# Load Balancing for Multiplexing Gains of BBU Pool in 5G Cloud Radio Access Networks

Ramaraju Chaganti

A Thesis Submitted to
Indian Institute of Technology Hyderabad
In Partial Fulfillment of the Requirements for
The Degree of Master of Technology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Computer Science and Engineering

June 2015

## Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

(Signature)

(Ramaraju Chaganti)

CS12M1002

(Roll No.)

# Approval Sheet

This Thesis entitled Load Balancing for Multiplexing Gains of BBU Pool in 5G Cloud Radio Access Networks by Ramaraju Chaganti is approved for the degree of Master of Technology from IIT Hyderabad

(Dr. Kiran Kuchi) Examiner
Dept. of Electrical Engineering
IITH

(Dr. Antony Franklin) Examiner
Dept. of Computer Science and Engineering
IITH

(Dr. Bheemarjuna Reddy Tamma) Adviser
Dept. of Computer Science and Engineering
IITH

(Dr. Kotaro Kataoka) Chairman
Dept. of Computer Science and Engineering
IITH

# Acknowledgements

I would like to express my gratitude to my adviser Dr. Bheemarjuna Reddy Tamma for the useful comments, remarks and engagement through the learning process of this master thesis. Furthermore, I would like to thank Mr. Anil Kumar Rangisetti for technical discussions and insightful comments of this master thesis. I take this opportunity to express my gratitude to all the members who directly or indirectly helped in carrying out my research. I also thank my parents for the unceasing encouragement, support and attention.

# Abstract

Cloud Radio Access Network (C-RAN) is an architecture for 5G cellular networks to improve coverage, increase data rates, enhancing signaling efficiency etc. In C-RAN architecture of 5G cellular networks, multiple Base Station (BS) Base Band processing Units (BBU) are centralized in the cloud. Remote Radio Heads (RRHs) that reside at cell sites will have only antennas and other radio frequency functions. The central cloud based system will provide higher layer protocols of LTE BS that process on a pool of BBUs on top of a pool of computing resources i.e., General Purpose Processors (GPPs). The centralized BBU pool and RRHs are connected with high speed optical fiber links. Each BBU maps to a GPP that has a specified processing capacity and processes Inphase Quadrature (IQ) samples received from Remote Radio Heads (RRHs) deployed at cell sites. A single BBU can serve multiple RRHs based on the limits imposed on processing capacity of GPP. C-RAN helps telecom service providers in cutting down their CAPEX and OPEX by reducing power consumption of BBUs.

In this work, we propose a new architecture for C-RAN where LTE BS composed of three parts and they are higher layer protocol stacks, BBU pool and RRHs. The higher layer protocol stack process on top of BBU and RRH can receive/send IQ samples from/to BBU. For realizing C-RAN architecture, we implemented a framework to exploit the potential of C-RAN in terms of computing resource utilization and power consumption. For estimating computing resource utilization, we profile protocol stack of a real-time LTE BS. Using the profiling statistics, we determine the maximum processing capacity of LTE BS in terms of Floating Point Operations per Second (FLOPS). For estimating power consumption, we used the power consumption statistics of the real-time LTE BS. Based on the analysis of computing resource utilization and power consumption statistics, we propose BS processing and BS power models. The BS processing model is created with parameters that influence the computing resource utilization and the BS power model is created with parameters that influence the power consumption of LTE BS. These models can be used in any LTE system simulator/emulator. We incorporated the proposed BS processing and BS power models in a packet level LTE network simulator (ns-3). For realistic simulation, we have created Radio over Fiber (RoF) model that allows us to simulate the latencies of C-RAN in NS-3.

From BS processing statistics, we can determine that most of processing in LTE BS happens in BBUs. So, using maximum processing capacity of LTE BS, we can determine the maximum processing capacity of BBU. Based on maximum processing capacity of BBU, we can form a cluster of RRHs that can be served by the BBU. Each RRH corresponds with a higher layer protocol processing on BBU. When BBU is serving cluster of RRHs, corresponding higher layer protocol stacks are multiplexed on BBU. The processing requirements of RRHs (i.e., traffic load) are subjected to spatio-temporal variations and to deal with these variations, we need to dynamically cluster the RRHs in such a way that their processing must be done with minimum BBUs (i.e., load should be balanced with minimum BBUs). In this work, we solve the dynamic clustering of RRHs problem as a linear optimization problem with the objective of minimizing the number of clusters. By reducing the number of clusters, C-RAN helps to reduce computing cost and saves power consumed at the cloud. We evaluated the proposed dynamic clustering of RRHs on the C-RAN framework by considering two scenarios for spatio-temporal variations. In one scenario, the processing demands are high during working hours i.e., office areas and in other scenario, the processing demands have very low variations i.e., residential areas. For comparison of the proposed approach, we used the traditional LTE BS

deployment. With our proposed approach, we have achieved 27% and 36% savings of computing resources in office and residential areas, respectively. In case of power consumption, we reduced 30% in office areas and 34% in residential areas.

# Contents

# Chapter 1

# Introduction

## 1.1 5G Evolution

The first generation (1G) mobile communication was based on analog signals. In the next generation, i.e., 2G, digital signals are used to offer voice services. Data services started with General Packet Radio Service (GPRS) on top of Global System for Mobile Communications (GSM), which is often referred to as 2.5G. With the increase of smart phones, demand for higher data rates has increased, which pushed telecom service providers to opt for innovative technologies. Thus third generation (3G) and fourth generation (4G) mobile communications came into existence with a goal of improving data rates. Though 3G and 4G cellular technologies are able to meet the data rate requirements posed by smart phones, but to achieve this more Base Stations (BSs) have to be deployed which in turn increases Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) of service providers. In a typical Base Station (BS) deployment, major cost is due to the specialized hardware for Base Band Processing Unit (BBU) at the cell site. Hence, we require novel network architectures for reducing CAPEX and OPEX of telecom service providers without effecting their service quality.

### 1.1.1 Challenges of Radio Access Networks

#### Increasing CAPEX and OPEX

Radio Access Network (RAN) is the most important asset of telecom service providers for offering high data rates, high quality, and $24 \times 7$ services to mobile users. In recent years, with high growth of mobile data consumption, service providers are increasing their network capacity for providing better services. As shown in Fig 1.1, cost to build (i.e., CAPEX), operate (i.e., OPEX) and upgrade the RAN is becoming more and more expensive while the revenue is not growing at the same rate. In competitive marketplace, rapid changes in technologies and decline in voice revenue pose challenges to service providers for deploying traditional BSs as the cost is high and the return Average Revenue Per User (ARPU) is not high enough. As shown in Fig. 1.2, mobile Internet traffic is rapidly increasing but ARPU is decreasing.

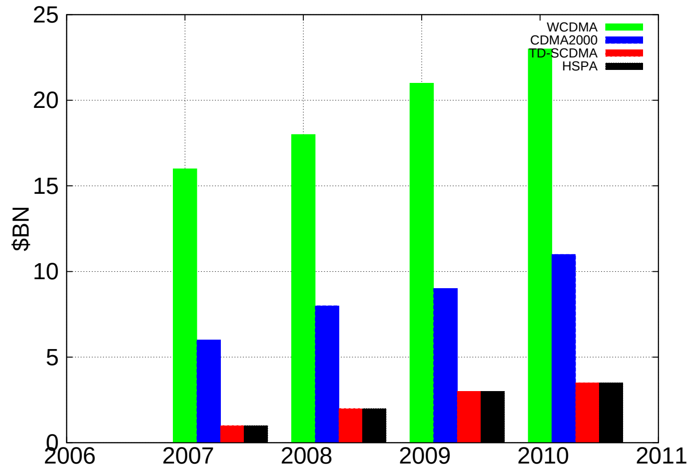


Figure 1.1: Increasing CAPEX of 3G Network Construction and Evolution [1]

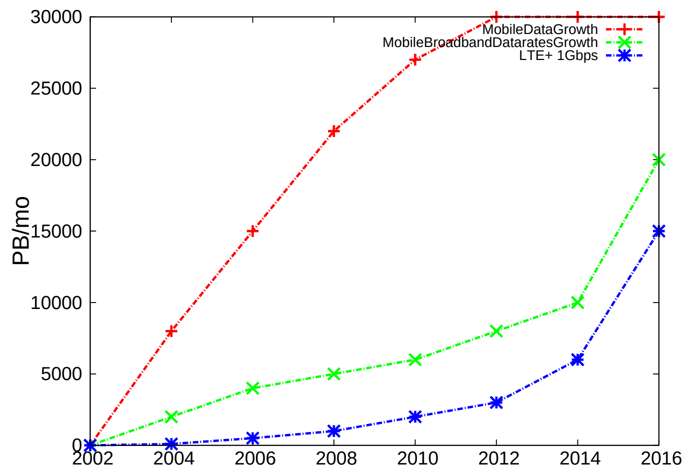


Figure 1.2: Mobile Broadband Data-rates/Traffic Growth [1]

### Dynamic mobile network load and low BS utilization rate

The mobile subscribers are frequently moving from one place to another. With data from real operation network [1], we can notice that the subscribers movement has strong relation with spatio-temporal patterns. If we notice, mobile subscribers move from home to their offices regularly at beginning of working hours period. When the working hours period ends, mobile subscribers will move from offices to their homes. Consequently, traffic load on network varies in the mobile network with patterns called "tidal effect". As shown in Fig. 1.3, during working hours period (i.e., busy hours) the load on BSs in the office areas is high, after working hours period (i.e., non-busy hours) the load on BSs in residential areas is high.

Generally, most of BS processing is utilized by active subscribers in its cell range, with variation of active subscribers in different areas causes BS processing utilization low in some areas and high in some areas. With mobility of active subscribers, the serving BS processing stays idle or goes low with lack of active subscribers. As operators must provide $24 \times 7$ coverage, the idle BSs consume almost same power as how they consume in busy hours. Often BSs are designed with higher processing

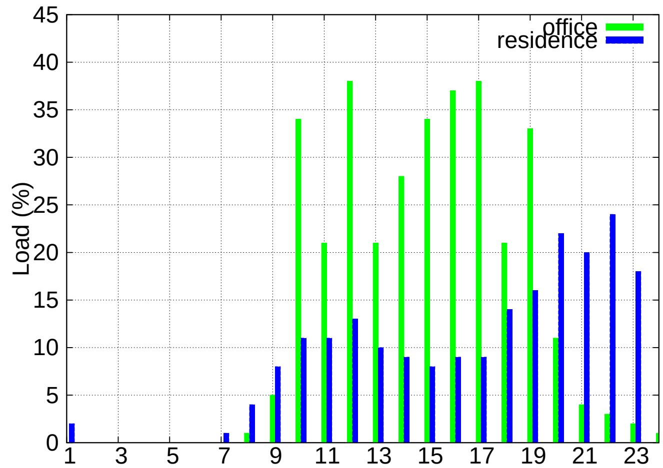


Figure 1.3: Mobile Network Load in Daytime [1]

capacity to handle maximum number of active subscribers, thus low processing utilization of BSs makes it worse, which means most of the processing power is wasted in non-busy hours. The possible solution can be sharing the processing capacity of BS and the power between different BS areas to utilize their processing resources more effectively.

In order to achieve the goal of providing higher data rate with high coverage and reduced CAPEX and OPEX, a novel architecture called Cloud Radio Access Network (C-RAN) has been proposed by China Mobile Research Institute [1].

### 1.1.2 C-RAN Architecture

To maintain profitability and growth, service providers are finding solutions to reduce cost as well as to provide better services to mobile subscribers. In order to reduce CAPEX and OPEX and maintain profitability and growth, operators are developing evolving network architectures. Advancements in broadband Internet and IT infrastructure provide an unique opportunity in developing new network architectures. Unlike in traditional RAN, where all reception and transmission functionalities reside as a single unit, a centralized approach has been proposed. In centralized approach, RAN functionalities are centralized for processing of all signals and Remote Radio Heads (RRHs) are spatially distributed for transmission. Centralized RAN provides [1]

- Reduced cost (CAPEX and OPEX)
- Low energy consumption
- High spectral efficiency
- Support of multiple standards
- Support for evolution
- Platform support for additional revenue generation

Based on centralizing the functionalities of RAN, we can classify the architectures into two types.

- Fully Centralized: Along with Base Band Unit (BBU) processing (known as L1), all higher layer (known as L2 and L3) functions of BS will be co-located in central unit.

- Partially Centralized: RRH has radio frequency functions along with some BBU functions where as the central unit has remaining BBU and higher layer functions co-located with it.

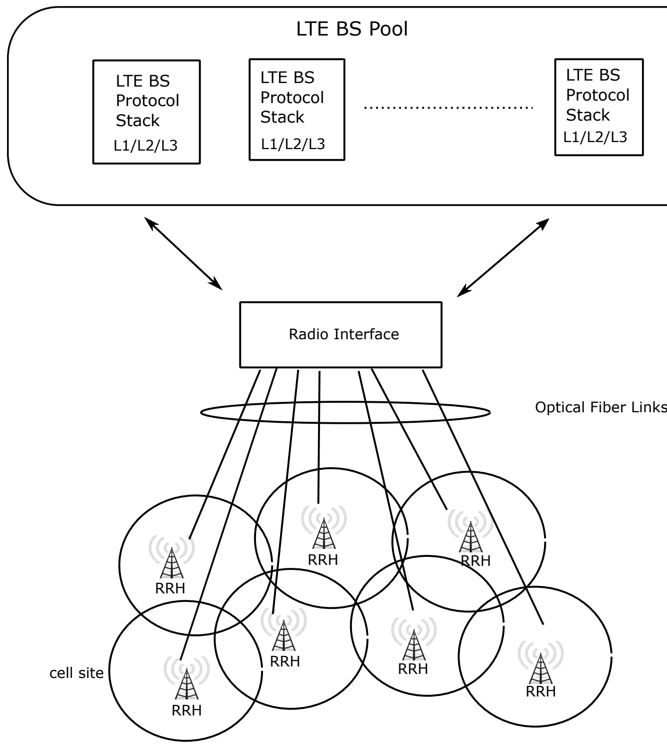


Figure 1.4: Fully Centralized C-RAN Architecture

A Fully centralized architecture C-RAN architecture has been shown in Fig 1.4 and partially centralized architecture is shown in Fig. 1.5. Based on these function splitting methods of C-RAN architectures, each of them have three entities [1].

- RRHs will have radio frequency functions and antennas.

- RRHs and BBUs are connected with high bandwidth low-latency optical fiber link that carries the cellular signals known as In-phase Quadrature (IQ) samples to centralized location [2].

- BBU which resides in the centralized location have base band signal processing functionality and other higher layers of LTE BS protocol stack.

With one of these C-RAN architectures, service providers can easily deploy and quickly make upgrades to their network. The service provider only needs to install new RRHs at cell site and connect them to the centralized location of the BBU pool for increasing the network coverage. If the network load grows, the service provider only needs to upgrade the BBU pool processing for accommodating the increased load processing. C-RAN breaks up the static relationship between RRHs and BBUs that exists in traditional distributed BS architecture. Each physical BBU does not serve to any specific RRH.

