Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

PhD Thesis

Enhanced PON Architectures for Converged Access Networks for 5G and Beyond

Author:
Sandip Das

Supervisor:
Prof. Marco Ruffini

16th February 2022

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed:

Sandip Das, 16th February 2022

Summary

In the past few years, the existing telecommunication networks are challenged to support the growing number of internet-reliant devices (such as smartphones and other smart devices), increasing penetration of mobile and residential broadband connectivity, and the escalating popularity of multimedia streaming applications. Besides, the need for supporting various emerging applications such as tactile internet, tele-medicine, telesurgery, Intelligent Transport Systems (ITS) and Machine Type Communication (MTC) calls unprecedented high data rate, ubiquitous availability, ultra-low latency, high reliability and network robustness. These multidimensional network requirements have created the drive for standardization of the telecommunication networks to Fifth Generation (5G). Ultra-high wireless data rate, ultra-low latency and high reliability are identified as some of the key network requirements in 5G networks. Cloud Radio Access Networks (Cloud-RAN) and progressive cell densification are currently being regarded as some of the most promising technology solution in 5G to support such network requirements. However, these technologies pose a ultra-high bandwidth and ultra-low latency requirements in the fronthaul transport access network which current transport access technologies cannot support cost-effectively. In this dissertation, we explore possible technological enhancements in the transport access technologies to support such requirements. The broad scope of this dissertation concerns cost-effective optical transport network design to support ultra-high bandwidth and stringent latency requirements in the converged access. As optical technologies for fronthaul transport in the access still remain as the most costly option due to its large scale of deployment, a reduction of cost in the optical access would have a multi-fold impact on the Capital Expenditure (CapEx). Therefore, this dissertation, focuses on Passive Optical Network (PON) as the main contender for the fronthaul transport and addresses some of the key challenges,

such as ultra-high bandwidth and ultra-low transport lateny.

We first address the ultra-high fronthaul transport bandwidth requirements in Cloud-RAN based next-generation converged access. We provide a solution to enable statistical multiplexing of cells for a fully centralized Cloud-RAN (CPRI split) over shared fronthaul media. This is achieved by dynamically adjusting the cell bandwidth depending on the cell load with the help of a software defined network control. Using theoretical analysis and discrete event simulation, we show that this scheme enables statistical multiplexing while substantially reducing the blocking probability in a TDM-PON based shared fronthaul media.

Next, we address the stringent latency requirements of fronthaul transport in Cloud-RAN based next-generation converged access. For this, we propose a virtualized EAST-WEST PON architecture supporting direct communication between PON endpoints to enable ultra-low latency fronthaul transport for Cloud-RAN cells with functional split over Multi Access Edge Computing (MEC) enabled converged access. Using experimental feasibility demonstration and discrete event simulation, we prove that our EAST-WEST PON architecture can maintain the transport latency below a given threshold by dynamically offloading functional split computation across MEC nodes using dynamic virtual PON slicing technique. This enables the convergence of mobile and MEC nodes, delivering deterministic low-latency performance under highly dynamic traffic scenarios.

Although, the above virtualized EAST-WEST PON architecture with dynamic virtual PON slicing technique can address the stringent latency requirements in the converged access, the optimal formation of such virtualized PON slices to maintain low latency while maximizing the statistical multiplexing of Cloud-RAN cells is a challenging task. This motivated us to explore optimization techniques to form such optimal virtual PON slices over TWDM-PON based fronthaul transport network. Toward this, we proposed a mixed-analytical iterative optimization method that computes optimal virtual PON slice configuration. With theoretical analysis (based on queuing theory and discrete optimization) and discrete event simulation, we show that the proposed mix-analytical iterative optimization method can compute optimal virtual PON slices in timescales suitable for real-time or near-real time network optimization.

Finally, this dissertation concludes with the summary of the contributory works and providing some of the open issues and future research challenges that emerged while carrying out the technical contributory works in this dissertation.

Acknowledgements

As I wind up this dissertation secluded in my home while the world is fighting with the second wave of Covid-19 pandemic, I feel deep gratitude to the people who were always there with me during my ups and downs and offered their help, support and companionship whenever I needed them.

Firstly, this research would not have been possible without the guidance and support of my supervisor Prof. Marco Ruffini. His faith in my abilities, continued trust and constant encouragement even at the low-times of my PhD years have been the prime factor behind the outcome of this dissertation. I feel fortunate to have him as my PhD supervisor, teacher, mentor and a good friend. I also recognize the great benefits obtained from the conversations with Prof. Nicola Marchetti, Prof. Luiz Da Silva and Prof. Linda Doyle.

I cannot help but recognize the role of CONNECT had in keeping me inspired during my PhD years and shaping my research career. Being a part of CONNECT and spending time with the awesome people at Dunlop Oriel House during the pre-covid times made me grow both as a researcher and a human being. Here, I may be able to just name a few but there are several others during my time who one way or other contributed towards developing my research spirit. First and formost, I express my gratitude to Prof. Linda Doyle and Prof. Luiz Da Silva for being such great leaders of CONNECT and providing fantastic workplace. I would also like to extend my gratitude to the following fantastic people of Dunlop Oriel House: Dr. Indrakshi Dey (for her irreplaceable company and support at all times), Nima (for being my go-to person for any queries), Frank (for always providing his technical expertise throughout my research), Joao and Erika (for being such a joyful friend), Atri and Fadhil (for being awesome table tennis buddies), Francisco, Pedro and Johnathan

(for inspiring me to fooseball), Alan, Ramy, Parna, Arman, Merim, Jernej, Sanwal, Tom, Andrea, Jacek, Yi, Boris, Andrei, Diarmuid, Maicon, Christian, George, Jean, Natal, Ana, Matias, Ramiro, Adam, Johan, Avishek, Dennis, and Aleksandra. I would also like to extend my thanks CONNECT's administrative staff Shirley, Catherine, Monica, Lassane, Andrew, Frank, Sean, Mark, and Martin for helping me out with whatever administrative issues I may have had during my time here.

Most importantly, all of this would not have happened if it had not been for the support of my family. I would espeially like to express my heartfelt gratitude to my parents Panchu and Doli, and my brother Sudip (also my best friend) for their unconditional love and support. I know you are proud of me and I would like to dedicate this thesis to you.

Contents

List of Figures

List of Tables

1 Introduction

Introduction

In the last decade, we have witnessed a massive growth in mobile data connectivity, which is expected to continue steadily for the foreseeable future [7, 8]. Therefore, in the current era of telecommunication, existing technologies are challenged to support the ever-increasing demand for mobile data traffic. In addition to this, mobile communication industries are now facing challenges to support diversified specialized services such as ultra-low-latency communication, Internet of Things (IoT), intelligent transport systems etc. using the existing Fourth Generation (4G)-Long Term Evolution (LTE) technologies. This led to the evolution of Fifth Generation (5G) networks that are expected to support this future demand for data traffic and diversified telecom services. Therefore, 5G networks are more complex with progressive densification of cells to meet the high user-throughput requirement.

Addressing Fronthaul Transport Capacity Requirements in Cloud-RAN based Converged Access

Cloud Radio Access Networks (Cloud-RAN) [9] has emerged as one of the major technology for 5G to address these network requirements. In Cloud-RAN, the baseband and protocol stack processing of several cells (generally divided into two parts: Distributed Unit (DU) and Central Unit (CU)) are centralized and virtualized at a cloud processing site, whereas the cell and antenna related processing is kept at cell sites called Radio Unit (RU), to better achieve cell densification. Cloud-RAN has evolved from the traditional centralized Radio Access Networks (RAN) system where the antenna sites were highly simplified with only the antenna-related processing and called Remote Radio Unit (RRU) whereas the rest of the protocol stack processing (called Baseband Unit (BBU)) is centralized at a central server. This

technology requires the raw I/Q data samples to be carried to the BBU via the fronthaul link standardized as the Common Public Radio Interface (CPRI) interface [10]. This traditional centralized RAN system with CPRI has the advantage of centralized resource allocation and ability to carry out all advanced coordination across multiple cells such as Coordinated Multipoint (CoMP) communication, distributed beamforming etc. [11]. However, as 5G brings progressive densification of cells with much higher cell bandwidth and up to 64 antennas per cell site, traditional centralized-RAN systems with CPRI no longer becomes sustainable. For example, a fronthaul capacity of more than 1 Gbps (1.228 Gbps) is required per Antenna Carrier (AxC) for a 20 MHz LTE-RRU [12]. When scaled to a 20-MHz bandwidth, 8-channel Multiple Input Multiple Output (MIMO) systems over three sectors (which is the minimum bandwidth and antenna configuration for 5G), the required capacity reaches 150 Gbps. Furthermore, CPRI transports data at a fixed rate regardless of the cell traffic. Therefore, it does not allow statistical multiplexing over a shared fronthaul medium. Cloud-RAN addresses these shortcomings of centralized-RAN by distributing the protocol stack processing between antenna unit RU and protocol processing unit (DU and CU) by providing few possible split points at the protocol stack (turns into centralized-RAN for the case with CPRI split) to reduce the fronthaul link capacity requirement. Though Cloud-RAN loses a few centralization benefits as compared to fully centralized-RAN, it offers great scalability and sustainability. Furthermore, as 5G networks are expected to co-exist with previous 4G networks to offer backword compatibility, any innovation to enable statistical multiplexing in Cloud-RAN with CPRI split would have a significant impact as it would enable statistical multiplexing of cells over shared fronthaul even with the CPRI link.

Addressing Fronthaul Transport Latency Requirements in Cloud-RAN based Converged Access

Another key challenge is to provide fronthaul services to support Ultra Reliable Low-Latency Communication (URLLC) for various 5G applications that require ultra-low latency with high reliability such as connected logistics, autonomous vehicles,

mission-critical control, remote surgery etc. Although Cloud-RAN has the potential to achieve high reliability by means of advanced CoMP and enhanced Inter Cell Interference Coordination (ICIC) schemes, it imposes strict latency requirements on fronthaul transport. For example, a Cloud-RAN with a 7.2 functional split, that can perform advanced CoMP schemes, imposes a stringent latency requirement of the order of several hundreds of microseconds. In order to meet these requirements, the CU and/or DU processing is brought closer to RUs by deploying Multi Access Edge Computing (MEC) servers at the access network to reduce the fronthaul propagation latency. In this regard, although, traditional point-to-point fiber-based fronthaul solution can still work to support MEC based Cloud-RAN, with the progressive cell-densification in the foreseeable future, the cost of using such point-to-point fiber-based solutions will soar to unsustainable levels. Time-Wavelength Division Multiplexing (TWDM)-Passive Optical Network (PON) therefore which was primarily used for residential access, has recently gained wide attention as a cost-effective alternative to provide fronthaul services in 5G, as it provides great scalability and can already re-use the existing fiber deployments. Figure. 1.1 shows an example of Cloud-RAN based Fixed/Mobile converged architecture enabled with TWDM-PON based optical transport.

The traditional TWDM-PON supports NORTH-SOUTH bound communication whereas MEC based Cloud-RAN creates EAST-WEST traffic flow across the fronthaul network for example, when DU processing has to migrate across different MEC nodes for load-balancing and handover scenarios. Therefore, enhancement of traditional TWDM-PON architecture to support EAST-WEST bound traffic together with NORTH-SOUTH bound traffic (thus MESH traffic pattern) becomes an attractive research challenge with several research works explored such architecture to enable the support for MESH traffic pattern over TWDM-PON based fronthaul [13], [14]. However, as the cell deployment gets increasingly densified with the growing number of MECs to support the processing at the cell edge, management and configuration of transport networks during scenarios such as DU migration or handover becomes increasingly complex and unsustainable. Virtualization of PON transport network is a poten-

tial candidate solution to address this challenge as using PON virtualization, one can create virtual interest sub-groups of Optical Line Terminal (OLT) and Optical Networking Units (ONUs) called virtual-PON within a physical PON deployment. This can greatly simplify and abstract the network at a higher level to dynamically reconfigure the network depending on the scenarios and network requirements.

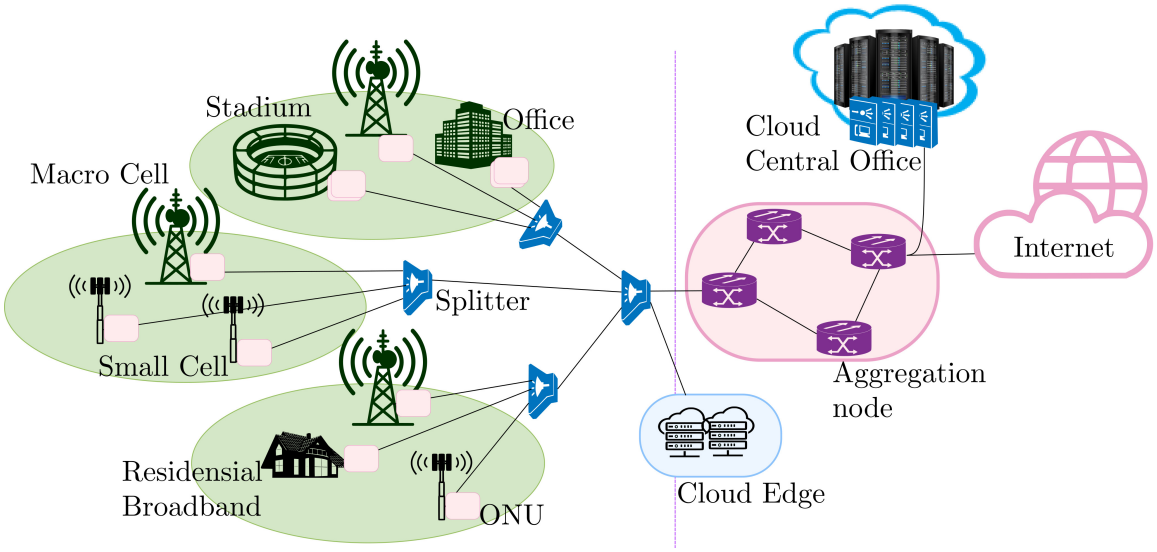

Figure 1.1: Cloud-RAN based converged access network for 5G and beyond

1.1 Research Questions

1. How to improve the statistical multiplexing of cells over Centralized Radio Access Network (C-RAN) to transport multiple fronthaul links effectively over a shared TWDM-PON fronthaul in order to meet the high bandwidth requirements of converged access in 5G and beyond?

2. How the ultra-low end-to-end latency requirements for 5G and beyond can be satisfied through PON architectural enhancements in transport network for MEC based Cloud-RAN? What architectural modifications are required to support them?

3. How can we use transport network virtualization to support multiple Cloud-RAN, MEC-based services, such as to meet the high fronthaul transport bandwidth required in Cloud-RAN based converged access by exploiting the stat-

istical multiplexing over a shared fronthaul while simultaneously meeting the target ultra-low end-to-end latency requirement?

1.2 Contributions

In this study of enhanced PON architectures for converged access networks for 5G and beyond, the following primary contributions are made.

1.2.1 A variable rate fronthaul for Cloud-RAN

This work proposes a novel Variable Rate Fronthaul (VRF) scheme for Cloud-RAN that enables statistical multiplexing of fronthaul links over shared PON media. The proposed mechanism provides a significant reduction of fronthaul bandwidth requirements while maintaining the simplicity and highest degree of centralization in Cloud-RAN.

1.2.2 PON virtualization with EAST-WEST Communication for Ultra-low latency converged MEC

In this work, we propose a novel Mobile Fronthaul (MFH) architecture based on PON architecture enhancements and virtualization that allows direct communications between MEC nodes (enabling EAST-WEST communication). This together with traditional NORTH-SOUTH communication in PON, enables the support for MESH traffic pattern across the fronthaul access network. The proposed virtualized EAST-WEST PON architecture can maintain the system latency below a given threshold by dynamically offloading functional split computation across MEC nodes using our proposed dynamic virtual PON slicing technique.

1.2.3 Optimal virtual PON slicing to support mesh traffic pattern in MEC based Cloud-RAN

In this work, we provide a mixed analytical-iterative optimization model to compute optimal virtual PON (vPON) slice allocation, providing MESH access connectivity

to next-generation MEC-based Cloud-RAN. Optimal slice allocation is provided in timescales compatible with real-time or near real-time operations.

1.3 Thesis Organization

The Organization of the dissertation is as follows.

- In Chapter 2, we give an introduction along with a state-of-the-art review of the underlying concepts addressed in this dissertation. It begins with an introduction to Cloud-RAN, functional split and Next Generation RAN (Ng-RAN) architecture (Section 2.1) as the access network technologies which is the main research area in a broader context of this dissertation. In Section 2.2, we provide an introduction to PON as a potential contender for fronthaul transport solution in converged access and discuss its related technological advancements. In Section 2.3, we provide a brief background on MEC and review its potential use cases, service scenarios and discuss the standardization aspects of its architectures. In Section 2.4, we provide a brief background on the network virtualization technologies showing how they can facilitate network slicing. We cover a brief background on SDN and NFV technologies to facilitate network slicing (in Section 2.4.1), and extend this discussion focusing towards software-defined optical transport networks, and PON virtualization to show how these technologies can facilitate end-to-end slicing of the transport network in TWDM-PON based optical transport networks.

- In Chapter 3, we introduce the simulation and hardware tools that are used to evaluate the performance of the proposed schemes in the successive technical contributory chapters and briefly introduce how these tools are used. It begins with the discussion of discrete event simulation in MATLAB (Section 3.1) using its SimEvents Toolbox. Next, in Section 3.2, we discuss OMNET++ based discrete event simulation which provides a more comprehensive network simulation experience through its C++ based discrete event libraries and discuss its pros and cons against the MATLAB's SimEvents. In Section 3.3, we discuss discrete

optimization using MATLAB through its optimization toolbox. Finally, in Section 3.4, we discuss some relevant details about the implementation overview and the testbed setup that we have used in the technical contributory chapters.

- In Chapter 4, we present a VRF scheme for Cloud-RAN. This chapter covers the following: introduction and the related state-of-the-art (Section 4.1), description of the proposed scheme and the system model (Section 4.2), theoretical analysis of the proposed scheme (Section 4.3), details of the simulation framework and the corresponding results (Section 4.4 and Section 4.5), and finally concludes in Section 4.6.

- In Chapter 5, we present an EAST-WEST PON architecture based on PON virtualization to enable ultra-low latency in MEC based converged access. This chapter covers the following: background and the related state-of-the-art (Section 5.1 and Section 5.2), description of the proposed architecture (Section 5.3 and Section 5.4), demonstration of the experimental feasibility (Section 5.5), performance evaluation through discrete event simulation (Section 5.6), and finally concludes in Section 5.7.

- In Chapter 6, we present a mixed-analytical iterative optimization method that computes optimal virtual PON slice configuration in a MESH-PON type fronthaul network to support ultra-low latency under dynamic traffic scenarios over a Cloud-RAN based converged access with heterogeneous functional split. This chapter covers the following: background and the related state-of-the-art (Section 6.1), system architecture (Section 6.2), description of the proposed scheme (Section 6.3), performance evaluation through theoretical analysis and discrete event simulation (Section 6.4), and finally concludes in Section 6.5.

- Finally in Chapter 7, we provide concluding remarks by providing a summary (Section 7.1) and discussing some of the open research challenges and possible future works (Section 7.2) that are inspired by the topics discussed on the technical contributory chapters.

1.4 Dissemination

1. **S. Das** and M. Ruffini, “A Variable Rate Fronthaul Scheme for Cloud Radio Access Networks," in IEEE Journal of Lightwave Technology, vol. 37, no. 13, pp. 3153-3165, 1 July 1, 2019, DOI: 10.1109/JLT.2019.2912127.
2. **S. Das** and M. Ruffini, “PON Virtualisation with EAST-WEST Communications for Low-Latency Converged Multi-Access Edge Computing (MEC)," 2020 Optical Fiber Communications Conference and Exhibition (OFC), San Diego, CA, USA, 2020, pp. 1-3.
3. **S. Das**, F. Slyne, A. Kaszubowska and M. Ruffini, “Virtualized EAST-WEST PON architecture supporting low-latency communication for functional split based on multiaccess edge computing." in OSA Journal of Optical communications and networking, vol.12, no. 19, pp. D109-D119, October 2020, DOI: 10.1364/JOCN.391929.
4. **S. Das** and M. Ruffini, “Optimal virtual PON slicing to support ultra-low latency mesh traffic pattern in MEC-based Cloud-RAN," in 25th International Conference on Optical Network Design and Modelling (ONDM), July 2021 (submitted)

The following works are the outcome of the collaboration with other researchers loosely related to the topic of the dissertation but are not described in detail in the dissertation.

