The simulation shows that the proposed joint beamforming and resource allocation decision scheme further improve the wireless sum-rates in our proposed partially-centralized C-RANs.

Lastly, we addressed the joint fronthauling method and beamforming configuration decision problem for a given fronthaul capacity assuming both IQ sample transport and separate transport of precoder and data symbol are possible. The numerical results show the performance gain of the adaptive fronthauling and beamforming configuration scheme.

6.2 Future Research Directions

Based on the results of this dissertation, there are several new research directions to be further investigated.

While we mathematically analyzed the performance of large-scale zero-forcing precoding for the cases of noncooperation and cooperation, future work is needed to analyze considering both ZF and MRT precoding schemes. It might be extension and combination of the second and third research items of this dissertation.

Another research direction is to extend the joint beamforming and resource allocation problem to joint beamforming and user selection problem considering users' different channel correlation time. By properly selecting users for each precoder and allocating adaptively pilot resources to them, system capacity can be further improved with efficiently allocated pilot resources and thus reduced fronthaul traffic for CSI report.

Bibliography

[1] Metis, "Scenarios, requirements and KPIs for 5G mobile and wireless system," May 2013.

[2] China Mobile Research Institute, "C-RAN: The road towards green RAN," *White Paper*, http://labs.chinamobile.com/article_download.php?id=104035.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Comm.*, vol. 9, pp. 3590–3600, Nov. 2010.

[4] B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, F. Rusek, D. Persson, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with large arrays," *IEEE Sig. Proc. Mag.*, vol. 30, pp. 40–60, Jan. 2013.

[5] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Comm. Mag.*, vol. 22, pp. 152–160, Apr. 2015.

[6] "CPRI specification v5.0,"

[7] "Open base station architecture initiative (OBSAI),"

[8] S. Park, C.-B. Chae, and S. Bahk, "Large-scale antenna operation in heterogeneous cloud radio access networks: A partial centralization approach," *IEEE Wireless Comm. Mag.*, vol. 22, Jun. 2015.

[9] "Visual networking index: Global mobile data traffic forecast update, 2012-2017," *Cisco, Tech. Rep.*, Feb. 2013.

[10] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5g heterogeneous cloud radio access networks," *IEEE Wireless Comm. Mag.*, vol. 29, no. 2, pp. 152–160, 2015.

[11] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Comm. Mag.*, vol. 21, pp. 126–135, Dec. 2014.

[12] F. Ponzini, L. Giorgi, A. Bianchi, and R. Sabella, "Centralized radio access networks over wavelength-division multiplexing: A plug-and-play implementation," *IEEE Comm. Mag.*, vol. 51, pp. 94–99, Sep. 2013.

[13] G. Wang, Q. Liu, R. He, F. Gao, and C. Tellambura, "Acquisition of channel state information in heterogeneous cloud radio access

networks: challenges and research directions," *IEEE Wireless Comm. Mag.*, vol. 22, Jun. 2015.

[14] K. Sundaresan, M. Y. Arslan, and S. Singh, "Fluidnet: A flexible cloud-based radio access network for small cells," in *Proc. ACM MobiCom*, pp. 99–110, 2013.

[15] M. Peng, S. Yan, and H. V. Poor, "Ergodic capacity analysis of remote radio head associations in cloud radio access networks," *IEEE Wireless Communications Letters*, vol. 3, pp. 365–368, Aug. 2014.

[16] C. Liu, L. Zhang, M. Zhu, J. Wang, L. Cheng, and G.-K. Chang, "A novel multi-service small-cell cloud radio access network for mobile backhaul and computing based on radio-over-fiber technologies," *J. Lightw. Techno*, vol. 31, pp. 2869–2875, Nov. 2013.

[17] C. Liu, J. Wang, L. Cheng, M. Zhu, and G.-K. Chang, "Key microwave-photonics technologies for next-generation cloud-based radio access networks," *J. Lightw. Techno*, vol. 32, pp. 3452–3460, Oct. 2014.

[18] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Comm. Surveys and Tutorials*, pp. 1–24, Sep. 2014.

[19] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Sig. Proc. Mag.*, vol. 31, pp. 69–79, Nov. 2014.

[20] A. Liu and V. Lau, "Joint power and antenna selection optimization in large cloud radio access networks," *IEEE Trans. Sig. Proc.*, vol. 62, pp. 1319–1328, Mar. 2014.

[21] A. Liu and V. Lau, "Phase only RF precoding for massive MIMO systems with limited RF chains," *IEEE Trans. Sig. Proc.*, vol. 62, pp. 4505–4515, Sep. 2014.

[22] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Communications Letters*, vol. 3, pp. 653–656, Dec. 2014.

[23] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," vol. 31, pp. 264–273, Feb. 2013.

[24] C. Shepard, H. Yu, N. Anand, L. E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: practical many-antenna base stations," in *Proc. of ACM MobiCom*, pp. 53–64, 2012.