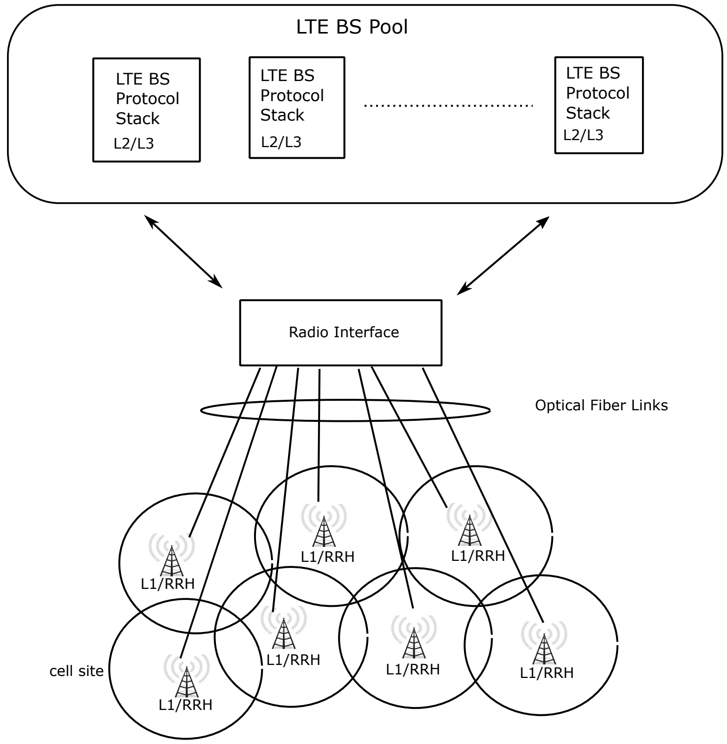


Figure 1.5: Partially Centralized C-RAN Architecture

Both full and partially centralized architectures described above are under development and evaluation. The further discussion will focus on the Fully Centralized architecture.

### 1.1.3 Advantages of C-RAN

- Energy Efficient and Green Infrastructure: With centralized architecture, we can reduce air conditioning and other site supporting equipment's power consumption. Along with these we can reduce power consumption for signal transmission by deploying more number of RRHs.

- Cost-saving on CAPEX and OPEX: With centralized BBUs and other site supporting equipment, it is easy to operate and maintain BSs. However, number of RRHs are not reduced in C-RAN, its deployment and maintenance is easier for service providers.

- Capacity Improvement: In C-RAN, multiple BSs work together in a large physical BBU pool, it is easier to implement joint processing and scheduling that reduces interference and improves spectral efficiency.

- Adaptability to Non-uniform Traffic: With load balancing the processing requirements among BBUs, C-RAN can easily handle the non-uniform traffic generated by active User Equipment's (UEs).

- Smart offloading traffic: We can adopt smart technologies like self organizing network, in C-RAN. This technologies helps C-RAN in handling the variations in traffic and can easily offload the traffic for processing among the BBUs.

## 1.2 Challenges of C-RAN

Vision of C-RAN poses a lot of research challenges in its architecture. Some of those challenges are listed below.

1. How to handle dynamic mobile traffic load with reduced power consumption?
2. How to make effective utilization of BBU processing?
3. How to support Multiple Radio Access Technologies (Multi-RAT) technologies?
4. How to increase capacity of networks with reduced power consumption?
5. How to support growing mobile traffic on core network?

In our work, we choose the following technical challenges posed by fully centralized C-RAN to solve.

1. Handling dynamic traffic load processing of BS.
2. Generating large scale and long term dynamic traffic loads based on spatio-temporal conditions.
3. Reducing BS power consumption with dynamic traffic loads.

## 1.3 Our Proposed Work

Based on the literature survey and research challenges of C-RAN architecture, we have made the following contributions for C-RAN.

1. Radio over Fiber (RoF) for C-RAN
   - Designing RoF model
   - Studying RoF impacts on LTE RAN
2. Proposed C-RAN Architecture
   - Control and data plane processing for estimating traffic load.
   - Estimating power consumption of BSs.
   - Generating large scale and long term traffic loads based on spatio-temporal conditions.
   - Adopting dynamic clustering of BSs based on estimated traffic load.
   - Scheduling the BSs for processing on computing platform.

Each problem has been solved and explained as a separate chapter that has been listed with their respective literature survey, motivation and experimental results. At the end of each chapter summary section will discuss the limitations and the applicability to other research problems.

In chapter 2, our research is aimed at the investigation of variation of performance in the physical (PHY) and Medium Access Control (MAC) layers of those cellular networks. We decided to study this performance variation with ns-3 since it provides more accurate PHY and MAC layer models of LTE among available open source simulators. As there is no RoF model in ns-3, we added new features to the PHY layer in ns-3 to enable the simulation of RoF. This work shows how ns-3 can be used for simulation of cellular RoF networks and compares simulation results with traditional LTE network.

In chapter 3, we study the load discrepancies in multiplexing the higher layer protocols on BBU pool of C-RAN. Each BBU has a specified processing capacity based on traffic load it is processing. With the maximum processing capacity of BBU and exploiting the variation in spatio-temporal variations of traffic loads, we can map multiple higher layer protocols on to a single BBU and the process of mapping is termed as multiplexing. For balancing traffic loads, we perform dynamic clustering of RRHs into sets and serving those sets with BBU pool. Each BBU will be scheduled to process on GPP. In this chapter, we define the processing capacity of BBU and processing requirement of RRHs. We also describe how dynamic clustering of RRHs will optimize the BBU resources by considering different traffic load scenarios. Chapter 4, has concluding remarks and future work.

# Chapter 2

# Radio over Fiber Model

## 2.1 Introduction

Recent trends in cellular network architectures have been widely focused on improving the data rate of the users. On the other hand, demand for data has been growing exponentially, where users demanding to connect wireless networks seamlessly. The radio-over-fiber (RoF) is one of the candidate solutions which look promising in satisfying user demands. This RoF is deployed with Distributed Antenna System (DAS) that improves coverage region and reduces energy consumption and also more flexible for mobility in wireless networks. With DAS, service provider can reduce the OPEX [1].

Wide range of wireless technologies like LTE, Wi-Fi, WiMAX can provide services with RoF based DAS. In Our research, we investigate the transmission of LTE signals using RoF based DAS. Fig. 2.1 shows the architecture of LTE system with RoF based DAS. The RoF model is basic building block in realizing the RoF based DAS networks.

The research community has shown a lot of attention on transmission of radio signals on optical fiber. The main research activities that need attention are investigating transmission techniques over optical networks in order to propagate a radio signal over long distance [3,4] and the analysis of impacts of the propagation delay on Physical/Medium Access Control (PHY/MAC) layer operations of wireless services. The former activity requires the utilization of networking simulators and tools in order to assess the performance of RoF based DAS networks under given conditions [5].

In existing source models of ns-3, there is no support for simulation of RoF based DAS networks. In order to examine the performance of the PHY/MAC layers of LTE in RoF based DAS, we modified the features of the PHY layer in LTE module of ns-3 [6]. Exiting PHY layer in LTE module has one-one to mapping with spectrum channel. We integrated the modified PHY layer of LTE module in the ns-3 simulator to model the transport of radio signals over optical fiber. We also developed a new device model to represent the RRH entities in the simulator in order to set their position in the RoF based DAS network. Our proposed RoF model for LTE is able to investigate the network performance for RoF based architectures with a single antenna as well as Distributed Antenna Systems (DAS).

In this work, we first explain the relevant advantages and the challenges linked to transmission of radio signals of LTE over a RoF based DAS network. Then, we discuss our implementation of

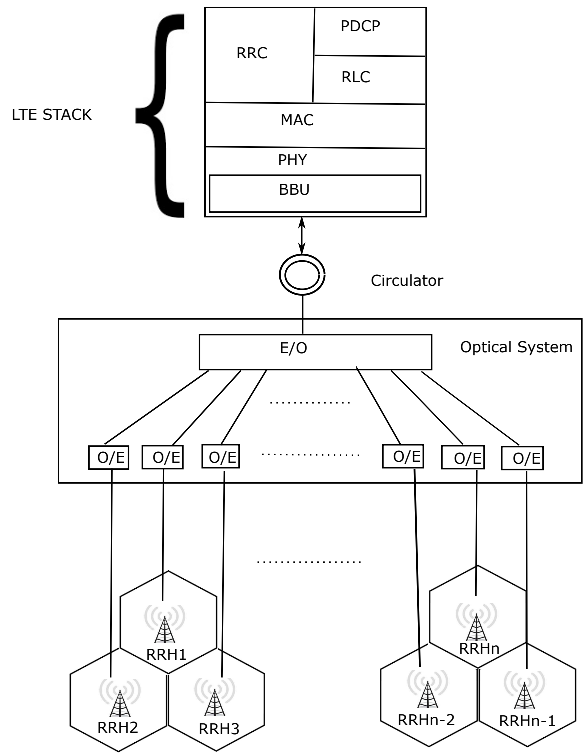


Figure 2.1: C-RAN based LTE system using a RoF DAS architecture

RoF model for LTE in ns-3. At the end, we validate our RoF model by comparing the simulation results obtained from ns-3 with results obtained with theoretical evaluations.

## 2.2 Motivation

RoF based DAS involves the transmission of radio signals over an optical fiber. In these systems, the radio signals arriving from the LTE BS are transformed into an optical signal and transmitted through optical fiber. This optical signal is transformed back to a radio signal using a photo detector so that output and input signals are copied to each other. RoF based DAS provides flexibility in providing several wireless technology services to the users. Among various wireless technologies, our work focus on the transmission of LTE radio signals using RoF based DAS.

In comparison to traditional LTE network architectures, the utilization of RoF based DAS infrastructure referred in Fig. 2.1 offers numerous advantages to centralized architectures. The centralized architecture enables network coverage by placing the antennas close to the users, which provide an opportunity to reduce the emitted power at each antenna in the cell site. Moreover, handling mobility of users becomes easier as there will be no handover since users will stay attached to the same BS. Furthermore, RoF based DAS enable the whole RAN functions to be processed at a central location, which reduces the complexity of the RRHs and maintenance of those networks [3, 4].

However, LTE PHY/MAC layers were not initially developed for RoF based DAS. Indeed, in traditional LTE system, there will not be any propagation delay involved between RAN and transmission antenna, whereas in RoF based DAS network the propagation delay between RAN and

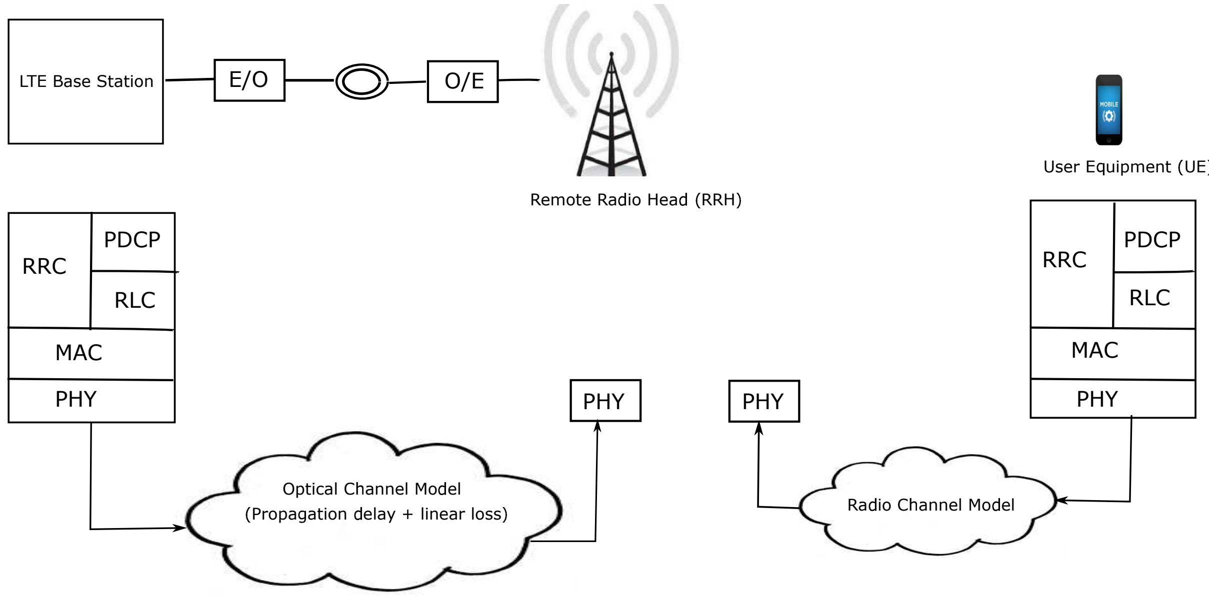


Figure 2.2: C-RAN RoF model

transmitted antenna (RRH) is around 5$\mu$sec per kilometer. The PHY/MAC layer parameters as defined in 3GPP standards may not be adopted to a RoF transmission. So, it is necessary to adopt the changes to the parameters of current PHY/MAC layers of LTE in order to improve the performances of the LTE systems using RoF based DAS.

In order to investigate the network performance of LTE system with a RoF (optical fiber) link, we implemented a new model in ns-3 for the simulation of RoF based DAS. In further sections, we describe how the existing LTE module provided by ns-3 has been modified to enable the simulation of LTE system with RoF based DAS network.

## 2.3 RoF Model for C-RAN

Introducing optical fiber links between the BS and the RRH will increase propagation time of radio signals. However, it does not only correspond to an air interface delay, but it linked with optical fiber link propagation delay i.e., time taken by a radio signal to travel over optical fiber.

The propagation delay of optical fiber will impact the performance of PHY/MAC layers. At the MAC layer, the additional delay will result in unnecessary re-transmissions. So, in order to study the performance of RoF based DAS in cellular networks, we considered a minimal functional optical channel component, which computes optical fiber transmissions as an extra delay equal to the time needed by the radio signal to travel along the optical fiber link and linear loss equivalent to attenuation that is introduced in optical fiber as shown in Fig. 2.2.

The existing interaction of PHY layer of LTE with spectrum channel in ns-3 is showed in Fig. 2.3. This interaction can still be used for LTE RoF based DAS simulations, but the LteEnbPhy and LteSpectrumPhy modules need to be modified. Since RoF transmissions need optical fiber, an optical channel component needs to be added in ns-3.