5. Y Li, M Bhopalwala, **S. Das**, J Yu, W Mo, M Ruffini and D C. Kilper, “Joint Optimization of BBU Pool Allocation and Selection for C-RAN Networks", in 2018 Optical Fiber Communications Conference and Exposition (OFC), San Diego, CA, USA, 2018, pp. 1-3.
6. J. Yu, Y. Li, M. Bhopalwala, **S. Das**, M. Ruffini and D. C. Kilper, "Midhaul transmission using edge data centers with split PHY processing and wavelength reassignment for 5G wireless networks," 2018 International Conference on Op-

tical Network Design and Modeling (ONDM), Dublin, Ireland, 2018, pp. 178-183, DOI: 10.23919/ONDM.2018.8396127.

2 Background

Background and State-of-the-Art

In this chapter, we set the background by introducing a few state-of-the-art technological concepts in a broader context of the research question that lays the foundation of the thesis. It covers a brief background of the standardization efforts as well as some important related technological concepts including Cloud Radio Access Networks (Cloud-RAN), Multi Access Edge Computing (MEC), Passive Optical Network (PON), virtualization, and network slicing. While discussing such concepts, we keep our discussion focused to answer the following question as a broader context of this dissertation: "What are the motivations and implications of using next-generation fronthaul/backhaul architecture for converged access networks in Fifth Generation (5G) and beyond?"

2.1 Cloud-RAN, Functional Split and Next Generation RAN (Ng-RAN) Architecture

2.1.1 Cloud-RAN

As the deployment of 5G networks is picking up pace and standardization for different 5G components is being finalized, Cloud-RAN has already been identified as a key technology driver. In a typical 4G deployment, the Radio Access Networks (RAN) is distributed where each base station (eNodeB) consists of two parts: an antenna unit which is responsible for the RF-related processing, and a protocol processing unit which is a custom-built hardware unit that processes the baseband telecom protocol stack. The protocol processing unit is placed in the equipment room and connected

with the antenna unit via optical fiber. As 5G is expected to see progressive cell densification, the RAN function calls for enhanced flexibility, ease of management, improved Inter Cell Interference Coordination (ICIC) through advanced Coordinated Multipoint (CoMP) etc. As the network grows denser, supporting such requirements using the traditional Distributed RAN (D-RAN) technology becomes unsustainable. Cloud technology offers a great alternative for the deployment of such RAN functions with much higher flexibility and scalability while also bringing in the centralization benefits [15]. Cloud-RAN refers to realizing 5G RAN processing functions over a generic compute platform in cloud compute resources instead of purpose-built hardware [16]. In the first generation Cloud-RAN, the protocol processing unit (called Baseband Unit (BBU)) of several cells are virtualized and centralized at the cloud processing facility while the antenna unit containing only the RF functionality (called Remote Radio Unit (RRU)) is kept at the cell site to exploit the centralization benefits of RAN such as BBU resource sharing, enhanced coordination between cells etc. The transport interface between the BBU and RRU is known as fronthaul transport interface. Common Public Radio Interface (CPRI) is the first transport interface protocol that is standardized and also deployed in the first generation Cloud-RAN.

2.1.2 Functional Split and Statistical Multiplexing of Fronthaul in Cloud-RAN

CPRI transports digitized I/Q baseband samples between RRU to BBU over the fronthaul link at a fixed rate. While CPRI based fronthaul gained a lot of attraction towards the initial realization and deployment of Cloud-RAN, it imposes a heavy bandwidth burden on the fronthaul link. For example, for a 20 MHz LTE RRU with only one antenna per unit, the CPRI fronthaul bandwidth requirement already surpasses 1 Gbps (1.2288 Gbps). When scaled to a future 5G system with 100MHz bandwidth, 8-channel MIMO systems over three sectors, the required fronthaul capacity reaches 150 Gb/s, which makes it unsustainable to provide fronthaul services even via point-to-point fiber based Wavelength Division Multiplexing (WDM) optical link. In order to ease the severe burden on the fronthaul, many solutions have been

proposed. On the link level, one direct way is to increase the fronthaul capacity, such as using point-to-point bidirectional fiber, wavelength-division multiplexing etc. The other is to reduce the required data rate on the fronthaul, by means of baseband signal compression (Compressed CPRI), Functional Split processing.

Compressed CPRI technology uses an efficient CPRI compression schemes to reduce the fronthaul rate and consequently relax the bandwidth burden in the fronthaul network [17]. Possible CPRI compression techniques include reducing signal sampling rate [18], use of non-linear quantization [19], frequency sub-carrier compression or IQ data compression techniques [20] etc.

Although compressed CPRI has the advantage of reducing the fronthaul rate, it nonetheless transports I/Q fronthaul data at a fixed rate, independent of the actual cell load, therefore doesn't allow statistical multiplexing over a shared fronthaul media. In order to relax the fronthaul bandwidth burden, *functional split* is introduced to split the baseband processing chain functionality between the cell site and the cloud processing site. There are more than 8 different potential split points defined in the Third Generation partnership Project (3GPP) TR 38.801 [21] and NGMN alliance ([1]) as shown in Figure. 2.1-A. Depending on the splitting point, functional splits are broadly classified into two categories High Layer Split (HLS) and Low Layer Split (LLS). Based on the functional split, in the Ng-RAN architecture defined on 3GPP TR-38.801, the BBU functionality is split into three functional units namely Central Unit (CU), Distributed Unit (DU) and Radio Unit (RU). CU is responsible for processing non-real-time L2 and L3 functions of the protocol stack, while DU is responsible for processing real-time L1 (PHY) and L2 scheduling functions. The split between the CU and DU is generally referred to as the HLS and the interface is defined and standardized as the F1 interface in [22]. RU may retain a part of the physical layer processing (L1) along with the antenna unit depending on the split point between DU and RU which is generally referred to as LLS. The standardization of the interface at the LLS between DU and RU is an ongoing activity in ETSI-3GPP [2], which is generally referred to as Fx in ITU-T G-Supplement-66 [6] and is being

standardized as Next Generation Fronthaul Interface (NGFI) by IEEE [3] and Open-Fronthaul interface by the O-RAN alliance [23]. In 3GPP TR 38.816 [2], split option 6-8 is recognized as the potential split options for LLS with split 7 further divided into three split points namely 7-1, 7-2 and 7-3 (or commonly referred to as 7.1, 7.2 and 7.3 respectively) in the protocol stack as can be seen from Figure. 2.2.

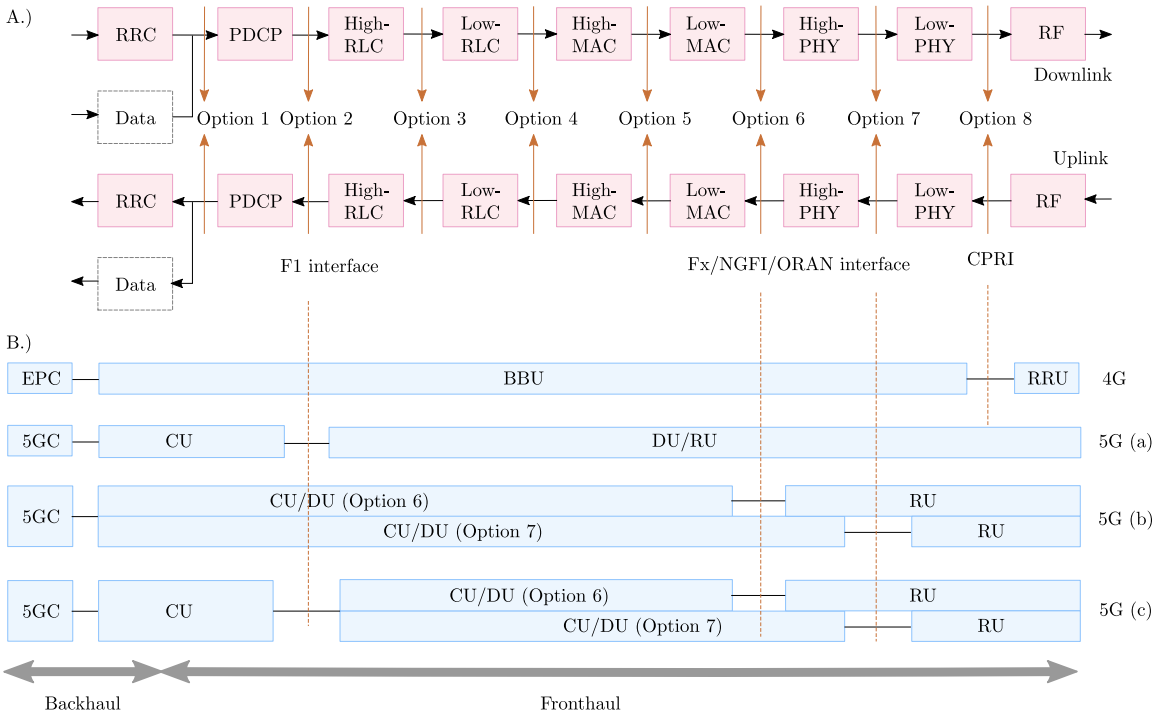

Figure 2.1: Functional Split architecture in next-generation Cloud-RAN (reproduced form [1])

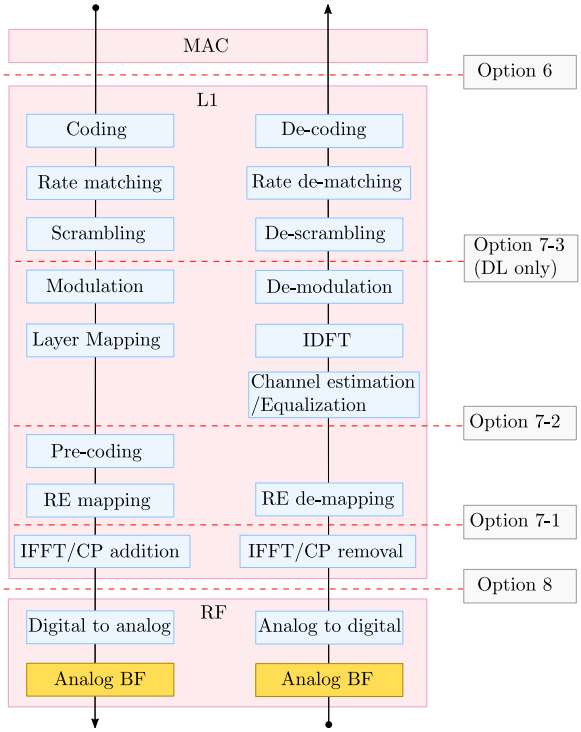

Figure 2.2: Possible, functional split options for LLS as defined in 3GPP TR38.816 [2] (Downlink (left) and Uplink (right))

The processing entities CU, DU, and digital processing part of RU (together is also the entire BBU) can be all collocated while leaving the RF processing part of RU (which is also the RRU) at the cell site (Figure. 2.1-B, row 4G), pairwise collocated (rows 5G (a) and 5G (b)), or all separate (row 5G (c)) depending on the 5G deployment phase and the use case. They are interconnected via external interfaces F1 and Fx, depending on which split option is considered. Table. 2.1 lists various functional splits and their corresponding fronthaul bandwidth and transport latency requirements. For the Cloud-RAN architecture in Figure. 2.1-A row 5G (c), where the deployment of both CU DU and RU are separated, the transport segment between CU and DU is sometimes referred to as the midhaul segment. Figure. 2.3 shows a physical realization of such network architecture (showing fronthaul, midhaul and backhaul segment) as proposed by the IEEE Next-Generation Fronthaul Interface (NGFI) group in [3] for converged network access in 5G and beyond.

Protocol split option	Required downlink bandwidth	Required uplink bandwidth	One way latency (order of magnitude)
Option 1	4 Gb/s	3 Gb/s	1-10 ms
Option 2	4016 Mb/s	3024 Mb/s	
Option 3	[lower than Option 2 for UL/DL]		
Option 4	4000 Mb/s	3000 Mb/s	100 to few 100 μsec
Option 5	4000 Mb/s	3000 Mb/s	
Option 6	4133 Mb/s	5640 Mb/s	
Option 7-1	10.1-22.2 Gb/s	16.6-21.6 Gb/s	
Option 7-2	37.8-86.1 Gb/s	53.8-86.1 Gb/s	
Option 7-3	10.1-22.2 Gb/s	53.8-86.1 Gb/s	
Option 8	157.3 Gb/s	157.3 Gb/s	

Table 2.1: The required fronthaul bandwidth and transport latency for different split options [6]

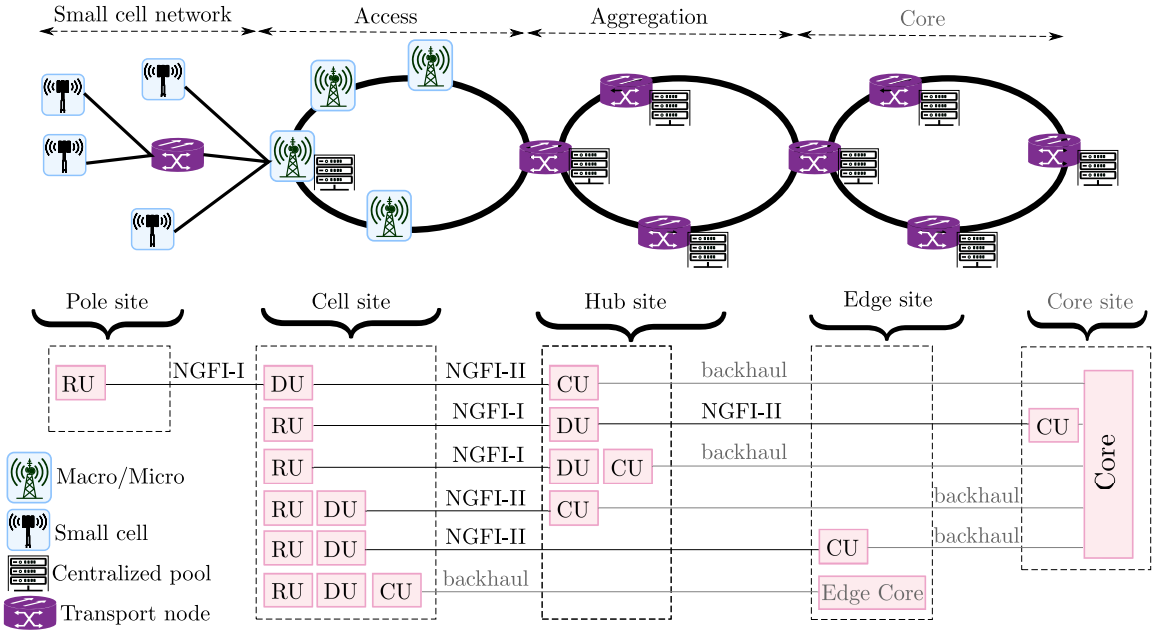

Figure 2.3: NGFI architecture for Cloud-RAN [3]

Statistical Multiplexing of C-RAN cells in shared fronthaul transport

On the network level, packet switching can provide hierarchical and flexible fronthaul networking, allowing multiple RRUs to form a cluster and share a fronthaul link,

which is adopted by the recent IEEE 1914 working group for next generation fronthaul interface (NGFI). Moreover, with appropriate RRU-BBU functional split, data rates in the fronthaul links can be traffic-dependent. As a result, the randomness of user traffic can be exploited to multiplex several RRU-BBU fronthaul links statistically by exploiting the variability of the aggregated fronthaul traffic. This would provide statistical multiplexing gain in the hope of reducing the fronthaul capacity demand. In Chapter 4, we define a Variable Rate Fronthaul (VRF) scheme over CPRI based fronthaul that enables the statistical multiplexing over shared PON based fronthaul transport to reduce the fronthaul capacity demand. This statistical multiplexing gain of C-RAN cells over a shared fronthaul media can be defined as follows [24].

$$MG_{FH} = \frac{\sum^{cells} Single\,Cell\,RRU - BBU\,Fronthaul\,Link\,Resources}{Aggregated\,Fronthaul\,Link\,Resources} \tag{2.1}$$

In Eq. (2.1), link resources specify sufficient bandwidth for a given deployment. They can be defined in several ways, as providing the peak throughput requested by users is costly: 1) 95th percentile of requested throughput, 2) 95th percentile confidence intervals of mean of requested throughput 3) peak requested throughput averaged over given time and 4) link data rate for which the application layer delay is acceptable.

2.2 PON for Fronthaul Transport

The transport access network has gradually become one the most important network entity in 5G and beyond. Consequently, there has been an intensive research effort in the last few years to address the challenges that are posed in the transport network in order to support the diversified network requirements in 5G and beyond. In the past decade, for 3G of 4G networks, the backhaul technology was mostly copper-DSL and point-to-point fiber Ethernet and microwave links. Among these, optical fiber-based backhaul delivers the best performance due to its superior transmission quality and long-reach capability without signal degradation. However, due to its high Capital Expenditure (CapEx), the deployment of fiber-based backhaul coverage was lesser

than its microwave or copper-DSL counterpart. However, with the recent evolution of RAN in 5G, calling for ultra-high transport bandwidth and communication reliability, the transport access networks are rapidly moving towards optical fiber-based fronthaul/backhaul, as the existing alternative transport access technologies (such as copper DSL or microwave links) fails to deliver such required performance. In the current 4G and initial phase of 5G deployment, fiber-based fronthaul networks are mostly WDM or point-to-point fiber-based in order to support the high bandwidth and low-latency requirements. As 5G is expected to foresee rapid cell densification, the cost of providing fronthaul services through WDM or point-to-point fiber would soar to unsustainable levels. Therefore, building a cost-effective fronthaul becomes a major research issue.

In this context, PON is recently being looked upon as a major technology enabler to provide cost-effective mobile fronthaul transport in the future 5G deployment. For over a decade, PON has been one of the most popular technology to provide residential broadband access via Fiber to the Home (FTTH) services due to its superior quality, high speed and future proofness. Because of its increasing availability and coverage, PON offers a great solution for 5G fronthaul transport as it can already use the existing deployed fibers with minimal network reconfiguration. Furthermore, over the past decade, research and development in PON technology have experienced rapid progress, as two standardization bodies (International Telecommunication Union (ITU) and Institute of Electrical and Electronics Engineers (IEEE)) have been working in parallel to standardize PONs. Consequently, two different standard PON systems are available in the market: ITU-T Gigabit Passive Optical Network (GPON) and IEEE Ethernet Passive Optical Network (EPON) with rapid evolution taking place in both.

EPON, first developed by the IEEE Ethernet in the First Mile (EFM) task force of 802.3 standards committee in 2004 [25], is based on the Ethernet protocol, modified to work over Point to Multi Point (P2MP) fiber. Starting from the first generation of EPON systems supporting a maximum throughput of 1 Gb/s, the later efforts

from the committee standardized support for 10 Gb/s (IEEE Std 802.3av-2009, 10G-EPON) in 2009 [26]. The most recent development in the EPON standardization is the added support for 25 Gb/s and 50 Gb/s in IEEE Std 802.3ca-2020 [27]. GPON on the other hand is a Quality of Service (QoS)-enabled variant of PONs that has been standardized by ITU. GPON enforces QoS by defining logical queues called Transmission Containerss (T-CONTs) for different service requirements and providing different transmission opportunities for various T-CONTs, for example, higher service frequency for delay-sensitive applications while longer transmission opportunities for bandwidth-hungry applications. Initial standardization of GPON is defined in G.984 series recommendations by ITU-T supporting a max data rate of 1 Gb/s [28]. This has evolved to support up to 10 Gb/s 10-Gigabit-capable Passive Optical Network (XG-PON) defined in G.987 series recommendations [4], finally state-of-the-art Next Generation Passive Optical Networks 2 (NG-PON2) systems supporting data rate up to 50 Gb/s defined in G.989 series recommendations [29].

While the evolution of PON systems to support high data rates up to 50 Gb/s can support most of the high bandwidth fronthaul requirements, one major bottleneck in using PON for fronthaul services is to support the low-latency requirements. This is especially an issue when the latency budget for fronthaul is very stringent, for example for a split between RU and DU below 4 (i.e split point between 4-8 inclusive), the fronthaul latency budget is few hundreds of microseconds as listed in Table. 2.1. This is because PONs inherently work on P2MP protocol, where the Optical Line Terminal (OLT) runs a Dynamic Bandwidth Allocation (DBA) algorithm to create uplink transmission opportunities for Optical Networking Units (ONUs) and conveys them in the downlink frame before ONUs can perform the uplink transmission.

