COORDINATED MULTIPOINT-AWARE BASEBAND UNIT PLACEMENT IN CLOUDRAN

by

Anıl Üzümcüoğlu

B.S., Computer Engineering, Boğaziçi University, 2014

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2019

COORDINATED MULTIPOINT-AWARE BASEBAND UNIT PLACEMENT IN CLOUDRAN

APPROVED BY:

Prof. Tuna Tuğcu ..... (Thesis Supervisor)

Assoc. Prof. Ali Emre Pusane .....

Assist. Prof. Atay Özgövde .....

DATE OF APPROVAL: 03.01.2019

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisor Prof. Tuna Tuğcu for his endless support and patience, his invaluable guidance and suggestions throughout the study.

I am grateful to Assoc. Prof. Ali Emre Pusane and Assist. Prof. Atay Özgövde for their participation in my thesis committee and for their precious feedback.

I would like to offer my deepest thanks to the NETLAB members and all the participants of our research group; Tuğrul, Sina, Salih, Kerim, Buğra, Yunus and Birkan for their hard work, support and refreshing ideas.

My friend Özgür also deserves a mention for always being there whenever I needed to take a small break. Without him, this thesis would probably have been completed six months earlier.

I cannot begin to express my thanks to my family for their unconditional love, care and support, especially during the most challenging parts.

Lastly, Zeynep; thank you for being my source of motivation through this journey.

This work is supported by the Turkish Ministry of Development under the TAM Project, number 2007K120610.

ABSTRACT

COORDINATED MULTIPOINT-AWARE BASEBAND UNIT PLACEMENT IN CLOUDRAN

Cloud Radio Access Network (CloudRAN) is proposed as a candidate architecture for future mobile networks in order to meet the challenging requirements presented in 5G specifications. The key rationale behind CloudRAN is two fundamental principles; the first one is splitting the base station functionality into two as radio and baseband processing, and the second one is grouping those baseband units (BBU) into centralized and virtualized data centers called BBU pools. CoMP (Coordinated Multipoint) is an advanced radio coordination technique taking advantage of using multiple base stations for a single user equipment to improve signal quality and reduce interference. Despite being effectively utilized in current mobile networks, integrating CoMP to the CloudRAN, especially for joint transmission and reception modes, reveals both conveniences and challenges. The most important problem exposes itself as increased pressure over bandwidth requirements due to exchange of raw IQ signals between virtual BBUs. In this study, the placement of virtual BBUs into physical nodes of BBU pool is explored to discover efficient allocation methods with respect to given CoMP sets. The method for creating an efficient allocation follows the intuition that data exchange between two BBUs causes less stress over the network resources of the BBU pool when they are hosted on the same physical machine, as opposed to exchanging data across different machines. The optimal placement around this insight is formulated and seeing the computational complexity of finding the optimal solution, a heuristic method is developed to produce close to optimal results in polynomial time.

ÖZET

BULUT RADYO ERİŞİM ŞEBEKELERİNDE KOORDİNE ÇOKLU NOKTA İLİŞKİLERİNE GÖRE TEMEL BANT ÜNİTESİ YERLEŞİMİ

5G tanımında sunulmuş, uyulması gereken zorlayıcı şartlara ayak uydurabilmek için tasarlanan gelecek nesil mobil şebekelere aday gösterilen mimarilerden biri de Bulut Radyo Erişim Şebekeleridir (CloudRAN). Bu mimari iki temel ilke üzerine kuruludur; ilki baz istasyonu fonksiyonlarının radyo ve temel bant işlemleri olarak ayrıştırılması, ikincisi ise şebekedeki bütün temel bant ünitelerinin (BBU) temel bant ünite havuzu (BBU pool) diye adlandırılan merkezi bir birimde sanallaştırılarak gruplanmasıdır. Koordine çoklu nokta tekniği (CoMP) gelişmiş derecede bir radyo koordinasyon tekniğidir. Sinyal kalitesini arttırmak ve hücreler arası girişimi azaltmak amacıyla tek bir mobil cihaz için birden fazla baz istasyonu kullanma prensibine dayanır. Şu anki mobil şebekelerde aktif bir şekilde kullanılmasına rağmen, ortaklaşa iletim ve alım modları Bulut Radyo Erişim Şebekelerine uyarlandığında hem kolaylıklar hem de sorunlar ortaya çıkmaktadır. En önemli problem de işlenmemiş IQ sinyallerinin sanal temel bant üniteleri arasında karşılıklı değişiminden kaynaklanan, bant genişliği üzerindeki baskının daha da artmasıdır. Bu çalışmada, daha verimli yerleşim metodlarına ulaşmak için, sanal temel bant ünitelerinin havuzdaki fiziksel makinelere birbirleri arasındaki koordine çoklu nokta ilişkileri gözetilerek paylaştırılması konusu araştırılmaktadır. Verimli bir yerleşme planı için varsayılan yaklaşım, aynı fiziksel makinedeki sanal üniteler arasındaki iletişimin, farklı fiziksel makineler arasında olana kıyasla havuzun şebeke kaynaklarını daha az kullanacağı yönündedir. Bu önseziden yola çıkarak en iyi yerleşim planı matematiksel olarak formüle dökülmekte, ardından en iyi sonuca ulaşmanın hesaplama zorluğunun görülmesiyle en iyiye yakın sonuçları daha makul zamanlarda üreten bir metod geliştirilmektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$C(t)$	Calculated cost of the placement for the given time interval t
$L_i(t)$	Total network load coming from cell site i for the given time interval t
t	A single timeframe in the baseband unit pool
U	Set of remote radio heads
V	Set of virtual baseband units
W	Set of physical nodes in the baseband unit pool
W_{idle}	Set of physical nodes in the baseband unit pool which are turned off
β_i	Bandwidth capacity of physical node i
$\gamma_i(t)$	Binary variable indicating an act of migration for virtual baseband unit v_i at given time interval t
θ	Experimental weight coefficient for number of migrations
λ	Signal exchange factor for the current baseband unit placement
$\mu_{ij}(t)$	Binary variable indicating if signals coming from remote radio head u_i are processed in virtual baseband unit v_j at given time interval t
$\rho_{ij}(t)$	Binary variable indicating if virtual baseband unit v_i is placed in physical node w_j at given time interval t
$\phi_{ij}(t)$	Signal exchange cofactor between virtual baseband units v_i and v_j at given time interval t
ψ	Experimental weight coefficient for ratio of idle physical nodes

LIST OF ACRONYMS/ABBREVIATIONS

5G	Fifth generation cellular mobile networks
BBU	Baseband Unit
BFD	Best Fit Decreasing
BS	Base Station
CB	Coordinated Beamforming
CloudRAN	Cloud Radio Access Network
CoMP	Coordinated Multipoint
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CRAN	Cloud Radio Access Network
CS	Coordinated Scheduling
DPS	Dynamic Point Selection
eNodeB	Evolved Node B
FFD	First Fit Decreasing
H-CRAN	Heterogeneous Cloud Radio Access Network
IoT	Internet of Things
IP	Internet Protocol
JP	Joint Processing
JR	Joint Reception
JT	Joint Transmission
LTE	Long Term Evolution
MIMO	Multiple Input Multiple Output
QoS	Quality of Service
RRH	Remote Radio Head
RX	Reception
TP	Transmission Point
TX	Transmission
UE	User Equipment

vBBU	Virtual Baseband Unit
VM	Virtual Machine

1. INTRODUCTION

Mobile Internet traffic is continuously rising, even faster than the fixed Internet traffic. The increasing number of smart phone and tablet users, coupled with the recent developments in the Internet of Things (IoT) segment suggests that the growth will continue with a soaring rate [2]. On the other hand, the average user expectation in terms of quality of experience from Internet applications has raised. The overall situation puts a constant pressure on the telecommunications industry to improve the network capacity and prepare for the challenging requirements.

The next generation of mobile networks, commonly referred as 5G, reflects on the future expectations by setting up high-level targets. The mostly settled requirements include sub-millisecond latency values, 100-fold increase in both average data rate and the number of connected devices compared to the current generation, and significant savings on energy and cost [3].

Cloud Radio Access Network (CloudRAN or CRAN) is a novel approach for a better mobile network architecture, in order to answer the mentioned challenges and requirements. The essence of CloudRAN as described in [4] is splitting the typical base station functionalities into radio and baseband processing, as remote radio heads (RRH) and baseband units (BBU), respectively. Moreover, it aims to centralize the baseband processing in a virtualized BBU pool and dynamically share it among the sites, consequently enabling the network operator to adapt to non-uniform traffic and utilize the BBU resources efficiently.

Since Long Term Evolution (LTE) Release 11, coordinated multipoint (CoMP) transmission and reception methods are a part of the specification [5]. It involves using multiple base stations for a user equipment (UE) utilizing advanced signal combining methods and scheduling decisions in order to reduce inter-cell interference, decrease the power consumption in signal transmission and organize a better resource allocation. However, CoMP leads to additional data cost in the form of signal exchange between

base stations, particularly in joint processing case. Compared to LTE architecture, CloudRAN provides a much more suitable environment for deploying different CoMP scenarios due to the centralization of the virtual BBUs, offering higher bandwidth and lower latency on the data path. Nonetheless, the volume of data that comes into view once CoMP is enabled causes a great strain on the system, especially when RRHs send demodulated IQ data over Common Public Radio Interface (CPRI) in CloudRAN. Therefore, there is still a lot of space to explore on the optimization of this data exchange within a BBU pool.

1.1. Contribution of Thesis

Given all the CoMP sets that need to be formed in a BBU pool for joint processing, the placement of the virtual BBUs can be arranged with respect to their CoMP memberships in such a way that virtual BBUs of the same set are placed into same physical machines in the data center, or at least in close proximity. An optimal placement minimizes the data traffic across different physical machines in the pool and most of the exchange happens within the machine, with superior bandwidth and latency. However, finding the optimal placement proves itself to be a variation of bin packing problem belonging to the set of NP-hard problems and cannot be applied in online manner. Therefore, this study tries to find a sub-optimal heuristic approach to the placement problem and compares it with the optimal solution.

Our key contributions in this thesis are summarized as below:

- Formulating the virtual BBU placement problem in BBU pools based on their CoMP set memberships
- Developing a sub-optimal heuristic method as a variation of best fit decreasing approach to the bin packing problem
- Implementing a discrete event based BBU pool simulator to analyze the performance of the heuristic method
- Conducting simulation based experiments to investigate the impact of different system parameters

2. RELATED WORK

Improving the resource utilization in a CloudRAN architecture is a popular subject, and various studies have been conducted not necessarily on the network resource, but also on processing power, bandwidth occupation and green energy consumption. In [6], the authors try to minimize the overall operational cost of a BBU pool in which BBUs, as virtual machines, are considered as the primary cost factor. Considering the maximum UE count served by each RRH and fronthaul capacity as constraints, they formulate the problem as a mixed non-linear integer problem to optimize active virtual machine count in the system. The study in [7] tackles on joint processing scenarios in CoMP enabled CloudRAN deployments and tries to propose efficient resource allocation methods. As well as optimizing the BBU processing in the BBU pool, radio spectrum allocation on the RRH side and wavelength bandwidth allocation on the fronthaul side are investigated.

A lot of research focuses on possibilities of dynamic mapping between the RRHs and BBUs as it is one of the most important implications of the centralized and virtualized architecture of CloudRAN. Khan *et al.* introduce a novel approach in [8] to dynamically update RRH-BBU mapping so that lightly loaded BBUs can offload their tasks to the others and can be turned off, resulting in less energy consumption. A genetic algorithm is used in [9] to find the RRH-BBU mapping that would produce the maximum QoS (Quality of Service) measured by the number of blocked calls in the network. Their solution is further examined by [10], concentrating on the side effect that load is unequally distributed over BBUs and offers a particle swarm optimization algorithm, which proves itself to be helpful. Another research [11] also works on the load balancing problem in CloudRAN but their approach involves the UE-RRH mapping instead of RRH-BBU mapping. Although RRHs are stationary, they show that by increasing/decreasing the transmission and reception power levels of the UEs and RRHs, they can manually trigger handoffs and regulate the number of UEs per RRH. Similarly in [12], the set of UEs and the beamforming decisions for each RRH are analyzed. Chen *et al.* endeavor with the concept of mapping one RRH into multiple BBUs

in [13]. They utilize multiple mappings in order to provide dual connectivity to the RRHs so that handoffs are easier and the network is more resilient to BBU failures.

In [14], a more thorough analysis is presented on resource utilization. In addition to the BBU-RRH mapping, their utilization framework also covers TX (transmission)/RX (reception) power level regulation and dynamic BBU sizes, thanks to the virtualized BBU pools in the CloudRAN. They propose both proactive and reactive measures for a complete dynamic provisioning to deal with the fluctuations in the network. A re-configurable fronthaul for CloudRAN is proposed in FluidNet [15]. It aims to handle dynamic traffic load by defining a one-to-many mapping algorithm from RRHs to BBUs. There is an intelligent controller in the BBU pool of FluidNet, which constantly re-configures the BBU pool allocation according to the network feedback. On the physical level, an optical switching element is used to enable/disable a RRH-BBU connection. For cache-enabled Cloud-RAN, a content-centric transmission design is presented in [16]. The proposed scheme clusters the base stations according to the content they are being requested to serve, and these clusters cooperatively serve the same group of users requesting the same content by multicasting.

The burden of CoMP by means of multiplied data load has always been a problem in a mobile network, even before CloudRAN. For instance Zhao *et al.* in [17] deal with the same problem: data exchange introduced by CoMP, but in a traditional LTE architecture. Baseband processing is still performed in the base stations in the LTE and the only way of exchanging collected signals is over the S1 links between eNodeBs (Evolved Node B). There is no way to optimize eNodeB placement unlike the case in this thesis, therefore they steer their attention to minimizing CoMP sets instead. They present a way to require much less data exchange as a result of narrowed CoMP sets at the cost of a marginal decrease in the CoMP performance. The study in [18] also deals with the same problem in LTE. The suggested solution is triggering calculated handoffs on the UEs to change their serving base station so that CoMP sets can be re-adjusted as efficient as possible.