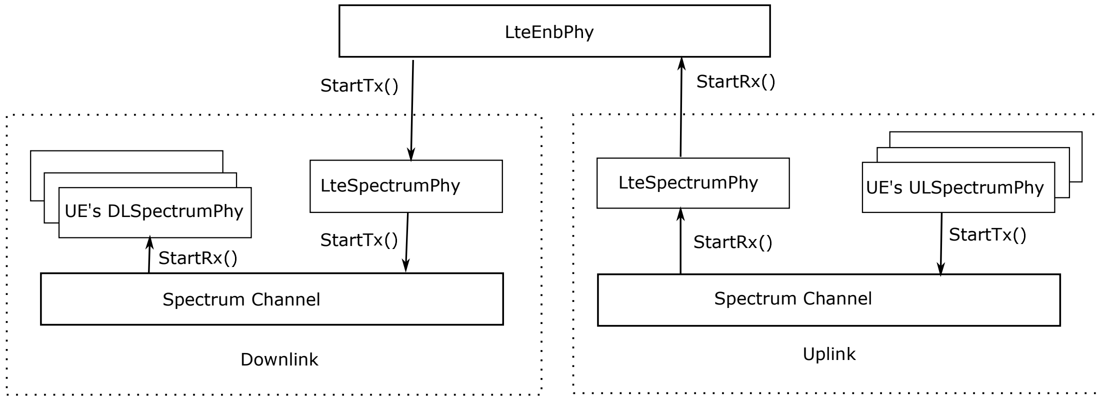


Figure 2.3: Existing PHY layer and Spectrum Channel Model

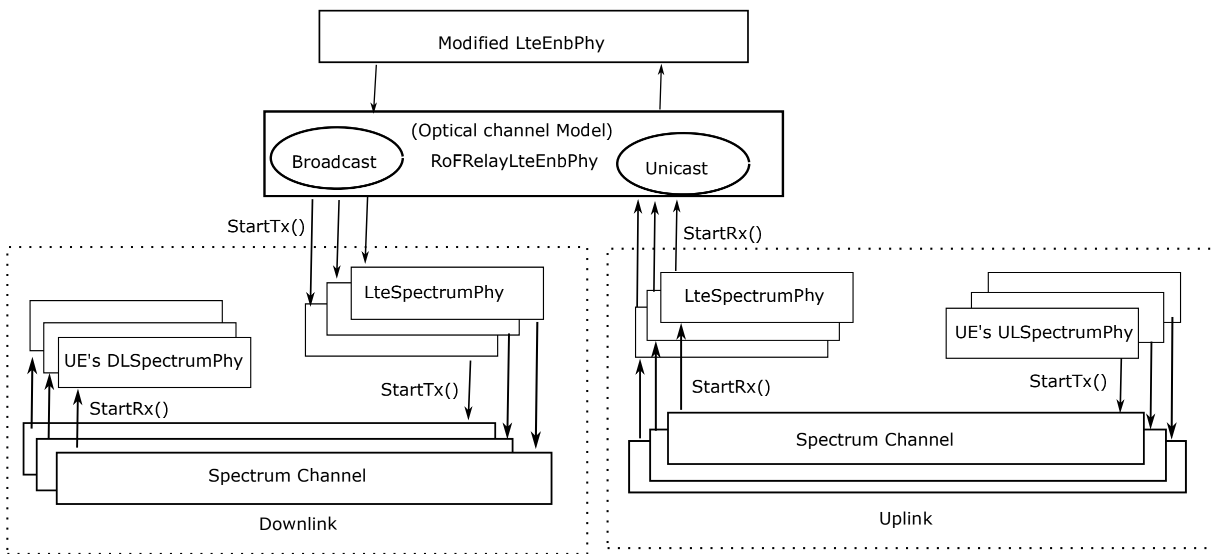


Figure 2.4: Proposed RoF model

In our proposed RoF model shown in Fig. 2.4, a new module called RoFRelayLteEnbPhy is introduced, which works together with optical channel component that computes the delay and loss due to the transport of the LTE radio signals over the optical fiber. As shown in Fig. 2.4, each LteEnbPhy instance is attached to RoFRelayLteEnbPhy instance which in turn connects to multiple LteSpectrumPhy instances. RoFRelayLteEnbPhy is responsible for forwarding the packets coming from LteEnbPhy to LteSpectrumPhy and receiving packets from LteSpectrumPhy to LteEnbPhy instances. Here LteSpectrumPhy mimics as RRH residing at cell site.

When the BS transmit signals, it is broadcast to all RRHs that are attached to RoFRelayLteEnbPhy (i.e., broadcasting in down link) and the signal received from RRH is transmitted back to a BS (i.e., unicasting in uplink). In ns-3, LTE standard implementation, when a BS (i.e., LteEnbPhy) has a packet to send, it is send to RRH (i.e., LteSpectrumPhy)and RRH will transmit the radio signals into air interface (i.e., Spectrum Channel). All receivers (i.e., UE) residing within the range of RRH will receive the transmitted signals through air interface. Being dependent on the received power of the incoming signal, the receiver needs to synchronize with the signal. As a result, a receiver located near to a RRH will receive the signals sent by all the other RRHs, but since the received power of those signals will be under the threshold it will correctly synchronize on the signal coming from its closest RRH.

In the OpticalChannel component, the calculation of the optical propagation time ($\tau$) is directly proportional to the distance (L) between the BS and the RRH and is inversely proportional to the velocity of the signal ($\nu$) in the optical fiber. Using equation 2.1, we can compute the $\tau$. Computing of $\nu$ is done using the fiber refractive index (n) and the light speed velocity in the vacuum (c) as shown in equation 2.2.

$$\tau = \frac{L}{\nu} \tag{2.1}$$

$$\nu = \frac{c}{n} \tag{2.2}$$

The refractive index of the fiber can be tuned to alter the velocity of the signal in the optical fiber. The linear loss is also configurable parameter. By default, parameter n is set to 1.5 and linear loss is set to 0.2 dB/km respectively. These parameters will result in a classical delay of 5 $\mu$s per kilometer of optical fiber.

The key point in our model, instead of simply adding an additional delay in the existing LTE module, resides in the ability to set the position of each individual node (BS, RRH and UEs) of the network. To do so, we implemented RoFRelayLteEnbPhy which has the responsibility to hold together all nodes (i.e., BS and UEs) used by a RRH. The optical channel calculates the length of the optical fiber link between the BS and each RRH using their respective positions. The distance between RRH and UE will determine the propagation loss of signal where as distance between BS and RRH will determine the propagation delay of signal.

Our RoF model in ns-3 can be utilized for any radio-over-fiber configuration using the LTE system. In the subsequent sections, we validate our model and we also provide an example scenario for utilizing our RoF based DAS implementation in ns-3.

## 2.4 Simulation Setup and Validation

In this section, we validate RoF model for throughput impact with optical fiber length. We also study the RoF model for power consumption and coverage region advantages.

### 2.4.1 Throughput

We have set up the simulation experiment with our RoF model implementation in ns-3 (version: ns-3.21) using the parameters listed in Table 2.1. We validate our implementation with a theoretical evaluation for RoF based DAS networks. To do so, we considered an ideal spectrum channel (no interference, no error, no shadowing etc.,). Once validated, our model could be used to investigate the performance of RoF based DAS networks when the spectrum channel is not ideal.

The scenario considered for simulation setup is a point-to-point optical fiber link between BS and RRH with single antenna and a single user, which is continuously transmitting UDP packets to the BS (Fig. 2.5). The simulation parameters are listed in Table 2.1.

The scenario considered is quite simple with a single user and no error models used in the channel,

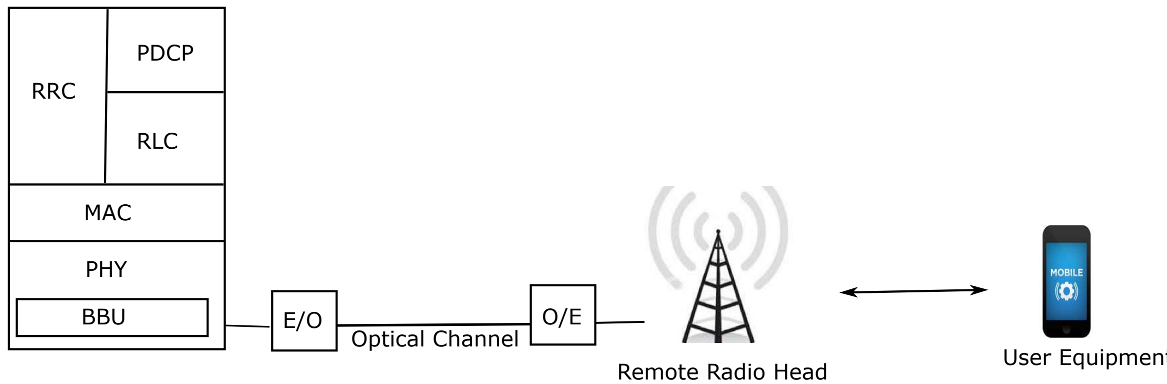


Figure 2.5: Scenario Considered for validation

Table 2.1: Simulation Parameters

| Parameter | value |
|---|---|
| EnodeB Tx Power | 17dBm |
| Spectrum model | Multimodel Spectrum model |
| Path Loss Model | Friis Propagation Loss Model |
| Optical fiber length | 0 - 20 km |
| Payload Size (UDP) | 1024 bytes |
| simulation time | 100sec |

so it is possible to estimate throughput theoretically. In this RoF case, it should be calculated as a function of the delay $d_F$ which is introduced by the optical fiber. Consider the payload size and the time required for complete transmission i.e., time taken to transmit one packet when there is no delay of fiber in the network ($d_F$=0) is compared with delay introduced by fiber in the network.

$$Throughput = \frac{payloadsize}{d_F + (2 * propagationdelay)} \tag{2.3}$$

$$Throughput = \frac{1024 * 8}{(d_F = 0) + 1285[\mu sec]} \tag{2.4}$$

In the equation 2.3, the factor 2 before propagation delay is used as the fact that both acknowledge packets and data packets are getting affected by the delay introduced by the optical channel. Using the simulation parameters of Table 2.1, we obtain the theoretical expressions that are plotted in Fig. 2.6, and are compared with the simulation results of ns-3. We can observe that the theoretical values are much closer to our simulation values. As both plots have nearly identical slopes and are on the brink of each other, it shows that the effect of the RoF transmission on the LTE network performance is well modeled in our ns-3 implementation.

### 2.4.2 Power Consumption

In this section, we explain how power consumption is related with achievable rates and later we show that RoF model gives advantages of power consumption in terms of coverage region.

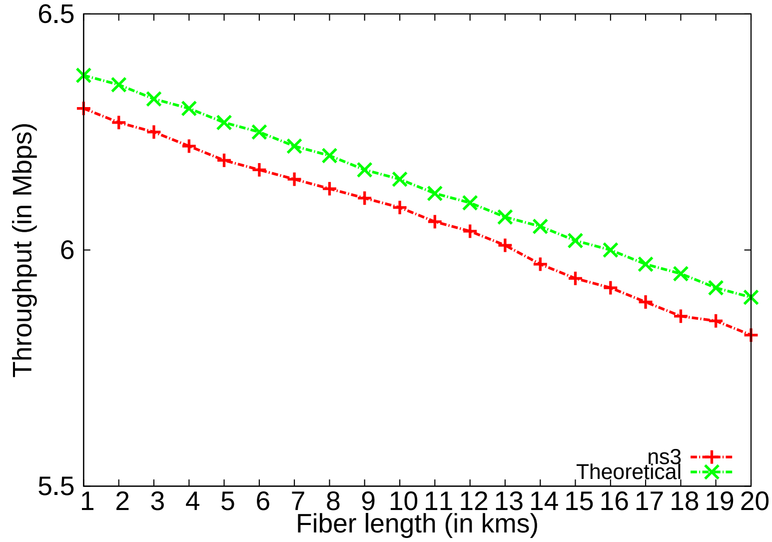


Figure 2.6: Comparison of results obtained with our model in ns-3 and theoretical model

#### Relationship of Achievable Data Rate and Received Power

In order to evaluate energy consumption between central and distributed antenna schemes, we need to use a statistical model rather than graphical data. So, we have used a simple path loss model for estimating the radio frequency power received by the receiver as a function of distance. The average power at receiver is expressed in dBm as

$$P_{recv} = P_{trans} + G_{trans} + G_{recv} - PL \tag{2.5}$$

In equation 2.5, $P_{trans}$ is the transmitter power in dBm, $G_{trans}$ and $G_{recv}$ are gains of the transmitter and receiver antennas respectively, and PL is the path loss expressed in dB. If we consider isotropic antennas, then $G_{trans} = G_{recv} = 0\ dB$ and the path loss as a function of distance is calculated by the following equation 2.6

$$PL = PL(d_0) + 10 * n * log10(\frac{d}{d_0}) \tag{2.6}$$

In equation 2.6, PL($d_0$) is the path loss at $d_0$ = 1 m, PL($d_0$) = 20*log10(c/4f$d_0$) = 26.4 dB at 5 MHz, d is the distance between transmitter and receiver and n is the power decay index. In a similar way, we can calculate the path loss at different frequencies.

Based on simulation environment, we need to adjust the power decay index. Suppose, if we consider a dense environment in a multi-floor office where each employee has his/her own office with separate rooms, the power decay index can vary between 4 and 6. In our simulations, we have used n = 5.

In LTE standard, the modulation scheme is adapted according to the received power level of user equipment. When a modulation scheme is selected, the maximum reachable data rate is set by the LTE standard is showed in Table 2.2. Data rates of LTE BSs provide the relationship between the modulation scheme used and the receiver sensitivity. As an example, the achievable data rates for a selected modulation scheme for respective receiver signal strength are presented in Table 2.2.

Table 2.2: MCS vs SINR vs Downlink throughput

| MCS | SINR(min)(dBm) | DL BS throughput (Mbps) |
|---|---|---|
| QPSK1/3 | -0.75 | 4 |
| QPSK1/2 | 1.5 | 6 |
| QPSK2/3 | 3.5 | 8 |
| 16QAM1/2 | 7.0 | 12 |
| 16QAM2/3 | 9.5 | 16.1 |
| 16QAM4/5 | 11.5 | 19.2 |
| 64QAM1/2 | 11.5 | 21 |
| 64QAM2/3 | 14.7 | 24.01 |

### Simulation of Power Consumption

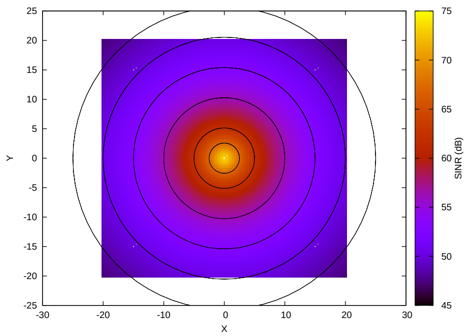


Figure 2.7: Radiated power of Central Antenna

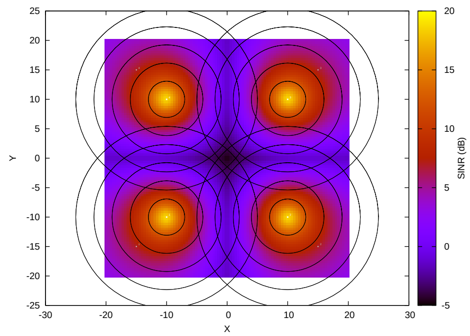


Figure 2.8: Radiated power of Distributed Antenna

As shown in Fig 2.7 the evolution of the signal strength that can be reachable at receiver is a function of the distance between the user and BS with centralized antenna with a transmitted power of 17 dBm covering a range of 30∼35m. However, the transmitted power variation can cause variation in data rates. So, the maximum achievable data rate that user can receive is calculated based on the quality of signal received. The achievable data rate of a user at a distance of 10m is given by the LTE standard. Even LTE standard specifies that user can connect to BS within 10 meters and beyond that range users may not get data rates as expected. At distance of 15m user may not even sense the signal of a BS. If we consider the scenario showed in Fig 2.7, user can experience good data rates till the range of 30∼35m, beyond that user may lose connection. Users can get good data rate within 15m i.e.,43% of total area. However, the distance greater than 35m no connection is possible i.e., 8% of the total area.