[25] C.-S. Park, Y.-S. Byun, A. M. Bokiye, and Y.-H. Lee, "Complexity reduced zero-forcing beamforming in massive MIMO sys-

tems," in *Information Theory and Applications Workshop (ITA)*, Feb. 2014.

[26] S. Park, C.-B. Chae, and S. Bahk, "Before/after precoded massive mimo in cloud radio access networks," in *Proc. IEEE ICC*, pp. 169–173, Jun. 2013.

[27] S. Park, C.-B. Chae, and S. Bahk, "Before/after precoding massive MIMO systems for cloud radio access networks," *IEEE/KICS Jour. Comm. Networks*, vol. 15, pp. 398–406, Aug. 2013.

[28] D. Lee, H. 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 Comm. Mag.*, vol. 50, pp. 148–155, Feb. 2012.

[29] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. W. H. Jr., "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Comm.*, vol. 8, pp. 1910–1921, Apr. 2009.

[30] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell mimo cooperative networks: A new look at interference," vol. 14, pp. 1380–1408, Dec. 2010.

[31] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," vol. 59, p. 31223134, Nov. 2011.

[32] M. Kountouris and J. G. Andrews, "Downlink sdma with limited feedback in interference-limited wireless networks," *IEEE Trans. Wireless Comm.*, vol. 11, pp. 2730–2741, Aug. 2012.

[33] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," vol. 30, pp. 550–560, Apr. 2012.

[34] H. S. Dhillon, M. Kountouris, and J. G. Andrews, "Downlink mimo hetnets: Modeling, ordering results and performance analysis," *IEEE Trans. Wireless Comm.*, vol. 12, pp. 5208–5222, Oct. 2013.

[35] H. Yang and T. L. Marzetta, "Modeling heterogeneous network interference using poisson point processes," *IEEE Trans. Sig. Proc.*, vol. 61, pp. 4114–4126, Aug. 2013.

[36] F. Baccelli and A. Giovanidis, "A stochastic geometry framework for analyzing pairwise-cooperative cellular networks," *IEEE Trans. Wireless Comm.*, vol. 28, pp. 794–808, Feb. 2015.

[37] C.-B. Chae, I. Hwang, R. W. H. Jr., and V. Tarokh, "Interference aware-coordinated beamforming in a multi-cell system," *IEEE Trans. Wireless Comm.*, vol. 11, pp. 3692–3703, Oct. 2012.

[38] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," in *Proc. IEEE ICC*, pp. 4033–4037, Jun. 2008.

[39] S. A. Ramprashad, G. Caire, and H. C. Papadopoulos, "A joint scheduling and cell clustering scheme for mu-mimo downlink with limited coordination," in *Proc. IEEE ICC*, pp. 1–6, Jun. 2010.

[40] J.-M. Moon and D.-H. Cho, "Inter-cluster interference management based on cell-clustering in network mimo systems," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, pp. 1–6, May. 2011.

[41] E. Pineda and V. Garrodo, "Domain size distribution in a poisson voronoi nucleation and growth transformation," *Physical Rev. E.*, vol. 75, p. Available: http://pre.aps.org/abstract/PRE/v75/i4/e040107., Sep. 2007.

[42] D. Cao, S. Zhou, and Z. Niu, "Optimal combination of base station densities for energy-efficient two-tier heterogeneous cellular networks," *IEEE Trans. Wireless Comm.*, vol. 12, pp. 4350–4362, Sep. 2013.

[43] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications, 2nd Edition.* 1996.

[44] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Comm. Mag.*, vol. 52, pp. 122–130, Feb. 2014.

[45] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Comm. Mag.*, vol. 52, pp. 186–195, Feb. 2014.

[46] C. Lee, C.-B. Chae, T. Kim, S. Choi, and J. Lee, "Network massive MIMO for cell-boundary users: From a precoding normalization perspective," in *Proc. IEEE Glob. Telecom. Conf.*, pp. 1–5, 2012.

[47] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE Jour. Select. Areas in Comm.*, vol. 31, pp. 172–179, Feb. 2013.

[48] R. Rogalin, O. Bursalioglu, H. Papadopoulos, G. Caire, A. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," vol. 13, pp. 1815–1831, Apr. 2014.

[49] Q. Yang, X. Li, H. Yao, J. Fang, K. Tan, W. Hu, J. Zhang, and Y. Zhang, "Bigstation: enabling scalable real-time signal processing in large mu-MIMO systems," in *Proc. of ACM SIGCOMM*, pp. 399–410, 2013.

[50] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "CloudIQ: A framework for processing base stations in a data center," in *Proc. of ACM MobiCom*, pp. 125–136, 2012.

[51] Y.-G. Lim, C.-B. Chae, and G. Caire, "Performance analysis of massive mimo for cell-boundary users," *IEEE Trans. Wireless Comm.*, to appear.