2.2.1 DBA

DBA is the most important part of the PON system that is responsible for maintaining QoS and Service Level Aggrements (SLAs). It processes the uplink transmission opportunities for ONUs in PON. Interleaved Polling with Adaptive Cycle Time (IPACT) [30] and GigaPON Access Network (GIANT) [31] are the two most

popular DBA algorithms for PON. While the DBA for EPONs is mostly derived from IPACT, GPON DBAs are mostly derived from the GIANT algorithm.

Status Report based DBA (SR-DBA)

In general, the DBA algorithm allocates uplink transmission opportunities for ONUs based on the knowledge of the packets queued on the ONU buffer or traffic intensity. these uplink transmission opportunities (or uplink bandwidth allocation) are then conveyed to individual ONUs before they can start uplink transmission. This is accomplished using the Status Reports (SRs) from the ONUs possibly embedded in the frame header during the previous uplink transmission, therefore generally known as Status Reports based DBA or SR-DBA algorithm. The SR-DBA algorithm defined in the transmission convergence layer of the ITU-T G.987.3 XG-PON standard [4] controls the QoS in uplink by introducing a strict priority hierarchy in the bandwidth assignment. Specifically, this DBA algorithm allocates bandwidth according to the following hierarchical traffic classes.

- Fixed bandwidth (highest priority)
- Assured bandwidth
- Non assured bandwidth
- Best-effort bandwidth

The DBA procedure of [4] is illustrated in Figure. 2.4. Fixed bandwidth is the amount of bandwidth that is assigned to all ONUs regardless of their traffic load and buffer status. Following this, assured bandwidth is allocated to ONUs in proportion to their traffic load, up to a certain provisioned allocation limit. The sum of fixed bandwidth and the assured bandwidth is the guaranteed bandwidth that is assigned to each ONU. The surplus bandwidth following this assignment is allocated to ONUs in a non-assured form. Finally, the amount of uplink capacity that remains available after all the ONUs eligible for the non-assured bandwidth assignment have been saturated, and all the other ONUs have been assigned their respective guaranteed bandwidth components, is assigned to the eligible ONUs as best-effort bandwidth assignment.

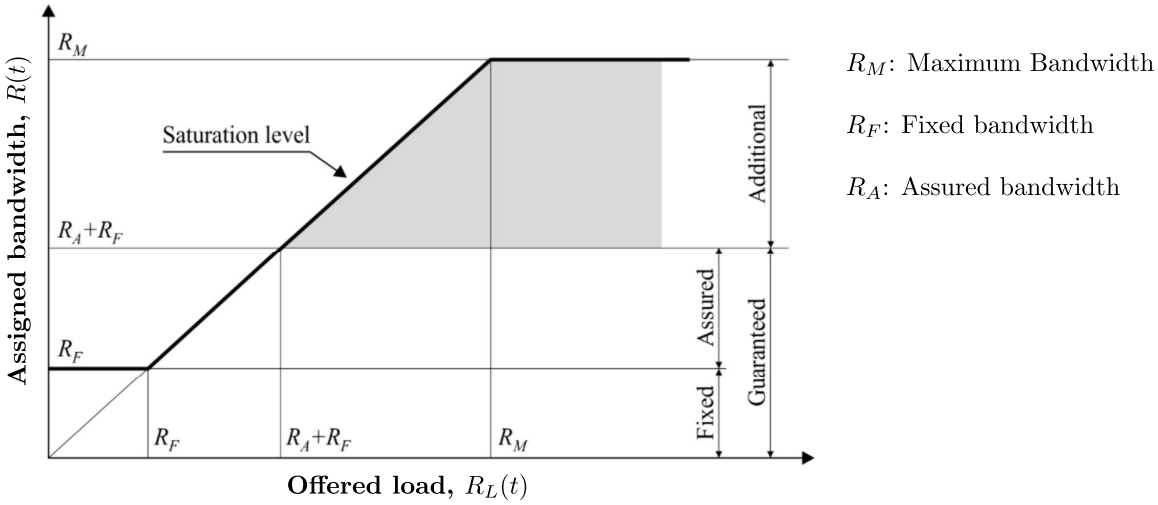

Figure 2.4: Assigned bandwidth components with respect to offered load (reproduced from [4])

Non-Status Report based DBA (NSR-DBA) or Traffic Monitoring based DBA (TM-DBA)

One of the main problems of SR-DBA is that the DBA algorithm heavily depends on the buffer/queue status reports from ONU. These buffer reports are sent to OLT during an uplink transmission cycle, in receipt of which the DBA in OLT assigns uplink bandwidth accordingly. This process introduces an inherent uplink delay due to status reporting and the packet queuing in the ONUs. TM-DBA is a non-status reporting variant of the DBA that predicts the buffer status in the traffic-bearing entities within the ONUs based on the observation of the uplink idle frames from the ONUs (called XG-PON Encapsulation Method (XGEM)) during the reception of the upstream bursts. There have been much research works on the NSR-DBA over the past few years. In [32], a traffic prediction technique is proposed based on the average of the traffic arrivals on past several arrival instants. The predicted traffic is then added to the queue occupancy report and sent to the OLT for processing traffic prediction-based bandwidth assignment. Further research on this area mainly focused on improving traffic prediction using data mining [33], fuzzy logic [34] and learning automation [35] etc. All of these works aim to achieve the same goal: to further reduce the uplink scheduling latency by improving the prediction accuracy, thus reducing the overall end-to-end latency and increasing the network throughput.

Coordinated DBA (CO-DBA) for Cloud-RAN and low latency fronthaul

Although NSR-DBA or TM-DBA can decrease the end-to-end latency and increase the network throughput depending on the accuracy of traffic prediction and the DBA algorithm at the OLT end, it heavily depends upon the prediction accuracy. Furthermore, the requirement of traffic prediction itself brings the issue of accurately getting ONU buffer status information at the OLT end. This would lead to over assignment or under assignment of uplink bandwidth in PON. In the case of over-assigned bandwidth, this would incur a penalty in uplink bandwidth efficiency, while in the case of under-assignment, the queuing effect in the ONU buffer would cause a significant delay in the successive packets. Therefore, the NSR-DBA or TM-DBA cannot meet the required latency of fronthaul transport in 5G with high reliability. For example, for a Cloud-RAN deployment with LLS split option 4-8 between DU and RU, the fronthaul has a deterministic transport latency requirement of the order of a few 100 μsec. To address this challenge, there have been much research works in recent years to reduce the latency due to DBA processing and buffer status reporting by coordinating the scheduling of PON transport with the wireless access in Cloud-RAN. The first solution (known as CO-DBA) was proposed in 2014 ([36]), where a coordination interface between the LTE BBU unit and the OLT DBA unit is set up to exchange the mobile scheduling information in advance. With this prior scheduling information, the OLT DBA can then proactively set the upstream bandwidth assignment for ONUs and sends the bandwidth grants ahead of the traffic arrival in the ONU buffer. This way, it can bypass the ONU buffer status reporting mechanism while also the packet buffering latency at the ONU gets eliminated. Figure. 2.5 illustrates this CO-DBA scheme as compared to the conventional SR-DBA mechanism. With the help of an experimental evaluation using a 10G EPON prototype, the authors show that an average of 100-150 μs fronthaul latency with jitter as low as 20μs can be achieved for 10-20 km fiber length. The first practical implementation of such coordination interface is discussed in [37] from the same research group, and in [38], a further improvement on CO-DBA is proposed which considers the data arrival

period of the packets in the ONU buffers to achieve a better jitter performance. This CO-DBA is recently being standardized as Cooperative Transport Interface (CTI) in the third amendment of ITU-T G.989.3 standard on the XG-PON Transmission Convergence (XGTC) layer of NG-PON2 [29] systems.

While this discussion on CO-DBA implicitly assumes that the PON bandwidth required for mobile scheduling is always available, in practice, this condition will not always be met, for example, if the uplink capacity is oversubscribed due to the fluctuation of the incoming traffic from RU-ONUs at certain time instances. Moreover, for certain scenarios, given a particular uplink PON bandwidth, CO-DBA may not meet the target latency due to the different arrival timings of the packets on the ONU buffer. Authors in [39] addresses this issue by devising an optimization framework to find the optimal packet forwarding order on a CO-DBA-based PON fronthaul transport so that an ultra-low target latency is met for all ONUs if feasible. Apart from this, in general, the average uplink latency in PON increases with the traffic load in the ONUs which should always be considered to achieve a certain end-to-end target latency.

2.3 MEC

With the widespread adoption and deployment of Cloud-RAN, commercial availability and the deployment of compatible cloud computing resources equipped with next-generation networking interface have suddenly become one of the most critical criteria in the deployment of 5G and beyond networks. From the end-user's point of view, 5G networks are expected to support applications that require ultra-low latency (e.g., less than 1 ms) with high reliability such as tactile internet, connected logistics, mission-critical control, etc. Due to this ultra-low end-to-end latency requirement, the propagation latency accounting for the fronthaul transport between the RU and the first processing unit DU needs to be cut down significantly. One obvious solution is to deploy the cloud processing units closer to the cell sites thus reducing the propagation distance between DU and RU. In fact, with the recent advancement in

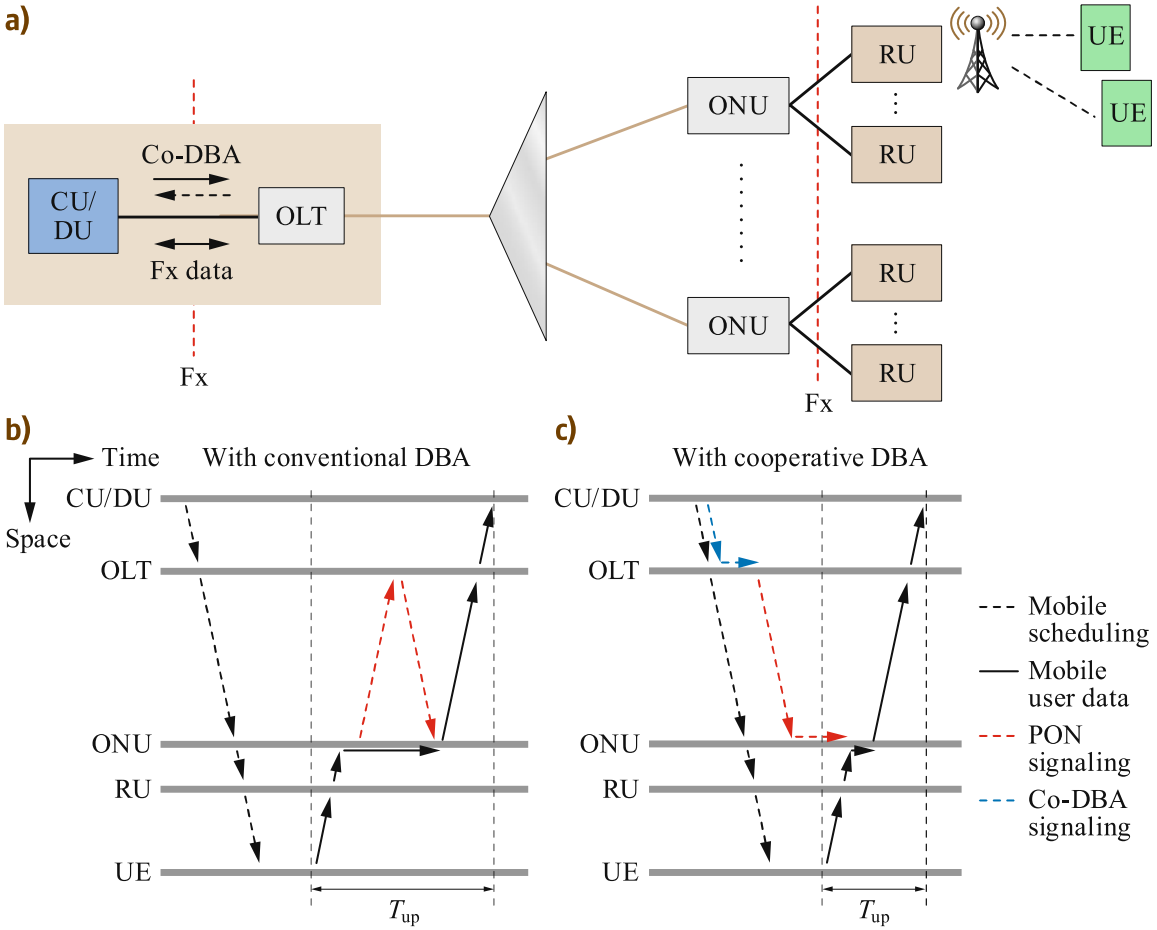

Figure 2.5: Illustration of Co-DBA based Fx fronthaul over TDM-PON network. a)-Network architecture. b)-The conventional SR-DBA scheme compared against c) Co-DBA scheme

cloud computing technologies, there has been a recent trend to increasingly deploy cloud computing units towards the network edges [40]. Harvesting this large amount of processing power and storage capacity distributed across the network edges can achieve the required performance for the computation-intensive and latency-critical RAN functions for low-latency 5G applications. This paradigm is called Mobile Edge Computing, which was first introduced by the European Telecommunications Standards Institute (ETSI) Industry Specification Group (ISG) in 2015 [41]. From 2017, the ETSI industry group changed its name to Multi-access Edge Computing (MEC) since its application reached way beyond computing mobile RAN functions, towards Wi-Fi and fixed access technologies. These use cases and the service scenarios are described as follows:

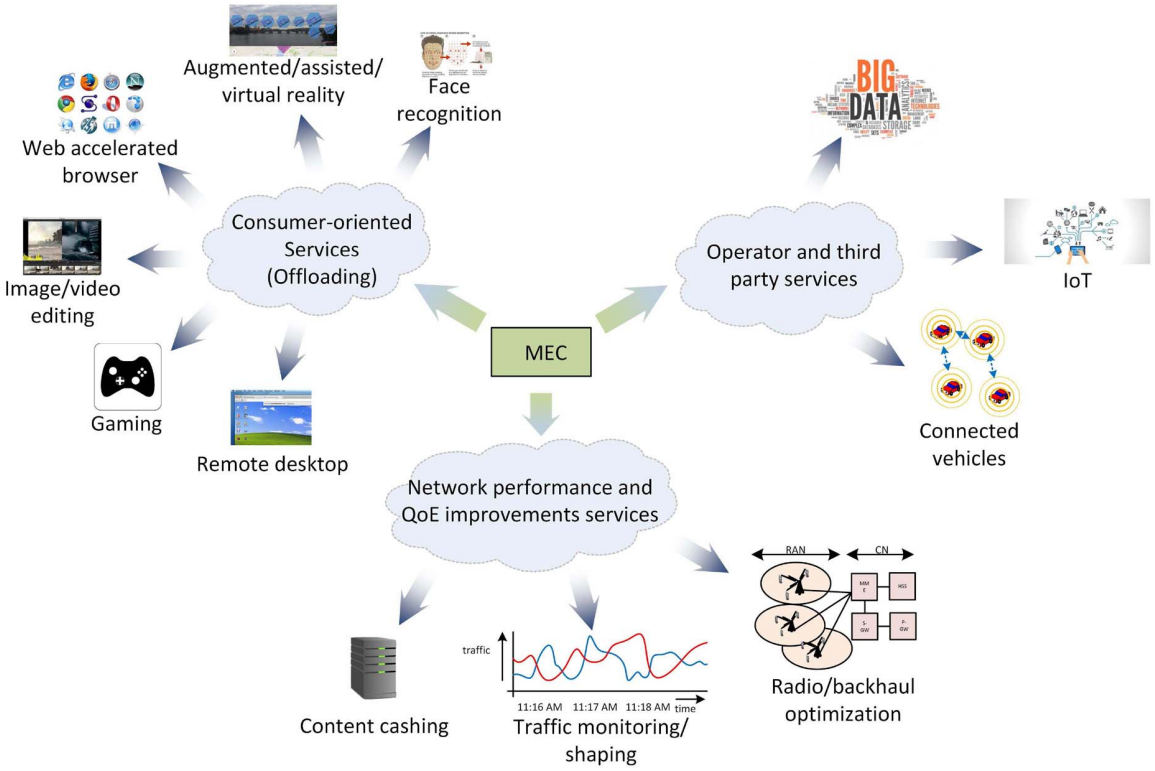

Figure 2.6: MEC use-cases and service scenarios

2.3.1 MEC Use Cases and Service Scenarios

MEC has a wide range of use cases and brings many advantages to various stakeholders. Depending on the use cases and service scenarios, this can be largely divided into three groups as described in [42], also depicted in Figure. 2.6. These are as follows:

Consumer-Oriented Services

This is where the end-users directly benefit from the MEC largely by means of computation offloading. Few application use cases are web-accelerated browsing [43], low-latency face/speech recognition or cloud-offloaded video editing [44], low-latency augmented or virtual reality using the cached information at the network edge [45], online gaming, tactile remote desktop, etc.

Operator and Third Party Services

This is where the network operators or the third-party service providers directly benefit from the MEC. A Few examples of such application are as follows:

- One use-case or application scenario is to exploit MEC as an aggregation point for Internet of Things (IoT) or use it as an IoT gateway [46]. As different IoT technology uses various radio technologies with diverse communication protocols, there is a need for an aggregation point or gateway for the distribution of messages and pre-processing before sending the processed data to the central IoT cloud.
- Another use case for MEC is in Intelligent Transport Systems (ITS) to extend the connected-car cloud into the network edge [47]. This enables low-latency exchange of messages between cars and roadside sensors for supporting latency-critical ITS applications.

Network Performance and QoE Improvement Services

This third category of the use case is mainly for optimizing the network performance and improving the Quality of Experience (QoE). Few examples of such application are as follows:

- One use case on this category is to exploit MEC for traffic shaping and re-routing of the traffic when the capacity of either the backhaul or radio links is degraded [48].
- Another way to improve the QoE and the network performance is to alleviate the backhaul load by proactively caching the contents at the network edge. processing and the storage of the MECs can be exploited to cache the most popular contents based on the content usage statistics on the geographical area containing the MEC. This would improve the network latency performance while also reducing backhaul congestion [49].

2.3.2 MEC Architecture

Since the introduction of Mobile Edge Computing in next-generation networks, there have been various solutions proposed in the literature to describe the architecture of MEC and its functionalities. Some notable solutions among them are Small Cell

Cloud (SCC) [50], Mobile Micro Cloud (MMC) [51], Fast moving personal cloud (MobiScud)[52], Follow me cloud [53], and CONCERT [54]. Besides all the above-mentioned architectural solutions, ETSI is also currently actively involved in the standardization of the MEC architecture in order to seamlessly integrate it into the mobile networks. Although the standardization of MEC is still in the early stage and not matured enough, the general and conceptual framework is already discussed and drafted in [55]. A reference architecture for MEC is also described by ETSI in [5]. Figure. 2.7 shows the reference architecture of ETSI reference MEC architecture which is composed of functional elements and reference points interconnection between. In practice, these functional blocks are Virtual Machine (VM) instances running on a virtualization infrastructure (for example data center).