In Fig. 2.8, the same area is covered with four distributed antennas with half of the transmitted power for each antenna (i.e., each antenna has transmitted power of 8 dBm). Radio environmental map 2.8 presented shows better coverage despite of having lower transmitted power. The user equipment can attach to BS in the whole area. The achievable data rate can be measured over $2826m^2$ area, which represents 73.4% of total area. This statistics clearly show the advantage of using DAS network for coverage and exposure of radio frequency. Not only in terms of coverage and exposure, DAS allows to reduce power consumption in the case of sparse density of users in large areas.

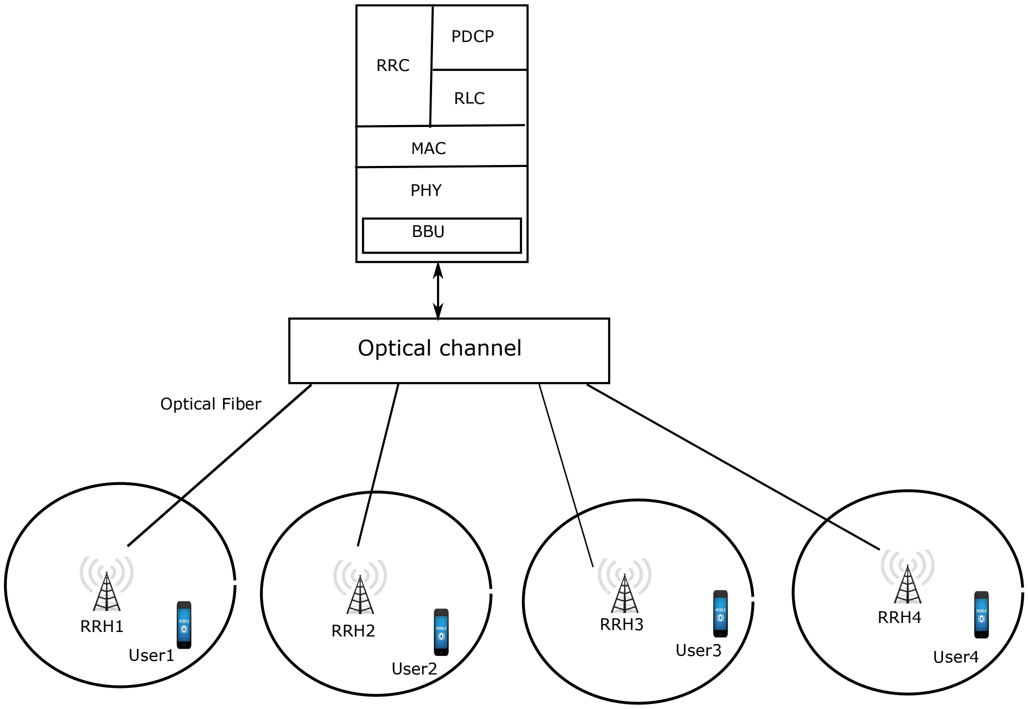


Figure 2.9: Scenario with non-overlapping coverage of RRHs

### 2.4.3 Coverage Region

RoF deployments are meant for using several distributed antennas with reduced power consumption. We provide an example scenario of simulation using a RoF based DAS with multiple antennas i.e., multiple RRHs and a single BS. When multiple RRHs are connected to a single BS, two scenarios can arise. First scenario occurs when RRHs have overlapping coverage regions with each other. In this overlapping case, we can reduce the interference in the overlapping region with coordinated scheduling of RRHs, which is out of scope of this work. We deal with the second scenario where RRHs do not overlap each other. In this scenario, the users are connected to each RRH are hidden from other users connected to another RRH.

For scenario where the coverage regions do not overlap, all user equipment's is concealed from each other since RRHs are located in clearly separate areas. We created such scenario in ns-3 using our RoF model with a DAS configuration where four RRHs are at a distance of 20 meters around a single BS and each RRH is being connected with one user (Fig. 2.9). As all four RRHs are situated far enough from each other, the received power of signals from other RRHs has much lower receiver sensitivity. Consequently, each user equipment receives all packets transmitted in down-link from the BS and do not hear the traffic sent by other users in up-link.

Since all the user equipment's is hidden from each other, we compare the simulation results in up-link obtained by considering a continuous traffic transmitted by all the user equipment's and the simulation parameters were set as in Table 2.1. Obtained simulation results with our implementation are compared in Table 2.3.

Table 2.3: Comparison of Aggregated Throughputs
of Central Antenna and Distributed Antenna Results

| Scheme | Aggre.Throughput(in kbps) |
|---:|---:|
| Central Antenna | 40.966 |
| Distributed Antenna | 40.863 |

## 2.5 RoF Model Impacts On LTE RAN

Most of the work discussed above concludes with proposing novel RoF model for LTE network in ns-3. But, it has adverse impacts on LTE HARQ transmissions of PHY/MAC layers. HARQ transmissions are like stop-n-wait protocol of wire-line networks where the server has to re-transmit packets to the client in case of time out/lost of Ack/Nack from client. 3GPP has provided the deadline for the timeout to be 3ms for LTE. If we introduce RoF model into LTE networks, the timeout reduces to 2.5∼2.6ms with a fiber length of 100km. With reduced timeout user will impacted with decreased throughput and network suffers with reduced spectral efficiency.

## 2.6 Summary

Our research evaluates the RoF model for LTE network performance. To evaluate RoF model, we work with LTE LENA model of ns-3 that includes a rather complete LTE EUTRAN model. However, the simulator has no support for RoF networks. In this work, we have proposed a RoF model in ns-3 which can be utilized to quantify the LTE network performance for any RoF configuration. Our proposed RoF model has been evaluated in terms of coverage and radio transmissions over a region and compared the results with theoretical calculations. Furthermore, we showed the working model of RoF based DAS when antennas are operating in non overlapping regions. However, we still need to focus on making RoF model to work on overlapping regions with advanced interference handling techniques. The main factor in supporting the simulation of overlapping regions of RoF based DAS is to handle the uni-cast conditions of the receiver at BS.

# Chapter 3

# Proposed C-RAN Architecture

## 3.1 Introduction

In traditional cellular networks, BS will have the processing of BBUs at its own location. However, in C-RAN architecture, base band unit process the signals in a centralized location. The C-RAN architecture can be realized by transporting the IQ samples through a high speed optical fiber from the antennas (also known as RRHs) at the cell site. The propagation delay of signals in optical fiber is a function of distance between a central location and cell site. So, the length of optical fiber can be limited based on performance requirements of C-RAN. The potential benefits with centralized approach include (i) savings of operational expenditure (OPEX) with reduction in the up gradation and maintenance i.e., less number of site visits, (ii) savings in capital expenditure (CAPEX) by reducing the cost of deploying a new BS and savings in energy expenditure i.e., exploiting the traffic load variations of BS allows us to use only required computing resources for processing and (iii) using coordinated multi-point transmission techniques for improving the overall network performance.

The OPEX is highly depends on density of mobile devices and its costs is a high as 50% of the total expenditure [1]. The cost of transporting the IQ samples to centralize location involves in OPEX. We need to factor the cost of transporting the IQ samples to the centralized location before evaluating the benefits of centralized processing. Budget analysis of using transporting medium is out of scope of this work. To address economic viability of using transportation medium, cellular industry has been considering other RAN split processing alternatives for evaluation i.e., user related signal processing will be done in a centralized location and only user independent portions of PHY layer (called as L1) will be processed at the cell site. This turns it into proportional to user load by reducing RAN processing requirements in a central location.

Our goal in this work is to evaluate the first and second benefits of the centralized architecture of LTE BS. We evaluated the benefits of centralized architecture by implementing a framework for multiplexing higher layer protocols on top of BBU pool processing on GPP computing platform. For multiplexing higher layer protocols on top of BBU pool, we have designed a scheduler for managing BBU resources and optimizing those resources.

## 3.2 Related Work

C-RAN architecture was proposed by China Mobile Research Institute [1], is a new type of LTE BS architecture that helps service providers to reduce CAPEX and OPEX. Even though, several C-RAN architectures have been proposed [7–9], different from these architectures, we propose a new fully centralized C-RAN architecture with centralized BBU pool processing on top of GPP computing platform.

For realizing C-RAN architecture, we have to implemented a framework where higher layer protocols of BSs will be processed on pool of BBU resources. For processing higher layer protocols of BS on pool of BBUs, we cluster the RRHs into sets. In general, clustering algorithms can be considered to be dynamic [10] and static. Static clustering is done with interfering sources in non-cooperative LTE network i.e., sectors of geographically closed RRHs with limited size of cluster. As the neighboring RRHs are only grouped with limited size, we can observe only limited gain of OPEX and CAPEX due to the fact that changing traffic demands are not taken into account for clustering RRHs. In order to address this problem, dynamic clustering approach has been introduced which leads to higher performance gains with a significant overhead. This leads us to the idea of forming clusters among RRHs that are not based on the position but rather the ones which cooperate to provide higher performance gains. This raises the question that how to dynamically cluster the RRHs for higher performance gains. The solution can be clustering RRHs that maximize the system coverage, minimize the multiple BSs co-operation and the power consumption. However, the goals are contradictory, we need to balance among those factors for achieving good results.

The authors in [11] has proposed two dynamic switching approaches, semi-adaptive and adaptive, where switching happens based on short time intervals and long time intervals. But authors in [11] did not mention how they have calculated and used the load information of BBU for taking the switching decision. In [12] authors have proposed a dynamic switching schemes based on multi-objective of maximizing data rates and minimizing interference. The authors in [13] has proposed BBU processing with the objective of minimizing power and RRH cluster formation. The authors in [14] has considered 50 users as upper limit for processing of super BSs. Different from these approaches, we have considered realistic BBU processing timings for defining processing load and used it in deciding the dynamic cluster formation of RRH. Based on the traffic demands simulated by creating realistic scenarios, we have validated our approach for power savings and computational efficiency.

## 3.3 Objectives and Contributions

Several field trials of the C-RAN architectures by different vendors demonstrates only transportation of the IQ samples to centralized architecture and remaining protocol processing will happen as in traditional BSs i.e., each BS has a BBU for processing signals and are not pooled on few BBUs for reducing the number of computing resources. In this work, we study the gains possible with resource pooling of BBUs and designed a framework to estimate the maximum pooling gains that can be possible with centralized architecture. We elaborate our contributions in further sections.

### 3.3.1 Benefits of Centralized Resource Pool

In traditional BSs, the BBU computing resources are capable of handling the peak/maximum traffic load processing while satisfying the provided deadlines. In our framework design, the traffic load processing variations in BSs are utilized to cluster the RRHs into sets and each set is scheduled to a serve by a single BBU processing on GPP. Variations in traffic load happen at very small time scales i.e., order of milliseconds. So, we are required to perform the clustering of RRHs into sets in the order of milliseconds, it is challenging and difficult to change the cluster of RRHs at such time scales. So, we estimated the possible time scale for changing the cluster of RRHs that allow to give maximum possible BBU computing resource gains. We generated traffic load variations through simulations for different scenarios to study the advantages and feasibility of proposed BBU computing resource pooling.

### 3.3.2 Base Stations Computing Resource Management

For potential BBU computing resource savings, we process multiple higher layer protocols of BSs utilizing minimum BBU computing resources ensuring their real-time deadlines. We refer the processing of higher layer protocols of BSs with BBU on GPP computing platform as virtual BSs. For processing of higher layer protocols of BSs on BBUs, we require a real-time schedulers for scheduling. Wealth of literature [15] [16] is found on real-time scheduling of computing resources, which include scheduling on multi-processor computing platforms. Most of the schedulers consider implicit deadlines for processing of tasks, but in centralized architecture we have scenarios where processing task can exceed the processing deadline. In our framework, we adopt the divide and conquer strategy where design of the system is segregated into sub problems. The problem of dynamic clustering of RRHs into sets is modeled as a linear optimization problem. Once clustering of RRHs into sets are solved. We have to design scheduling algorithm for scheduling each set of RRH onto BBU. Designing scheduling algorithm with adaptive processing deadlines for each task is addressed as a separate problem. Combining clustering and scheduling problems enable us to optimize the resource pooling of BBUs. The computing platform comprises of multiple processors with identical cores and each core is capable of processing and executing all tasks equally.

### 3.3.3 Framework Design and Evaluation Strategy

We designed the C-RAN framework as shown in Fig. 3.1, with centralized BBU pool processing on GPP computing platform. For downlink, we assume that each BBU is mapped to a single GPP. LTE BS higher layer protocol stack will process on pool of BBU resources. Mapping between LTE BS higher layer protocol stack and BBU will be done by Central Resource Controller (CRC). Based on processing requirements of RRHs, CRC clusters the RRHs and higher layer protocols to form virtual BSs for processing on single BBU. CRC uses BS processing and power models. BS Processing model is derived from profile statistics of real-time BS where as BS power model was utilized from project EARTH [17]. For ensuring real-time latencies, we designed a RoF model in ns-3. For generating traffic load data of a large geographical region, we choose traffic generating models in ns-3. We use the generated traffic load data as input data to the CRC for clustering the RRHs into sets. A scheduler has been designed to schedule each set of RRHs on to a single processor for computing based on their processing requirements. Using the BS processing, BS power and RoF models in simulation of LTE system, we can estimate the possible CAPEX and OPEX gains.