The CloudRAN architecture on the other hand, opens a whole new realm of possibilities for research in CoMP problems. For the cause of a more green energy footprint, Zeng *et al.* in [19] develop a method for creating CoMP sets in CloudRAN so that green energy powered RRHs (e.g. via solar panels) are prioritized over brown energy powered ones. In order to alleviate the effect of additional network load CoMP brings, the analysis in [20] focuses on reducing the amount of coordination traffic between BBUs of same CoMP set using a particle swarm optimization, without sacrificing too much. The authors of [21] choose to work on the same problem, but on a H-CRAN (Heterogeneous Cloud Radio Access Network) environment where heterogeneous networks and CloudRAN are integrated. In the interest of reducing the load, they exploit the joint processing opportunities at the centralized BBU pool to employ a distributed compression algorithm.

Khorsandi *et al.* delve into the location aspect to improve the availability in CloudRAN [22], but they focus on the whole radio access network dealing with the location of the BBU pools, as opposed to looking at how BBUs are distributed within a single pool. Similarly in [23], the optimal location for the centralized BBU pool is calculated with the help of a k-means centroid finder algorithm, and in [24], the same problem is solved using a particle swarm optimization to enhance spectral efficiency. When placement of BBUs, virtualized or not, within the BBU pools is considered, a study focusing on energy efficiency investigates three different architectures to minimize aggregated infrastructure power [25]. Their proposals are related to the network architecture and focus on the physical perspective, determining the way IP (Internet Protocol) and CPRI traffic are routed within the BBU pool. In [26], a set of BBU placement rules are developed to provide higher resiliency for processing and network failures.

To the best of our knowledge, there has not beed a former study about how BBUs should be allocated into the physical nodes in the BBU pool with respect to their CoMP set relations. Having said that, if this problem is projected into another domain, some research can be found about cloud data centers that deal with virtual machine placement into physical hosts based on their mutual traffic [27–30]. However,

in those studies, expected communication traffic between two virtual machines is based on either traffic patterns derived from the prior network footprints of the data center, or ambiguous inclinations of virtual machines that are likely to exchange traffic. The main advantage of exploring this problem in BBU pools is having access to clearly defined set of traffic exchange rules in the form of CoMP relations.

3. CLOUDRAN AND COMP

3.1. CloudRAN Overview

As the mobile network demand grows, current LTE architecture does not stand a chance without a significant paradigm shift as long as the 5G requirements are concerned. Improving the spectral efficiency with more advanced multiple input multiple output (MIMO) and beamforming technologies is not as effective, as it is getting close to the theoretical Shannon limit and now in the diminishing returns phase [31]. Exploiting the holes in the time and frequency spectrum with cognitive radio techniques is always an option, but still unreliable. The number of cells could be increased while making them smaller as microcells or picocells, but the interference issue remains to be solved. CloudRAN is a paradigm shift in the mobile network architecture that tries to solve these issues.

In the traditional LTE layout, a typical base station is responsible for both radio and baseband processing. The baseband processing is completed in the station, and S1/X2 interfaces connect the station into the core network and other base stations, respectively. CloudRAN drammatically changes this design by separating the radio unit and the signal processing unit as remote radio head and baseband unit and then putting all the BBUs together in a centralized, virtualized pool shared between all RRHs as can be seen in Figure 3.1. These two units are connected with optical fiber or microwave and the distance could be as far as 40km [1]. BBUs can be housed in a more convenient location, like a data and operations center with lower rental, power, cooling and maintenance costs to provide easier operation for the operator. Similarly, RRHs can be deployed on poles and rooftops with better coverage and cooling solutions. A BBU pool is typically located in a remote site and shared between cell sites. The resource allocation between heavily and lightly loaded base stations could be efficiently optimized thanks to the virtualization.

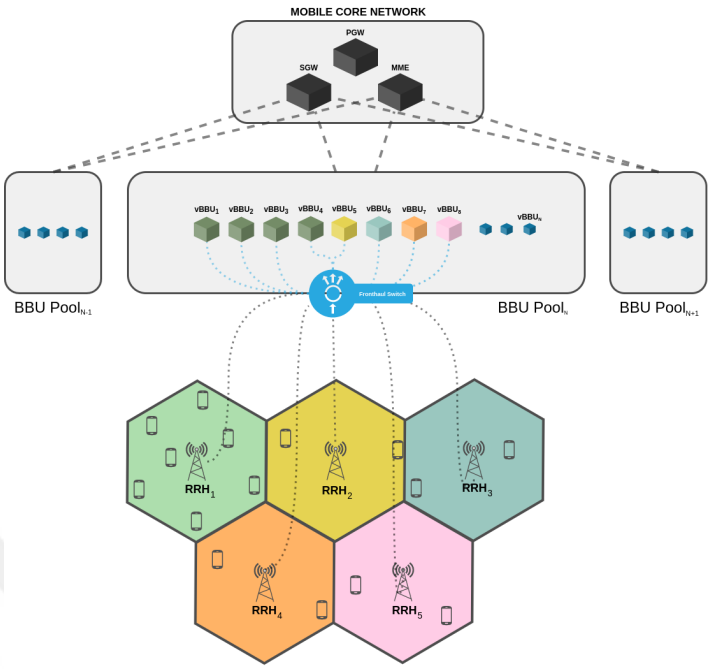


Figure 3.1. CloudRAN architecture

CloudRAN is a promising technology for laying out the infrastructure for the next generation mobile networks. The concepts introduced by the CloudRAN can be utilized in any level, regardless of the underlying technology. Its most valuable advantages are adaptability to non-uniform traffic and scalability. BBUs can be dynamically allocated in such a way that when the population density moves between two points, such as from residential are to commercial area, BBU resources assigned to the responsible RRHs will reflect the change in the load. Following the population shift from Figure 3.1 to Figure 3.2, it can be seen that virtual BBU resources previously allocated for RRH_1 can easily be reallocated for RRH_5 when the network demand calls for it. Designing the infrastructure for the peak hours of each area will no longer be necessary in the CloudRAN as the resource utilization will be maximized. Another advantage is energy and cost consumption. With processing units located in a data center, a significant amount of savings can be obtained in power consumption and cooling resources. Also, BBUs being so close to each other means that the average delay on the X2 interface will be extremely low. This would increase the performance of cross-station operations such as handoff management and CoMP. Last but not least, the maintenance and upgrade tasks in this architecture mean a lot less work. Not only having all BBUs in the same place makes these tasks physically easier, but also a clear separation between the radio

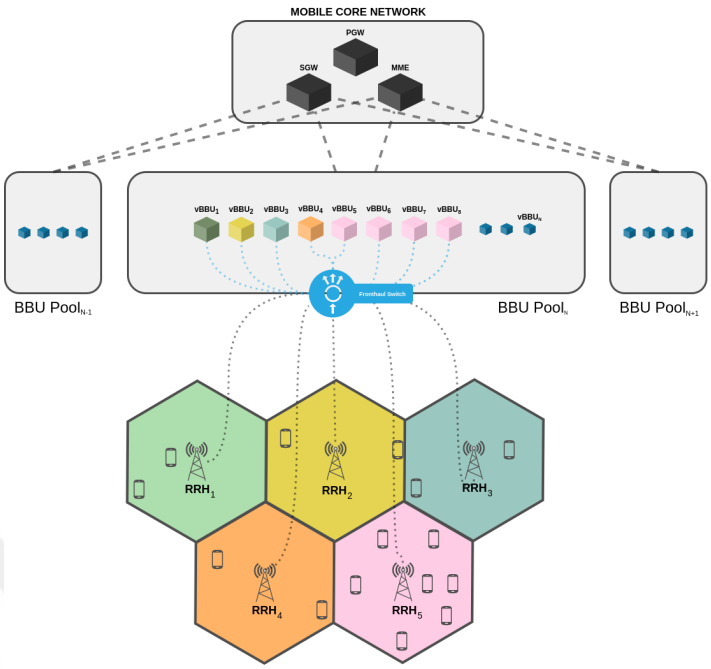


Figure 3.2. CloudRAN architecture after a population shift

and the baseband layer dramatically helps in partial upgrades.

3.2. RRH/BBU Split Problem

Defining the roles of RRH and BBU and where they should be separated has not completely been settled yet, resulting in two prominent models: full centralization and partial centralization [32, 33]. As shown in the Figure 3.3, base station functions start from receiving the radio signal and end in L3 with sending the signal to the core network. The initial architectures are fully centralized. They limit the RRH with only radio processing, leaving L1/L2/L3 processing to the BBUs. However, this means that there will be high bandwidth IQ data transmission between RRH and BBU over the CPRI protocol [34], which starts to become a bottleneck on the physical layer, especially when 5G requirements are considered. This challenge has led to the partial centralization, where L1 processing is moved to the RRH, thus reducing the bandwidth on the optical transport links. This solution provides 20 to 50 times decrease in bandwidth [4] but is less optimal since the resource sharing is reduced and information loss prevents performing advanced features on the BBU, such as CoMP. On the other hand, different alternatives are being worked on which allows flexible functional splits over

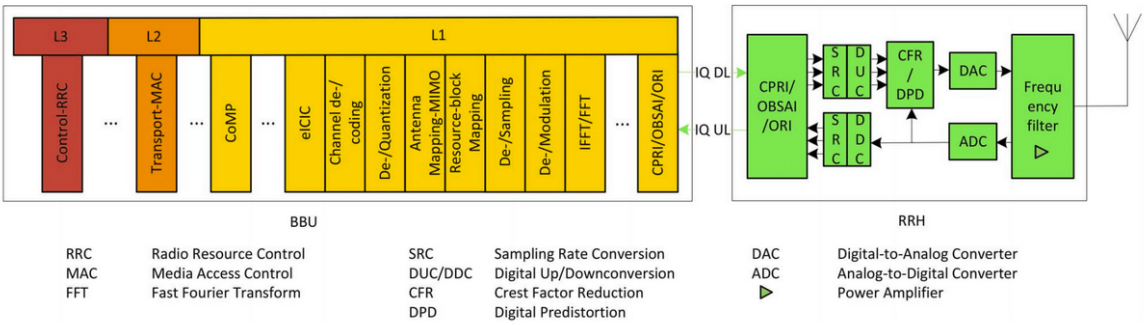


Figure 3.3. Base station functionalities [1]

the fronthaul links depending on the deployment state [35].

3.3. Coordinated Multipoint

During the evolution towards the LTE-Advanced in 4G cellular systems, various new technologies are introduced to battle with the exponential growth in the network demand. Utilizing spatial multiplexing with the help of MIMO, or deploying relay nodes are two examples to such improvements. As LTE-Advanced continues to evolve, another feature is discovered in the form of CoMP, focusing on the performance of the cell edge users in particular. The key aspect of CoMP is taking advantage of the natural consequence that the signals emitted from a single UE can be received by multiple base stations (BS) other than serving BS for that specific UE, especially if it is located close to a cell edge. Although these signals are discarded as noise in the previous generations, there is significant potential in processing these signals by more than one base station and trying to create multiple transmission and reception channels. The first specification set about this new paradigm is introduced to the LTE-Advanced with the Release 11 [5].

In a CoMP enabled cellular architecture, each user equipment is associated with a set of base stations, or transmission points (TP) in general. In order to create and maintain these TP sets, each user equipment is responsible for periodically reporting signal conditions for the TPs in its associated set. The content of these reports includes the channel state information, precoding matrix indicator and the channel quality

indicator [5]. This data is processed for each UE and the set of TPs are continuously updated as signal conditions change.

The methods for exploiting multiple transmission points can be divided into three categories. Coordinated scheduling (CS) and coordinated beamforming (CS/CB) is the most primitive type since only a single TP is responsible for transmission and reception of a UE. However, neighboring TPs in the radio access network communicate with each other to schedule their transmission period to minimize the interference. Based on these scheduling decisions, they can dynamically activate/deactivate transmission, or they can generate nullifying signals towards the area where they cause a serious interference. In dynamic point selection (DPS), the CoMP set for a UE includes multiple TPs, but only one of them is eligible to transmit or receive at any given time. This approach is useful for selecting the TP with the best signal condition, and the serving TP can be changed even in a subframe level [5].

The third CoMP category is joint transmission (JT) and joint reception (JR) where the dynamic nature of CoMP technology is utilized to the full extent. Multiple TPs are actively used and in constant coordination to provide a combined singular interface for UEs. In the downlink, one approach for joint transmission is transmitting the same signal simultaneously from multiple TPs, and consequently amplifying the signal quality on the receiver end, also called as non-coherent combining. Another approach is partially precoding the data in each TP so that they construct the final form in the wireless channel, which is aptly called coherent combining. Joint transmission can also be used for sending interference nullifying signals from the secondary TPs while primary TP sends the actual signal. In the uplink, joint reception is combining multiple TPs in the CoMP set as a unified receiver and it dramatically changes the way received signals are interpreted. In the traditional architecture, any signal other than those that belongs to the primary UEs are considered noise and immediately dismissed. In the joint reception scheme on the other hand, this noise represents a meaningful data if the TP is in the CoMP set of the signal source. A clear motivation for joint receiving is taking advantage of signal combining algorithms to cancel interference based on the footprint of the same signal received from non-primary TPs in the CoMP set [36]. A

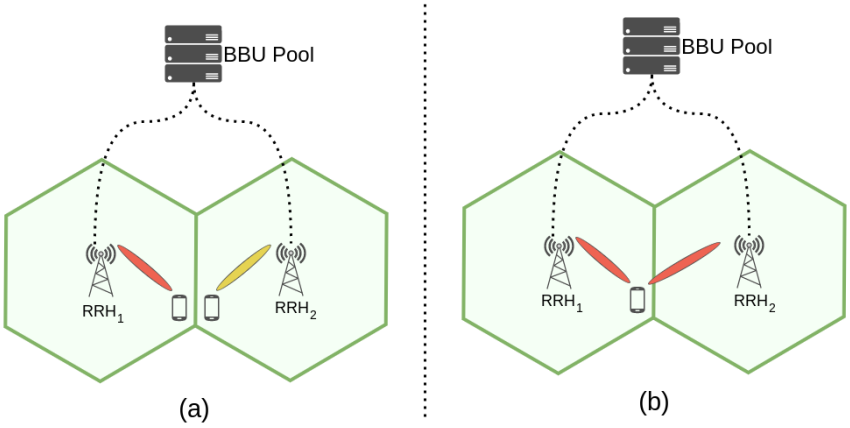


Figure 3.4. CoMP operation modes
a) Coordinated Beamforming/Coordinated Scheduling (CB/CS)
b) Joint Transmission/Joint Reception (JT/JR)

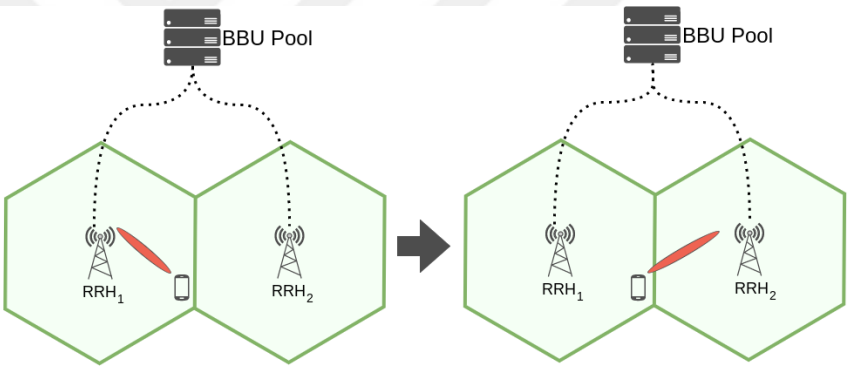


Figure 3.5. CoMP operation modes: Dynamic Point Selection (DPS)

different use case is, despite being in a slightly different domain —low power wide area networks, combining all weak signals from receivers to superimpose a single signal of higher quality [37]. Figures 3.4 and 3.5 summarize these three different CoMP operation modes.