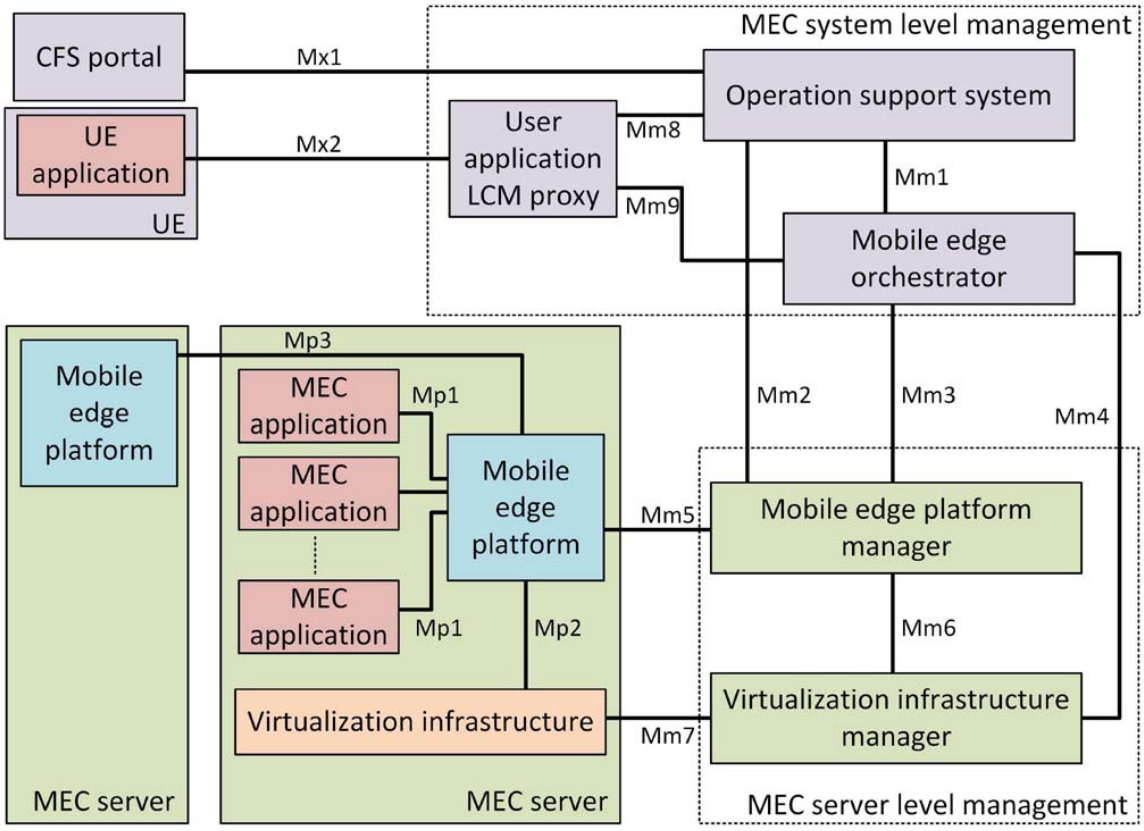

Figure 2.7: MEC reference architecture as provided by ETSI [5]

As illustrated in the Figure. 2.7, a MEC can be accessed directly by a user application on a UE or can be accessed by a third party customer or operator via a Customer Facing Service (CFS) portal. It manages the lifecycle of the application,

service authorization, traffic rules, etc. The end-user application interacts directly with MEC system level management which includes a Lifecycle Management (LCM) proxy, Operation Support System (OSS) and Mobile Edge Orchestrator. Upon reception of application requests, the LCM proxy mediate the connection requests to OSS which then decides if the request is to be granted or not. If granted, these requests are forwarded to the mobile edge orchestrator which provides the core functionality of the MEC server-level management as it maintains the overview of the network/compute/storage resources of the MEC server.

The MEC server-level management is responsible for allocation, management, and release of the network, storage and computation resources in the MEC server. Finally, the MEC server is the most important part of the ETSI MEC architecture as it represents the virtualized resources the application is using typically on a VM (such as a Linux Container (LXC), Docker, Kubernetes, etc.) on top of virtualization infrastructure.

In Chapter 4, we present a pool of BBU processing stacks initiated at the cloud server. Further, in Chapter 5 and Chapter 6, we present the notion of functional split processing instances of such protocol stack (DU and CU) hosted at the cloud edge (MEC server) in order to reduce the end-to-end latency. These BBU, CU and DU instances can be realized as individual MEC application instances (VMs, docker or containers as depicted in Figure. 2.7) which is orchestrated through a centralized controller. Therefore, in context of this dissertation, without the loss of generality we define MEC as simply a cloud edge server with limited storage and processing instances hosting dynamic instances of CU, DU and BBU processing.

2.4 Slicing and Access Network Virtualization

One of the key drivers behind the emergence of 5G systems is the need to support a diverse set of vertical industries such as manufacturing, healthcare, automotive, media and entertainment. Use cases originating from these verticals are very different from one another and therefore imposes a wide range of network requirements. Tra-

ditional 4G network architecture with a "one-size-fits-all" approach cannot support such a diverse set of network requirements. This calls for a next-generation network architecture that is highly scalable, flexible, reliable and easily maintainable.

Virtualization and softwarization of the network is an emerging trend that seeks to accomplish these network requirements by transforming the networks with software-based solutions. With the widespread availability of the Whitebox switches, programmable network cards and highly efficient general-purpose processors, technologies like Software Defined Networking (SDN) and Network Function Virtualization (NFV) are providing the required flexibility and programmability to dynamically create multiple virtual networks, each tailored to a particular use case over a common physical network. These separated virtual networks providing end-to-end communication with different QoE corresponding to different use case scenarios are called End-to-End (E2E) network slices.

In the remainder of this section, we briefly review the research efforts to adopt SDN and NFV to facilitate the convergence in the 5G and beyond networks.

2.4.1 SDN and NFV

In the pre SDN era, every minor feature change in the network end components would incur a significant cost to the network operators as the control plane of network equipment was generally vendor locked and tightly integrated with the data plane. SDN tackles this inflexibility by separating the control plane from the data plane and providing a unified control plane which is typically placed in a cloud central office. The ratio of the control-plane functions that are to remain at the network components or to be moved to the centralized controller forms a trade-off and a vast research domain by itself [56]. Although the separation of control and the data plane existed even before the emergence of SDN [57], SDN brings the cloud-centralization of the control plane functions using a unified framework, while leaving behind the data plane functionalities in the network devices. The advantages of this control plane centralization are most obvious in the 5G converged network architecture. In

the case of optical-wireless converged networks, the optical and the wireless components can communicate with each other through the unified centralized control plane. For example, in 5G, part of the hardware radio systems is rapidly being replaced by Software Defined Radio (SDR) to implement most of the protocol stack processing in software where any change of radio parameters can be performed dynamically with a significant level of user abstraction [58]. With the help of SDN Optical Transport Networks (OTN) [59], and control plane centralization of SDR systems ([60], [61]), end-to-end dynamic network reconfiguration can be facilitated seamlessly with minimal complexity.

The programmability of network components is another key technology that SDN offers. Although the concept of programmable network switches is not new, SDN brings it one step forward by providing a unified framework for the control of such programmable switches from a common central office. Specifically, SDN provides the user abstraction of the programmable network switches using a human interpretable format such as JSON or XML which is then processed and converted to programmable switch readable format such as Open-Flow [62] and NETCONF [63] to hide the complexity of network reconfiguration from the end-user [64].

NFV [65] is another concept that aims at virtualization and softwarization of network functions in commodity servers instead of running them in specialized hardware. Examples include the virtualization of Internet Protocols (IPs), Virtual Network Functions (VNFs) (e.g., load balancing, firewall, security) [66], or transport control functions such as path computation [67], etc.

SDN controlled RAN is an emerging technology driver for 5G and beyond aimed to leverage SDN functionalities to control RAN functions. Open Networking Foundation (ONF)'s SD-RAN [68] is one prominent example of such architecture (compatible with O-RAN [23]) which aims to control RAN functionalities from ONF's well established ONOS SDN controller. In Chapter 4, we define an SDN controlled variable rate fronthaul scheme which dynamically adjusts the fronthaul data rates by dynamically changing the cell bandwidth depending on the cell load, with the support of a Software

Defined Network (SDN) controller. This type of SDN based RAN controls can be achieved using SD-RAN architecture. Therefore, in the context of this dissertation SDN is referred to as a centralized entity for dynamic control RAN functions (for e.g., cell bandwidth in Chapter 4) without loss of any generality.

2.4.2 Software-Defined Optical Networks and Network Slicing

As optical networks such as Dense Wavelength Division Multiplexing (DWDM) or Time-Wavelength Division Multiplexing (TWDM)-PON is being considered as a promising solution for the transport access networks, there is a growing need to support various services (e.g. residential broadband, mobile fronthaul, and IoT gateway data transport). Incorporating such diverse services with current inflexible Optical Transport Network (OTN) with proprietary hardware components would incur huge CapEx and Operational Expenditure (OpEX) as different sets of proprietary devices would be required along with the interaction of a broad range of technologies.

Central Office Re-architected as Datacenter (CORD) [69] seeks to resolve this issue by replacing the proprietary purpose-built hardware with software running on commodity servers and off-the-self white-box switches and access devices. The heart of CORD uses XOS [70], a service orchestration layer built on top of OpenStack [71] and Open Network Operating System (ONOS) SDN controller that manages the scalable services running over CORD. The ONOS SDN controller [72] is responsible for enabling the communication and exchange of control messages between VNFs and the Whitebox hardware over the southbound interface. The communication protocol on the southbound interface includes Open Flow, NETCONF, Border Gateway Protocol (BGP) Simple Network Management Protocol (SNMP), etc. On the other hand, OpenStack is responsible for managing the network, compute and storage resources that are to be assigned to each VNF.

The CORD project, currently led by the ONF has gained wide attention globally with multiple industry partners contributing to the development and improvement of CORD [73]. At present, the CORD project has three categories: M-CORD, R-

CORD and E-CORD each corresponding to a particular service scenario. Mobile CORD (M-CORD) project is aimed to enable 5G services on CORD by virtualizing the BBU (CU and/or DU) and hosting VNFs for Open and Disaggregated Transport Network (OTDN) with SDN control [74]. On the other hand, Residential CORD (R-CORD) is primarily aimed at providing virtualization of the access technologies for last-mile residential broadband services. A key component of R-CORD is Virtual OLT Hardware Abstraction (VOLTHA), which provides virtualization and abstraction of OLT and provides ONOS SDN controllability via the southbound interface. The latest development on the CORD project is Converged Multi-Access and Core (COMAC) a multi-access variant of CORD aimed at bringing convergence to Operators' mobile, broadband access and core networks using a unified platform.

PON virtualization

As the development of CORD with support for OTDN is becoming the most promising technology to enable virtualization and bring convergence in access networks, there has been growing attention to make use of multiple logically separated Virtual Optical Networks (VONs) to provide specialized services over a common physical optical transport architecture. In a TWDM-PON architecture, this can be achieved by creating multiple distinct virtual PON slices (called virtual PONs (vPONs) slices) containing an OLT and a subset of ONUs over a physical TWDM-PON deployment. However, the idea of vPON is not a new concept as a similar concept of creating virtual interest groups over a physically deployed PON dates back to 2006 [75]. However, the SDN compatibility and the centralized control of such virtual PON slices from the central office makes it more attractive than ever. Especially, in 5G cloud-RAN systems, vPON over TWDM-PON can provide solutions to many use-cases and improve QoE such as improving connection reliability by reducing handovers [76], providing energy-efficient CU DU placement [77], joint allocation of radio and optical resources in Cloud-RAN [78], etc. In Chapter 5 and Chapter 6, we introduce the concept of dynamic virtual PON slicing approach where each TWDM OLT sharing an Optical Distribution Network (ODN) creates virtual interest group with a subset of ONUs

over different wavelengths. These virtual interest groups can then be reconfigured by dynamically migrating ONUs across different interest groups through dynamic tuning of ONU transceivers onto different OLT operating wavelengths. Therefore, we call this dynamic virtual interest groups as virtual PON slices (vPON slices) in this context of the thesis without the loss of any generality.

2.5 PON based Next-Generation Virtualized Fronthaul/Backhaul Architecture for converged access: Motivations and Implications

At this point, it is important to put the term "Next-generation fronthaul/backhaul architecture for converged access" in the context of this dissertation to explicit the scope of this thesis. As discussed in the past few sections on state-of-the-art technologies, the next generation access network (in 5G and beyond) should bring in massive cell densification, access network virtualization and fixed-wireless convergence to provide diversified services for 5G and beyond. However, current transport network technology cannot cope-up to support such evolution of access network. Therefore, the transport network for access should also evolve to support such converged access in 5G and beyond. The discussion on the past few sections in this chapter indicates that transport network virtualization, SDN controllability, and MESH connectivity are some of the candidate technologies for this evolution of transport networks. In the context of this dissertation, we broadly refer to this transport network evolution as next-generation fronthaul/backhaul. It should be noted that throughout this dissertation we confine ourselves to only optical fibre transport when discussing such fronthaul/backhaul architectures. Although in the later technical part of this thesis we narrow down our focus to PON, this dissertation addresses optical transport network architectures in a broader context.

One of the substantial implications of using TWDM-PON in the access is the cost benefit that it brings due to the sharing of fiber infrastructure. Especially, for the massive densification of cells in 5G C-RAN where a large number of RUs are needed to

be connected to a centralized cloud processing unit where DU and/or CU processing is hosted, TWDM-PON can greatly reduce the fiber cost as compared to the point-to-point solution. Although, the end-equipment costs for TWDM-PON based optical fronthaul solutions are higher than the point-to-point based solution, studies in the literature has shown that the overall cost of TWDM-PON is significantly lower [79]. A recent study from 5G PICTURE project [80] shows that the usage of TWDM-PON in the access can lead to a quantified cost benifits of at least 10%. Further savings can be exploited by sharing the TWDM PON optical distribution network fiber infrastructure also with the residential traffic running over legacy G-PON and XGS-PON as indicated in this study.

Therefore, after reviewing the related technologies and the state-of-the art literature relevant to the scope of the thesis, we are now in a position to provide an abstract answer to the following question to set the background of the dissertation: "What are the motivations and implications of using PON based next-generation fronthaul/backhaul architecture for converged access networks in 5G and beyond?"

As per the above discussion, the identified motivations for PON based next-generation fronthaul/backhaul architecture for converged access are as follows:

- Potential to enable diversified services such as residential, mobile, IoT, etc. over a common physically deployed transport access network.
- Potential to support multiple QoS and QoE pertaining to various use cases and service scenarios.
- Reduced CapEx and OpEX achieved through SDN controllability in the transport access networks
- Potential to enable Ultra Reliable Low-Latency Communication (URLLC), one of the key requirements for 5G by enhancing the connectivity between the edge computing nodes while maintaining the low-cost footprint.

We also recognize the following implications of next-generation fronthaul/backhaul architectures for converged access:

- Architectural enhancement of optical transport networks to support such diversified services of converged networks.
- Criteria of sustained fronthaul services with the progressive cell-densification implies a low-cost alternative of current optical transport technologies (e.g WDM or TWDM-PON) to be the main contender for optical transport in the converged access.
- The need of supporting various use cases and service scenarios with end-to-end network slicing would require virtualization of optical transport networks over a shared physical deployment. This would require careful analysis with network optimization to form optimal capacity allocation and maximize the statistical multiplexing gain.

3 Simulation and Hardware Tools

Simulation and Hardware Tools

In this chapter, we introduce briefly the simulation and hardware tools used to evaluate the performance of the proposed schemes in the technical contributory chapters. It covers a brief discussion around the modeling and discrete event simulation using MATLAB and OMNET++. Following this, we briefly introduce MATLAB's discrete optimization framework. Finally, we end this chapter with a brief description of the Hardware Testbed development framework that we have augmented to evaluate the proposed PON architecture enhancement in Chapter 5.

3.1 Discrete Event Simulation using MATLAB

MATLAB provides a great programming platform for numeric computation, behavioral simulation of systems, implementation of algorithms and plotting of functions and data for visual analysis and illustration. Simulink is an additional package distributed with MATLAB which adds graphical multi-domain simulation and model-based design capability to simulate dynamic and embedded systems.

MATLAB provides a diverse set of toolboxes each tailored to enable efficient computation in specific application areas, for example signal processing, communications, statistics, machine learning, etc. SimEvents is one such toolbox that can be used to simulate any discrete event process or model a message-based communication thanks to its highly efficient discrete-event processing engine. SimEvents toolbox provides integrated libraries in Simulink to enable graphical model-based simulation of discrete event systems [81]. The libraries in SimEvents provide queues, servers, switches and other pre-defined blocks to enable modeling and performance analysis of various net-

work utilities such as routing, processing delays, packet scheduling and prioritization, end-to-end communication, etc. In Chapter 4, we have used SimEvents to analyze the performance of the proposed variable rate fronthaul scheme in Cloud Radio Access Networks (Cloud-RAN) (detailed simulation framework is provided in Section 4.4).

3.2 Discrete Event Simulation using OMNET++

OMNeT++ is an object-oriented modular discrete-event framework primarily used for building network simulators [82]. OMNET++ itself is not a simulator for any specific industry-standard networks or protocols. It rather has a generic architecture and provides infrastructure and a set of generic libraries and tools for writing simulations for specific applications. OMNET++ is distributed for free for non-commercial use such as academic research or teaching purpose. However, an extended version of OMNET++ (called OMNEST [83]) is available for commercial use.

Due to its generic framework and a vast set of libraries, OMNET++ can be used to build network simulation of a multitude of different domains as follows:

- modeling of wired and wireless networks
- modeling and performance analysis of various network protocols
- modeling of queuing networks
- modeling of multiprocessor systems and inter-processor communications
- modeling of any system and communication protocols where a discrete event approach is suitable.

Although MATLAB and OMNET++ both provide platforms for discrete event simulation, there are various use cases where one provides significantly more advantages than the other. For example, the SimEvents with its model-based design (with modular functional blocks) over Simulink's graphical interface provides a fast implementation of various simulation models. Furthermore, MATLAB and Simulink's inbuilt graphical plotting functions provide ease of performance analysis of the simulation model. However, due to its heavy user-level abstraction in functional library

modules, it is difficult to perform packet-level analysis on discrete event simulation models such as packet latency, jitter, etc. On the other hand, OMNET++ provides more fine-grain control on functionalities at each discrete event, as all the functional modules in the simulation model have to be written by the user in C++. Therefore, it facilitates a significant in-depth packet-level analysis such as end-to-end packet latency and jitter.

In Chapter 5, we have used OMNET++ to analyze the latency performance of the proposed virtualized EAST-WEST PON scheme. For this, we have built a simulation model to imitate the XG-PON transmission convergence layer. The details of the simulation model and performance analysis of our proposed virtualized EAST-WEST PON scheme is presented in Section 5.6

3.3 Discrete Optimization using MATLAB

Optimization problem arises in almost every research problem. In many cases of the network optimization problems, we often have to minimize a variable (e.g., latency) or maximize it (e.g., Spectral efficiency) while satisfying certain network constraints (e.g., bandwidth or Quality of Service (QoS)). MATLAB's optimization toolbox is an additional package that provides a set of library functions to solve optimization problems [84]. These functions find parameter values that minimize or maximize a given objective function while satisfying constraints. The toolbox includes solvers for Linear Programming (LP), Mixed-Integer Linear Programming (MILP), Quadratic Programming (QP), Second-Order Cone Programming (SOCP), Nonlinear Programming (NLP), etc. These solvers can be used to find optimal solutions to both continuous and discrete optimization problems.

In many real-world use cases, the optimization problem contains multiple maxima or minima. Therefore, depending on the start point, the traditional optimization algorithm (algorithms from MATLAB's optimization toolbox) may land onto local minima or maxima. Therefore finding an optimal solution to these problems requires finding the globally optimal solution (global minima or maxima). MATLAB's Global

Optimization Toolbox facilitates this by providing the library functions that can perform the search for the global optimal solution of a given problem [85]. Solvers from this Global optimization toolbox include surrogate optimization, pattern search, genetic algorithm, particle swarm, simulated annealing, global search, etc.

In Chapter 6, we have used MATLAB's optimization toolbox and Global optimization toolbox to find the optimal virtual PON slices in a MESH-PON network. In this work, we used Intlinprog, an Integer Linear Programming (ILP) framework along with an iterative algorithm to tackle integer non-linear optimal virtual PON slice allocation problem. We have also used MATLAB's dedicated non-linear genetic algorithm solver from Global optimization toolbox to compare against our proposed optimization method.

3.4 Hardware Testbed Development Overview

A testbed provides a great way to demonstrate the proof-of-concept of any proposed scheme and perform real-world performance analysis. In the following, we provide some related details about the testbed set-up and implementation that we have used in the technical contributory chapters.

3.4.1 Burst Mode Transmission and Reception at 10G

One of the most important physical layer requirements in Passive Optical Network (PON) is the burst mode reception in the uplink. In Chapter 5 we experimentally demonstrate the feasibility of the proposed EAST-WEST PON scheme. The experimental setup uses a Xilinx FPGA to operate burst mode transmission and reception over optical SFP+. This setup is an augmented version of the following burst mode transmission and reception testbed for 10Gb/s XGS-PON systems.