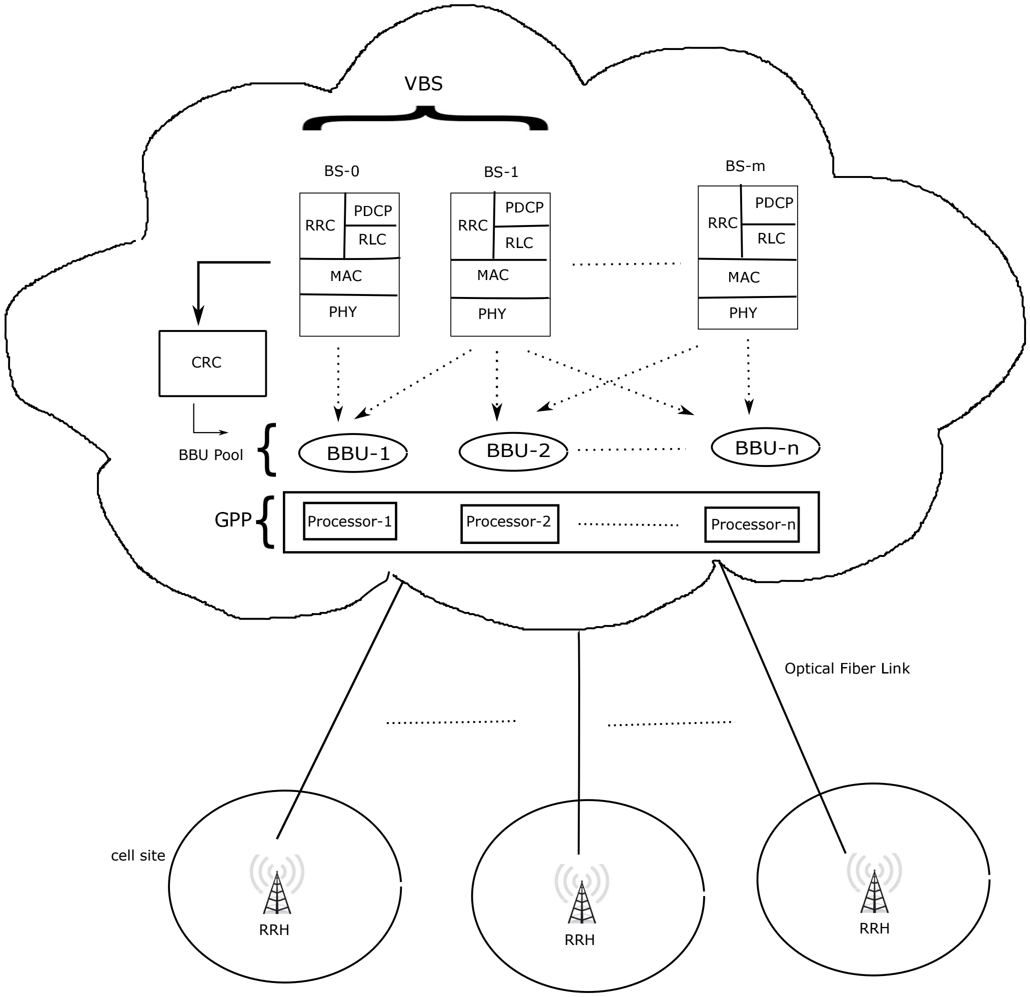


Figure 3.1: Cloud Radio Access Network Architecture

## 3.4 Analyzing Load Processing of Real-time Base station

In this section, we analyze the processing capacity offered by each procedure in the control and user planes of LTE BS then we characterize the parameters that influence the traffic load.

### 3.4.1 Real-time LTE Base Station Processing

For analyzing the processing of real-time LTE BS, we have used Open source implementation of LTE standard OpenAirInterface [18]. For making analysis easier, we segregate the control and user plane procedures and analyze their processing capacity along with parameters influence the processing of traffic load. The LTE BS with control and user planes is shown in Fig. 3.2.

The Control plane handles the radio access bearers and connection between the UE and network. Control Plane processing involves radio specific procedures that depend on two states of user equip-

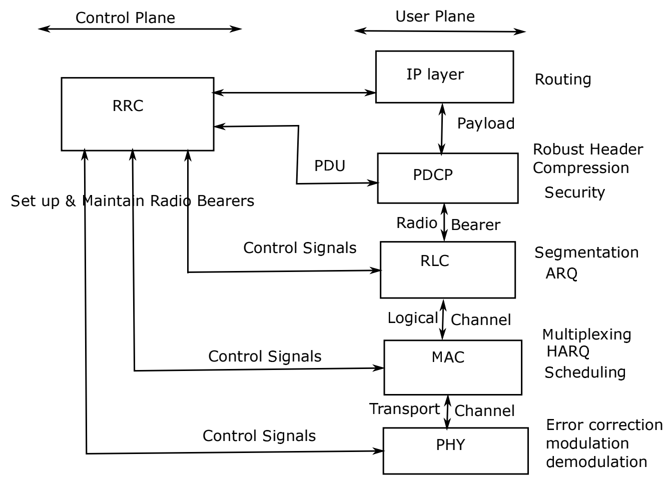


Figure 3.2: Control Plane and User Plane in LTE BS

ment (UE) i.e., idle and connected. In LTE standard this states of UE are referred as "RRC_IDLE" and "RRC_CONNECTED". Control plane processing involves latencies that is measured as time taken for performing the transitions from one state to another state.

In RRC_IDLE state, existence of UE is known in evolved packet core (EPC) and has an IP address. In this state, UE can receive broadcast or multi-cast data, it can monitor a paging channel for detecting incoming calls, it can perform BS measurements and does BS selection or re-selection. In this state of UE, there won't be any user specific processing related to data. So, very low overhead in processing traffic load will incur on the control plane of BS with UE RRC_IDLE state. Once UE has changed its state to RRC_CONNECTED, its existence is known in EPC and Evolved Universal Terrestrial Radio Access Network (EUTRAN). The location of UE is known at the BS level and its mobility is assisted and network controlled. In this state, UE monitors control channels that are associated with the shared data channel for determining whether data is scheduled for it. Along with monitoring, UE provides channel measurements and its condition as feedback information. In this state also, very less UE related processing will be seen in control plane processing.

Supported capacity of LTE system is dependent on the throughput and also on number of UEs that can simultaneously connect to the BS by control plane signaling. For supporting at least 400 UEs per BS for wider spectrum and at least 200 UEs [19] per BS for 5MHz spectrum, we may have only a few of those UEs in active state by transmitting or receiving data at any given instant of time. Most of the connected UEs will be in non-active state and can start transmitting or receiving data with low latency or can able to be paged. Presence of large number of non-active UEs doesn't have much impact on control plane processing which concludes that control plane processing of BS doesn't vary much for with a large number of UEs (i.e., $> 100$) attaching to BS.

In LTE BS, user plane deals with actual data transmissions which referred as UE data transmission or reception states. UE with transmission/reception of data states are treated as active UEs. User plane processing is a crucial performance factor for interactive and real-time services. User plane processing/latency is fully dynamic based on number of active UES. On the air interface, the UE latency can be calculated based on analysis of signaling of an unloaded LTE system i.e., very minimal user plane processing. UE latency can be defined as the average time between the initial transmission of a packet and reception of an acknowledgement. In LTE standard, the typical calculation includes 8 Hybrid Automatic Re-transmission Request (HARQ) rates. Thus definition of user plane latency defines the capability of the designed system for re-transmissions, without being distorted by other delays (i.e., scheduling delays) that would appear in loaded LTE system i.e., very high user plane processing.

LTE BS has to operate one way packet transmission latency across the RAN as low as 3ms. In practical systems actual delay experienced is dependent on system traffic load and propagation delay. For example, HARQ plays a key role while re-transmissions take place, it maximizes the spectral efficiency at the expense of increased delay, whereas spectral efficiency maximization may not be essential in situations when minimum latency is required.

In LTE BS the control plane processing doesn't vary much with increase in number of UEs camping on the BS, where as the user plane processing varies proportionally with increase in number of active UEs i.e., majority of processing is done in user plane of BS with increase in number of active UEs. In further sections, we consider UE specific functions related to data transmission or reception that are processed in user plane for estimating traffic load on BS.

### 3.4.2 OpenAirInterface Profiling

In this section, we analyze the processing of each user specific functions offered by each layer in user plane of LTE BS. Below is the list the major functions of each layer of LTE BS.

- Radio Resource Control (RRC) layer handles functions related to control and maintenance of EUTRAN and UE configuration information.

- Packet Data Convergence Protocol (PDCP) layer handles data request and data indication functionalitites for transmitting and receiving data.

- Radio Link Control (RLC) layer handles related to data Acknowledgment, Unacknowledged and Transparent modes.

- Medium Access Control (MAC) layer, handles wide-band multi-user scheduling and Hybrid Automatic Re-transmission Request (HARQ).

- Physical (PHY) layer handles symbol-level processing of up-link and down-link functionalities of OFDM modulation/demodulation, scrambling, encoding/decoding,rate-matching/rate-dematching, turbo decoding etc.

The major user plane processing includes PDCP,RLC,MAC and PHY layer functionalities. For computing the processing timings of each layer, we have used OpenAirInterface software [18] as LTE BS protocol stack running with gcc 4.7.3, x86-32 (2.49 GHz Intel Xenon octa processor) on RTAI Kernel version 3.12 of Ubuntu 12.04. In Openair LTE BS, we have physical layer procedures with BBU running on GPP in emulation mode i.e., Openair software is running as emulator. Using static positions for BS, we move the UE with 20m/s speed for getting varying modulation coding index. We measure the processing times of each layer in LTE BS for every sec of UE movement. For example, if we can observe the timings for down-link and up-link are showed in Table 3.1 for LTE BS processing with MCS index 4.

| Name | Time(ms) | Average/Frame($\mu$ s) | Trials |
|---|---|---|---|
| Total phy proc | 524.355 | 873.924 | 600 |
| Total phy proc tx | 316.590 | 753.785 | 420 |
| Mac scheduler | 8.272 | 19.695 | 420 |
| Downlink SI | 0.139 | 2.312 | 60 |
| Downlink RA | 0.089 | 1.483 | 60 |
| Downlink/Uplink DCI | 0.366 | 1.017 | 360 |
| Downlink preprocessor | 1.392 | 7.908 | 176 |
| Downlink schedule tx dlsch | 2.051 | 11.653 | 176 |
| Downlink pdcp data req | 0.075 | 18.858 | 4 |
| Total phy proc rx | 205.125 | 1139.583 | 180 |
| Uplink Shared Channel | 4.109 | 34.238 | 120 |
| Uplink rx ulsch sdu | 2.434 | 27.977 | 87 |
| Uplink pdcp data ind | 2.072 | 518.084 | 4 |

Table 3.1: LTE FDD BS profiling results

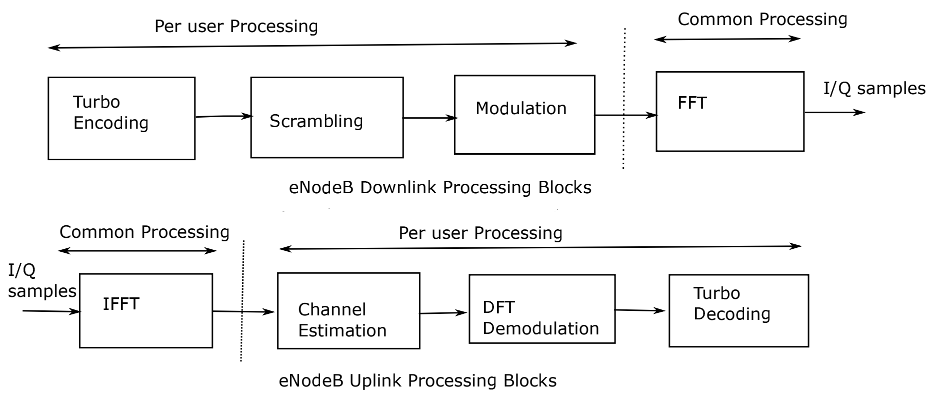


Figure 3.3: Open air LTE PHY processing blocks

If we observe the processing timings of LTE BS, up-link reception requires more processing capacity than down-link transmission. If we analyze the processing timings of each layer in down-link, physical layer processing for transmission clearly dominates the processing offered by other layers. For transmission of data in physical layer, we use Physical Downlink Shared Channel (PDSCH). In the same way for reception the Physical Uplink Shared Channel (PUSCH) and data indicator function in PDCP layer takes majority of processing time.

To characterize the factors that influence the processing load, we use the down-link processing of physical layer which has majority of processing. In physical layer, the functions of PDSCH are showed in Fig. 3.3. The key observation from Fig. 3.3 is that the functions Inverse Fast Fourier transform (IFFT) and Fast Fourier Transform on the up-link and down-link blocks are processed for every sub-frame and are not dependent on MCS or the number of PRBs allocated. So, the IFFT and FFT functions processing are independent of user specific processing and can be treated a constant processing on the BS. Other subsequent functions to the IFFT or the FFT depends on allocated MCS and PRBs to users and are treated as user specific processing i.e., dynamic processing.

To understand dynamic processing load of BS based on number of active UEs, we profile the physical layer of openair code. For understanding relationship of dynamic processing, we scheduled a single UE for transmission or reception. The profiling results obtained can be generalized to multiple UEs based on the PRBs and MCS allocated for every TTI. For profiling the PHY layer, we vary the following parameters (i) number of PRBs to be used (ii) the MCS value and (iii)the SINR for a given value of MCS. We calculate the time taken for execution of PHY layer by varying the parameters for the up-link and down-link settings.

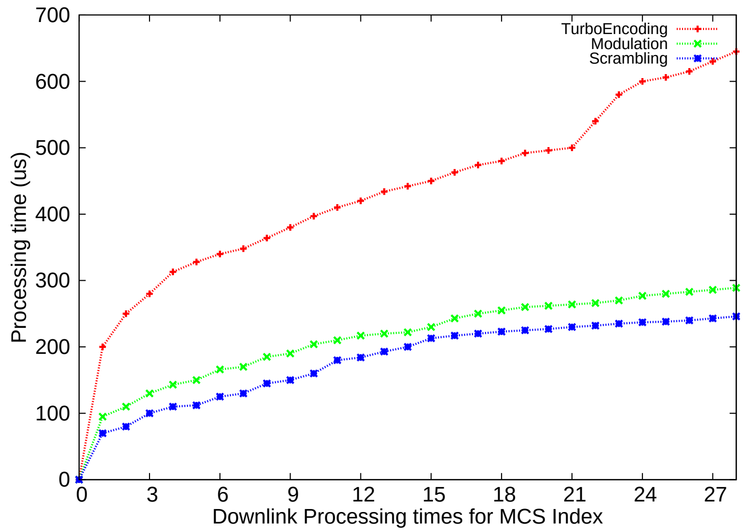


Figure 3.4: Downlink processing times

We have done the profiling with PRBs varying from 5 to 25 Resource Blocks (RBs) i.e., with frequencies 1MHz, 2MHz, 3MHz, 4MHz and 5MHz. The MCS index is varied from 1 to 28 with changing SINR value for RB allocated. For each MCS index, we measure the timings of PHY layer procedures Turbo Encoding, Scrambling and Modulation in down-link whereas in up-link, we measure the timings of procedures Turbo Decoding, IFFT and Demodulation. We have the following observations from profiling results,

1. In Fig 3.5, we observe that the turbo decoder in up-link processing of a subframe take higher processing time compared to other functions that offered a constant processing time. The turbo decoder [20] is an iterative algorithm where each iteration will improve the decoders estimation of the transmitted code word. From Fig. 3.4, we can observe that the turbo encoder have higher processing than remaining functions.

2. From Fig. 3.4, 3.5,and 3.6, the Processing load can be approximated as linear functions of PRBs and MCS. The total load can divide into two parts as a constant processing load that is not dependent of SINR, MCS and PRB and a dynamic part that depends on PRBs and MCS.

3. With linearity of MCS and PRBs in processing dynamic load, we can extend the load processing statistics to multiple user case. If multiple users are scheduled then the total processing load is the summation of individual processing loads that cannot exceed the load processing of a single user who is allocated all PRBs at highest MCS (refer section 3.4.3).