JT/JR type CoMP offers larger gains compared to other types of CoMP implementations [5, 38] but with the cost of higher computation complexity and more sophisticated problems. Incorporating multiple TPs into a unified TP is not straightforward, especially considering the fact that the set of TPs is a different combination for each UE. Coordinating such a complicated behavior and scheduling these operations across the whole network introduces immensely complex computation tasks. Apart from these, joint reception inherently requires the exchange of received signals among

all the TPs in the CoMP set, with the purpose of locally combining them. Also, the most critical challenge to address in joint processing is providing low latency and high bandwidth if the objective is producing promising gains [39]. In the context of 5G expectations, the amount of information exchange at this scale increases the pressure on bandwidth and latency requirements even further. In addition, it should be kept in mind that JT/JR is only meaningful when CloudRAN is deployed with full centralization model in which RRHs are sending unprocessed IQ data without loss of information in the demodulation process.

3.4. Applying CoMP in CloudRAN

CloudRAN is particularly more convenient for CoMP, because the architectural shift it brings completely aligns with the essentials needs of efficient and potent CoMP applications. First of all, decoupling baseband processing and radio units and placing all BBUs next to each other in a centralized pool means a drastic change in the latency and capacity of the X2 interface, which is a much needed feature for CoMP because of the time critical nature of constant information exchange and coordination between BBUs. Furthermore, it is been shown in [40] that the centralized structure of CloudRAN lends itself to operating with bigger CoMP clusters which results in better gains. Another possibility is being able to easily extend the computation capabilities of BBU pools to make room for CoMP control unit functions [41].

One major issue with the CloudRAN and CoMP synergy is the increasing load on fronthaul network in partial centralization model. Even when 5G requirements are already challenging enough for this model, CoMP brings out more data into the picture, which were previously destined to dismiss as noise, but now has become meaningful. Aware of this burden, some research has been done to ease the load on the fronthaul network, proposing different joint processing schemes. For instance the study in [20] tries to reduce the number of feedback signals used for coordination by offering a partial version of joint processing with minimum impact loss.

4. BBU-RRH MAPPING AND BBU PLACEMENT PROBLEM

4.1. BBU-RRH Mapping

Since the main focus of this research is BBU placement within physical nodes in a fully virtualized CloudRAN deployment, the reference architecture on CloudRAN is modified to include a higher resolution in BBU pools. Following the conventional data center architectural designs in modern cloud systems, it can be safely agreed upon that virtualization of the BBUs in the BBU pool decidedly resembles virtual machines in a fixed number of physical machines in a cloud data center. A BBU pool could have ranging number of virtual BBUs (vBBU) based on the current network load, energy consumption criteria, or quality of service requirements, but the actual physical machines and the total amount of available physical resource is limited and constant, unless there is an infrastructural change in the BBU pool physically adding or removing resources.

In the reference architecture in Figure 4.1, five RRHs are connected to a single BBU pool with high throughput optic cables, sending non-processed IQ data over CPRI protocol [34]. The fronthaul switch is responsible for distributing these IQ data to the target vBBUs, guided by the current RRH-BBU mapping of the system. The placement of virtual BBUs within physical nodes on the other hand, depends on the resource management procedure of the BBU pool. Centralized and virtualized structure of CloudRAN allows high dynamism and the configuration within the pool can change at any instant. The changes in the RRH-BBU mapping only takes a single forwarding rule modification in the fronthaul switch, because all the cables are already laid. If the load is decreasing and the demand can be met with fewer vBBUs, then some of the physical nodes can be emptied and shut down, resulting in significant energy savings.

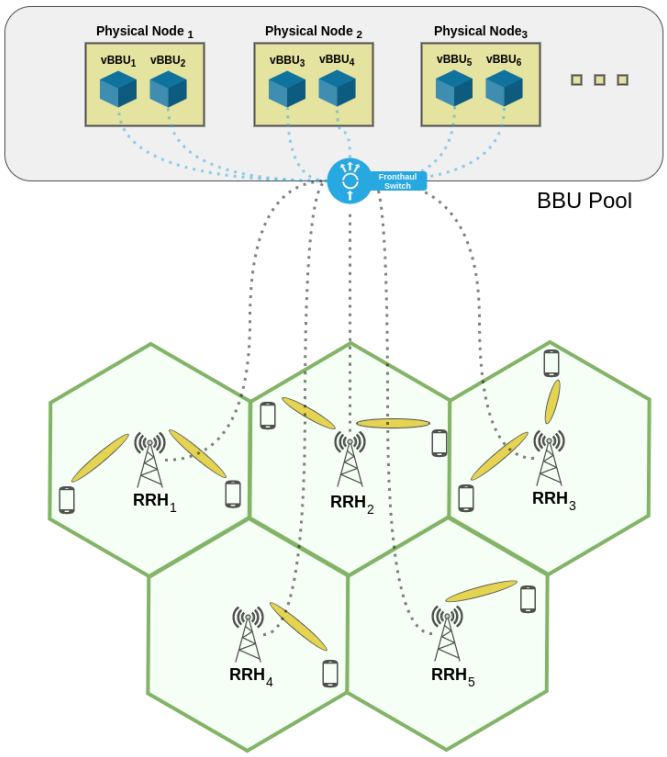


Figure 4.1. CloudRAN architecture with virtual BBUs in the BBU pool

4.2. Effect of CoMP on mapping

Due to signal exchange implications, the complications introduced by CoMP operation modes over this mapping system do not perfectly fit into the picture. Assume a CoMP enabled deployment, where CoMP measurements suggest that there is a substantial gain potential if JT/JR could be applied to RRH_1 and RRH_2 for cell edge users as displayed in Figure 4.2. This means that the signals received by RRH_2 from UE_B are treated as noise in alternative approaches, but now they can be combined with the signals RRH_1 collects. Similarly for UE_E, the signals received by RRH_2 are now meaningful data if they can be combined with those received by RRH_3. Now for the sake of simplicity, assume there is a one-to-one mapping between RRHs and BBUs such that RRH_i sends its IQ data to BBU_i for all i in the system. After CoMP decisions are introduced to vBBUs, new signal exchange routes emerge between them. $vBBU_2$ needs to receive the IQ data from $vBBU_1$ in order to perform a signal combining operation for UE_B. Likewise, it also needs the IQ data sent to $vBBU_3$ for UE_E. On the downlink channel, if joint transmission is to be utilized, both vBBU pairs must coordinate with

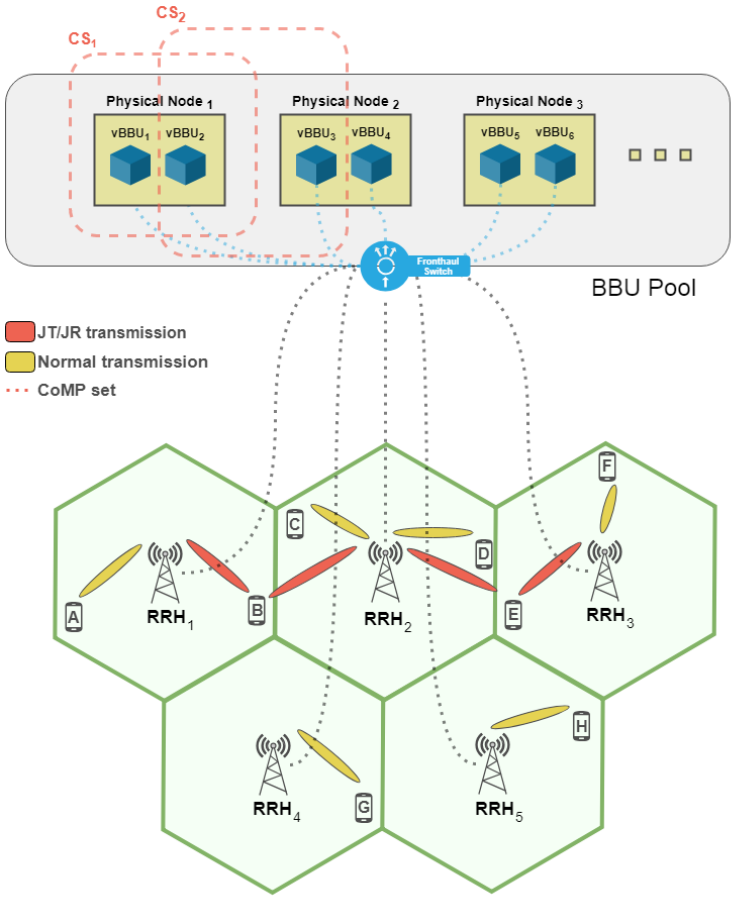


Figure 4.2. CoMP activity for two user equipment

each other to make the transmission in full synchronization. This situation reveals two CoMP sets consisting of vBBUs in which there will be intensive signal exchange and coordination, denoted as CS_1 and CS_2 in the Figure 4.2. Revisiting how the vBBUs are placed in the physical nodes of the BBU pool, it can be seen that all data transfer within CS_1 occurs in the same physical node, whereas $vBBU_2$ and $vBBU_3$ in CS_2 need to communicate across different physical nodes.

It has been elaborately analyzed in [27] and [28] that in traditional data centers, placing VMs with high inter-traffic on the same physical node or at least in close proximity, has considerable effect on the network performance. In Figure 4.2, the placement of vBBUs in CS_2 suffers from the same problem; the communication distance between the members of this set is higher compared to the members of CS_1. IQ data exchange between two vBBUs consumes more network resources if all this data needs to be transferred from one physical node to another. Furthermore, it becomes more

challenging to meet the latency demands for scheduling a joint transmission if the coordination messages have to travel longer distances in CS_2. Arranging the overall vBBU placement within a BBU pool in such a way that vBBUs of the same CoMP set are placed in the same physical node, is an important objective that helps to improve CoMP efficiency in CloudRAN architectures.

4.3. BBU Placement within Physical Nodes

When this subject is investigated in the modern data centers domain, virtual machine placement is a well-known and seasoned problem. As a matter of fact, there are already full-fledged, industrial tools for solving the placement problem for various size of cloud system deployments, including VMware Capacity Planner [42] and Lanamark [43]. However, such an ordinary data center allocation tool would prioritize the overall utilization of the physical nodes in terms of processing units, bandwidth, memory, or power consumption, while the positions of the CoMP sets would be far from optimal and the CoMP performance of the system would be impaired. Virtual machine placement solutions need a CoMP adaptation before it could be used in BBU pools. Nonetheless, it is still an active research area as there are many perspectives to look at this problem. In its very simplistic form, VM placement is a variation of the famous bin packing problem with physical nodes being considered as bins and VMs as elements. The capacity of the bins are usually the central processing unit (CPU) utilization of the VMs, but trying to solve the problem with a multi-dimensional approach by including disk, memory, bandwidth and so forth to the equation proves to be worthwhile [44].

Finding the optimal solution to the bin packing belongs to the memorable class of NP-hard problems [45]. Therefore, using different heuristic methods to the bin packing problem is a common way to approach the placement problem. One of the most straightforward approximation algorithm is first fit decreasing (FFD). This algorithm starts with ordering the elements by their sizes in decreasing order and then puts each element into the first bin it fits. Another variation with a minor modification is best fit decreasing (BFD), in which elements are again sorted by their sizes in decreasing order, then each element is put into the bin that leaves the minimum space. In other

words, each put operation tries to find the bin that has the minimum capacity among those which the current element can fit. Both heuristic methods can be implemented in $O(n \cdot log\ n)$ time complexity, and their distance to the optimal solution are usually in the acceptable range. On the average, the solution found by FFD method is the actual optimal solution in 94.694% of the time, while BFD produces the optimal solution in 94.832% of the time [46]. Due to its ability to obtain a near optimal solution in a quite low computational cost, BFD or its variations with slight modifications are prevalent in virtual machine placement problems when the energy consumption and maximum utilization is of concern [47, 48]. Overall, BFD looks reasonable enough to utilize in the vBBU adaptation of the placement problem, and it will be a part of the proposed heuristic solution of this study.

5. COMP-AWARE BBU PLACEMENT

5.1. Problem Definition

Without loss of generality, we consider a single BBU pool serving a set of RRHs, denoted as U. Each RRH has its own associated cell site, consequently the number of cells are also $|U|$. The set of vBBUs is represented by V, which is determined by a central entity according to the current load and QoS protocols. Although the members of V could change at any moment due to the dynamic nature of CloudRAN, it is regarded as a fixed value in the scope of this study. The set of physical nodes in the BBU pool is denoted by W. The number of physical nodes is again constant, but keep in mind that some physical nodes could be turned off if there are no vBBUs on it at a given time for the sake of green networking. Since the load of each cell fluctuates over time, overall timespan of the analysis is divided into T discrete intervals.

Since the main concern is to reduce the usage of network resources, the capacity of a physical node $w_i \in W$ is defined as total bandwidth capacity of the node, denoted by β_i. The bottleneck for the bandwidth capacity could be dictated by the network interface of the physical node, or the internal switch that distributes the IQ packets to tenant vBBUs.

$L_i(t)$ represents expected overall IQ data load in a cell connected to the RRH $u_i \in U$ during the time interval t.