The motivation towards building this setup is to realize the optical physical layer of the next-generation PON systems. In a typical PON network (e.g., XGS-PON or NG-PON2), the upstream transmission operates in bursts while downstream operates over continuous transmission(as shown in Figure. 3.1). Because multiple ONUs transmit

using the same wavelength, each ONU must send data only during its permitted time slot to avoid collisions. Data from different ONUs arrive at the OLT at a phase that is uncontrolled and varies significantly over time and changes in temperature. In order to receive the data accurately, the phase and frequency of the local clock should be matched (or locked) with each incoming ONU burst to avoid sampling at the edges of the incoming data symbols. To achieve this, a Burst mode Clock and Data Recovery (BCDR) circuit is required. Such circuit can derive the local clock of same frequency and phase from the individual received optical packet within a short locking time. For the BCDR to function properly, each burst should allocate adequate time for the BCDR circuit to:

- Acquire the sampled phase.
- Identify the start-of-packet and end-of-packet to determine the packet boundaries.
- Identify the start-of-packet and end-of-packet to determine the packet boundaries.
- Allow guard time for each ONU to power on and power off their laser source.
- Allow the automatic gain equalizer in the OLT to settle.

Figure. 3.2 shows the data flow for both downstream and upstream transmissions. Note that all upstream bursts have a preamble, which is required only for upstream transmission. The preamble is a periodic repetition of a pattern. Figure. 3.3 shows the experimental setup of the system. This setup uses two Xilinx-Virtex Ultrascale Evaluation Board (VCU-108), equipped with a 4-channel QSFP-28 optical transceiver module to set up the link for burst mode transmission and reception. The FPGA design uses Xilinx's GTY transceiver cores at the QSFP interface to transmit and receive burst bata streams at a rate of 10 Gbps. The Clock and Data Recovery (CDR) state machine of GTY transceivers uses the sampled incoming data from both the edge and data samplers to determine the phase of the incoming data stream and to control the phase interpolators (PIs) for modifying the phase of the local sampling clock accordingly. For a successful lock with the incoming data burst, the phase

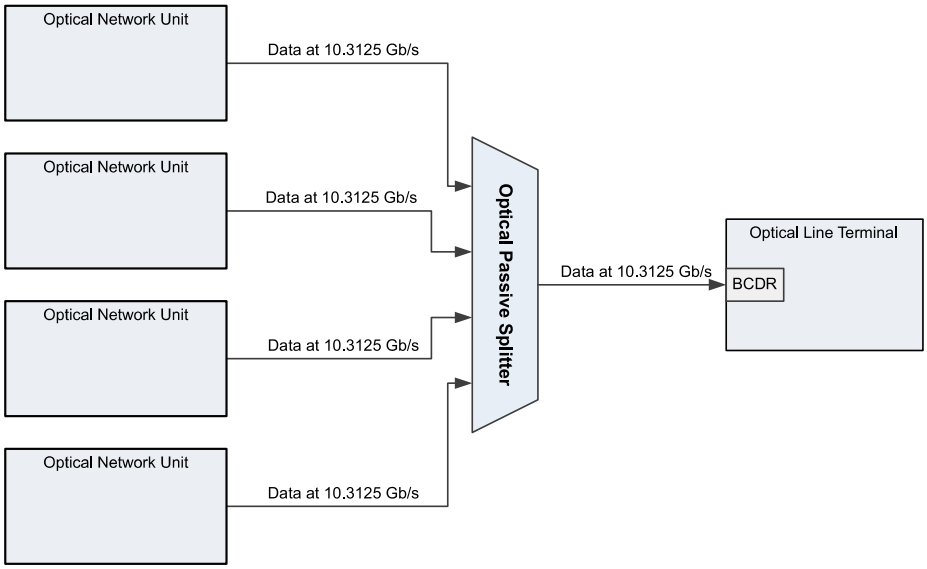

Figure 3.1: 10Gb/s XGS-PON Upstream Transmission Architecture

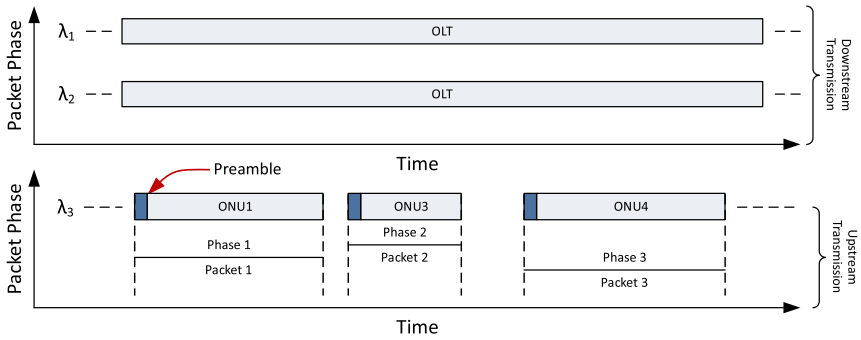

Figure 3.2: 10Gb/s XGS-PON Upstream and downstream data transmission

for the edge sampler is locked to the transition region of the data stream while the phase of the data sampler is positioned in the middle of the data eye. The current native CDR process of the GTY transceivers consumes a long time to get CDR lock with the incoming data. The BCDR circuit augments this native CDR of the GTY transceivers to quickly lock with the incoming burst within a deterministic time. This is achieved by dynamically tracking the incoming phase (by reading the PI's value) and making CDR adjustments accordingly to achieve such quick and deterministic lock time. Figure. 3.4 provides the performance result of the BCDR obtained from the test setup. We can observe a low and deterministic clock and phase-locking time (around 60 ns) with the incoming burst in our implemented testbed setup. In Chapter 5, we re-use this burst mode transmitter from this setup to experimentally

validate our proposed EAST-WEST PON architecture.

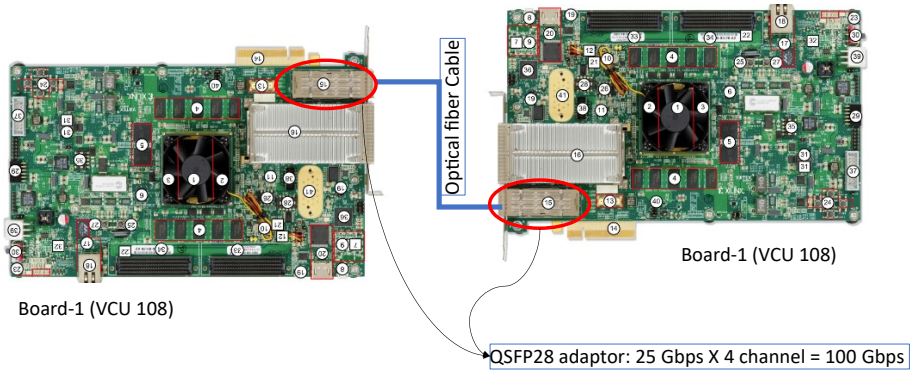

Figure 3.3: Experimental setup for burst mode reception in 10Gb/s XGS-PON Upstream

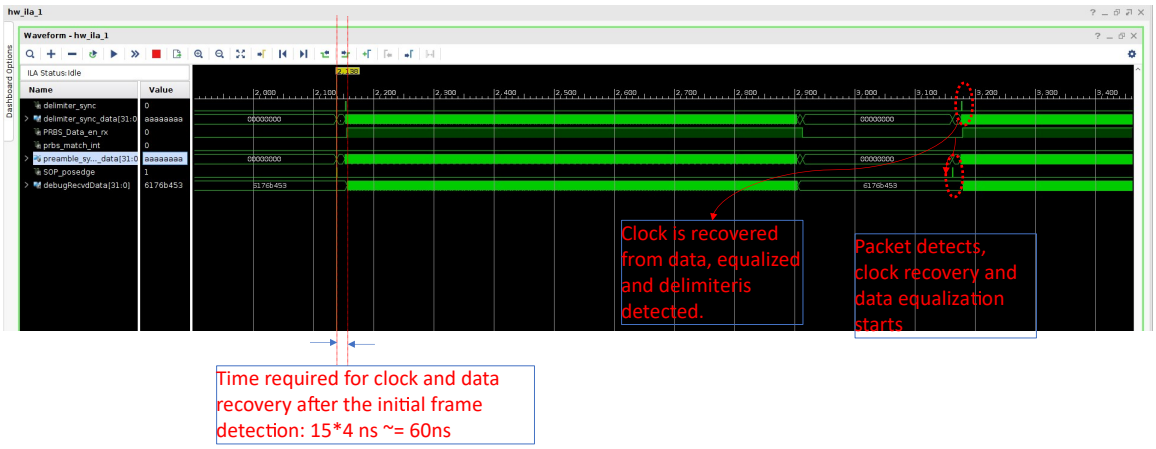

Figure 3.4: Performance of the implemented BCDR, showing the locking time with the incoming burst.

3.4.2 An Ethernet-to-PON bridge interface for Open Disaggregated PON systems

The disaggregation of the Optical transport system has already been identified as a disruptive technology for the next-generation transport networks. Towards PON, disaggregation and SDN control has achieved quite a lot of attention in recent years. Especially, with R-CORD, VOLTHA, virtualized DBA, and PON virtualization, the research towards PON disaggregation and high layer abstraction for SDN control is rich with use cases and a diverse set of challenges.

In Chapter 5 and Chapter 6, we propose dynamic formation of virtualized PON slices which is controlled from the Central Office (CO) (possibly with SDN). In real-world

systems, this would require the OLT to be capable of forming a dynamic virtual PON over a shared physical PON. With the possibility of softwarization of the protocol stack with low-latency data processing software such as DPDK, the realization of such virtualized PON system is feasible provided a generic compatible optical physical layer with burst transmission and reception capability is available.

Figure. 3.5 shows an introductory experiment that takes the very first step towards the realization of such virtualized PON systems. In this setup, we use a DPDK based softwarized XGS-PON OLT box, capable of forming virtualized PONs over Ethernet-based SFP+ physical layer interface. We introduce the burst mode capability (as discussed in the previous section) to complete the physical layer and realizing towards a disaggregated PON with software protocol stack processing. This setup realizes the PON uplink with two ONUs and one OLT. As the virtual OLT is implemented in a server with Ethernet NIC, it does not support burst mode transmission and reception. Therefore, we implement a bridge interface between the Ethernet side implementing the MAC and the burst mode interface implementing the PON physical layer. Figure. 3.6, shows the corresponding experimental setup.

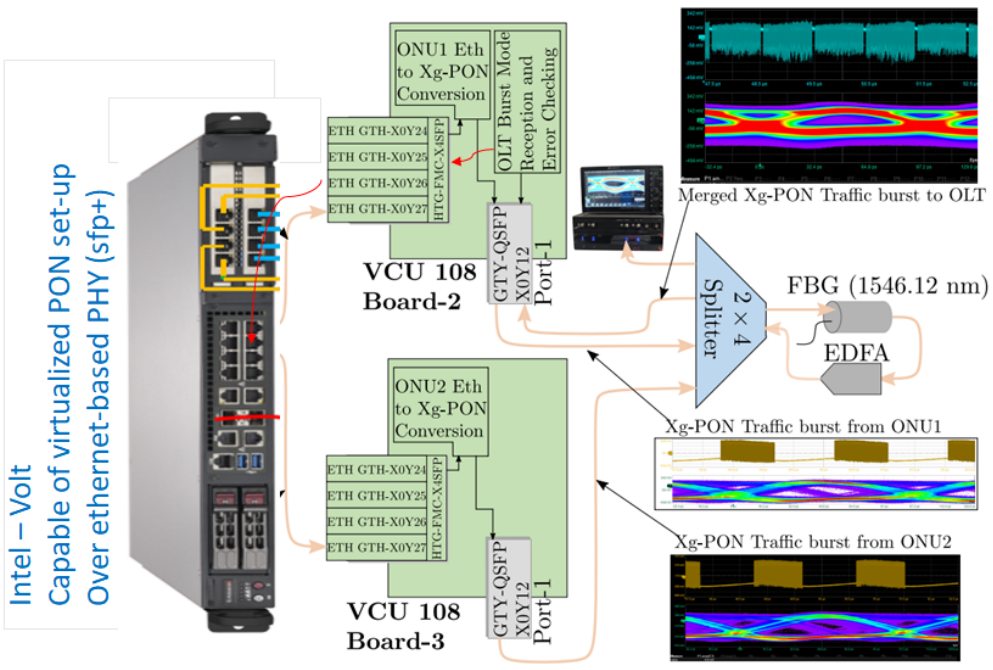

Figure 3.5: Architecture of the Ethernet-to XgPON Bridge Interface setup

The working principle of the experiment is as follows. Two continuous Ethernet streams (carrying uplink data for two ONUs from the virtual OLT) are received into two other VCU-108 boards at the GTH 10G interface of the HTG-FMC-X4SFP module (a 4-port SFP+ 10G adapter module). In each of these boards, an ONU packet processing module receives and decodes the Ethernet frame into a packet, and then generates XG-PON bursts by adding XG-PON preamble (0x05560556) and delimiter (0xB2C50FA1) pattern on top of the packet.

These XG-PON bursts are then sent over through GTY-10G interface of the QSFP module. This is the PON interface side. A 2x4 splitter is used, two bursts are fed onto two ports of the 4-port side of the splitter. On the other side of the splitter, The multiplexed output is observed on the port-4, of the 4-port side of the splitter, and multiplexed output on port-3 is fed onto the OLT.

The multiplexed output of the ONU bursts that is obtained from the port-3, is inspected in the FPGA after the Burst mode clock and data recovery. The received XG-PON bursts are then converted back to continuous ethernet frames and fed into the OLT port of the virtual-OLT module.

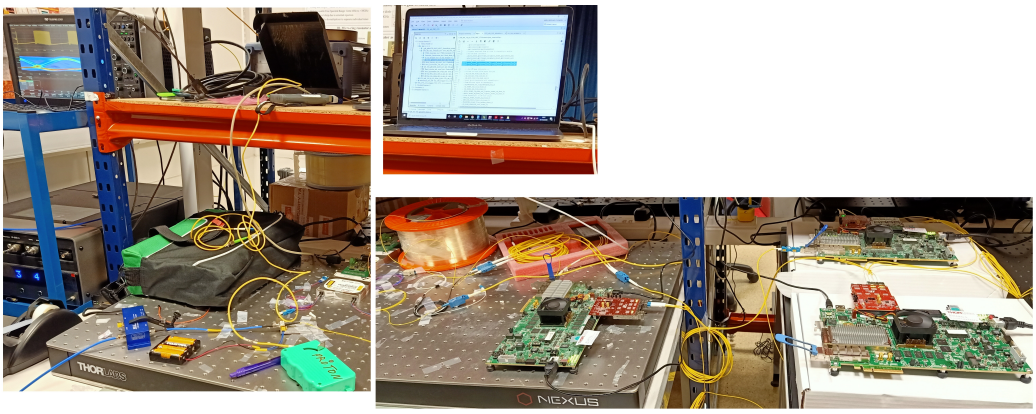

Figure 3.6: Test Set up.

4 Variable Rate Fronthaul for C-RAN

A Variable Rate Fronthaul for Cloud Radio Access Networks (Cloud-RAN)

In this Chapter, we introduce the concept of Variable Rate Fronthaul (VRF) for Cloud Radio Access Networks (Cloud-RAN). This scheme operates on a Common Public Radio Interface (CPRI) type of interface (e.g., one that transmits I/Q data samples) with the novelty of varying the fronthaul data rates by dynamically changing the cell bandwidth depending on the cell load, with the support of a Software Defined Network (SDN) controller. This scheme allows for a more efficient transport of Cloud-RAN cells' data over a shared fronthaul media. We first propose a mathematical analysis modeling the performance of VRF using a queuing theory approach based on the Markov model. We then provide the results obtained from the discrete event simulation framework to validate and support our mathematical model. Our results show that the proposed VRF scheme enables statistical multiplexing while providing significant reduction in blocking probability over a shared fronthaul than the standard CPRI.

4.1 Introduction

As the deployment of Cloud-RAN picks up pace in order to support cell densification and diversified services for 5G, it has also become prominent that the RAN configuration will be immensely heterogeneous as the various split configurations and consequently the fronthaul interface will be realized in few 5G deployment phases [1]. Cloud-RAN has evolved from traditional fully decentralized RAN by introducing the

virtualization of mobile protocol stack processing and including 8 possible split points (Split-1 to Split-8) in the mobile protocol stack [1] (refer Chapter 2, Section 2.1). Among this, split-8 which also corresponds to the fully centralized RAN, was also the first one currently under deployment in the field through a CPRI interface [10]. In Split-8 with CPRI, the entire mobile protocol stack processing is centralized and virtualized (called Baseband Unit (BBU)) in a central processing server, whereas only the analog antenna processing and RF up/down conversion remained at cell site and is called Remote Radio Unit (RRU) [9]. Later in the next phases of 5G deployment the possibility of different functional splits in Cloud-RAN and wide adoption of edge processing has produced an evolved Cloud-RAN architecture. In this next generation Cloud-RAN (as defined in ITU-T [6] or in O-RAN alliance [23]), the mobile protocol stack processing is distributed into three parts: The Radio Unit (RU) which processes the cell and antenna related processing and kept at the cell site, followed by the Distributed Unit (DU) which is the first processing site that performs the Low Layer Split (LLS) processing of the few cells and is usually kept at an edge processing site near to cell site, and finally the Central Unit (CU) which performs the High Layer Split (HLS) processing of several cells in a RAN and is usually kept at a cloud Central Office (CO). Therefore, owing to the different phases of 5G deployment and rapid evolution of Cloud-RAN, it is likely that different functional splits might co-exist in a heterogeneous 5G access network deployment.

Although lower split points (split 6 or 7) have been in the much consideration to leave some of the baseband processing part at the cell site, they makes the antenna unit more complex. The advantages of the split-8 are those of a simpler RRU, fully centralized resource sharing and the ability to carry out advanced coordinated transmission across multiple cells, such as Coordinated Multi-Point (CoMP) communication [16]. However, it requires much higher data rates than the legacy distributed RAN (or than a functional split option equal or less than 4, i.e., above the MAC) [11]. CPRI supports transport of baseband samples at a fixed rate only and the per Antenna Carrier (AxC) fronthaul capacity demand is higher than 1Gbps (1.2288 Gbps) for a 20MHz LTE RRU [12]. When scaled to a 5G 100MHz bandwidth, 8-channel MIMO

systems over three sectors, the required fronthaul capacity reaches 150 Gb/s. In order to overcome these bottlenecks, several solutions have been proposed in recent years, among which compressed CPRI [86] and lower functional splits [87] are two most popular ones. The compressed CPRI scheme applies compression in the I/Q baseband samples transported through fronthaul. Though these schemes have the advantage of reducing the fronthaul rate, they nonetheless transport data at a fixed rate, which is independent of the actual cell usage. On the other hand, lower functional splits reduce the fronthaul capacity demand and enables statistical multiplexing as the fronthaul capacity varies depending on the number of users the RRU is currently serving. However, lower splits require a more complex, expensive and power-hungry RRU. In addition, any processing resource installed in the RRU remains local to the cell and cannot be shared with other cells. In [88], the advantage of using centralized processing in Cloud-RAN with a split-8 is analyzed through teletraffic theory and queuing systems. In this work, the authors analyze the improvement of blocking probability in RRU and BBU processing due to Cloud-RAN architecture. Similar work is carried out for Cloud-RAN employing functional split in [89].

The insight of our work is based on the observation that in the future, the progressive densification of mobile cells will dramatically change their traffic patterns. Smaller cells will serve a smaller number of users, leading to much larger statistical fluctuation in cell traffic. Especially when using next generation high-capacity multi-media application, for example, just a few users in one given cell could easily drive it to its maximum rate, while the nearby cells might have no users. If such small cells are multiplexed over a shared fronthaul transport system such as a Passive Optical Network (PON), statistical multiplexing can be used to reduce the overall backhaul requirement. However, today statistical multiplexing of Cloud-RAN streams is only possible if functional decomposition is applied with a split-PHY level equal or lower than split-7 [1].