4. We can generalize our observations to other cellular standards, whose physical layer share similar processing components as LTE PHY layer. This applies to WCDMA where similar turbo encoding/decoding blocks exist.

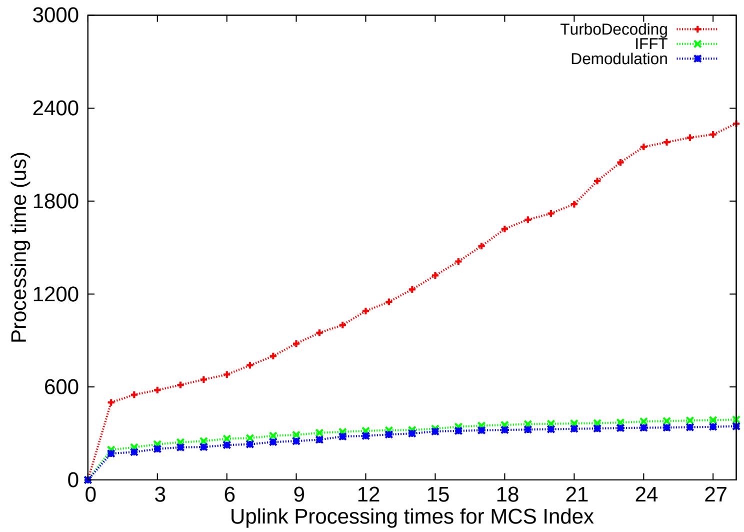


Figure 3.5: Uplink processing times

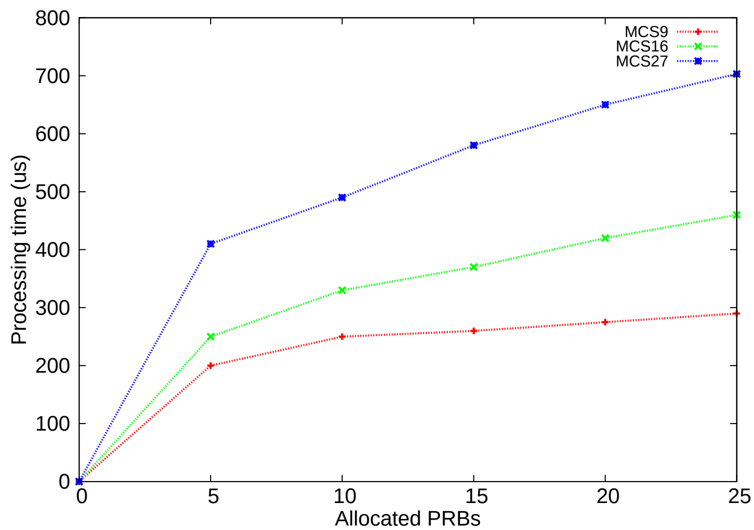


Figure 3.6: Downlink Processing times as function of Allocated PRBs

### 3.4.3 Peak Processing Load

Based on profiling results of openair code, we are able to separate the dynamic and constant processing loads of LTE BS. There has been a wealth of literature [21] in defining load of BSs, but it is not applicable in case of real-time processing of traffic load.

$$PeakLoad(L_{max}) = L_{constant} + L_{dynamic} \tag{3.1}$$

In equation 3.1, $L_{constant}$ is BS base load that includes constant processing times of control and user plane processing irrespective of number of active and non-active UEs, $L_{dynamic}$ is dynamic load processed based on number of active UEs who are generating traffic.

The dynamic processing of traffic load is computed based on active UEs who are transmitting/receiving data that will be a function of the actual channel resource used for sub-frame and value of MCS used for assigning each channel resource. For computing peak processing load, we will assume that highest coding rate is used for each modulation scheme. The Table 3.2 shows mapping of modulation and coding rates. 9, 16 and 27 are highest MCS values for coding rates QPSK, 16QAM and 64QAM.

Table 3.2: MCS index and Modulation order

| MCS Index | Modulation order | TBS index |
|---:|---:|---:|
| 0 to 9 | 2 | 0 to 9 |
| 10 to 16 | 4 | 9 to 15 |
| 17 to 28 | 6 | 15 to 26 |
| 29 | 2 | reserved |
| 30 | 4 | reserved |
| 31 | 6 | reserved |

The QPSK, 16QAM and 64QAM transmit 2, 4 and 6 bits per channel respectively. If we compare the loads imposed by making observations from Fig 3.4 that 16QAM at MCS16 takes double processing time than QPSK at MCS9 and 64QAM with MCS27 takes three times the processing time of QPSK at MCS9.

The units of measurement for peak load processing are floating point operations per second (FLOPS) that depends on processing speed of computing platform. So, computing constant processing load and dynamic processing load will give us peak load processing of BS in FLOPS. If we consider a single-core 2.5 GHz processor, it has a theoretical performance of 10 billion FLOPS = 10 GFLOPS i.e., 1000 million floating point operation per second.

For deriving processing load we use a unit of FLOPS for processing time. From Fig. 3.6, we can compute that the QPSK modulation scheme takes 2.8 MFLOP (i.e., million floating point operations). Similarly, for 16QAM it is 5.4 MFLOP and for 64QAM it is 6.9 MFLOP.

Further, to simplify we refer the 2.8 MFLOP as 1 unit i.e., QPSK takes 1 unit of processing for transmitting 1 code over channel, 16QAM takes approximately 2 units of processing for sending 4 codes that means 8 units of processing and 64QAM takes 3 units of processing for sending 7 codes that means 21 units of processing load.

If we apply the above observations for computing the base load of 16QAM with 15 codes i.e, $15 \times 2 = 30$ units of processing (i.e., $30 \times 2.8 = 84$ MFLOP). For dynamic load, we need to use the equation 3.2 that computes based on the modulation scheme and the number of codes used for transmission.

$$L_{dynamic} = \sum_{i=1}^{nPRBs} x_i \tag{3.2}$$

$$x_i = processing\_time\_for\_ith\_PRBs\_MCS\_index \tag{3.3}$$

In equation 3.2, nPRBs is the number of physical resource blocks allocated and x is the processing unit for MCS value of $i^{th}$ physical resource block allocated.

In FDD LTE with 5MHz bandwidth, we can compute the peak processing load of BS with a single user who has allocated all 25 resource blocks with 64QAM and highest MCS value (i.e., 28). So, the dynamic load processing is computed as shown in equation 3.4.

$$L_{dynamic} = 25 \times (3 \times 7) = 525 units = 525 \times 2.8 = 1470 MFLOP \tag{3.4}$$

We can use the equations 3.1, 3.2, for calculating processing of traffic load in BS. We use MCS index and its corresponding modulation scheme for capturing dynamic processing load. With openair profiling results, the peak processing of traffic load of the BS can be defined as the processing load of a single UE who is allocated all the Physical Resource Blocks (PRBs) at highest modulated coding scheme (MCS) at one TTI. In order to generalize, we need to consider scheduling of multiple UEs for transmission or reception of data. As UEs will spread randomly in the cell site, the total load of the BS is given by the summation of individual processing loads of UEs at one TTI. We can conclude that the total processing load of any BS is given by the summation of individual loads. If the summation of individual load exceeds the peak processing load of the BS ($L_{max}$), then it is overloaded.

### 3.4.4 Power Consumption of LTE Base Station

Generally LTE BS site consists of multiple transceivers (TRXs). Each transceiver consists of following parts

- Antenna
- Power Amplifier
- RF transceiver
- Base band interface
- Power supply cooling

Each component power consumption is shown in Table 3.3 which has been used in [22]. The relation between BS power consumption $P_{base}$ is nearly linear with RF output power $P_{dyn}$. Therefore, the linear relation of power model can be given by equation 3.5.

$$P_{consume} = N_{TRx} * (P_{base} + \delta * P_{dyn}) \tag{3.5}$$

$$0 \leq P_{dyn} \leq P_{max} \tag{3.6}$$

Table 3.3: LTE Base Station Power Consumption Statistics

| | MacroBS | Femto |
|---|---|---|
| PA Tx power(dBm) | 46 | 17 |
| Total Tx power(Watts) | 128.2 | 1.1 |
| RF Tx (Watts) | 6.8 | 0.2 |
| Rx (Watts) | 6.1 | 0.3 |
| Total RF(Watts) | 13 | 0.6 |
| BB Radio (inner Rx/Tx)(Watts) | 10.8 | 1.0 |
| Turbo coder(Watts) | 8.8 | 1.2 |
| Processors(Watts) | 10.0 | 0.3 |
| Total BB(Watts) | 29.5 | 2.5 |
| DC-DC,Cooling | | |
| Mains supply | | |
| Total TRx chain(Watts) | 225 | 5.2 |
| # sectors | 3 | 1 |
| # Antennas | 2 | 2 |
| # Carriers | 1 | 1 |
| Total $N_{TRX}$ chains, $P_{in}$(Watts) | 1350 | 10.4 |

In equation 3.5, $N_{TRx}$ are number of transceivers in BS site, $P_{base}$ is constant power consumption irrespective of transmissions, where as in equation 3.6, $P_{dyn}$ power that depends on transmission range of BS and $P_{max}$ is the maximum power consumption at highest transmission range.

The Table 3.4 shows the parameters used for power consumption of LTE Macro and Femto BS.

Table 3.4: Power Model Parameters

| BS Type | $N_{TRx}$ | $P_{max}$ | $P_{base}$ | $\delta$ |
|---|---|---|---|---|
| Macro | 6 | 40.0 | 118.7 | 2.66 |
| Femto | 2 | 0.05 | 4.8 | 7.5 |

## 3.5 System Model and Problem Formulation

To realize the proposed framework for C-RAN, we need a system model that can integrate the partitioning i.e., clustering of RRHs and scheduling the higher layer protocols of BSs on top of BBU computing resources. The Fig. 3.7 shows where computing resources resides in cloud data center and we can process the upper layer protocols (partial PHY, MAC, RLC, PDCP and RRC) of LTE BS. With inter-layer communication and multi-threading capacity, we can carry out the layer wise protocol processing. With modern advancements in hardware, we have availability of systems with multiple processors on which we can process layer wise protocols of multiple BSs.

There will be a central resource controller (CRC) which does computing resource management for dynamically clustering the RRHs and scheduling processing of each cluster with a single BBU. Based on periodic requests for traffic load processing, central resource controller manages the BBU resources on computing platform. Each cluster of RRHs served by a single BBU with higher layer protocols running on top of it, which we termed as a virtual BS (VBS). As each cluster is composed of several RRHs and higher layer protocols as a set, so it is VBS associates with multiple RRHs i.e., each BS will be connected with one RRH and multiple higher layer protocols of BSs combined to

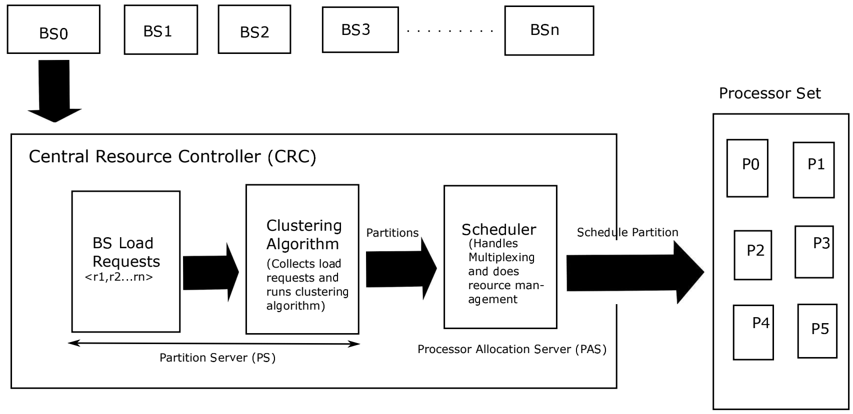


Figure 3.7: System Model for C-RAN framework

form VBS.

As showed in Fig. 3.7, CRC has two components Partitioning Server (PS) and Processor Allocation Server (PAS). The Partitioning Server runs the dynamic clustering algorithm based on BS processing model and create cluster of RRHs. Each cluster is treated as set of VBS and send to Processor Allocation Server for scheduling on to a computing platform. Processor Allocation Server (PAS) that keeps track of VBS scheduling on each processor. Based on each VBS traffic load processing, PAS calculates the load that can be adjusted on each VBS and stores in its database for taking decisions on resource optimization and allocation. In current work, each BS signal processing relies on BBU pool and rest of processing is associated with its higher layer processing. Each VBS associated with one BBU and offers services to multiple RRHs. As clustering algorithm may always shifts each RRH into different clusters, its associated higher layer processing has to associate with corresponding BBU of VBS and we termed this mechanism as dynamic switching of RRH and BBU with respective formation of VBS.

While multiple BS are running in a single system, each BS can coordinate with other BS to take better decisions in radio resource management (RRM) if they are geographically adjacent. If clustering algorithm is subjected to the condition of geographical adjacency, we may not reap full benefits of centralized architecture. Based on our objectives on design of centralized architecture, we can imply different conditions on clustering algorithm. But for coordination, among VBS that are running on different processors, we need to form a cluster of VBS that are processing in adjacent processors. Coordination of different VBS may require in case of network wide optimization. The cost of exchanging information during coordination of VBS must be minimized to avoid impacts on real-time constraints of protocol processing. However, current work focus on reaping the benefits of dynamic clustering of the RRHs and processing those clusters using BBU pool.

To formulate dynamic clustering problem, we need to define the peak traffic load on LTE BS. By using peak load processing in section 3.4.3, we define load on LTE BS. The CRC uses the parameters defined in the Table 3.5 on each VBS, where N is the total number of VBS (i.e., number of clusters formed) available in system that each VBS are mapped to one processor and T is the total number of time stages that one day can be divided to perform load calculations. Dividing a day into time stages give regular load variations from 0 to 24 hours. Based on periodic load variations of each

Table 3.5: Description of System Parameters

| Parameter | Description |
|---|---|
| Max Load ($L_{max}$[i]) | Maximum load that $i_{th}$ VBS can handle where i $\in$ 1,2,..N |
| Load ($vbs_{load}$[i,j] ) | Processing load on $i_{th}$ VBS at time stage j$\in$ 1,2..T |
| scheduled VBS (vbs_num[i]) | $i_{th}$ scheduled VBS |

VBS, PS can aggregate or segregate VBS and sends to PAS for mapping on the processors. After aggregation or segregation of VBS, three types of resource consumption tasks are considered by PAS i.e., load processing on each VBS along with associated RRHs, adjusting load of VBS and scheduling new VBS based on traffic load requirements.

### 3.5.1 Base Station Processing Load

The computing resources that each VBS has been allocated based on their network load (vbs_load) which can be calculated as described in section 3.5. The PAS calculates and maintains statistics of load on each VBS at each time stage as follows:

$$vbs_{load}[i,j] = \sum_{k=1}^{N} vbsload_{jk} \tag{3.7}$$

In equation 3.7, vbs_load[i,j] is the load of $i^{th}$ VBS at time stage j. The load can be calculated based on the actual processing of up-link and down-link protocols of $i^{th}$VBS. For example, in the practical LTE BS protocol processing the procedures including up-link and down-link control information, up-link user data receiving, down-link user data preparing, user data scheduling between physical/transport/logical channels in $i^{th}$ VBS will occupy processors time at every sub-frame. The summation of these up-link and down-link frame processing time in time stage j forms vbs_load[i,j].