RRH and vBBU mapping at a given time t is assigned by a central entity. The IQ data collected by an RRH is meaningful for a given vBBU if either that vBBU is the primary serving station, or it is in the CoMP set used for JT/JR. Following this relation, the mapping variable can be defined as follows:

$$\mu_{ij}(t) = \begin{cases} 1, & \text{if IQ data of } u_i \in U \text{ is processed in } v_j \in V. \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

The placement variable ρ is another boolean variable that is used to show which vBBU $v_i \in V$ is placed on which physical node $w_j \in W$ at a given time t:

$$\rho_{ij}(t) = \begin{cases} 1, & \text{if vBBU } v_i \in V \text{ is placed in } w_j \in W. \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

The next step is formulating the signal exchange cost between two vBBUs. If they never appear in the same CoMP set for all RRHs in the topology, then there is no signal exchange between them. Otherwise, the formulation needs two distinct cofactors; the first cofactor is used when they are placed in the same physical node, and the second one is used when they are on different physical nodes. We can define these cofactors as 0 and 1, respectively, since the real objective here is to minimize the placements that would result in signal exchange scenarios of the latter. The final equation for signal exchange factor ϕ as a function of t is:

$$\phi_{ij}(t) = \begin{cases} 0, & \text{if } i = j \\ 0, & \text{if } \mu_{ki}(t) \cdot \mu_{kj}(t) = 0, \ \forall k \in U \\ 0, & \text{if } \sum_{k=1}^{|U|} \mu_{ki}(t) \cdot \mu_{kj}(t) > 1 \text{ and } \rho_{iw}(t) = \rho_{jw}(t) = 1, \ \exists w \in W \\ 1, & \text{if } \sum_{k=1}^{|U|} \mu_{ki}(t) \cdot \mu_{kj}(t) > 1 \text{ and } \rho_{iw}(t) \neq \rho_{jw}(t), \ \forall w \in W \end{cases} \tag{5.3}$$

The problem can be finalized as a minimization problem for the signal exchange between vBBUs on different physical nodes, which produces the following objective function:

$$\text{minimize } \lambda = \frac{\sum_{i=1}^{|V|}\sum_{j=1}^{|V|}\sum_{k=1}^{|U|} L_k(t)\cdot\left[1+\phi_{ij}(t)\big(\mu_{ki}(t)+\mu_{kj}(t)\big)\right]}{\sum_{i=1}^{|U|} L_i(t)} \tag{5.4}$$

Upon further inspection, it can be seen that the computed λ value has a quite intuitive interpretation: It represents how many times an IQ packet has to travel to a physical node on average if each IQ packet is assumed to be of equal size. This simple interpretation is also the main reason behind deciding cofactors as 0 and 1 in definition of ϕ. In a perfect assignment, λ would be equal to 1; in other words, all IQ packets would go into a single physical node in which all the vBBUs in the CoMP set would be conveniently placed together. If the placement gets worse, an IQ packet has to be exchanged between more physical nodes on the average, and λ will increase.

The first constraint to the optimization problem is the physical node capacity. Each physical node should have a fixed capacity that decides how much load it can take, otherwise the obvious solution would be placing all the vBBUs into a single physical node keeping $\lambda = 1$ always. The bandwidth capacity of the physical nodes are used for this metric. The following constraint should be satisfied at each time interval t:

$$\sum_{i=1}^{|U|}\sum_{j=1}^{|V|} L_i(t)\cdot\mu_{ij}(t)\cdot\rho_{jk}(t) \;<\; \beta_k,\ \forall k \in W \tag{5.5}$$

Another constraint demands that all vBBUs are placed in one and only one physical node in order to produce a valid placement:

$$\sum_{i=1}^{|V|}\sum_{j=1}^{|W|}\rho_{ij}(t) = |V| \tag{5.6}$$

and

$$\sum_{i=1}^{|V|}\rho_{ij}(t) = 1,\ \forall j \in W \tag{5.7}$$

The placement algorithm should output the decision variable ρ for each time interval t as a $|V| \times |W|$ binary matrix indicating which vBBU $v_i \in V$ should be placed in which physical node $w_j \in W$ during the interval t.

The decision variable ρ can also be used to extract two more meaningful interpretation. The first one reveals the set of physical nodes $W_{idle} \subseteq W$ which could be shutdown because there are no vBBUs on them for the given interval t:

$$W_{idle} = \left\{ w_i \mid \sum_{j=1}^{|V|}\rho_{ji}(t) = 0 \text{ and } w_i \in W \right\} \tag{5.8}$$

Another result that can be obtained from ρ is the number of vBBUs that needs to be migrated into a different physical node because of the difference between $\rho(t-1)$ and $\rho(t)$. The act of migration can be defined as follows for a single vBBU when

transitioning from $t-1$ to t:

$$\gamma_i(t) = \begin{cases} 0, & \text{if } t = 0 \\ 0, & \text{if } \rho_{ij}(t-1) = \rho_{ij}(t),\ \forall\ w_j \in W \\ 1, & \text{if } \rho_{ij}(t-1) \neq \rho_{ij}(t),\ \exists\ w_j \in W \end{cases} \tag{5.9}$$

Keep in mind that the number of idle physical nodes and number of migrations are not part of the optimization function, its actual objective is reducing signal exchange factor (λ). However, a final cost function can still be defined using experimentally decided normalized weights for signal exchange factor, migration count and the idle physical node proportion:

$$\mathcal{C}(t) = \lambda \cdot \sum_{i=1}^{|U|} L_i(t) + \Psi \cdot \frac{|W_{idle}|}{|W|} + \Theta \cdot \sum_{i=1}^{|V|} \gamma_i(t) \tag{5.10}$$

where Ψ is the weight coefficient for idle physical node proportion and Θ is the weight coefficient for number of migrations.

5.2. Proposed Heuristic Method

The optimization problem defined in Section 5.1 can be solved with a simple implementation, but it takes too long to converge to a solution due to its non polynomial computation complexity as a variation of the bin packing problem. Computation time rises exponentially as the network gets larger to the point that a typical personal computer cannot find the optimal placement in a feasible time for only 20 virtual BBUs. This hindrance prevents it to be the de facto algorithm in BBU pools, where

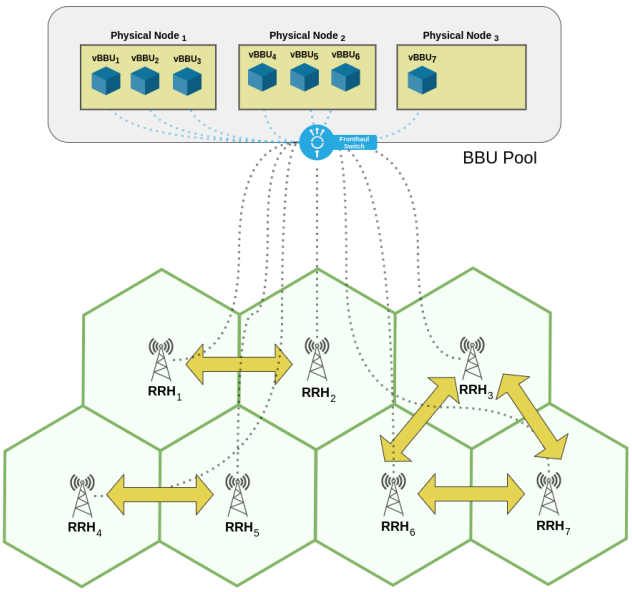


Figure 5.1. CoMP pairings between RRHs

In this configuration, current CoMP sets are: $\{vBBU_1,\ vBBU_2\}$, $\{vBBU_3,\ vBBU_6,\ vBBU_7\}$ and $\{vBBU_4,\ vBBU_5\}$

the evaluation of the placement is a constant need due to environment characteristics. Cumulative network load coming from mobile users, vBBU-RRH mapping, CoMP set configurations, power consumption criteria are all highly dynamic components that could change in any moment, triggering a possible revision in the vBBU placement. This disadvantage leads to developing alternative solutions to the placement problem that provides a decision in polynomial time even if that decision is sub-optimal.

5.2.1. Bin packing adaptation

A natural tendency to find a sub-optimal approach to the placement problem is looking at the prominent heuristics used for the bin packing problem. However, when physical nodes are considered as bins and vBBUs are considered as elements, the placement of vBBUs reveals an essential distinction: vBBUs are not independent. The success of the placement heavily depends on the CoMP set relations between the vBBUs.

The definition of the element could be changed to compensate for the dependencies across vBBUs. Each CoMP pairing between two cells indicate a signal exchange across the relevant serving vBBUs. If these vBBU sets defined by the CoMP pairings are selected as elements for the bin packing problem, then the dependency issue is remedied. Consider the topology given in Figure 5.1 with seven cells and seven vBBUs with one-to-one RRH-vBBU mapping for simplicity. There are five CoMP pairings in the given system, assigned by a central entity according to the UE positions, signal qualities, load balancing schemes and so forth. Those five assignments ($RRH_1 \leftrightarrow RRH_2$, $RRH_3 \leftrightarrow RRH_6$, $RRH_3 \leftrightarrow RRH_7$, $RRH_4 \leftrightarrow RRH_5$ and $RRH_6 \leftrightarrow RRH_7$) form three vBBU clusters: $\{vBBU_1,\ vBBU_2\}$, $\{vBBU_3,\ vBBU_6,\ vBBU_7\}$ and $\{vBBU_4,\ vBBU_5\}$. So, a vBBU cluster could be defined as the set of vBBUs such that each vBBU has at least one other vBBU in the cluster to exchange data. A formulation for membership of a vBBU $v_i \in V$ to a cluster C is as follows:

$$v_i \in C \leftrightarrow \sum_{k=1}^{|U|} \mu_{ki}(t) \cdot \mu_{kj}(t) \geq 1,\ \exists v_j \in C - \{v_i\} \tag{5.11}$$

Now that vBBUs are transformed into clusters as elements in the bin packing algorithm, the next step is defining how their size is represented. If the elements were individual vBBUs, the size of an element would be the expected load arriving to that vBBU and those values would be compared with physical node switch bandwidth capacities during the execution. When sets of vBBUs are used instead, a cluster size would be the cumulative network load for all vBBUs in that set. The size of a cluster C then becomes:

$$\sum_{i=1}^{|U|} \sum_{j=1}^{|C|} L_i(t) \cdot \mu_{ij}(t),\ v_j \in C \tag{5.12}$$

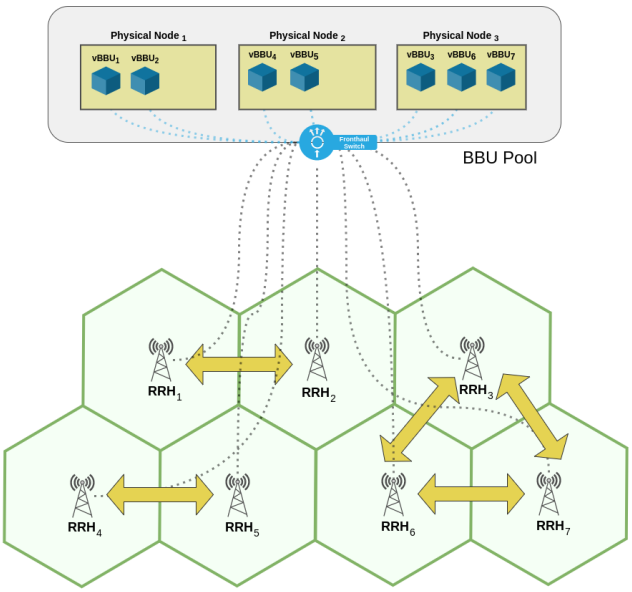


Figure 5.2. vBBU placement within physical nodes are adjusted according to CoMP sets

On top of this foundation, a heuristic placement solution could be built by modifying the best fit decreasing approach to the bin packing problem. By using clusters of CoMP paired vBBUs instead of individual vBBUs as elements, efficient placements could be obtained.

Assuming equal network load coming from each cell and each physical node can handle the total bandwidth requirements of three cells, if best fit decreasing algorithm is executed with modifications mentioned above for Figure 5.1, the resulting output would place the vBBUs as in Figure 5.2. After calculating the λ value of these two different placements using the Equation 5.4, it can be seen that despite having $\lambda \approx 1.86$ in the initial configuration, the new placement readjusts the data exchange routes in an optimal way resulting in $\lambda = 1$. Figure 5.3 depicts the changes in data exchange routes in detail.

5.2.2. Splitting vBBU clusters

In a typical vBBU placement problem, it is usually safe to assume that no single vBBU is larger than the capacity of a physical node, otherwise the problem would

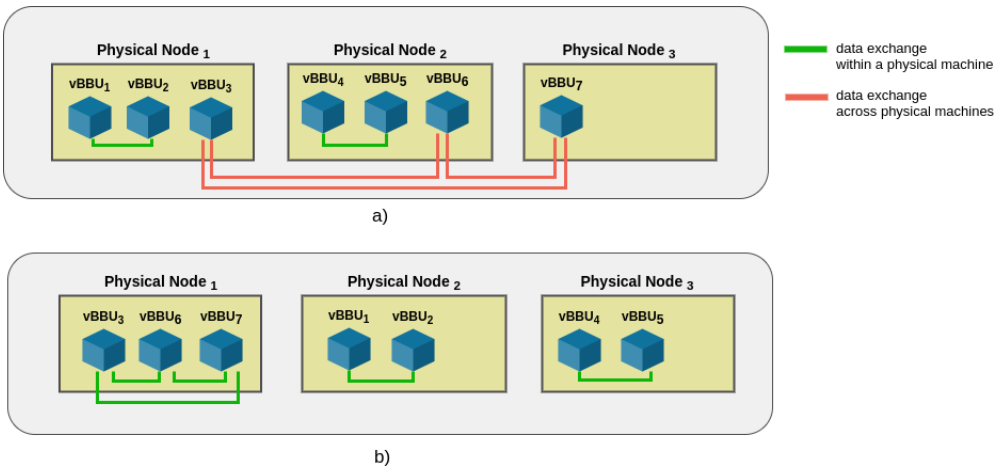


Figure 5.3. Data exchange traffic in the BBU pool for Figure 5.1 and 5.2
In placement (a), CoMP sets are disregarded, resulting $\lambda \approx 1.86$
Placement (b) is the optimal assignment with $\lambda = 1$

become infeasible. However, after the replacement of vBBUs with vBBU clusters in the proposed algorithm, the same assumption could no longer be made, because it is perfectly reasonable for vBBUs to form a cluster that exceeds the capacity of the physical nodes.