In this work, we propose a Variable Rate Fronthaul (VRF) scheme that transports the raw baseband I/Q samples as the traditional CPRI transport mechanism (i.e., at

split-8). The possibility for commercial TDM-PONs to support CPRI through the use of fixed upstream capacity allocation was experimentally demonstrated in [90]. However, in this work the upstream capacity is fixed, independent of the actual cell usage. **In our scheme instead, a Software Defined Network (SDN) controller monitors the cell usage, and dynamically adapts the cell wireless bandwidth to meet the traffic demand**. A reduction in traffic demands will thus trigger a reduction in wireless spectrum usage, which in turn will decrease the I/Q sampling rate and consequently the fronthaul transport rate. The advantage is that the RRU remains simple while restoring statistical multiplexing even for a fronthaul transport system shared between multiple cells (e.g., such as a PON). By adjusting the sampling rate and consequently the wireless bandwidth of the cell according to the cell-traffic load, this scheme can also be exploited for wireless spectrum re-use. It should be noted that split-7.1 [1], which puts the Inverse Fast Fourier Transform (IFFT) at the RRU, can also be used with our VRF scheme, since it produces a constant transmission rate (although at a lower rate than split-8). Using both mathematical analysis and event-driven simulation, we evaluated the performance of our proposed scheme for a typical cloud-RAN scenario. The results show that our VRF scheme can achieve significant reduction of traffic congestion in shared fronthaul medium reducing blocking probability of end-user services.

The rest of this chapter is organized as follows. Section 4.2 provides the system model for the work considered here. Section Section 4.3 provides a detailed description of the mathematical analysis of the system based on queuing theory. The details of the simulator framework and the simulation parameters are discussed in Section Section 4.4. In Section Section 4.5, we compare and discuss the results obtained from the analytical method with those obtained through simulation. Finally, in Section 4.6 we conclude this article by briefly discussing the outcomes of this work.

4.2 System Model

Let us consider the system illustrated in Figure. 4.1. The mobile User Equipments (UEs) are connected to RRUs via an LTE network. Each RRU transmits data to its BBU through CPRI links for centralized processing. We refer to these as fronthaul links. A Fronthaul Aggregator (FHA) aggregates several such links from a cluster of RRUs and creates a Fronthaul Aggregated Link (FAL) which then connects to a centralized BBU pool, forming a tree-like network topology [91]. This abstract configuration could be implemented in practice with a Passive Optical Network (PON), by using a power splitter and Optical Line Terminal (OLT) as the fronthaul aggregator, with each RRU having an Optical Networking Unit (ONU) for communicating to the OLT [92]. It should be noticed that in general, a shared backhaul can provide substantial cost savings in dense cell deployments, as it allows to take advantage of statistical multiplexing across base stations. However, statistical multiplexing cannot be exploited by traditional CPRI, which produces constant data rate over the fronthaul link. Our Variable Rate Fronthaul is a solution to this issue: by dynamically changing the wireless bandwidth of each cell and consequently the associated fronthaul rate, it enables both reuse of the wireless spectrum and statistical multiplexing over the shared fronthaul link.

In our proposed architecture, the SDN controller interacts with the BBUs to constantly monitor the cell usage of each RRU. As a result, it coordinates operations with the BBU and RRU to dynamically adapt the wireless bandwidth and fronthaul rate. Recently, in [93], [94], we have experimentally demonstrated the practical feasibility of our variable rate fronthaul concept using a software LTE BBU connected to its RRU (implemented using a USRP board [95]) via a fronthaul link operating over a PON. The details of the experimentation and the corresponding results are omitted from this chapter to retain the originality of this dissertation as the work has been carried out by other co-authors.

In this work, we consider that the UEs are connected to the nearest RRU. There-

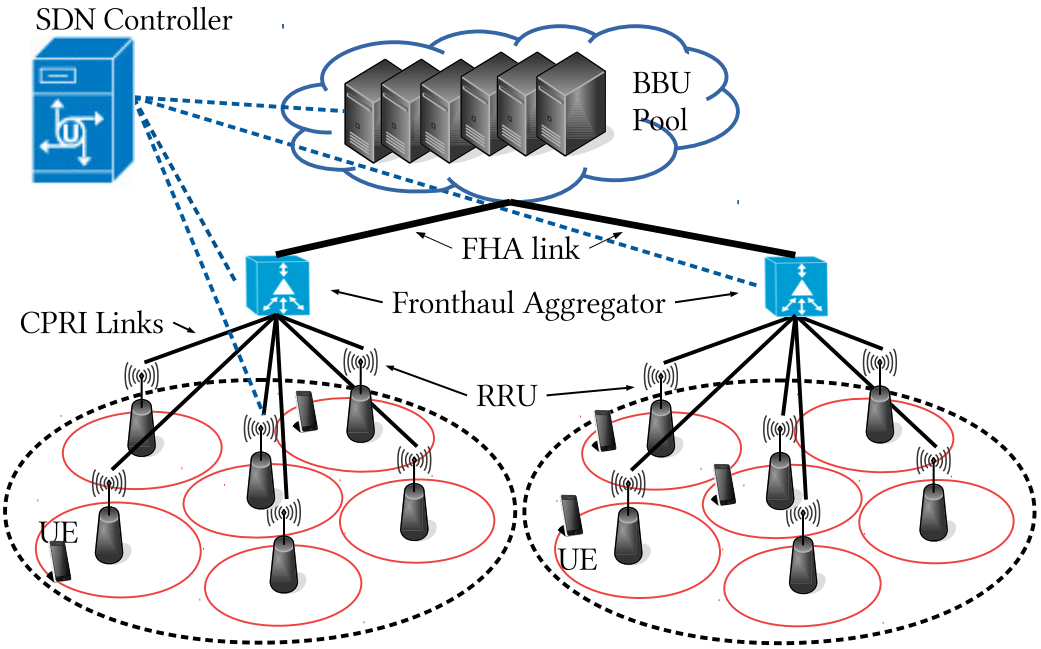

Figure 4.1: System architecture of cloud-RAN

fore, the cell coverage area per RRU can be estimated analytically using Voronoi diagrams as performed in [89]. However, unlike [89], where authors assume arbitrary transmission rates depending on user's behavior, and make use of a functional split architecture, we assume a fully centralized scheme with standard CPRI rates [10], listed in the Table. 4.1. Each line of the table indicates the wireless bandwidth supported, the size of the Fast Fourier Transform (FFT) required, the number of Physical Resource Blocks (PRBs) supported, the transmission rate at the CPRI interface and the maximum number of users that can be supported. In practice, the number of supported users depends on multiple factors such as traffic type, data rate requirement per user, per user service fairness index, etc. However, in this work, we assume for simplicity that each user is allocated two PRBs (or one Resource Group (RG)) when an application data stream (which we refer to as "call", following typical queuing theory terminology) is accepted for service. From this, we derive the maximum number of supported users per bandwidth, listed in Table. 4.1.

In order to explain how the variable rate fronthaul system operates, let us consider a scenario where an RRU is serving already the maximum number of users for the allocated bandwidth. If a new user arrives, which cannot be handled within the

Table 4.1: Standard CPRI data rates for LTE

LTE bandwidth configuration (in MHz)	FFT size	Number of PRBs	CPRI data rate (in Mbps)	Max supported users
1.25	128	6	76.8	3
2.5	256	12	153.6	6
5	512	25	307.2	12
10	1024	50	614.4	25
15	1536	75	921.6	37
20	2048	100	1228.8	50

current bandwidth, the SDN controller triggers a request to increase the allocated wireless spectrum. As a consequence, the fronthaul CPRI rate also increases to support the higher bandwidth configuration (as listed in Table. 4.1). Similarly, when a call departs and the number of remaining users can be supported by the next lower bandwidth configuration, both wireless spectrum and fronthaul rate are decreased accordingly. In the following paragraph, we provide a mathematical interpretation of this model.

Let ϕ_i be the homogeneous Poisson random process with intensity Λ^i, modeling the call arrival process corresponding to the i^{th} RRU. In this work, we consider the simple case where $\Lambda^i = \lambda$, $\forall i \in \{1, 2, \ldots, N\}$, where N denotes the number of RRUs connected to the same aggregator. Therefore, at any given time instant t, the capacity at the FHA link is represented as,

$$C(t) = \sum_{i=1}^{N} \alpha\{u_{i,t}\} \tag{4.1}$$

In Eq. (4.1), $u_{i,t} \in \phi_i$ represents the number of users the i^{th} RRU is serving at time instant t. $\alpha : \phi_i \rightarrow D$, where $D = \{d_1, d_2, \ldots, d_M\}$ represents the data rates available for the fronthaul interface. In this work, we also consider the latency introduced by the SDN controller and BBU-RRU system to reconfigure the system on a different bandwidth, which, as shown in [96], can be estimated in a few hundred milliseconds.

This work aims to find the probability that a new customer call request is blocked, for a given network configuration. Here, we define the call request as the request from an UE to set up a connection with its nearest RRU for transmission and reception of

data over one RG. Different network configurations are achieved by varying N, M, D and λ. Therefore, if B_C represents the capacity of the FHA link, then the steady state blocking probability at the aggregator is given by Eq. (4.2)

$$P_b = \lim_{t \to \infty} P\{C(t) \geq B_C\} \tag{4.2}$$

In a typical Cloud-RAN deployment supported with a PON based fronthaul, blocking probability as expressed by Eq. (4.2) can be interpreted as follows: When an UE makes a connection requests with eNodeB, it is usually allocated an uplink transmission resource in the form of a PRB (or a group of PRBs, also called grouped Resource Block (gRB)). In case if the cell is already at its highest bandwidth configuration with all PRB resources allocated, the UE connection request would be blocked due to the unavailability of the uplink RAN resource. On the other hand, if in order to allocate the PRBs, the cell bandwidth might need to be increased which would lead to higher fronthaul rates and due to the increased fronthaul rate, the aggregated fronthaul rate surpasses the uplink PON capacity, then also the UE connection request would be blocked. However, this time the blocking is due to the unavailability of uplink fronthaul capacity due to the increased fronthaul rates caused by the new UE connection request to the cell. In either of this cases, the UE connection request to the cell would be blocked which is captured by the above expression. Therefore, in this context of this work, we call this as the UE blocking probability or simply the blocking probability which is referred throughout the rest of this chapter.

4.3 Theoretical Analysis

In this section, we provide an analytical solution to the calculation of the blocking probability of a variable rate fronthaul system, based on queuing theory. The process can be divided in two phases. In the first phase, we find the probability for each RRU in a cluster to use a specific CPRI rate. In the next phase, we find the probability that the aggregator ends up in a blocking state, i.e., where an increase in CPRI rate cannot be supported as the FHA link is already at full capacity.

The queuing analysis for an RRU in a cluster using a given CPRI rate at any time instant requires the use of a threshold-based queuing system. Let F_i represent the forward threshold for an RRU to transition to a higher CPRI rate ($d_i \rightarrow d_{i+1}$) and R_i be the reverse threshold for an RRU to transition to a lower CPRI rate ($d_i \rightarrow d_{i-1}$). Let S_i denote the state of an RRU using CPRI rate d_i (or in bandwidth configuration B_i) with F_i and R_i being the forward and reverse thresholds, respectively. This means that if the RRU is in state S_i and currently serving F_i number of users, then any incoming call request to the RRU will trigger the adoption of the next higher CPRI rate $d_i \rightarrow d_{i+1}$ (i.e., the adoption of the next higher bandwidth configuration). Similarly, if the RRU is in state S_i, and currently it is serving R_i number of users, then any call departure from the RRU will trigger the adoption of the next lower CPRI rate $d_i \rightarrow d_{i-1}$. This state transition is represented in Figure. 4.2.

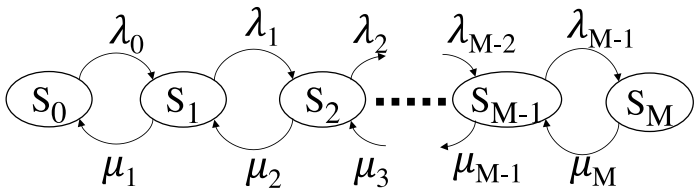

Figure 4.2: State transition diagram for individual RRUs

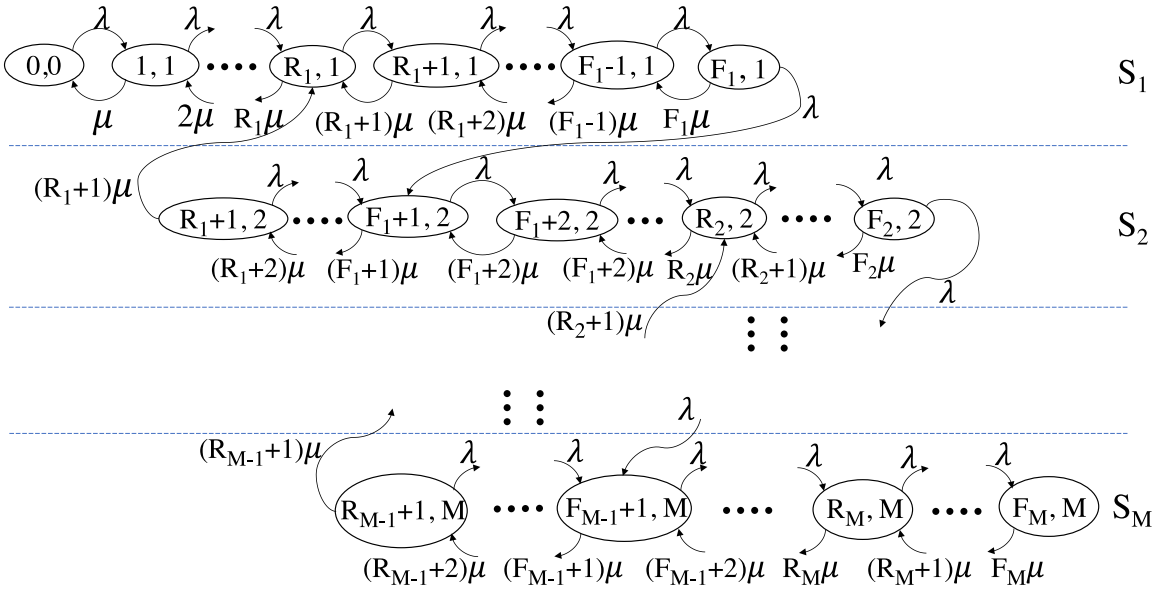

Figure 4.3: State transition diagram of the RRU with partition

Our aim in the first phase of the analysis is to find the steady-state probabilities

of S_i and the corresponding transition rates λ_i and μ_i. We can clearly notice that the steady-state probability of S_i, and the transition rates from S_i to S_{i+1} and S_{i-1} depend on the sub-states of S_i. Thus, in order to provide a solution for this type of system, we use the concept of stochastic complementation technique. This technique has been studied extensively in the field of computer science for adaptive processor allocation in computer servers [97], [98]. In the remainder of this section, we briefly introduce the stochastic complementation technique, before applying it to our model to obtain the steady-state probability and transition rates.

4.3.1 Background of Stochastic Complementation

The concept of stochastic complementation is based on the classic theory of decomposability of Markov chains which was introduced in 1977 by P. J. Courtois [99]. The stochastic complementation technique originally introduced by C.D. Meyer in 1989 [100] provides a way of decoupling a Markov chain by means of partitioning. Let P be the transition probability matrix of a discrete space, discrete time Markov chain $\mathcal{M}$ with state space S. Let us partition this state space into two disjoint sets $\mathcal{L}$ and $\mathcal{R}$. The one step transition probability matrix of $\mathcal{M}$ becomes:

$$P = \begin{bmatrix} P_{\mathcal{L},\mathcal{L}} & P_{\mathcal{L},\mathcal{R}} \\ P_{\mathcal{R},\mathcal{L}} & P_{\mathcal{R},\mathcal{R}} \end{bmatrix}$$

The steady-state probability vector of $\mathcal{M}$ is $\pi = [\pi_{\mathcal{L}}, \pi_{\mathcal{R}}]$, where $\pi_{\mathcal{L}}$ and $\pi_{\mathcal{R}}$ are the steady-state probability vectors of $\mathcal{L}$ and $\mathcal{R}$, respectively. The stochastic complement of $P_{\mathcal{L},\mathcal{L}}$ which is denoted by $C_{\mathcal{L},\mathcal{L}}$, is given by:

$$C_{\mathcal{L},\mathcal{L}} = P_{\mathcal{L},\mathcal{L}} + P_{\mathcal{L},\mathcal{R}}[I - P_{\mathcal{R},\mathcal{R}}]^{-1}P_{\mathcal{R},\mathcal{L}} \tag{4.3}$$

Let $\pi_{|\mathcal{L}}$ denote the steady-state probability vector corresponding to the states of $C_{\mathcal{L},\mathcal{L}}$. We can write $\pi_{|\mathcal{L}} = \frac{\pi_{\mathcal{L}}}{\pi_{\mathcal{L}}.e}$, where e is the column vector with all entries equal to 1. $\pi_{|\mathcal{L}}$ can be interpreted as the conditional state probabilities of the associated states of the original Markov chain $\mathcal{M}$.

We can rewrite Eq. (4.3) as

$$C_{\mathcal{L},\mathcal{L}} = P_{\mathcal{L},\mathcal{L}} + diag(P_{\mathcal{L},\mathcal{R}}e)Z \tag{4.4}$$

In Eq. (4.4), $diag(v)$ is a diagonal matrix whose i^{th} diagonal element is the i^{th} element of vector v and $Z = P^n_{\mathcal{L},\mathcal{R}}[I - P_{\mathcal{R},\mathcal{R}}]P_{\mathcal{R},\mathcal{L}}$. $P^n_{\mathcal{L},\mathcal{R}}$ is essentially the matrix $P_{\mathcal{L},\mathcal{R}}$ with all rows normalized. If r_i is the i^{th} element of $P_{\mathcal{L},\mathcal{R}}e$, and z_i is the i^{th} row of Z, then Eq. (4.4) can be re-written as:

$$C_{\mathcal{L},\mathcal{L}} = P_{\mathcal{L},\mathcal{L}} + \begin{bmatrix} r_1 z_1 \\ r_2 z_2 \\ : \\ r_n z_n \end{bmatrix} \tag{4.5}$$

Expression in Eq. (4.5) can be interpreted as follows. Due to the partitioning of the original Markov process to the sub-processes $\mathcal{L}$ and $\mathcal{R}$, any transition from $\mathcal{L}$ to $\mathcal{R}$ in the original Markov process becomes a transition to some states in $\mathcal{L}$ instead (i.e., it folds back to itself). This process is well known as decoupling of Markov chain. Finding Z can be computationally intensive, although some special cases exist where Z can be easily computed. The following describes one such special case, which we use in our analysis.

Theorem (1): Let Q be the transition rate matrix of a given irreducible Markov process with state space S. If we partition the state space into two disjoint sets $\mathcal{L}$ and $\mathcal{R}$, then we can write

$$Q = \begin{bmatrix} Q_{\mathcal{L},\mathcal{L}} & Q_{\mathcal{L},\mathcal{R}} \\ Q_{R,\mathcal{L}} & Q_{\mathcal{R},\mathcal{R}} \end{bmatrix}$$

In the representation above, $Q_{j,k}$ is the transition rate sub-matrix corresponding to the transitions from partition-j to partition-k. If $Q_{\mathcal{R},\mathcal{L}}$ has all zero entries except for some non-zero entries in the i^{th} column, then the conditional steady-state probability vector (corresponding to the states in $\mathcal{L}$), given that the system is in partition $\mathcal{L}$, is

denoted by $\pi_{|\mathcal{L}}$ and is the solution to the following system of linear equations.

$$\pi_{|\mathcal{L}}[Q_{\mathcal{L},\mathcal{L}} + Q_{\mathcal{L},\mathcal{R}} e e_i^T] = \mathbf{0}, \qquad \pi_{|\mathcal{L}} e = 1$$

where e_i^T is a row vector with a 0 in each component except for a 1 in the i^{th} component.

Proof: The proof of this theorem can be obtained by following the arguments of the stochastic complementation. We need to apply the simple transformation between a continuous time Markov chain with rate matrix Q and a discrete time Markov chain with probability matrix P which can be obtained via uniformization [101] as provided in Eq. (4.6). In this equation, $\zeta = \max\{|q_{ii}|\}$, where q_{ii} is the i^{th} diagonal element of Q. For a more detailed proof of this theorem, the reader can refer to [97], [98].

$$P = I + Q/\zeta \tag{4.6}$$

4.3.2 Steady-state Probability Analysis of the RRU using Stochastic Complementation

In our work, we adopt the stochastic complementation method mentioned above to analyze the probability of each RRU using a particular CPRI rate and subsequently its transition rates. We consider the case where all RRUs belonging to the same cluster are characterized by similar arrival rate λ and service rate μ. When a UE call request is accepted, a user session is created at the BBU by allocating an RG to the corresponding UE, over which it transmits and receives data. From this point, we will refer to this user session as 'server', following the terminology used in queuing theory. Therefore, the maximum number of servers '$\mathcal{K}$' per RRU is constant and is determined by the maximum number of RGs in the highest bandwidth configuration. Increments and decrements of the CPRI rates or bandwidth configuration are governed by the forward and reverse threshold vectors $F = [F_1, F_2, \ldots, F_{M-1}]$ and $R = [R_1, R_2, \ldots, R_{M-1}]$, respectively, where $R_i \leq F_i$.