### 3.5.2 Dynamic Clustering of Base stations

Each BS is associated with RRH at a cell site in C-RAN. We combine multiple BSs to form a VBS with their respective RRHs. When the process of forming VBSs becomes dynamic, the respective association of RRHs also becomes dynamic. Each VBS has to serve UEs connected through RRH. So, we formulated the clustering of BS problem as clustering of RRHs that need to be served with VBS.

Let us consider a C-RAN system with r RRHs each with single antenna. In our work, we focus on down-link scenario in which x RRHs form a cluster and each cluster will associate with single VBS. There is a central unit called CRC which manages VBS pool and connectivity of RRHs to VBS. Let C and R represent finite sets of clusters and RRHs, respectively. |C| represents cardinality of set C. The RRHs are indexed from 1 to |R|. Let $L_{max}$ be the load that each VBS can process at maximum.

Let $C_j$ represents $j^{th}$ cluster in set C and n represents the processing requirement of each RRH. When CRC has all the information of each RRH processing requirement, it can decide how many RRHs can form as cluster to connect to each VBS with the objective of minimizing the number of clusters.

The mathematical formulation is,

$$min \sum_{j=1}^{|C|} C_j \tag{3.8}$$

where $C_j$ is $j^{th}$ cluster with associated $i^{th}$ RRH

$$n_i = \sum_{k=1}^{m} t_k \tag{3.9}$$

$$\sum_{j=1}^{|C|} (\sum_{i=1}^{|R|} n_i * R_{ij} \leq L_{max} * C_j) \tag{3.10}$$

$$\forall R_i = \{0, 1\} \tag{3.11}$$

$$\forall C_j = \{0, 1\} \tag{3.12}$$

Equation 3.8 shows the objective of minimizing the clusters, its corresponding constraints are shown in equation 3.9, 3.10. The value $n_i$ on equation 3.9, is summation of the individual task processing of BBU in terms of MFLOPS and t represents a task in BBU and a BBU has m tasks. The selection of RRH is represented with integer values {0,1} as shown in equation 3.11. The cluster limits is represented in equation 3.12.

The CRC iteratively uses this approach until all RRHs are exhausted in the set R. Each iteration forms a cluster of RRHs, which are going to associate with VBS.

### 3.5.3 Scheduling of Base stations

The CRC is responsible for collecting the traffic load statistics on each VBS and communicating the load statistics information with PAS. The PAS usually maps each VBS on to the processors, based on load statistics. Whenever the load adjustment is not able to do on one of the running VBS, PAS can be switch-on i.e., schedules the new VBS for handling the traffic load. However, there can be situations where PAS can switch-off the existing VBS on using load adjustment factors. Scheduling of VBS is tracked with status of vbs_num[i], where i is the identifier of VBS.

## 3.6 Spatio-temporal variations

In C-RAN, traffic demands are based on user density and subjected to temporal and geographical conditions like commercial/office areas, rural areas etc. Based on these temporal and geographical conditions [17], combinations of VBS-RRH can change because of limit on processing capacity of VBS. For example, Fig 3.8 shows an example condition of VBS-RRH switching. In Fig 3.8, {RRH1,RRH2,RRH3},{RRH4,RRH5,RRH6} and {RRH7,RRH8} are connected to VBS1, VBS2 and VBS3, respectively. On the other hand, in Fig 3.9, {RRH1,RRH2},{RRH3,RRH4},{RRH5,RRH6} and {RRH7,RRH8} are connected to VBS1,VBS2,VBS3 and VBS4 respectively. With increase in traffic load processing, re-association of VBS-RRHs will happens.

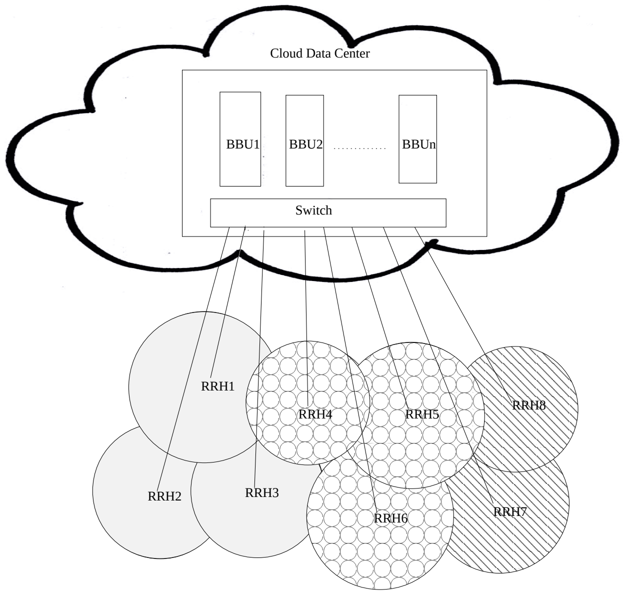


Figure 3.8: cluster of BBU-RRH in C-RAN based LTE system

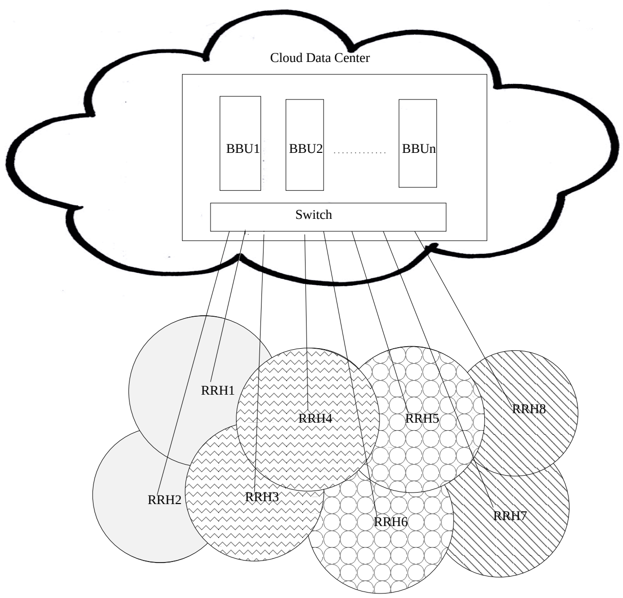


Figure 3.9: Change of BBU-RRH cluster

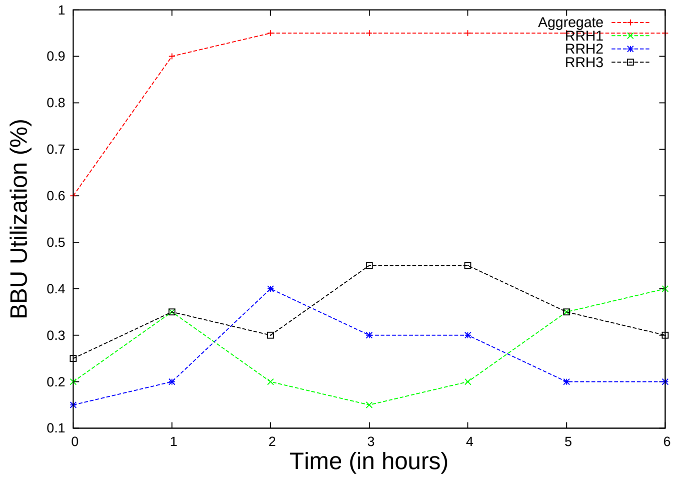


Figure 3.10: Example of radio resource utilization in time series for RRH1, RRH2, RRH3 and its aggregation

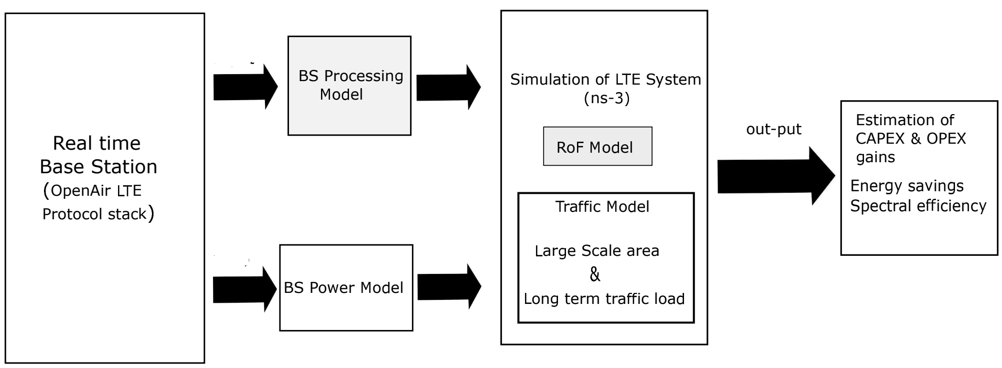


Figure 3.11: C-RAN System Implementation

In Fig. 3.8 shows temporal variations in the coverage areas of RRH1, RRH2 and RRH3 along with resource aggregation usage of VBS. If we observe the resource usage doesn't exceed the full capacity of all 3 RRHS. So, we can aggregate the processing of RRH1, RRH2 and RRH3 into one VBS. By reducing the individual processing of BSs for each RRH and aggregating RRHs into one VBS, we reduce the CAPEX costs by allocating the processing resources efficiently.

## 3.7 System Implementation

To achieve multiplexing BBU pool gains, we have implemented the system showed in Fig. 3.11. For defining the peak processing load we have profiled the openair code and estimated the maximum processing capacity. We used the processing capacity statistics of openair code for clustering algorithm.

Based on the processing capacity statistics, we designed the simulation system with a large geographical area and long term traffic profiles. In simulation system, we use clustering algorithm for clustering BSs into sets (i.e., formation of virtual base stations). The periodicity of running the clustering and scheduling algorithms is decided on rate of increase in traffic demands. We often use terms partition and cluster interchangeably.

Once cluster formation has been done, each cluster is assigned to one VBS at time stage t. Let us say t=0. As time changes t > 0, some events like existing users move onto other VBS coverage region, new users joining the network etc, can happen. As the result of events traffic demands in the network can increase or decrease. If traffic demands increase, we need to find another VBS which can adjust the load with existing. If we cannot find a VBS to adjust load, then a new VBS has to switch-on for accommodating traffic load. In-case of traffic load decrease, we need to adjust the load on neighbor VBS and if any VBS load processing becomes empty in these process, we can switch-off the empty VBS.

The most important factor is defining the time interval for traffic demands monitoring. The algorithm behavior is based on defining the time interval of traffic monitoring. However, defining the time interval is depending on the geographical conditions. So, we have to use different time intervals the algorithm to deal with different geographical conditions. For example, we want to define the time interval for commercial areas, we should use very small intervals like 10 min, 2 min etc., as the dynamics of traffic will be higher. Where as if we need to define time interval for traffic monitoring in rural areas, we can use larger time intervals like 5 hours, 10 hours etc., as there won't be much frequent change in traffic demands are observed. However, considering past traffic demands in defining the time interval will increase its accuracy.

The work flow of C-RAN framework is shown below:

- Step 1: Specify pseudo traffic load requests on each BS
- Step 2: Perform clustering of RRHs.
- Step 3: Each cluster of RRHs and corresponding higher layer processing is treated as VBS
- Step 4: Schedule each VBS onto processor.
- Step 5: Track the load processing on each VBS and update.
- Step 6: Specify the periodicity of load reporting on traffic loads
- Step 7: If there is increase in traffic load
  - Verify load adjustment on each VBS to accommodate
  - If a VBS can accommodate then select the VBS
  - Else schedule a new VBS
- Step 8: If there is decrease in traffic load
  - Verify load adjustment for aggregation on neighbor VBS
  - If a neighbor VBS can accommodate then aggregate VBS and switch-off running VBS

## 3.8 Experimental Results

We evaluate the performance of the proposed dynamic clustering algorithm with two different scenarios of traffic loads and compared the results of proposed approach with conventional BS deployment for BBU utilization and power consumption. Simulation parameters are shown in Table 3.6. The number of RRHs used are 40 and coverage area of each RRH is a circular layout. Therefore, 40 BSs are needed for the conventional BS deployments.

Table 3.6: Simulation Parameters

| | |
|---|---|
| Number of Base stations | 40 |
| Number of RRHs | 40 |
| RRH coverage region | Circular |
| Geographical regional | Urban/sub-urban |
| Simulation time | 240sec |
| Traffic profiles | a) office b) residence |
| Traffic profiles periodicity | 1 hour |
| Time interval for multiplexing | 1 ms, 5 ms and 10 ms |
| Upper limit on BBU load | 0.8 |
| Constant BBU load | 0.2 |

We generate two typical traffic load profiles as input data. One is commercial/office area as showed in Fig 3.12 considered as a frequent traffic varying area. Another one is a residential area where traffic variations are not frequent as show in Fig 3.13. The traffic load profile for residential areas becomes peak during night times at 22-24 hours. In Fig 3.12, 3.13 the vertical axis represented as traffic load. The traffic profile has been produced based on random traffic generations of traffic.

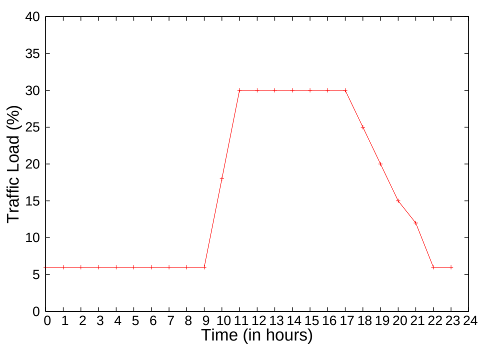


Figure 3.12: Traffic profile data for office

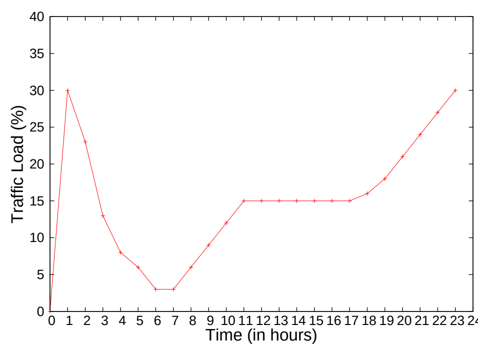


Figure 3.13: Traffic profile data for residence

In our simulations, time interval for multiplexing BS and generating traffic profiles in the proposed algorithms is fully dynamic with short intervals (1 hour). The time intervals are chosen in such a way that the proposed approach will provide higher performance gains. The upper and constant limit for utilization of computing resource usage is 0.8 and 0.2, respectively. That means the proposed schemes are applied for optimizing computing resource utilization when processing of each BS is less than the upper limit and greater than constant load processing.