Figure 5.4 explains this situation by introducing a new CoMP pairing between RRH_5 and RRH_6. The addition of this new CoMP configuration merges two of the previous ones into a single cluster with five vBBUs. Following the simplifying assumption that a physical node could handle the load of at most three vBBUs in this scenario, there is no way to find an assignment for this new cluster. The obvious solution is to split this cluster into smaller parts even though it means an inevitable increase in λ, but how this split is performed can completely change the outcome of the algorithm. Figure 5.5 demonstrates two different vBBU placements for the same CoMP sets from Figure 5.4 but different cluster splits. The difference in λ is solely due to the way larger CoMP set is split in two.

The question of how to split a vBBU cluster with CoMP pairings into two can be transformed into the following question: How to find the minimum cut in an unweighted and undirected graph in which the vertices are the vBBUs in the cluster and the edges are the CoMP pairings between vBBUs? Once the vBBU cluster is represented as the

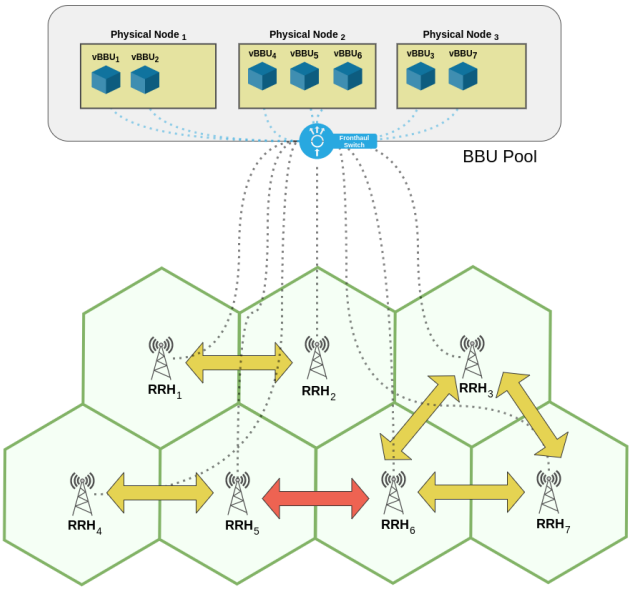


Figure 5.4. A new CoMP configuration changes vBBU clusters vBBU clusters become $\{vBBU_1,\ vBBU_2\}$ and $\{vBBU_3,\ vBBU_4,\ vBBU_5,\ vBBU_6,\ vBBU_7\}$

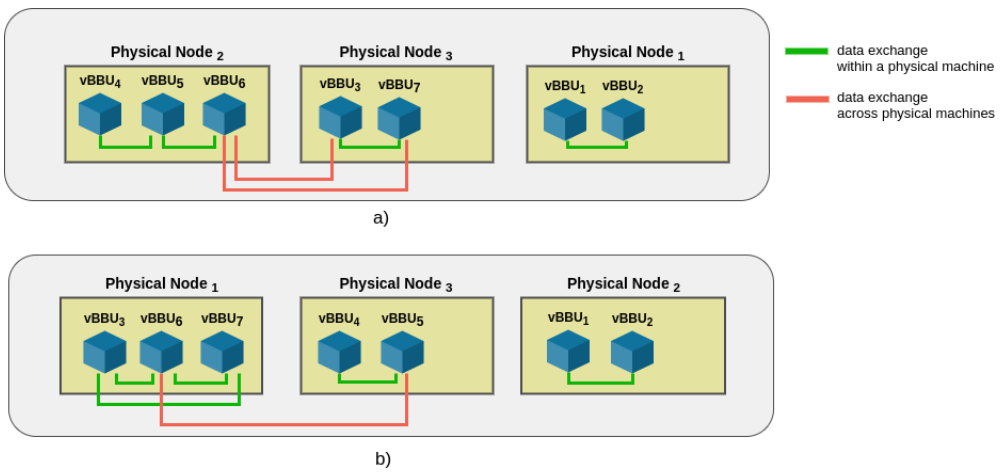


Figure 5.5. Data exchange routes for two different placements after clusters split In placement (a), an inefficient cluster split results in $\lambda \approx 1.43$. Placement (b) has $\lambda \approx 1.28$ with a better split

described graph, each edge crossed by the minimum cut line corresponds to total data exchange, which has to travel across different physical nodes due to split operation. If the examples from Figure 5.5 are revisited, the cut that would separate the cluster as in the first placement would have passed through two pairing edges while the second cut would only have passed through a single pairing edge. However, notice that cutting that cluster through the edge between RRH_4 and RRH_5 would be equally powerful with only a single edge, but the outcome would not be as useful because of the imbalance between the members of two individual parts. First cut divides the vBBUs into two sets of size three and two, while the second cut divides them as four and one. Therefore, the balance of the remaining parts is also an important factor in finding the best possible cut.

Minimum cut and maximum flow algorithms are intensely studied in the literature [49]. The problem space in vBBU placement deals with undirected and unweighted graphs, so the minimum cut algorithm proposed and improved by Karger [50, 51] is a quite simple and promising randomized algorithm that can solve the cluster split problem in polynomial time. The algorithm is built on top of a fundamental operation called edge contraction, in which a random edge from the graph is selected and contracted by conjoining the vertices at the both ends together. Edge contraction continues until only one edge remains in the graph, and at that point that particular edge symbolizes the decisive cut. It has been shown in [50] that if this operation is repeated at $\binom{n}{2} \cdot log\ n$ times, the minimum cut of the graph is found with high probability. In fact, all the minimum cuts are found after this operation, and this is especially useful because cluster splitting requires not only finding the minimum cut, but also finding the minimum cut with the most balanced partitioning.

Karger's minimum cut algorithm is meant to be executed when the first phase of the algorithm could not find a place for a vBBU cluster after an iteration of best fit decreasing bin packing heuristic. The residual clusters are split into two and bin packing algorithm iterates again with partitioned clusters until all clusters are successfully positioned into physical nodes.

```
Require Clusters = clusters of vBBUs
Require Bins = physical nodes in pool
Return the Bins in which Clusters are placed conveniently;
while |Clusters| > 0 do
  Bins ⇐ SortIncreasing(Bins)
  Clusters ⇐ SortDecreasing(Clusters)
  for i = 1 to |Clusters| do
    for j = 1 to |Bins| do
      if Bin_j.capacity > Cluster_i.size then
        Bin_j.put(Cluster_i)
        Bins ⇐ SortIncreasing(Bins)
        Clusters ⇐ Clusters - {Cluster_i}
      end if
    end for
  end for
  for i = 1 to |Clusters| do
    Clusters ⇐ Clusters - {Cluster_i}
    Clusters ⇐ Clusters ⋃ Split(Cluster_i)
  end for
end while
Return Bins
```

Figure 5.6. Placement algorithm for the heuristic utilizing best fit decreasing method

5.2.3. Pseudocodes

Simplified pseudocodes for the algorithms used in the experiments are presented below. Placement algorithm in Figure 5.6 takes the physical nodes in the BBU pool and vBBU clusters as defined in Equation 5.11 as parameters and carries out best fit decreasing bin packing method. When a cluster needs to be split because it could not fit into any bins, the procedure in Figure 5.7 is called. This subroutine takes the given cluster and returns two partitions divided by the minimum cut found after Karger's minimum cut method.

Require Cluster
Return two new clusters divided by minimum cut;
Create graph G = (V, E) from Cluster
Initialize Cuts = {}
Initialize Partitions = {}
for $i = 1$ to $\binom{|V|}{2} \cdot log|V|$ **do**
  $G' \Leftarrow G$
  while $|E'| > 1$ **do**
    $v \Leftarrow$ randomly selected from V'
    $w \Leftarrow$ randomly selected from V'
    EdgeContract(G', v, w)
  end while
  Cuts $\Leftarrow$ Cuts $\bigcup$ G'
end for
BestCut $\Leftarrow$ Cuts[0]
for $i = 1$ to $|Cuts|$ **do**
  if CutLength(Cut_i) < CutLength(BestCut) **then**
    BestCut $\Leftarrow$ Cut_i
  else if CutLength(Cut_i) = CutLength(BestCut) **then**
    if PartitionBalance(Cut_i) > PartitionBalance(BestCut) **then**
      BestCut $\Leftarrow$ Cut_i
    end if
  end if
end for
Partitions $\Leftarrow$ Partition Cluster wrt BestCut
Return Partitions

Figure 5.7. Cluster splitting algorithm by using Karger's randomized minimum cut method

6. EXPERIMENTS AND RESULTS

In order to compare our heuristic approach with the optimal solution and analyze it further, we have developed a discrete event simulator based on the Python simulation framework SimPy [52]. The simulator is designed with a very high resolution in mind, trying to include every component of a CloudRAN architecture as an individual unit. Physical nodes in the BBU pool, virtual BBUs, RRHs and their associated cell sites are modeled separately. Along with the fronthaul switch of the BBU pool, each physical node has its own switch implementation where bandwidth constraints are introduced in the form of bitrate and packet queue limits. The fronthaul switch is responsible for distributing the IQ packets to the target physical nodes, then physical node switches take over to deliver those packets to target vBBUs. Each IQ packet is a distinct entity in the simulation, duplicated along the way when it is necessary for signal exchange. Their journey over the course of their lifetime is continuously monitored to distinguish data exchange within physical nodes and across physical nodes.

The transmission of an IQ packet is modeled after Equation 5.3. The simulator computes aggregated transmission cost collected from each IQ packet at every time interval t to calculate the λ value. IQ packets are created with a Poisson process for the given arrival rate and each packet has a different payload size generated by a normal distribution for the given mean and standard deviation. These load generation metrics (arrival rate, packet size mean and packet size variation) can be changed at any time to simulate dynamic load. Note that the load function $L_i(t)$ defined in Section 5.1 is not only about the number of packets coming from a cell, but also their packet size.

RRH-BBU mapping is not explored in these experiments; so for the sake of simplicity, it is assumed there is a one-to-one mapping between RRHs and BBUs, without loss of generality.

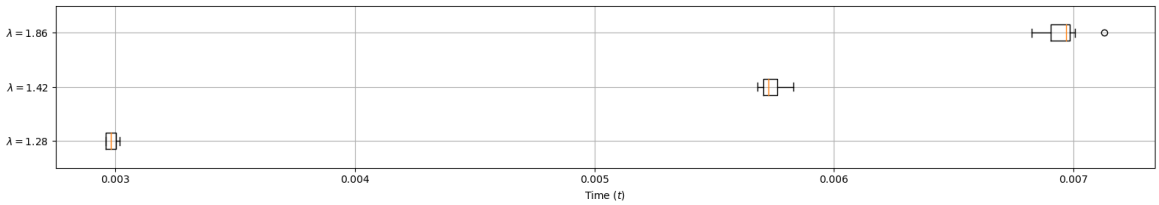


Figure 6.1. Average wait times on switch queues with different λ values

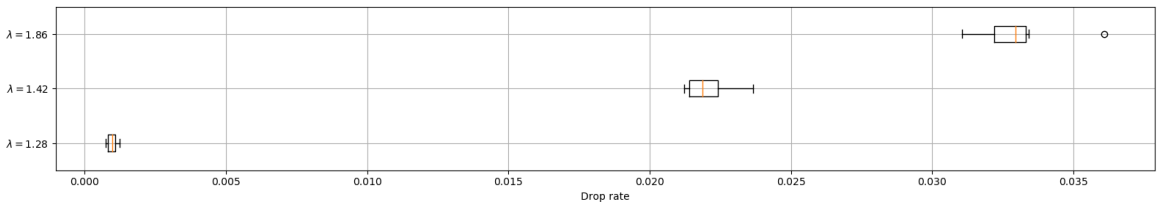


Figure 6.2. Average packet drop rate with different λ values

6.1. Effect of λ over network utilization

The first experiment is to verify that our intuition about reducing λ by tweaking vBBU placement has an actual positive effect over network metrics. In order to test this, three different placement configurations from Figure 5.3-b, 5.5-a and 5.5-b are prepared in the simulation resulting in λ values 1.86, 1.43, and 1.28, approximately. Afterwards, each configuration is simulated under dynamic load for 25 times and at the end two network metrics are calculated: average wait time an IQ packet spends in physical node switch queues and average packet drop rate at those switches. Both metrics are in line with our anticipation as can be seen in Figure 6.1 and Figure 6.2. As λ increases, an IQ packet is more likely to have to travel across physical node switches and producing artificial load, as opposed to travel within the same physical node with no cost. Therefore, it is no surprise that queue wait times and drop rates are negatively affected.

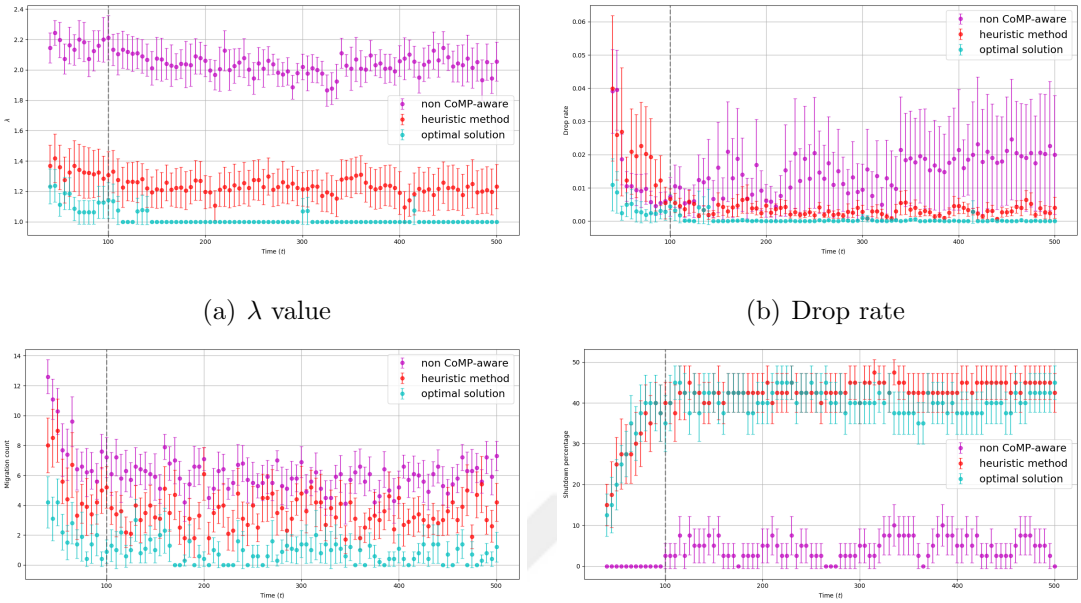
6.2. Comparing placement algorithms

6.2.1. Comparing placement performance

In the interest of evaluating the performance of the proposed heuristic method, three different algorithms are implemented in the simulator. The first one is an exhaustive solution to the placement problem depicted in Section 5.1 using recursive backtracking to find the minimum λ value out of all possible placement combinations. It is denoted as *optimal solution* in the figure legends. The second algorithm corresponds to what would have happened if CoMP sets had no effect over how vBBUs are allocated into physical nodes. Following the traditional practices in virtual machine placement in data centers, it uses a simplified best fit decreasing bin packing heuristic to put virtual BBUs into physical nodes with respect to their individual bandwidth demands. However, having no knowledge about the data traffic caused by the signal exchange due to CoMP, it generally performs inadequately. This one is denoted as *non CoMP-aware* in the legends. The last algorithm is our heuristic implementation, denoted as *heuristic method*, which uses Karger's min cut algorithm to cluster vBBUs into groups and then utilizes best fit decreasing bin packing heuristic to place them into physical nodes.