We consider a homogeneous threshold-based queuing system with hysteresis. Let us

construct a Markov process $\mathcal{M}$ with the following state space $\mathcal{S}$ according to our system model.

$$\mathcal{S} = \{(N_u, s) \,|\, N_u \geq 0, (s \,|\, d_s \in D)\}$$

Where N_u represents the number of users currently served by the given RRU and d_s is the CPRI rate associated with the current bandwidth configuration. Figure. 4.3 illustrates the transition diagram of the Markov process of each RRU. The horizontal lines in the figure are used to partition the whole state space $\mathcal{M}$ into different CPRI spaces. All states (i, j) in any partition S_j are states where the RRU uses the same CPRI rate d_j. Horizontal transitions, within the same space S_j are those where the arrival of a call does not trigger a transition to the next higher CPRI rate. This occurs until the number of users reaches the F_j value, after which a further call arrival triggers the transition to the next CPRI space S_{j+1}. A similar process occurs for the transition to lower rates, with the difference that these occur when the number of users, for a given CPRI state S_j, decreases below the value R_j.

The transition structure of the Markov process $\mathcal{M}$ can be mathematically expressed as follows:

$$\begin{aligned}
&(0,0) \to (1,1) && \lambda \\
&(i,j) \to (i+1,j) && \lambda \cdot \mathbf{1}\Big\{ j \in \{1,2,\ldots,M\} \wedge \Big((i \notin F) \\
&&& \quad \vee \big((i = F_z \in F) \wedge (j \neq z) \big) \Big) \Big\} \\
&(i,j) \to (i+1,j+1) && \lambda \cdot \mathbf{1}\Big\{ j \in \{1,2,\ldots,M\} \\
&&& \quad \wedge (i = F_z \in F) \wedge (j = z) \Big\} \\
&(i,j) \to (i-1,j) && \mu \cdot \mathbf{1}\Big\{ (i \geq 1) \wedge \big((i,j) \neq (1,1) \big) \wedge \\
&&& \quad \big(j \in \{1,\ldots,M\} \big) \wedge \big((i-1) \notin R \big) \\
&&& \quad \vee \Big(\big(i-1 = R_z \in R \big) \wedge \big(j \neq z+1 \big) \Big) \Big\} \\
&(i,j) \to (i-1,j-1) && \mu \cdot \mathbf{1}\Big\{ \big(j \in \{1,\ldots,M\} \big) \wedge \\
&&& \quad \big((i-1) = R_z \in R \big) \wedge \big(j = z+1 \big) \Big\} \\
&(1,1) \to (0,0) && \mu
\end{aligned} \tag{4.7}$$

In Eq. (4.7), $\mathbf{1}\{x\}$ is an indicator function such that $\mathbf{1}\{x\} = 1$ if the condition x is

true and 0 if the condition is false. The operators $\wedge$ and $\vee$ represent logical "AND" and "OR", respectively.

Let us partition the state space $\mathcal{S}$ into M disjoint sets $\mathcal{S}_l$ $\big(l \in \{1, 2, \dots M\}\big)$, where

$$\mathcal{S}_l = \{(i,j)| \quad (i,j) \in \mathcal{S} \; and \; j = l\}$$

Therefore, $\mathcal{S}_l$ represents the state where the RRU is using CPRI rate d_l (or in bandwidth configuration B_l). For $2 \leq l \leq M-1$ we can order the states in $\mathcal{S}_l$ as follows:

$$\{(R_{l-1}+1, l), \dots, (F_{l-1}+1, l), \dots, (R_l, l), \dots, (F_l, l)\}$$

We define another Markov process $\mathcal{M}_l$ for $l \in \{2, \dots, M-1\}$ which corresponds to the state space $\mathcal{S}_l$. The transition structure of $\mathcal{M}_l$ resembles $\mathcal{M}$ for the states in $\mathcal{S}_l$ with the following adjustments (shown in Figure. 4.4).

1. A transition from $(R_{l-1}+1, l)$ to $(R_{l-1}, l-1)$ in the original process $\mathcal{M}$ is replaced by a transition from $(R_{l-1}+1, l)$ to $(F_{l-1}+1, l)$.

2. A transition from (F_l, l) to $(F_l+1, l+1)$ in the original process $\mathcal{M}$ is replaced by a transition from (F_l, l) to (R_l, l).

For $l = 1$, adjustment (1) simply does not apply (see Figure. 4.5) as there are no states to transition to from the state $(0,0)$ in the original Markov chain $\mathcal{M}$, because it is the initial state. Similarly, for $l = M$, adjustment (2) does not apply (see Figure. 4.6) as (F_M, M) is the terminal state of the original Markov chain $\mathcal{M}$.

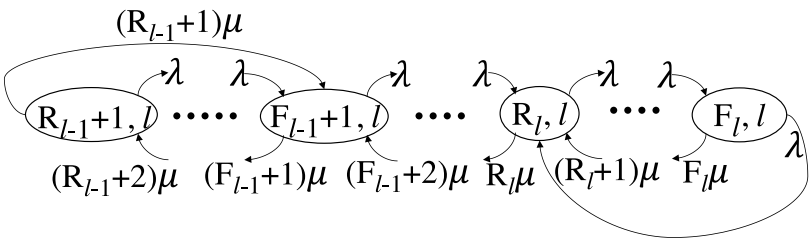

Figure 4.4: State transition diagram of $\mathcal{S}_l$ for $l \in \{2, \dots, M-1\}$

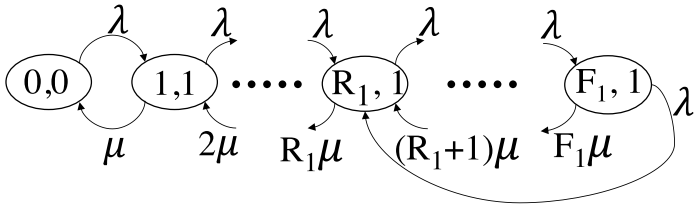

Figure 4.5: State transition diagram of $\mathcal{S}_1$

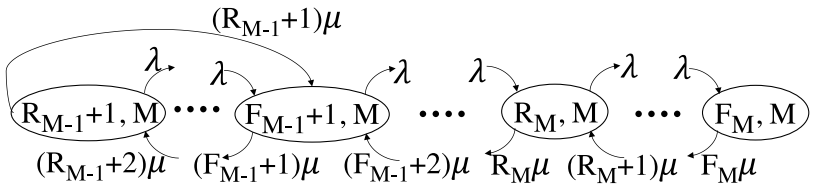

Figure 4.6: State transition diagram of $\mathcal{S}_M$

Given that we construct the partitioned Markov process $\mathcal{M}_l$ according to the procedures mentioned above, we can prove the following:

The steady-state probabilities derived from Markov process $\mathcal{M}_l$ are the conditional steady-state probabilities for the states in $\mathcal{S}_l$ of the original Markov process $\mathcal{M}$, provided that the system is in partition S_l.

This statement can be proved using the argument of stochastic complementation together with the help of *Theorem (1)*. The reader can refer to [97] and [98] for a detailed proof.

Analysis of $\mathcal{M}_l$:

Let us now derive the steady-state probability vector for the states in $\mathcal{S}_l$, namely $\pi_l(n)$, where $l \in \{1, \ldots, M-1\}$. Based on the flow balance equation of $\mathcal{M}_l$ for $l \in \{2, \ldots, M-1\}$, we can express the steady-state probabilities $\pi_l(i)$ for $(R_{l-1}+1) \leq i \leq F_l$ in terms of $\pi_l(R_{l-1}+1)$ in the following form :

$$\pi_l(i) = C_i^l \pi_l(R_{i-1}+1) \quad \text{for} \quad l = 2, 3, \ldots, M \tag{4.8}$$

The unified expression of C_i^l for $(R_{l-1}+1) \leq i \leq F_l$ is provided in Eq. (4.9), where $\rho = \lambda/\mu$. This can be obtained as follows:

Considering the flow balance equation of $\mathcal{M}_l$ for $(R_{l-1}+1) \leq i \leq (F_{l-1}+1)$, we express $\pi_l(i)$ in terms of $\pi_l(R_{l-1}+1)$. Therefore, we obtain the expression of C_i^l as provided in Eq. (4.9a), for the corresponding sub-states of $\mathcal{M}_l$.

Next, we consider the flow balance equation of $\mathcal{M}_l$ for $(F_{l-1}+2) \leq i \leq R_l$ and express $\pi_l(i)$ in terms of $\pi_l(R_{l-1}+1)$. Thus, obtaining the expression of C_i^l for the corresponding sub-states of $\mathcal{M}_l$ as shown in Eq. (4.9b).

Similarly, based on the flow balance equation of $\mathcal{M}_l$ for $(R_l+1) \leq i \leq F_l-1$, we get the expression of $\pi_l(i)$ in terms of $\pi_l(R_{l-1}+1)$ and $\pi_l(F_l)$. Therefore, we obtain the expression of C_i^l for the corresponding sub-states of $\mathcal{M}_l$ as provided in Eq. (4.9c). In Eq. (4.9c), the expression of $C_{F_l}^l$ can be found from the from the steady-state probability of equation: $\pi_l(F_l) = C_{F_l}^l \pi_l(R_{l-1}+1)$

Finally, considering the flow balance of $\mathcal{M}_l$ for $i = F_l$, we obtain the expression of $\pi_l(F_l)$ in terms of $\pi_l(R_{l-1}+1)$. Therefore, we obtain the expression of $C_{F_l}^l$ as shown in Eq. (4.9d).

We follow a similar process and apply it to Markov chains $\mathcal{M}_1$ and $\mathcal{M}_M$ to derive the steady-state probabilities for π_1 and π_M, shown in Eq. (4.10) and Eq. (4.11), respectively. The unified expression of C_i^l $(\forall\, l \in \{1,2,\ldots,M\})$ is given in equations Eq. (4.9) - Eq. (4.11).

$$C_i^l = \Big[\frac{(R_{l-1}+1)!}{i!}\Big] \sum_{j=0}^{i-(R_{l-1}+1)} \rho^j \Big\{\frac{(i-j-1)!}{R_{l-1}!}\Big\} \qquad \text{for } (R_{l-1}+1) \leq i \leq (F_{l-1}+1) \tag{4.9a}$$

$$= \Big[\frac{(R_{l-1}+1)!}{i!}\Big] \sum_{j=0}^{i-(R_{l-1}+1)} \rho^j \Big\{\frac{(i-j-1)!}{R_{l-1}!}\Big\} - (R_{l-1}+1)\Big[\frac{(F_{l-1}+1)!}{i!}\Big] \sum_{j=0}^{i-(F_{l-1}+2)} \rho^j \Big\{\frac{(i-j-1)!}{(F_{l-1}+1)!}\Big\}$$

$$\text{for} \quad (F_{l-1}+2) \leq i \leq R_l \tag{4.9b}$$

$$= \Big[\frac{(R_{l-1}+1)!}{i!}\Big] \sum_{j=0}^{i-(R_{l-1}+1)} \rho^j \Big\{\frac{(i-j-1)!}{R_{l-1}!}\Big\} - (R_{l-1}+1)\Big[\frac{(F_{l-1}+1)!}{i!}\Big] \sum_{j=0}^{i-(F_{l-1}+2)} \rho^j \Big\{\frac{(i-j-1)!}{(F_{l-1}+1)!}\Big\}$$

$$- \Big[\frac{R_l!}{i!}\Big] \sum_{j=1}^{i-R_l} \rho^j \Big\{\frac{(i-j)!}{R_l!}\Big\} C_{F_l}^l \qquad \text{for} \quad (R_l+1) \leq i \leq (F_l-1) \tag{4.9c}$$

where,

$$C_{F_l}^l = \Big[\Big\{1+\frac{F_l}{\rho}\Big\} + \frac{R_l!}{(F_l-1)!} \sum_{j=1}^{(F_l-1)-R_l} \rho^j \Big\{\frac{((F_l-1)-j)!}{R_l!}\Big\}\Big]^{-1} \Bigg[\Big[\frac{(R_{l-1}+1)!}{(F_l-1)!}\Big] \sum_{j=0}^{(F_l-1)-(R_{l-1}+1)} \rho^j \Big\{ \frac{((F_l-1)-j-1)!}{R_{l-1}!}\Big\} - (R_{l-1}+1)\Big[\frac{(F_{l-1}+1)!}{(F_l-1)!}\Big] \sum_{j=0}^{(F_l-1)-(F_{l-1}+2)} \rho^j \Big\{\frac{((F_l-1)-j-1)!}{(F_{l-1}+1)!}\Big\}\Bigg]$$

$$\forall \quad l \in \{2, 3, \ldots, M-1\} \tag{4.9d}$$

$$C_i^1 = \frac{\rho^i}{i!} \qquad \text{for} \;\; 1 \leq i \leq R_1 \tag{4.10a}$$

$$= \frac{\rho^i}{i!} - \Big[\frac{R_1!}{i!}\Big] \sum_{j=1}^{i-R_1} \rho^j \Big\{\frac{(i-j)!}{R_1!}\Big\} C_{F_1}^1 \qquad \text{for} \quad (R_1+1) \leq i \leq (F_1-1) \tag{4.10b}$$

$$\text{where,} \quad C_{F_1}^1 = \Big[\Big\{1+\frac{F_1}{\rho}\Big\} + \frac{R_1!}{(F_1-1)!} \sum_{j=1}^{(F_1-1)-R_1} \rho^j \Big\{\frac{((F_1-1)-j)!}{R_1!}\Big\}\Big]^{-1} \Big[\frac{\rho^{F_1-1}}{(F_1-1)!}\Big] \tag{4.10c}$$

$$C_i^M = \Big[\frac{(R_{M-1}+1)!}{i!}\Big] \sum_{j=0}^{i-(R_{M-1}+1)} \rho^j \Big\{\frac{(i-j-1)!}{R_{M-1}!}\Big\} \quad \text{for} \;\; (R_{M-1}+1) \leq i \leq (F_{M-1}+1) \tag{4.11a}$$

$$= \Big[\frac{(R_{M-1}+1)!}{i!}\Big] \sum_{j=0}^{i-(R_{M-1}+1)} \rho^j \Big\{\frac{(i-j-1)!}{R_{M-1}!}\Big\} - (R_{M-1}+1)\Big[\frac{(F_{M-1}+1)!}{i!}\Big] \times \sum_{j=0}^{i-(F_{M-1}+2)} \rho^j \Big\{\frac{(i-j-1)!}{(F_{M-1}+1)!}\Big\} \qquad \text{for} \quad (F_{M-1}+2) \leq i \leq F_M \tag{4.11b}$$

Using Eq. (4.8), together with the newly obtained equations for π_1 and π_M, we can express $\pi_l(i)$ for $(R_{l-1}+1) \leq i \leq F_l$, $\forall\, l \in \{1, 2, 3, \ldots, M\}$ $(R_0 = 0)$ in the following

form:

$$\begin{aligned} &\pi_l(i) = C_i^l \pi_l(R_{i-1}+1) \quad \text{for} \quad l = 2,3,\ldots,M \quad \text{and,} \\ &\pi_1(i) = C_i^1 \pi_1(0) \end{aligned} \tag{4.12}$$

where the expression for $\pi_l(R_{i-1}+1)$ can be obtained by summing over the state probabilities of Markov chain $\mathcal{M}_l$ and set it equal to 1.

$$\pi_l(R_{l-1}+1) = \Big[\sum_{i=0}^{F_l} C_i^l\Big]^{-1} \quad \text{for} \quad l = 1,2,\ldots,M \tag{4.13}$$

Now, let us consider the aggregated process to calculate the transition rate between the CPRI states S_l. Figure. 4.2 shows the transition diagram of the resulting process. These are the transition rates of the RRU across the different CPRI rates. They can be computed as follows:

$$\begin{aligned} &\lambda_i = \lambda \pi_i(F_i) \qquad \text{for} \quad i = 1,2,\ldots,M-1 \quad \text{and,} \\ &\lambda_0 = \lambda \end{aligned} \tag{4.14}$$

$$\mu_i = (R_{i-1}+1)\mu\pi_i(R_{i-1}+1) \quad \text{for} \quad i = 1,2,\ldots,M \tag{4.15}$$

4.3.3 Steady-state Analysis of the Fronthaul Aggregator using Multidimensional Queuing Model

After we complete the computation of the state transition rates between different CPRI configurations, we need to analyze the steady-state probabilities at the aggregator process ($\mathcal{M}_A$). Let us consider a system with a cluster of N RRUs connected to an aggregator, where each RRU adopts a set of M CPRI rate configurations $(d_1,\ldots,d_M)$. The maximum number of RRU that can be supported with this configuration is $N_{\text{RRU}}^{\max} = \lfloor B_c/d_1 \rfloor$. The reason is that even for a single active user in the RRU, a CPRI rate of d_1 will be adopted. This causes the aggregated capacity to surpass the FHA link capacity, beyond a cluster size of $N_{\text{RRU}}^{\max}$ RRUs. Therefore, we will always get blocking probability $P_b \approx 1$. Let k_m represents the fact that RRU k is using the CPRI rate configuration d_m. Therefore we can consider an M-dimensional vector $\mathbf{k} = (k_1,\ldots,k_m,\ldots,k_M)$ to represent any steady-state of the aggregator.

Depending on call arrival and departure at the RRU in the corresponding cluster, the following events can occur at the aggregator:

1. An RRU transitions to a higher CPRI rate due to a call arrival, thus creating a transition from $d_m \rightarrow d_{m+1}$ for that RRU with a transition rate λ_m (λ_m is the transition rate computed in Section 4.3.2). This creates an aggregator event for a state transition from $(k_1, \ldots, k_m, k_{m+1}, \ldots, k_M)$ to $(k_1, \ldots, k_m - 1, k_{m+1} + 1, \ldots, k_M)$. A similar event occurs for a call departure, triggering an event at the aggregator with rate μ_m.

2. A call request arrives to an RRU belonging to a cluster which had no previous user, thus creating a transition from $(d_0 = 0) \rightarrow d_1$. This is also a wake-up RRU transition with rate λ. This creates an aggregator event for a state transition from $(k_1, \ldots, k_M)$ to $(k_1+1, \ldots, k_M)$. A similar event occurs for a call departure, triggering an event at the aggregator with a rate μ_1.

An incoming call request triggering any of these events is accepted if the new aggregated capacity, which accounts for the increase of the CPRI rate, is less than or equal to the FHA link capacity. Therefore, we get the following possible states for the aggregator:

$$\mathbb{K} = \{\mathbf{k} \mid k_1, k_2, \ldots k_M \geq 0, \sum_{i=1}^{M} k_i \leq N \text{ and } \sum_{i=1}^{M} d_i k_i \leq B_c\} \tag{4.16}$$

As there can only be one possible event at any instant of time, either due to call arrival or departure at an RRU, only a single entry of $\mathbf{k}$ can change at any epoch. Therefore, we can consider the aggregator as a multi-dimensional birth and death process. In the following text, we provide the mathematical description of the transition from the i^{th} to the j^{th} state of the aggregator process $\mathcal{M}_A$.

$$
\begin{aligned}
&Q_{\mathbf{k}^{(i)}\mathbf{k}^{(j)}} \\
&= \mathbf{k}_m^{(i)}\lambda_m && \text{if} \quad \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = e_m^{(1)},\ \mathbf{k}_m^{(i)} \neq 0 \\
&= \Big(N - \sum_{m=1}^{M} \mathbf{k}_m^{(i)}\Big)\lambda && \text{if } \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = e_1^{(2)}, \sum_{m=1}^{M} \mathbf{k}_m < N, \\
& && \qquad \text{and } N \leq N_{\text{RRU}}^{\max} \\
&= \mathbf{k}_m^{(i)}\mu_m && \text{if} \quad \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = -e_m^{(1)},\ \mathbf{k}_m^{(i)} \neq 0 \\
&= \mathbf{k}_1^{(i)}\mu_1 && \text{if} \quad \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = -e_1^{(2)},\ \mathbf{k}_1^{(i)} \neq 0 \\
&= 0 && \text{Otherwise}
\end{aligned}
\tag{4.17}
$$

where states $\mathbf{k}^{(i)}, \mathbf{k}^{(j)} \in \mathbb{K}$, $\mathbf{k}_m^{(i)}$ is the m^{th} entry of $\mathbf{k}^{(i)}$.

$$
\begin{aligned}
& & & m^{\text{th}} & (m{+}1)^{\text{th}} & \\
{e_m^{(1)}}^T &= \{0, 0, \ldots, -1\ , & 1, \ldots, 0, 0\} & \qquad \text{and} \\
{e_1^{(2)}}^T &= \{1, 0, \ldots, 0, 0\}
\end{aligned}
$$

An example of the aggregator process is shown in Figure. 4.7.