The simulations of traffic loads are done offline for a period of 240 sec and every 10 seconds is mapped as 1 hour traffic. The results are used for plotting the trend of traffic load in office/residence regions. To avoid intentional imbalance in traffic loads for multiplexing, we ran the experiments with

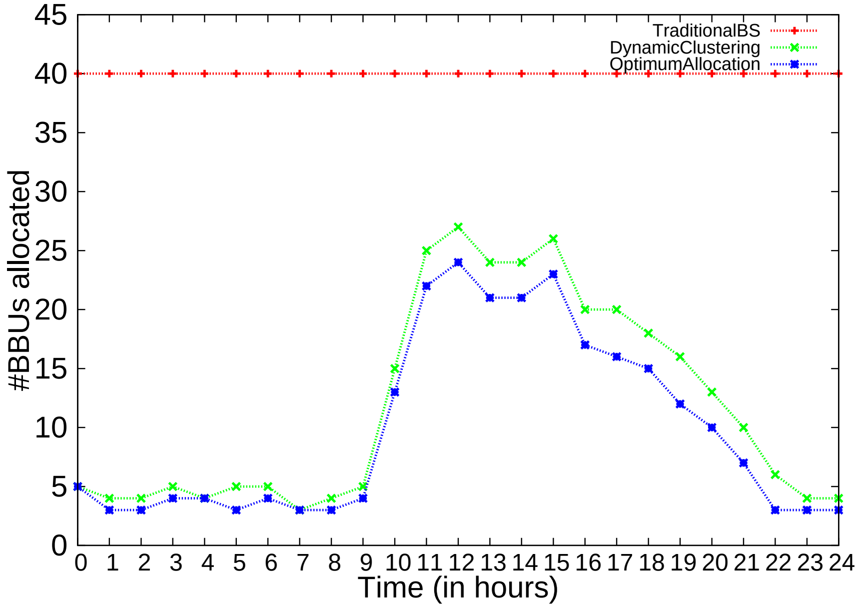

Figure 3.14: The number of BBUs assigned for high load areas (office).

varying user distributions in cell sites.

For comparison of dynamic clustering, we implemented optimum clustering algorithm. In Optimum clustering algorithm, all BSs processing requirements are optimized to full capacity of each VBS and fractional processing of BSs in different VBS is allowed. For example, in optimum clustering algorithm if a BS requires 20 MFLOP, we can process 17 MFLOP on one VBS and 3 MFLOP on another VBS.

Fig 3.14 shows the assigned number of BBUs for dynamic clustering, optimum clustering and traditional schemes for traffic loads of office and residential areas. The aggregated number of BBUs to the dynamic clustering scheme is 27 for office. In case of peak load processing for office areas, maximum number of BBUs has been utilized. The proposed clustering approach takes only 5% more processing for BBUs, but when compared to traditional cellular deployment, the proposed reducing nearly 27%. The 5% low processing of proposed approach compared with optimum approach is because of higher limit on processing of BBU.

Fig 3.15 shows the BBUs assigned for processing in residential areas using dynamic clustering, optimum clustering and traditional schemes. The aggregated number of BBUs to the dynamic clustering scheme is 23 for residence and in optimum it is 21. Both optimum and dynamic clustering approach used the maximum number of BBUs for peak load processing. The proposed clustering and optimum approaches are nearly performing same as beginning hours of the day. The proposed approach is showing nearly 36% reduced number of BBUs processing compared to traditional LTE deployment.

As showed in Fig. 3.16, power saving of dynamic clustering will also vary with the number of assigned of BBUs. For calculating power savings of BBU, we have used the same scenarios of office and residential traffic loads. Comparing optimum approach, our proposed dynamic clustering approach performs nearly 7% $\sim$ 8% less. When compared to traditional BS layout, our approach gives an average savings of 30% in office areas. In case of residential areas as showed in Fig. 3.17,

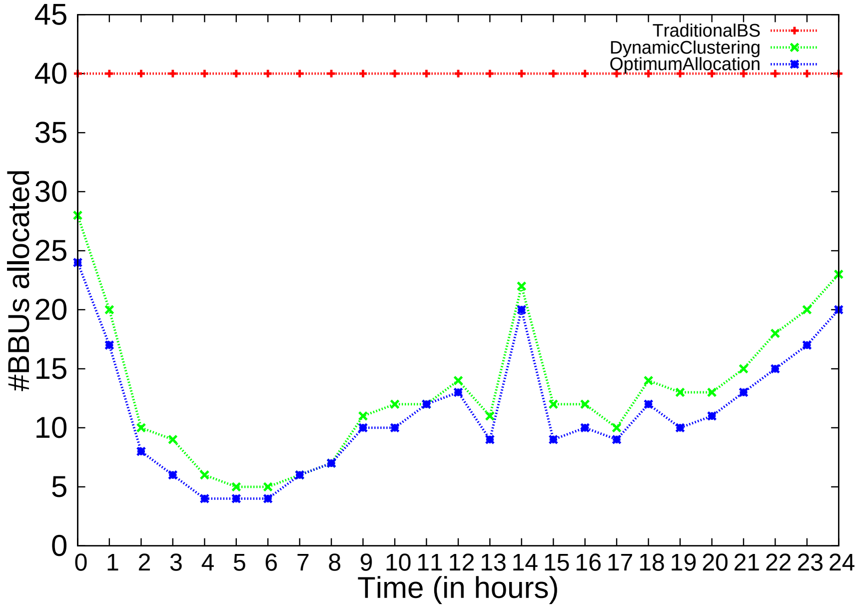


Figure 3.15: The number of BBUs assigned for low traffic load areas (residence).

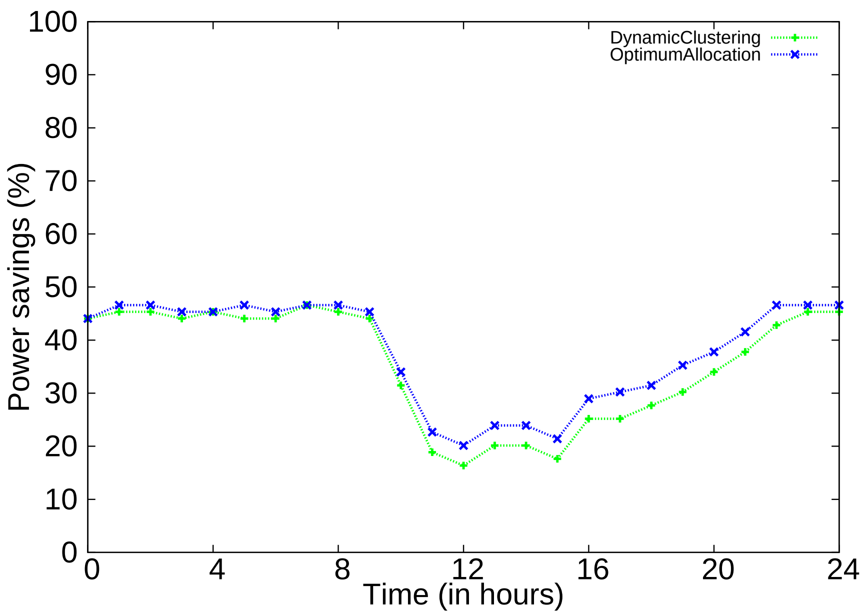


Figure 3.16: Power savings in office areas with Centralization in C-RAN

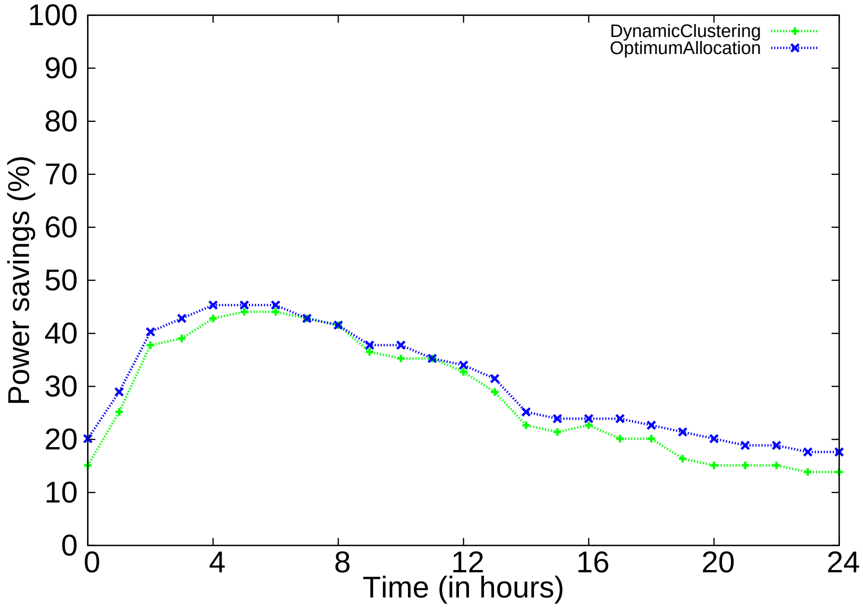


Figure 3.17: Power savings in residence areas with Centralization in C-RAN

our approach gives an average of 34% savings and is only 6% $\sim$ 5% less than optimum approach.

## 3.9 Summary

In this chapter, we proposed a new C-RAN architecture. For realizing C-RAN, we implemented a framework and on top of framework, we intended to evaluate the BBU pooling gains by using BS processing model with dynamic clustering of BSs in centralized architecture of radio access networks. Along with BS processing model, we used BS power model to evaluate the power consumption in C-RAN. With the implemented framework, we have demonstrated that our dynamic clustering approach gives pooling gains of 27% in office areas and 36% in residential areas for CAPEX and OPEX savings. Regarding power consumption, our approach gives an average of 30% in office areas and 34% in residential savings. As RoF has been designed for simulating latencies in C-RAN, its impact on performance of C-RAN has to be studied as separate problem.

# Chapter 4

# Conclusion and Future Work

- We have designed and evaluated the RoF model for simulating the latencies in C-RAN in chapter 2.
- We have designed C-RAN framework with Processing traffic Load model and Power model. Processing traffic Load model simulates the real-time processing requirements of BSs for traffic loads where as Power model simulates real-time power consumption of BSs.
- We evaluated the centralized resource pooling gains using C-RAN framework by proposing dynamic clustering of BSs in chapter 3. Our proposed dynamic partitioning approach gives pooling gains of 27% in office areas and 36% in residential areas for CAPEX and OPEX savings. Our proposed approach, achieves an average of 30% and 34% power consumption savings in office and residential areas.
- Future work
  - RoF model has been designed for emulating the latencies in C-RAN. Existing RoF model is implemented in ns-3, which can be useful for creating models for centralized architectures like creating Heterogeneous network cloud, wifi cloud etc. But, it is not useful for validating the scheduling of BSs on GPPs for validating the real-time latencies on C-RAN. So, we need to use the existing RoF model as prototype for designing RoF on emulator of LTE system.
  - The C-RAN framework allows us to implement various types of clustering algorithms. We just have to replace the dynamic clustering component used in CRC of C-RAN framework. For example, we can implement dynamic clustering algorithm by considering geographical adjacency of BSs and compare with existing clustering algorithms.
  - Different clustering algorithms of BSs are implemented with different objectives. One interesting clustering of BSs approach is evaluating the trade off between power consumption and spectral efficiency.

# References


[1] China Mobile Research Institute. C-RAN: The Road Towards Green RAN. http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf.

[2] 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 Proceedings of the 18th annual international conference on Mobile computing and networking. ACM, 2012 125–136.

[3] N. J. Gomes, M. Morant, A. Alphones, B. Cabon, J. E. Mitchell, C. Lethien, M. Csörnyei, A. Stöhr, and S. Iezekiel. Radio-over-fiber transport for the support of wireless broadband services[Invited Paper]. volume 8. OSA, 2009 156–178.

[4] A. Koonen, M. Larrode, A. Ng'oma, K. Wang, H. Yang, Y. Zheng, and E. Tangdiongga. Perspectives of Radio over Fiber Technologies. In Optical Fiber communication/National Fiber Optic Engineers Conference, 2008. OFC/NFOEC 2008. Conference on. 2008 1–3.

[5] B. Kalantari-Sabet, M. Mjeku, N. Gomes, and J. Mitchell. Performance Impairments in Single-Mode Radio-Over-Fiber Systems Due to MAC Constraints. volume 26. 2008 2540–2548.

[6] Misc. 3GPP LTE standard. https://www.nsnam.org/.

[7] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. di Girolamo, and G. Giuliani. RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network. In Future Network and Mobile Summit (FutureNetworkSummit), 2013. 2013 1–8.

[8] C.-L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li. Recent Progress on C-RAN Centralization and Cloudification. volume 2. 2014 1030–1039.

[9] G. Karagiannis, A. Jamakovic, A. Edmonds, C. Parada, T. Metsch, D. Pichon, M. Corici, S. Ruffino, A. Gomes, P. Secondo Crosta, and T. Bohnert. Mobile Cloud Networking: Virtualisation of cellular networks. In Telecommunications (ICT), 2014 21st International Conference on. 2014 410–415.

[10] P. Marsch and G. P.Fettweis. Coordinated Multi-Point in Mobile Communications From Theory to Practice. In Cambridge University. Press, 2011 139–159.

[11] S. Namba, T. Warabino, and S. Kaneko. BBU-RRH switching schemes for centralized RAN. In Communications and Networking in China (CHINACOM), 2012 7th International ICST Conference on. 2012 762–766.

[12] X. Chen, N. Li, J. Wang, C. Xing, L. Sun, and M. Lei. A Dynamic Clustering Algorithm Design for C-RAN Based on Multi-Objective Optimization Theory. In Vehicular Technology Conference (VTC Spring), 2014 IEEE 79th. 2014 1–5.

[13] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic. Baseband Processing Units Virtualization for Cloud Radio Access Networks. *Wireless Communications Letters, IEEE* 4, (2015) 189–192.

[14] G. Zhai, L. Tian, Y. Zhou, and J. Shi. Load diversity based optimal processing resource allocation for super base stations in centralized radio access networks. volume 57. Science China Press, 2014 1–12.

[15] K. Lakshmanan, S. Kato, and R. Rajkumar. Scheduling Parallel Real-Time Tasks on Multi-core Processors. In Real-Time Systems Symposium (RTSS), 2010 IEEE 31st. 2010 259–268.

[16] G. Borriello and D. Miles. Task scheduling for real-time multi-processor simulations. In Real-Time Operating Systems and Software, 1994. RTOSS '94, Proceedings., 11th IEEE Workshop on. 1994 70–73.

[17] G. Auer, V. Giannini, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, C. Desset, and O. Blume. Cellular Energy Efficiency Evaluation Framework. In Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd. 2011 1–6.

[18] EURECOM. Open Air Interface. `http://www.openairinterface.org/`.

[19] S. Sesia, I. Toufik, and M. Baker. LTE, The UMTS Long Term Evolution: From Theory to Practice. Wiley Publishing, 2009 .

[20] C. Berrou and A. Glavieux. Near optimum error correcting coding and decoding: turbo-codes. volume 44. 1996 1261–1271.

[21] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan. Load Balancing in Downlink LTE Self-Optimizing Networks. In Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st. 2010 1–5.

[22] G. Auer, V. Giannini, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, C. Desset, and O. Blume. Cellular Energy Efficiency Evaluation Framework. In Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd. 2011 1–6.