[52] P. Pardalos and M. Resende, *Handbook of applied optimization*. Oxford University Press, 2002.

[53] E. Telatar, "Capacity of multi-antenna gaussian channels," *Europ. Trans. on Telecomm*, vol. 10, no. 6, pp. 585–595, 1999.

[54] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," vol. 6, pp. 311–335, 1998.

[55] D. Gesbert, M. Kountouris, R. W. Heath, Jr., C.-B. Chae, and T. Salzer, "Shifting the MIMO paradigm: From single user to multiuser communications," *IEEE Sig. Proc. Mag.*, vol. 24, pp. 36–46, Oct. 2007.

[56] Q. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Sig. Proc.*, vol. 52, pp. 462–471, Feb. 2004.

[57] C.-B. Chae, D. Mazzarese, N. Jindal, and R. W. Heath, Jr., "Coordinated beamforming with limited feedback in the MIMO broadcast channel," *IEEE Jour. Select. Areas in Comm.*, vol. 26, pp. 1505–1515, Oct. 2008.

[58] T. L. Marzetta, "How much training is required for multiuser MIMO?," in *Proc. of Asilomar Conf. on Sign., Syst. and Computers*, pp. 359–363, 2006.

[59] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Comm.*, vol. 61, pp. 1436–1449, May. 2013.

[60] C.-B. Chae and J. R. W. Heath, "On the optimality of linear multiuser MIMO beamforming for a two-user two-input multiple-output broadcast system," *IEEE Sig. Proc. Lett.*, vol. 16, pp. 117–120, Feb. 2009.

[61] T. L. Marzetta, J. Jose, A. Ashikhmin, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Comm.*, vol. 10, pp. 2640–2651, Aug. 2011.

초 록

거대 다중안테나 기법은 끊임없이 증가하는 셀 용량 수요와 사용자 QoS를 만족시키기 위한 차세대 무선 네트워크의 유망 기술로 주목받고 있다. 대량의 안테나는 주파수 효율성을 더욱 향상시키거나 보다 많은 사용자를 다중화하여 네트워크 용량을 증대시키는 효과가 있다. 하지만, 원격무선장비에서 대량의 안테나를 사용하는 것은 원격무선장비와 중앙의 서버 간 프론트홀 링크의 트래픽 또한 대폭 증가시킨다. 본 논문에서는, 거대 다중안테나를 이용하여 무선 성능을 향상시키면서도 프론트홀 자원을 효율적으로 사용하기 위한 클라우드 기지국 구조와 운용 기법을 다룬다.

첫째, 클라우드 기지국에 거대배열 다중안테나를 적용함에 있어 당면과제를 살펴보고 거대배열 안테나 운용에 적합한 부분적으로 집중화된 클라우드 기지국 구조를 제안한다. 제안한 클라우드 기지국 구조에서는 프론트홀을 통해 안테나 별 IQ 샘플이 아닌 사용자 별 데이터 심볼, 프리코더, 채널상태정보를 분리하여 전송하고, 기존의 클라우드 기지국 구조의 무선 성능을 유지하면서도 프론트홀 트래픽을 크게 줄일 수 있다.

둘째, ZF 기반의 거대배열 다중안테나 시스템의 성능을 수학적으로 분석한다. 클라우드 기지국 시스템에서 RRH 간 다중안테나 협력을 고려하여 무선성능과 프론트홀 트래픽을 분석한다. 시뮬레이션을 통해 정확도를 검증하며, 프론트홀 트래픽과 무선성능 간에 트레이드오프에 대한 직관을 제시한다.

셋째, 프론트홀 용량이 제한되어있는 클라우드 기지국에서의 조인트 빔포밍 및 자원할당 기법을 문제를 다룬다. 빔포밍 기술로는 ZF과 MRT를 고려하며, 각 빔포밍의 전송 설정과 할당되는 무선 대역폭을 결정하는 휴리스틱 알고리즘을 제안한다. 시뮬레이션을 통해

제안하는 기법이 제안하는 클라우드 기지국의 성능을 더욱 향상시키며 최적에 가까운 무선 전송률 합을 가지는 것을 확인한다.

넷째, 프론트홀 용량이 제한된 클라우드 기지국에서 두 가지 프론트홀링 기법(IQ 샘플 데이터를 전송하는 방식, 프리코더와 데이터 심볼을 나누어 전송하는 방식)에 따른 ZF과 MRT의 성능을 살펴본다. 프론트홀 용량이 주어졌을 때, 무선 전송률 합을 최대화 시키기 위한 최적의 프론트홀링 기법과 빔포밍 전략 결정 알고리즘을 제안한다.

주요어 : 클라우드 기지국, 클라우드 무선 접속망, 프론트홀, 거대 다중안테나, 다중안테나

학번 : 2009-20803