Table 6.1. Constant simulation parameters.

Parameter	**Value**
Physical node count $\|W\|$	4
Virtual BBU count $\|V\|$	20
RRH and cell site count $\|U\|$	20
Physical node switch queue length	1000
Physical node switch bitrate β_i	40 Gbps
Target switch utilization	75%
Average IQ packet size	1 Kb
Timeframe length t/T	0.01



(a) λ value

(b) Drop rate

(c) Migration count

(d) Shutdown percentage

Figure 6.3. Performance comparison of three different placement algorithms over the course of 25 simulations

Evaluations are performed by running 25 simulations for each algorithm. Constant simulation parameters are listed in Table 6.1. Within each simulation, 18-24 CoMP pairs are chosen randomly to create differentiation. In other words, topology stays same but how CoMP sets are assigned is different in each simulation. Also, even though each cell starts with a fixed Poisson arrival rate for packet generation, their assigned arrival rates could change randomly during the simulation to create fluctuations in the load simulating UE mobility, without changing the overall load in the ecosystem. Otherwise, once a good placement is found, there would be no reason to change it until the end of simulation.

Figure 6.3 shows the simulation results exploring four different metrics. Each panel has simulation time as the x axis and y values are the averages of 25 data points at that specific simulation time coupled with the 95% confidence interval whiskers. From the first panel, it could be seen that the optimal solution almost always thrived on finding a placement with $\lambda = 1$. While our heuristic approach is quite close, non CoMP-aware method falls significantly behind. The impact of efficient placement clearly reflects on the drop rate in panel 6.3(b). Disregarding CoMP relations causes

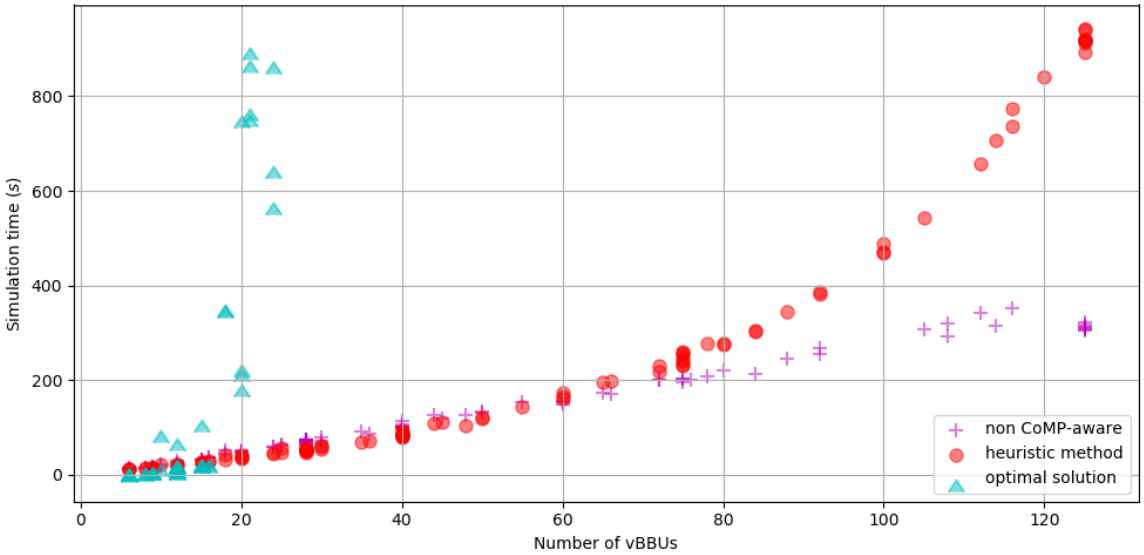


Figure 6.4. Simulation time with respect to vBBU size

much higher variation due to the randomness of the vBBU placement, and always with higher drop rate after the warm-up period. Number of migrations by Equation 5.9 and proportion of empty physical nodes by Equation 5.8 are not part of the objective in neither optimal solution nor heuristic method but still the simulation outputs for these metrics could be seen in panel 6.3(c) and 6.3(d).

6.2.2. Comparing computation complexity

Although the optimal solution seems like the clear winner after looking at the simulation results from the last section, computational complexity is a massive disadvantage as we anticipated. In fact, in a typical personal computer, simulation times exponentially jump into infeasible scales only after 25 vBBUs, as can be seen in Figure 6.4. Non CoMP aware method displays a linear trend as expected with $O(n \cdot log\ n)$ computation complexity. Although placement part of the heuristic algorithm has the same bound, cluster splitting part causes a visible deviation due to $O(n^2 \cdot log\ n)$ complexity.

6.3. Further analysis on heuristic method

6.3.1. Impact of topology size

As far as the conditions that could affect the algorithm performance are concerned, the first thing to look at is the change in the topology size. For this set of simulations, number of physical nodes is kept proportional to the number vBBUs such that each phsyical node could host four vBBUs on average. Different topologies are created up to 60 vBBUs (and 15 physical nodes) with randomized CoMP pairings but again with a fixed CoMP density of two; meaning each vBBU is expected to exchange signals with two other vBBUs on average. Three simulations are run for each vBBU size to not to be misdirected by the outliers because more challenging CoMP sets could occur when trying to randomize them.

The results of this study are depicted in Figure 6.5. The simulations for the optimal solution are limited to 20 vBBUs because of the exponential rise in the simulation time. Data indicates that topology size does not affect the heuristic performance at all. Even though calculation times increases proportional to the vBBU size, it could still find efficient placements with satisfying λ values regardless of the vBBU size. Non CoMP-aware method on the other hand, has a certain upwards trend as topology gets larger.

6.3.2. Impact of overall network load

Overall network load is another concern that could influence the algorithm performance. When the cumulative load coming to the BBU pool is well below the capacity, expected load per vBBU starts to decrease as the number of vBBUs is always constant in the scope of our study. This leads to an increase in the physical nodes' ability to host more vBBUs as the bandwidth constraints are more relaxed. As a consequence, it becomes easier to put all the vBBUs in the CoMP set together as there are plenty of available space in physical nodes compared to scenarios under normal load. Conversely, if the expected load per vBBU starts to escalate, then the bandwidth constraints gets

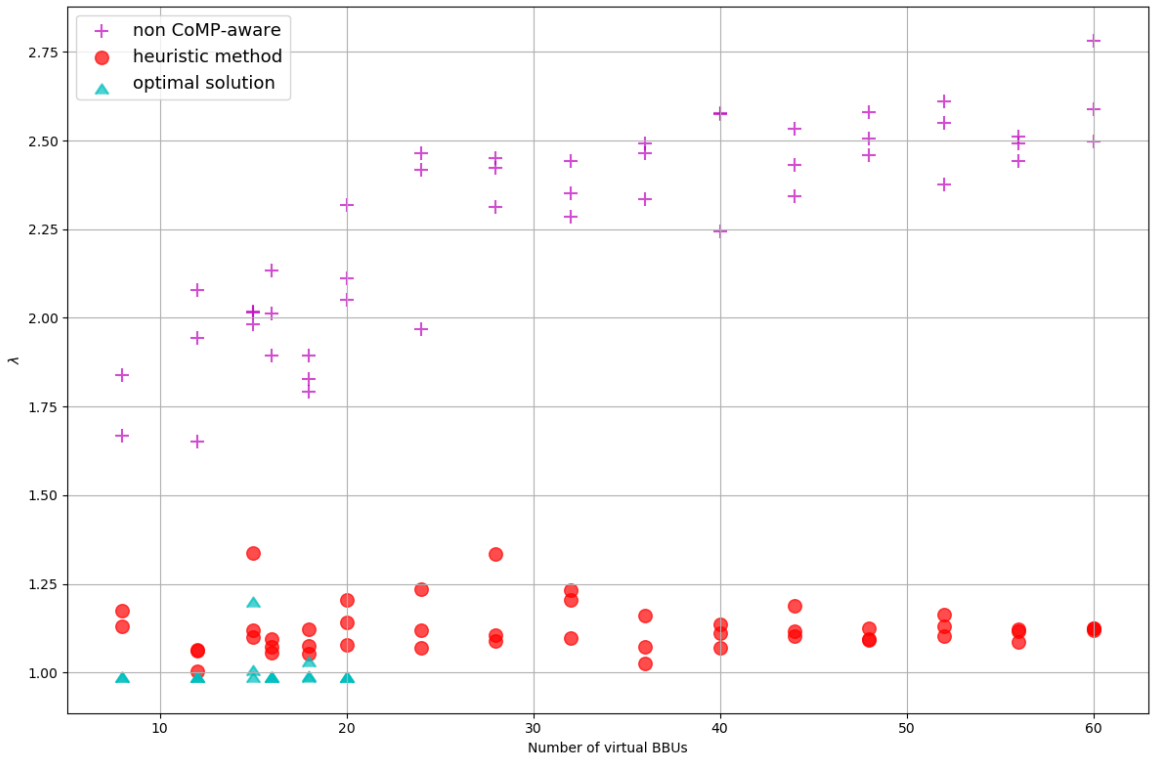
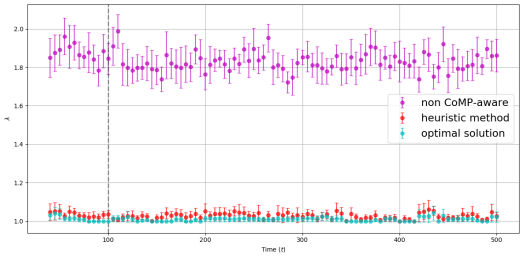


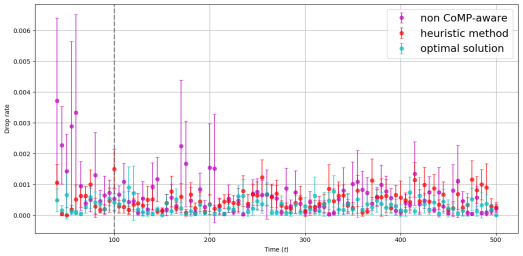
Figure 6.5. Change in λ value as topology gets larger

stricter and it becomes harder to group vBBUs with mutual traffic together.

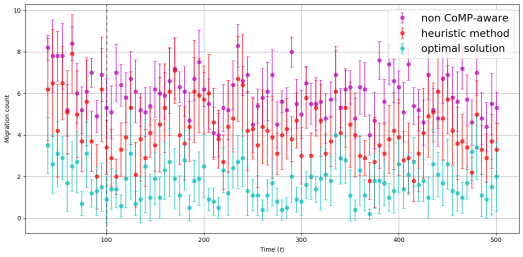
Two additional sets consisting of 25 simulations for each are prepared to test the impact of overall network load. Both sets have the same simulation parameters with the set from Section 6.2.1, only differing in average arrival rate for IQ packets as heavy load and light load variations. Simulation results are presented in Figure 6.6 and Figure 6.3, coupled with the previous Figure 6.3 from the last section. The first observation is all three algorithms indeed yield lower λ values as the relative network load decreases. As mentioned earlier, this could be attributed to being able to put more vBBUs into a single physical node, even in non CoMP-aware case. As a direct consequence of this, shutdown percentage panel can be examined to see how high the ratio of turned off physical nodes. But when the physical nodes are highly utilized because of the heavy load, it becomes too difficult to turn them off as in Figure 6.7(d). Drop rate figures are also in parallel with the expectations, notice the scaling of y axis in 6.6(b) and Figure 6.7(b).



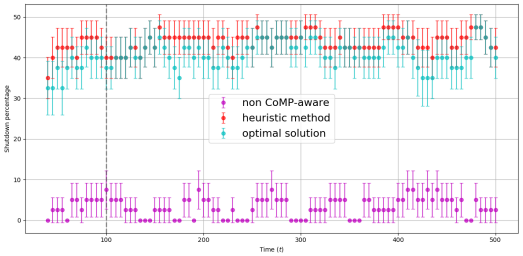
(a) λ value



(b) Drop rate

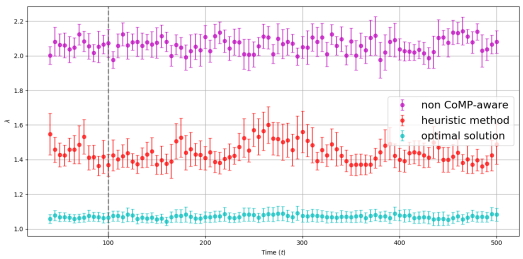


(c) Migration count

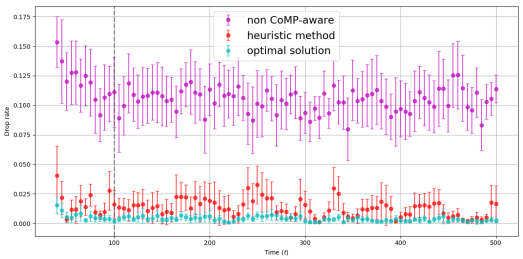


(d) Shutdown percentage

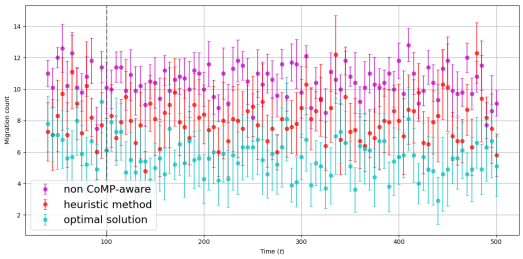
Figure 6.6. Algorithm performances under light load



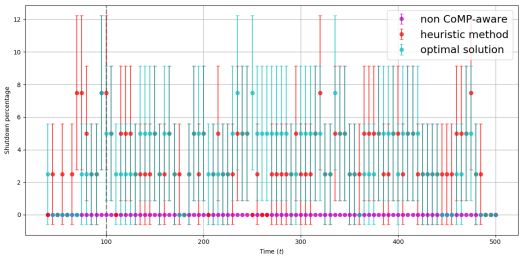
(a) λ value



(b) Drop rate



(c) Migration count



(d) Shutdown percentage

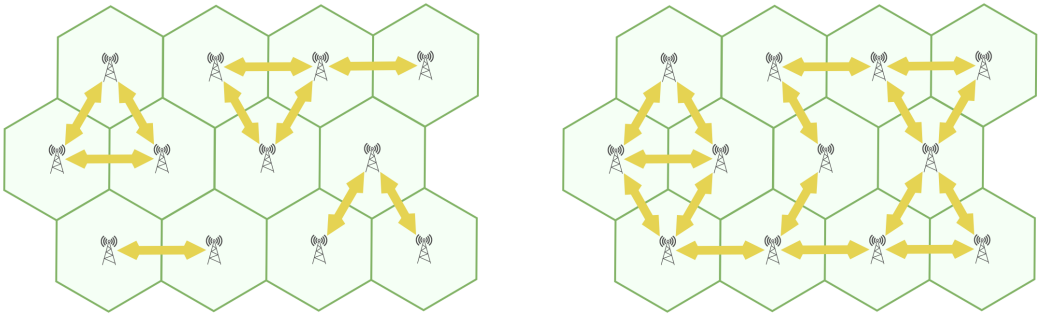
Figure 6.7. Algorithm performances under heavy load

After carefully examining λ time series plots, another critical observation is that as the load increases, the performance of our heuristic algorithm distances itself from the optimal solution. The difference between λ values of heuristic method and the optimal solution is substantially larger in Figure 6.7(a). This shows that our proposed approach falls a little behind under heavy utilization scenarios. Because cluster splitting will be much more frequent in those scenarios, it is only natural that our intuitive minimum cut approach cannot compete with the exhaustive search of the optimal solution.

6.3.3. Impact of cluster split algorithm in different CoMP density

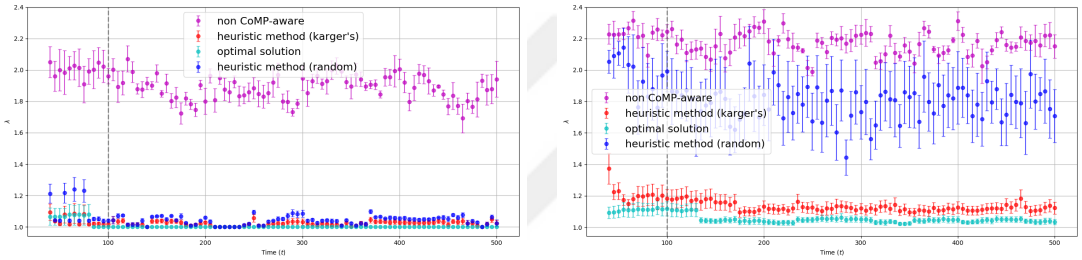
The next topic is examining how effective our split algorithm is, as opposed to just randomly splitting the clusters into two or more when needed. In order to measure this more elaborately, cell topology is classified into two groups as sparse and dense according to CoMP density, e.g. the number of CoMP pairings per cell. An example for each is provided in Figure 6.8. The number of CoMP pairings per cell is 1.67 on average for the sparse topology on the first panel, while it is 2.67 in the dense topology on the second panel.

When CoMP sets are extracted from Figure 6.8, it could be seen that sparse topology presents itself as a disconnected graph with four isolated sets that have no data exchange between them. However dense topology is a connected graph with a single CoMP set consisting of all vBBUs. It becomes apparent that there is no vBBU placement for the dense topology that yields $\lambda = 1$ unless all the vBBUs could fit into a single physical node. A cluster split operation is bound to happen in this scenario, and the effectiveness of the split could dramatically change λ. On the other hand, splits are much less frequent in a sparse topology and even if they occur, the difference between an efficient split and an inferior one is relatively smaller. The objective in this experiment is to investigate the impact of the split algorithm under these two classifications.



(a) Sparse CoMP pairing with 1.67 density (b) Dense CoMP pairing with 2.67 density

Figure 6.8. Example CoMP pairings for sparse and dense scenarios



(a) Sparse topology (b) Dense topology

Figure 6.9. Comparing four different algorithms under sparse and dense topology

Again, two sets of simulations are run for scenarios depicted in Figure 6.8 for the usual three algorithms, coupled with one extra in which the same heuristic method is used but clusters are split randomly instead of being guided by the Karger's min-cut algorithm. Two conclusions could be derived from the simulation results in Figure 6.9. The first one is that the algorithm performance is indeed negatively affected by the CoMP density of the topology. All methods produced higher λ values when density increased. The second conclusion is that using a specialized min-cut algorithm to split clusters is not as effective when the CoMP density is low, as there is only a miniscule difference between the two in Figure 6.9(a). However as density increases, the heuristic with the random splits starts to produce inadequate results, almost as poor as non CoMP-aware placements, while the heuristic with the Karger's min-cut algorithm stays close to the optimal solution thanks to the efficient cluster splits.

Finally, four different algorithms are compared on a larger scale using up to 60 cell networks with dense CoMP pairings, looking alike extended versions of Figure 6.8(b). Optimal solution simulations are limited to 25 cells once more but certain

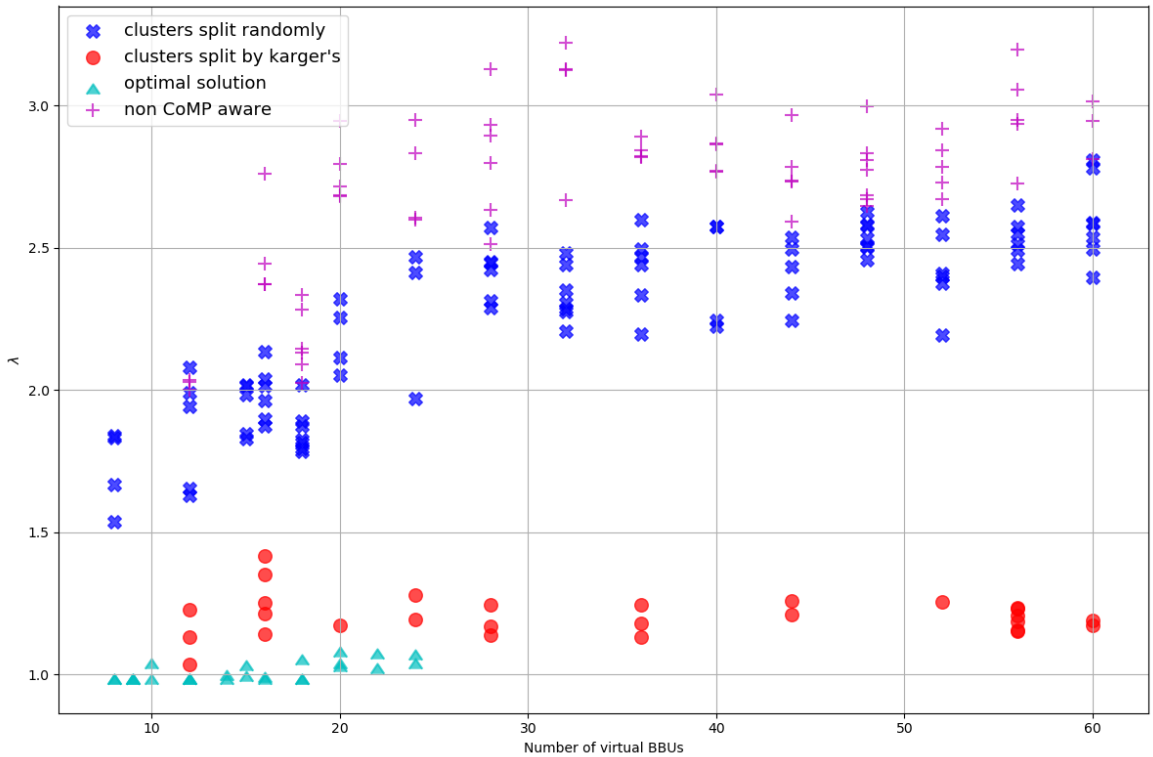


Figure 6.10. Impact of split algorithm in larger topologies

trends could be easily seen in Figure 6.10. It is safe to say that split method becomes more important as the topology gets larger in dense CoMP pairings because random splits exhibits an upwards trend while min-cut splits stay somewhat linear.

6.3.4. Impact of bandwidth threshold

Bandwidth threshold in physical nodes could be defined as what percentage of the switch capacity is targeted when deciding if that switch could handle more data. This threshold is relevant for all three algorithms as each of them have to check for the switch capacity at some point. In the heuristic approach and traditional non CoMP-aware solutions, this threshold is directly represented as the bin capacity. In other words, if the switch bitrate capacity is 40 Gbps in a simulation and the bandwidth threshold is 75%, then the bin object symbolizing a physical node is at full capacity when the expected load for that time frame is at 30 Gbps. The need for such a threshold comes from previous experimental results. When switches in the simulation work with 100% utilization, packet drop rate skyrockets due to random nature of arrivals and time spent

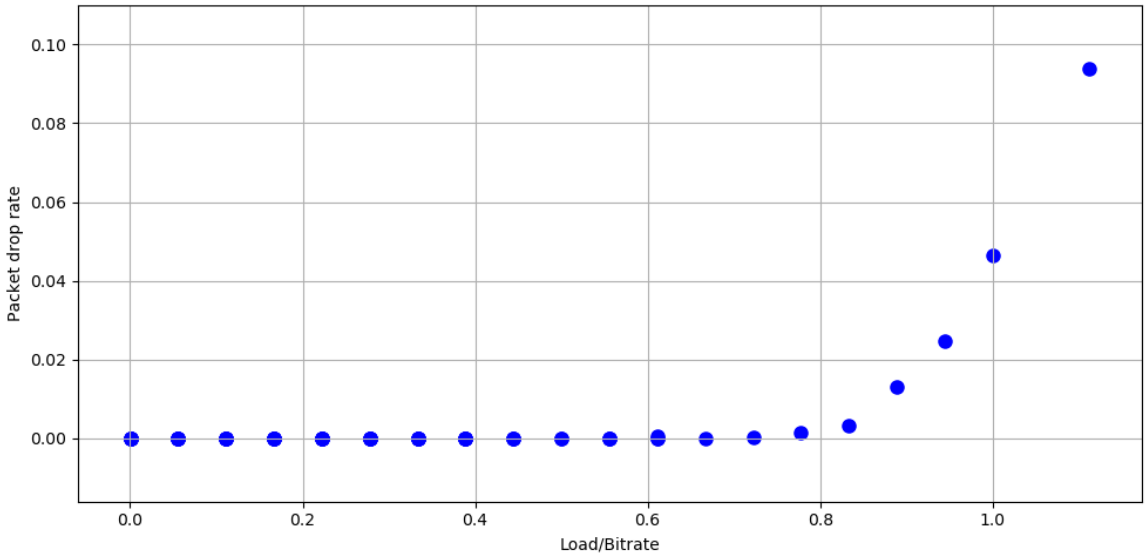
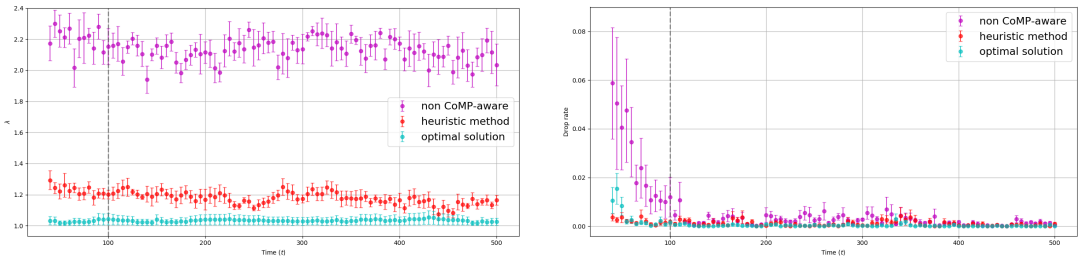


Figure 6.11. Packet drop rate vs switch utilization

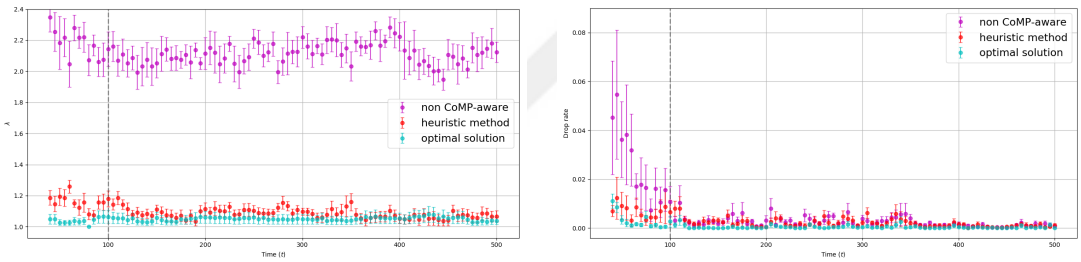
processing the packets. A brief test shows us that 75% is a good point to aim to keep the packet drop rate under 0.1% confidently, as can be seen in the Figure 6.11.

If the bandwidth threshold is raised, then we are falsely advertising for how much more a physical node could take. This is expected to decrease λ as it will become easier to find available rooms in the physical nodes, but we are overloading them which in turn causes more packet drops. If utilization ratio is decreased, then this time bandwidth capacities become more strict. The placement algorithms will have harder time to group vBBU clusters together, which will cause clusters to be split more, and λ will increase. But drop rate still decreases from the same reason: physical nodes are under-utilized. The picture is quite clear in the results when Figure 6.12, 6.13 and 6.14 is examined. There is an apparent difference between λ values across three iterations, but drop rates are still close to each other. Because the reason more efficient placements could be found is the same reason drop rate will be more than expected.



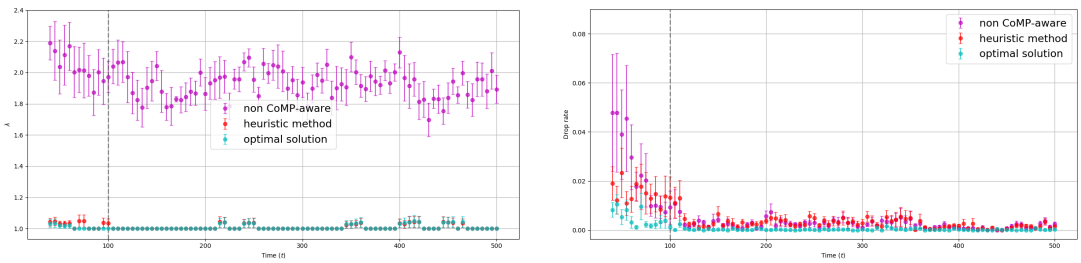
(a) λ value (b) Drop rate

Figure 6.12. Comparison of algorithms with 60% switch bandwidth utilization



(a) λ value (b) Drop rate

Figure 6.13. Comparison of algorithms with 75% switch bandwidth utilization



(a) λ value (b) Drop rate

Figure 6.14. Comparison of algorithms with 90% switch bandwidth utilization

7. CONCLUSION

Ever-growing demand on the mobile Internet traffic and extremely challenging 5G requirements targeted to meet this demand, lay insurmountable obstacles in front of the current LTE-Advanced architecture. Addressing this need for an architectural shift, CloudRAN is proposed as a candidate design for future networks. It is based on two fundamental principles, splitting the base station functionality into two as radio and baseband processing and then grouping all the baseband units into a centralized and virtualized data center called BBU pool.

CoMP is one of the recent technologies already being used in LTE-Advanced that exploits the realization that more than one base stations could be used for each UE to improve signal quality and reduce inter-cell interference through various operation modes. Although CloudRAN seems to be working in full harmony with CoMP at the first glance due to centralized BBU pool model, the implications of RRH-BBU split comes with a serious disadvantage. In a fully centralized model, RRHs have to send raw IQ data since baseband processing is performed in the BBU pool. The challenges with the transmission of this data gets multiplied with the involvement of the massive data exchange requirements of CoMP.

In this thesis, the placement of the virtual BBUs are explored to discover efficient vBBU allocations within physical nodes in BBU pools. There are number of studies working on issues related to CoMP deployments in CloudRAN architecture but the effect of vBBU placement inside the BBU pools is yet to be explored thoroughly. In the scope of our study, the optimal placement is formulated to create a mathematical background and metrics needed to describe the effectiveness of a placement are defined. Seeing the computational complexity in finding the optimal solution, a heuristic approach is proposed to find sub-optimal solutions in polynomial time. Finally, a discrete time event based simulator is implemented in order to compare the heuristic approach with the optimal solution under different scenarios.

Simulation results indicate that the heuristic approach to the vBBU placement problem proposed in this thesis is a reliable alternative that could produce close to optimal assignments with substantial improvements in computation time. Further experiments also display the heuristic method keeps working as intended under different environments when topology size, density of the CoMP pairings between cells or network load characteristics are changed.

The future plans for this research include developing a testbed environment that could resemble a BBU pool as close to real as possible, with the help of various virtualization software used in small scale data centers. Another focus is optimizing the heuristic approach around vBBU migration. This topic branches into two categories, the first one is considering the number of migrations as a key performance indicator so the algorithm favors the outputs with less number of migrations. The second part is devising optimum migrations paths given a new vBBU assignment. Applying the new placement vector generated by the algorithm is not always straightforward due to limitations of physical nodes, it transforms into a variation of Tower of Hanoi problem. One last topic in the roadmap is creating feedback about the impact of each CoMP pairing within the BBU pool. The placement algorithm has the knowledge of which CoMP pairings distort the vBBU clustering most, causing data exchange across physical nodes. The central entity in charge of CoMP decisions could be informed with this knowledge to reevaluate the CoMP sets and tweak them if necessary.

REFERENCES

1. Checko, A., H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras and M. S. Berger, "Cloud RAN for Mobile Networks — A Technology Overview", *IEEE Communications Surveys & Tutorials*, pp. 405–426, 2014.

2. Cisco Mobile, V., "Visual networking index: Global mobile data traffic forecast update, 2016-2021", *Tech.Rep.*, February 2017.

3. NTT DOCOMO, I., "5G Radio Access: Requirements, Concept and Technologies", *DOCOMO White Paper*, July 2014.

4. China Mobile Research, I., "C-RAN: the road towards green ran", *Tech. Rep.*, October 2011.

5. Lee, J., Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao and Y. Zhou, "Coordinated multipoint transmission and reception in LTE-advanced systems", *IEEE Communications Magazine*, Vol. 50, No. 11, pp. 44–50, 2012.

6. Tang, J., W. P. Tay, T. Q. S. Quek and B. Liang, "System Cost Minimization in Cloud RAN With Limited Fronthaul Capacity", *IEEE Transactions on Wireless Communications*, Vol. 16, No. 5, pp. 3371–3384, 2017.

7. Wang, X., C. Cavdar, L. Wang, M. Tornatore, Y. Zhao, H. S. Chung, H. H. Lee, S. Park and B. Mukherjee, "Joint Allocation of Radio and Optical Resources in Virtualized Cloud RAN with CoMP", *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, December 2016.

8. Khan, M., R. S. Alhumaima and H. S. Al-Raweshidy, "Reducing energy consumption by dynamic resource allocation in C-RAN", *2015 European Conference on Networks and Communications (EuCNC)*, pp. 169–174, June 2015.

9. Khan, M., R. S. Alhumaima and H. S. Al-Raweshidy, "Quality of Service aware dynamic BBU-RRH mapping in Cloud Radio Access Network", *2015 International Conference on Emerging Technologies (ICET)*, pp. 1–5, December 2015.

10. da Paixão, E. A. R., R. F. Vieira, W. V. Araújo and D. L. Cardoso, "Optimized load balancing by dynamic BBU-RRH mapping in C-RAN architecture", *2018 Third International Conference on Fog and Mobile Edge Computing (FMEC)*, pp. 100–104, April 2018.

11. Ran, C., S. Wang and C. Wang, "Optimal load balancing in cloud radio access networks", *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1006–1011, March 2015.

12. Chen, Y., Z. Lu, X. Wen and H. Shao, "User-centric Clustering and Beamforming for Energy Efficiency Optimization in Cloud-RAN", *Mobile Networks and Applications*, Vol. 23, No. 3, pp. 503–517, 2018.

13. Chen, Y.-S., C.-S. Hsu and F.-Y. Liao, "A bandwidth adaptation mechanism for Cloud Radio Access Networks", *Pervasive and Mobile Computing*, Vol. 40, pp. 639–659, 2017.

14. Pompili, D., A. Hajisami and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN", *IEEE Communications Magazine*, Vol. 54, No. 1, pp. 26–32, 2016.

15. Sundaresan, K., M. Y. Arslan, S. Singh, S. Rangarajan and S. V. Krishnamurthy, "Fluidnet: a flexible cloud-based radio access network for small cells", *IEEE/ACM Transactions on Networking*, Vol. 24, pp. 915–928, 2016.

16. Tao, M., E. Chen, H. Zhou and W. Yu, "Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN", *IEEE Transactions on Wireless Communications*, Vol. 15, No. 9, pp. 6118–6131, 2016.

17. Zhao, J. and Z. Lei, "Clustering methods for base station cooperation", *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 946–951, April 2012.

18. Biermann, T., L. Scalia, C. Choi, H. Karl and W. Kellerer, "Improving CoMP cluster feasibility by dynamic serving base station reassignment", *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1325–1330, September 2011.

19. Zeng, D., J. Zhang, L. Gu, S. Guo and J. Luo, "Energy-Efficient Coordinated Multipoint Scheduling in Green Cloud Radio Access Network", *IEEE Transactions on Vehicular Technology*, Vol. 67, No. 10, pp. 9922–9930, 2018.

20. Lakshmana, T. R., C. Botella, T. Svensson, C. U. of Technology, S. of Electrical Engineering, D. of Signals, C. S. Systems, C. tekniska högskola and K. Institutionen för signaler och system, "Partial joint processing with efficient backhauling using particle swarm optimization", *EURASIP Journal on Wireless Communications and Networking*, Vol. 2012, No. 1, pp. 1–18, 2012.

21. Qi, Y., M. Z. Shakir, M. A. Imran, A. Quddus and R. Tafazolli, "How to solve the fronthaul traffic congestion problem in H-CRAN?", *2016 IEEE International Conference on Communications Workshops (ICC)*, pp. 240–245, May 2016.

22. Khorsandi, B. M. and C. Raffaelli, "BBU location algorithms for survivable 5G C-RAN over WDM", *Computer Networks*, Vol. 144, pp. 53–63, 2018.

23. Mahapatra, B., A. K. Turuk, S. K. Patra and R. Kumar, "Optimal Placement of Centralized BBU (C-BBU) for Fronthaul and Backhaul Optimization in Cloud-RAN Network", *2017 International Conference on Information Technology (ICIT)*, pp. 107–112, December 2017.

24. Alhumaima, R. S. and H. S. Al-Raweshidy, "Optimising the BBU pool placement in cloud radio access networks based on power allocations", *2017 Computing Con-*

ference, pp. 1312–1316, July 2017.

25. Carapellese, N., M. Tornatore and A. Pattavina, "Energy-Efficient Baseband Unit Placement in a Fixed/Mobile Converged WDM Aggregation Network", *IEEE Journal on Selected Areas in Communications*, Vol. 32, No. 8, pp. 1542–1551, August 2014.

26. Colman-Meixner, C., G. B. Figueiredo, M. Fiorani, M. Tornatore and B. Mukherjee, "Resilient cloud network mapping with virtualized BBU placement for cloud-RAN", *2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 1–3, November 2016.

27. Meng, X., V. Pappas and L. Zhang, "Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement", *2010 Proceedings IEEE INFOCOM*, pp. 1–9, March 2010.

28. Zhang, B., Z. Qian, W. Huang, X. Li and S. Lu, "Minimizing Communication Traffic in Data Centers with Power-Aware VM Placement", *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp. 280–285, July 2012.

29. Ilkhechi, A. R., I. Korpeoglu and O. Ulusoy, "Network-aware virtual machine placement in cloud data centers with multiple traffic-intensive components", *Computer Networks*, Vol. 91, pp. 508–527, 2015.

30. Chen, T., X. Gao and G. Chen, "Optimized Virtual Machine Placement with Traffic-Aware Balancing in Data Center Networks", *Scientific Programming*, Vol. 2016, pp. 1–10, 2016.

31. Wu, J., Z. Zhang, Y. Hong and Y. Wen, "Cloud radio access network (C-RAN): a primer", *IEEE Network*, Vol. 29, No. 1, pp. 35–41, January 2015.

32. Agrawal, R., A. Bedekar, S. Kalyanasundaram, T. Kolding, H. Kroener and

V. Ram, "Architecture Principles for Cloud RAN", *IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, 2016.

33. Agrawal, R., A. Bedekar, T. Kolding and V. Ram, "Cloud RAN challenges and solutions", *Annals of Telecommunications*, Vol. 72, No. 7, pp. 387–400, 2017.

34. Ericsson AB, L. N. C., Huawei Technologies Co. and Nokia, *CPRI Specification V7.0 (2015-10-09)*, 2015, `http://www.cpri.info/downloads/CPRI_v_7_0_2015-10-09.pdf`, accessed at November 2018.

35. Chang, C., N. Nikaein, R. Knopp, T. Spyropoulos and S. S. Kumar, "FlexCRAN: A flexible functional split framework over ethernet fronthaul in Cloud-RAN", *2017 IEEE International Conference on Communications (ICC)*, pp. 1–7, May 2017.

36. Chand, P., R. Mahapatra and R. Prakash, "Energy Efficient Coordinated Multipoint Transmission and Reception Techniques-A Survey", *International Journal of Computer Networks and Wireless Commun. (IJCNWC)*, Vol. 3, No. 4, 2013.

37. Dongare, A., R. Narayanan, A. Gadre, A. Luong, A. Balanuta, S. Kumar, B. Iannucci and A. Rowe, "Charm: Exploiting Geographical Diversity through Coherent Combining in Low-Power Wide-Area Networks", *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 60–71, April 2018.

38. Huiyu, Y., Z. Naizheng, Y. Yuyu and P. Skov, "Performance evaluation of coordinated multipoint reception in CRAN under LTE-Advanced uplink", *7th International Conference on Communications and Networking in China*, pp. 778–783, August 2012.

39. Irmer, R., H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results", *IEEE Communications Magazine*, Vol. 49, No. 2, pp. 102–111,

2011.

40. Davydov, A., G. Morozov, I. Bolotin and A. Papathanassiou, "Evaluation of Joint Transmission CoMP in C-RAN based LTE-A HetNets with large coordination areas", *2013 IEEE Globecom Workshops (GC Wkshps)*, pp. 801–806, December 2013.

41. Bassoy, S., H. Farooq, M. A. Imran and A. Imran, "Coordinated Multi-Point Clustering Schemes: A Survey", *IEEE Communications Surveys Tutorials*, Vol. 19, No. 2, pp. 743–764, April 2017.

42. *VMware CP*, `vmware.com/latam/products/capacity-planner.html`, accessed at November 2018.

43. *Lanamark*, `https://www.lanamark.com/features.html`, accessed at November 2018.

44. Xu, A.-p. X. C., "Energy Efficient Multiresource Allocation of Virtual Machine Based on PSO in Cloud Data Center", *Mathematical Problems in Engineering*, Vol. 2014, pp. 1–8, 2014.

45. Garey, M. R. and D. S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness*, W. H. Freeman, San Francisco, 1979.

46. Korf, R. E., "A New Algorithm for Optimal Bin Packing", *Eighteenth National Conference on Artificial Intelligence*, pp. 731–736, American Association for Artificial Intelligence, Menlo Park, CA, USA, 2002.

47. Beloglazov, A. and R. Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers", *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, pp. 577–578, May 2010.

48. Beloglazov, A., J. Abawajy and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", *Future*

Generation Computer Systems, Vol. 28, No. 5, pp. 755–768, 2012.

49. Boykov, Y. and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 9, pp. 1124–1137, 2004.

50. Karger, D. R., "Global Min-cuts in RNC, and Other Ramifications of a Simple Min-out Algorithm", *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '93, pp. 21–30, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1993.

51. Karger, D. R. and C. Stein, "A New Approach to the Minimum Cut Problem", *J. ACM*, Vol. 43, No. 4, pp. 601–640, July 1996.

52. *SimPy - Process based discrete event simulation framework for Python*, `https://simpy.readthedocs.io/en/latest/`, accessed at December 2018.