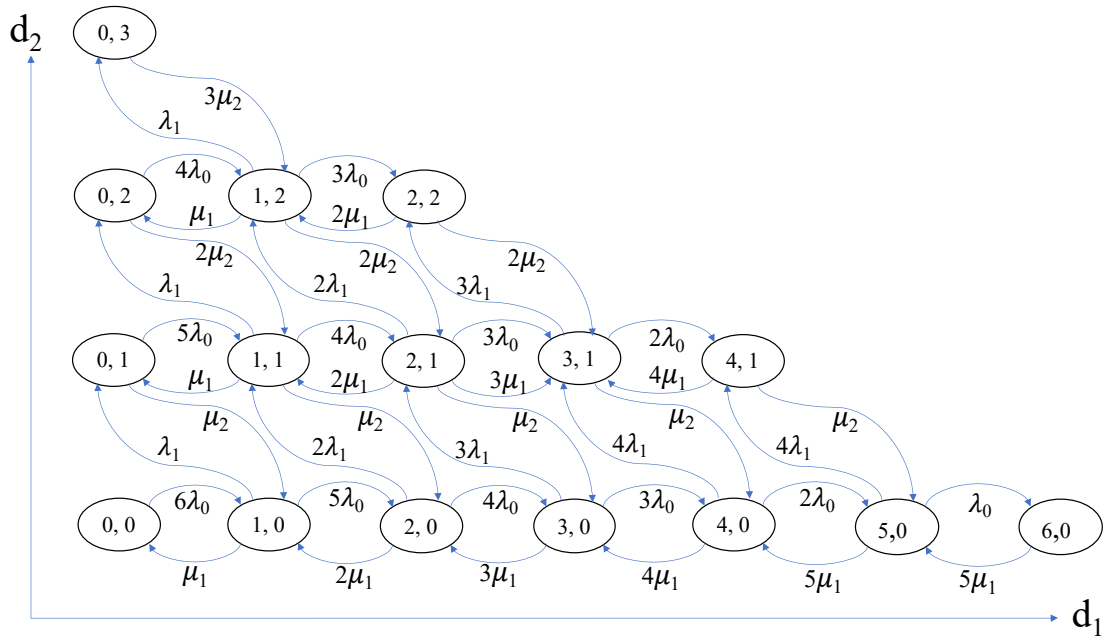

Figure 4.7: An example of the aggregator process with each RRU using two CPRI rate configurations (d_1 and d_2), assuming $d_2 = 2d_1$, FHA link capacity $= 6d_1$ and a N-RRU cluster connected to the aggregator ($N = 6$ for this example).

Analysis of $\mathcal{M}_A$:

It can be shown that $\mathbf{k}$ satisfies the reversibility property by following the method described in [102]. We use this property to derive the steady-state probabilities at the aggregator as follows: Since $\mathbf{k}$ is reversible, the local balance equation

$$P(\mathbf{k}^{(i)}).Q_{\mathbf{k}^{(i)}\mathbf{k}^{(j)}} = P(\mathbf{k}^{(j)}).Q_{\mathbf{k}^{(j)}\mathbf{k}^{(i)}} \tag{4.18}$$

holds at the steady state. Without loss of generality, let

$$\mathbf{k}^{(i)} = \{k_1, \ldots, k_m, k_{m+1}, \ldots, k_M\}^T,$$

$$\mathbf{k}^{(j)} = \{k_1, \ldots, k_m - 1, k_{m+1} + 1, \ldots, k_M\}^T$$

substituting Eq. (4.17) into Eq. (4.18), we get

$$\begin{aligned} &P(k_1, \ldots, k_m, k_{m+1}, \ldots, k_M)k_m\lambda_m = \\ &\quad P(k_1, \ldots, k_m-1, k_{m+1}+1, \ldots, k_M)(k_m+1)\mu_{m+1} \end{aligned} \tag{4.19}$$

Expression in Eq. (4.19) can be re-written as

$$\frac{P(k_1, \ldots, k_m-1, k_{m+1}+1, \ldots, k_M)}{P(k_1, \ldots, k_m, k_{m+1}, \ldots, k_M)} = \frac{k_m}{(k_m+1)} \frac{\lambda_m}{\mu_{m+1}} \tag{4.20}$$

After little manipulation of Eq. (4.20), we obtain Eq. (4.21).

$$\frac{P(k_1, \ldots, k_{m-1}, k_m, \ldots, k_M)}{P(k_1, \ldots, k_{m-1}+1, k_m-1, \ldots, k_M)} = \frac{k_{m-1}+1}{k_m} \frac{\lambda_{m-1}}{\mu_m} \tag{4.21}$$

Clearly, Eq. (4.21) is iterative, therefore we can iterate this equation to obtain the following:

$$\begin{aligned} &P(k_1, \ldots, k_{m-1}, k_m, \ldots, k_M) = \\ &P(k_1, \ldots, k_{m-1}+k_m, 0, \ldots, k_M) \left\{\frac{\lambda_{m-1}}{\mu_m}\right\}^{k_m} \left\{\frac{(k_{m-1}+k_m)!}{k_{m-1}!\ k_m!}\right\} \end{aligned} \tag{4.22}$$

Using Eq. (4.22), we can write:

$$P(k_1,\ldots,k_{M-1},k_M) = P(k_1,\ldots,k_{M-1}+k_M,0)\left\{\frac{\lambda_{M-1}}{\mu_M}\right\}^{k_M}\left\{\frac{(k_{M-1}+k_M)!}{k_{M-1}!\ k_M!}\right\} \tag{4.23}$$

Then, starting with Eq. (4.23) and iterating over all the entries except the first position we obtain:

$$P(k_1,k_2,\ldots,k_M) = P(K_s,0,0,\ldots,0)\left[\frac{K_s!}{\prod_{i=1}^{M}k_i!}\prod_{i=2}^{M}\left(\frac{\lambda_{i-1}}{\mu_i}\right)^{\sum_{j=i}^{M}k_j}\right] \tag{4.24}$$

In Eq. (4.24), $K_s = \sum_{i=1}^{M} k_i$. Now, with the help of Eq. (4.17), we can write the following flow balance equation for $N \leq N_{\text{RRU}}^{\max}$:

$$P(K,0,\ldots,0)(N-K)\lambda = P(K+1,0,\ldots,0)(K+1)\mu_1 \tag{4.25}$$

Expression in Eq. (4.25) can be simplified using the same process followed in Eq. (4.22) to obtain the following expression:

$$P(K,0,\ldots,0) = \binom{N}{K}\left(\frac{\lambda}{\mu_1}\right)^{K} P(0,0,\ldots,0) \tag{4.26}$$

Substituting Eq. (4.26) into Eq. (4.24) and after some manipulation, we get the following final deduction:

$$P(k_1,k_2,\ldots,k_M) = P(0,0,\ldots,0)\left[\binom{N}{K_s}\frac{K_s!}{\prod_{i=1}^{M}k_i!}\prod_{i=1}^{M}\left(\frac{\lambda_{i-1}}{\mu_i}\right)^{\sum_{j=i}^{M}k_j}\right] \tag{4.27}$$

Following the similar procedure, as discussed obtain Eq. (4.25) to Eq. (4.27), we can

obtain the expression for $N > N_{\text{RRU}}^{\max}$ as provided in Eq. (4.28)

$$P(k_1, k_2, \ldots, k_M) = \\ P(0, 0, \ldots, 0) \left[\binom{N_{\text{RRU}}^{\max}}{K_s} \frac{K_s!}{\prod_{i=1}^{M} k_i!} \prod_{i=1}^{M} \left(\frac{\lambda_{i-1}}{\mu_i} \right)^{\sum_{j=i}^{M} k_j} \right] \tag{4.28}$$

Eq. (4.27) and Eq. (4.28) can be combined to get the following form in Eq. (4.29).

$$P(k_1, k_2, \ldots, k_M) = \\ P(0, 0, \ldots, 0) \left[\binom{N_{\text{RRU}}}{K_s} \frac{K_s!}{\prod_{i=1}^{M} k_i!} \prod_{i=1}^{M} \left(\frac{\lambda_{i-1}}{\mu_i} \right)^{\sum_{j=i}^{M} k_j} \right] \tag{4.29}$$

In Eq. (4.29), $N_{\text{RRU}} = N$, for the cluster size $N \leq N_{\text{RRU}}^{\max}$ and $N_{\text{RRU}} = N_{\text{RRU}}^{\max}$, for the cluster size $N > N_{\text{RRU}}^{\max}$. We can derive the expression for $P(0, 0, \ldots, 0)$ as

$$P(0, 0, \ldots, 0) = \\ \left[\sum_{\mathbf{k} \in \mathbb{K}} \left[\binom{N_{\text{RRU}}}{K_s} \frac{K_s!}{\prod_{i=1}^{M} k_i!} \prod_{i=1}^{M} \left(\frac{\lambda_{i-1}}{\mu_i} \right)^{\sum_{j=i}^{M} k_j} \right] \right]^{-1} \tag{4.30}$$

by using the fact that

$$\sum_{\mathbf{k} \in \mathbb{K}} P(k_1, k_2, \ldots, k_m, \ldots, k_M) = 1$$

Therefore, the steady state probabilities of the aggregator can be obtained from Eq. (4.29). In this expression, the transition rates between different CPRI configurations λ_i and μ_i can be obtained from Eq. (4.12) - Eq. (4.15).

4.3.4 Blocking Probabilities at the Fronthaul Aggregator

With the expression of steady-state probabilities derived in the previous section, we can evaluate the blocking probability at the aggregator. In order to accomplish this, we can decompose the calculation of blocking probability into M-parts with M being the number of CPRI rates (or bandwidth configurations) available at the RRUs

belonging to a cluster. Let $P_B^{\lambda_m}$ be the probability that the aggregator goes into a blocking state due to a transition of CPRI rate from d_m to d_{m+1}. We take $d_0 = 0$ and $\lambda_0 = \lambda$ as we have described previously. Then $P_B^{\lambda_m}$ can be determined as follows:

$$P_B^{\lambda_m} = \frac{k_m \lambda_m \sum\limits_{\mathbf{k} \in \mathbb{K}^{\lambda_m}} P(\mathbf{k})}{\sum\limits_{i=0}^{M-1} \lambda_m} \qquad \text{for } i = 1, 2, \ldots, M-1$$

and (4.31)

$$P_B^{\lambda_0} = \frac{\left(N_{\text{RRU}} - \sum\limits_{i=1}^{M} k_i\right) \lambda \sum\limits_{\mathbf{k} \in \mathbb{K}^{\lambda_0}} P(\mathbf{k})}{\sum\limits_{i=0}^{M-1} \lambda_m}$$

where,

$\mathbb{K}^{\lambda_m} = \{\mathbf{k} \,|\quad \mathbf{k} \in \mathbb{K}, k_m \neq 0, \text{and} \sum\limits_{i=1}^{M} (k_i + e^{(1)}_{m\_i}) d_i > B_c\}$, $\mathbb{K}^{\lambda_0} = \{\mathbf{k} \,|\quad \mathbf{k} \in \mathbb{K}, \sum\limits_{i=1}^{M} k_i < N, \text{and} \sum\limits_{i=1}^{M} (k_i + e^{(2)}_{1\_i}) d_i\}$,

$e^{(1)}_{m\_i}$ is the i^{th} entry of $e^{(1)}_m$ and $e^{(2)}_{1\_i}$ is the i^{th} entry of $e^{(2)}_1$.

Finally, the overall blocking probability at the aggregator can be found as

$$P_B = \sum_{m=0}^{M-1} P_B^{\lambda_m} \tag{4.32}$$

4.4 Simulation Model

In addition to providing an analytical expression for the blocking probability at the fronthaul aggregator, we have also simulated the above model in Matlab using the discrete-event-blocks called 'Sim Events'. The model is shown in Figure. 4.8. For each RRU belonging to an RRU-cluster, the call arrival and service is simulated as an $M/M/S(0)$ loss system queuing process. In this work, we have taken the number of servers S as 50, as shown in Table. 4.1, which corresponds to the number of

RGs available at the 20 MHz LTE bandwidth configuration. We use an event-based function call module to translate the current state of the RRU to its corresponding fronthaul rate. In this simulation, an incoming call is dropped whenever the sum of all fronthaul rates at the aggregator is higher than the aggregator capacity. In the case of an XG-PON system, this is equivalent to 10 Gbps. The events in the aggregator are triggered by the events in any RRU. Therefore, an incoming call at any RRU can be blocked for two reasons:

1. If there is no idle server available at the call arrival instance to serve the incoming request, i.e., all 50 servers are busy. This means no RG is available to serve the call, which is thus blocked and dropped by the RRU.

2. If due to a call arrival, a higher bandwidth configuration needs to be adopted but there is not enough capacity available in the FHA link. Therefore the call request is blocked due to unavailable bandwidth at the FHA link.

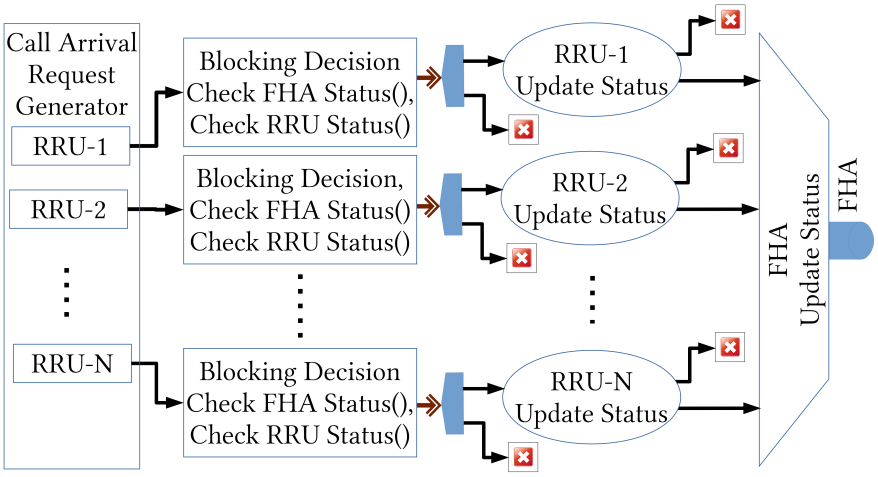

Figure 4.8: Model of the variable rate fronthaul simulator implemented in MATLAB

Tables 4.2 and 4.3 provide the parameters used in our simulations. If the holding time of a service session (or call duration of an accepted call) is τ time units, the service rate (μ) corresponding to each service session of a particular RRU is: $\frac{1}{\tau}$ (calls/time units). If the maximum number of service sessions per RRU is '$\mathcal{K}$' then the condition for the call arrival rate (λ) in each RRU so that the system maintains a steady state blocking probability is $\frac{\lambda}{\mathcal{K}\mu} < 1$ [103]. The traffic load (in Erlang) for each RRU can

be represented as $\rho = \lambda/\mu$. Hence, we can say the traffic load can essentially go up to $\mathcal{K}$ Erlang. Therefore, we denote $a = \frac{\lambda}{\mathcal{K}\mu}$ in order to obtain a normalized traffic load across RRUs. We use different values of 'a' to control the traffic load per RRU.

Table 4.2: Fronthaul rates used for different VBF configuration

$N_d = 1$	$N_d = 2$	$N_d = 3$	$N_d = 4$
1228.8 Mbps	{614.4, 1228.8} Mbps	{307.2, 614.4, 1228.8} Mbps	{153.6, 307.2, 614.4, 1228.8} Mbps

Table 4.3: Simulation parameters

Service Rate (per service session)	0.5
Max no. of service sessions per RRU	50
No. of available fronthaul rates (N_d)	1,2,3,4
Capacity of FHA link	10 Gbps
Max no. of RRU per Aggregator	20
Number of servers in Aggregator	20

4.5 Results

Here we report and compare the results of both our theoretical and simulation analysis of our variable rate fronthaul system, showing the blocking probability for a number of different configurations and scenarios. We run the event-driven simulations for approximately 10^7 events for each specific system configuration (i.e., for a combination of M, N and a) and capture the blocking probability.

Figures 4.9 to 4.11 illustrate the blocking probability (P_b) w.r.t the RRU cluster size for traffic intensity ('a') of 0.2, 0.3, and 0.5, respectively. In these figures, N_d represents the number of different data rates that can be used for the fronthaul transport (i.e., the parameter M, as defined in the theoretical analysis section). Therefore, $N_d = 1$ corresponds to the traditional fronthaul scheme. These figures also demonstrate the improvement of blocking probability when our proposed variable rate fronthaul scheme is used over traditional fronthaul (i.e., for $N_d = 1$). In these results, we have taken the forward thresholds (F_l) as the maximum number of supported users corresponding to the LTE bandwidth / CPRI rate configuration (as listed in Table. 4.1) and the reverse thresholds (R_l) as one step less than the F_l. Therefore,

the difference between the forward and reverse threshold is equal to 1.

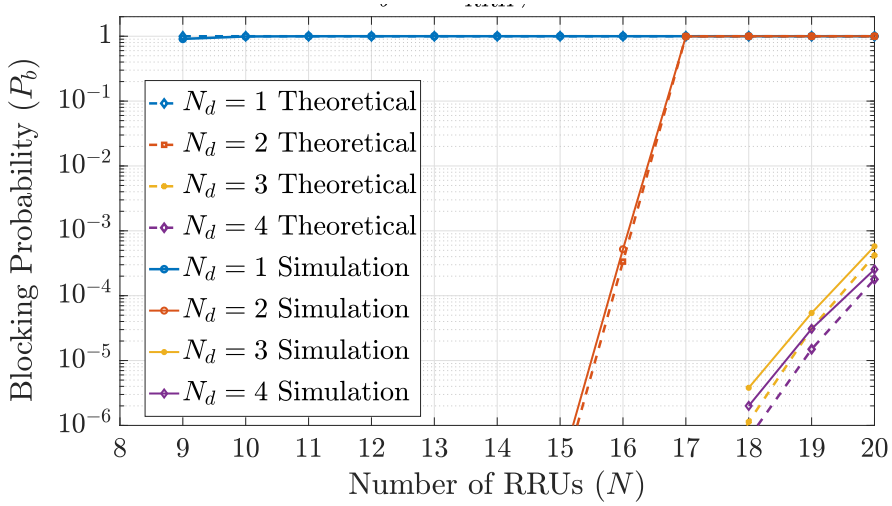

Figure 4.9: Blocking probability (P_b) *vs.* number of RRUs for $a = 0.2$

Firstly, we can see a close match between the analytical and simulated results, which corroborates the validity of our analysis. We do notice a slight discrepancy at the low end of the blocking probability ($\approx 10^{-4} - 10^{-5}$), which we believe is a statistical deviation due to the low number of instances where blocking occurred. In Figure. 4.9, it can be observed that for the traditional fronthaul scheme the blocking probability is zero up to a cluster of size 8. After that, it sharply increases to 1. This is because, in traditional fronthaul ($N_d = 1$), each of the RRU adopts a static rate of 1228.8 Mbps even for a single active user in the system, causing the aggregated rate to surpass FHA link capacity of 10 Gbps. In addition, for a cluster size of 9, we notice that P_b is slightly less than 1 (≈ 0.9). The reason for this is that at moderate traffic intensity ($a = 0.2$), it is relatively likely that there are no users in at least one RRU at any time instant. In that case, we have assumed that the RRU does not send any data over the fronthaul link. Thus, only 8 RRUs would be active, which is a non-blocking situation.

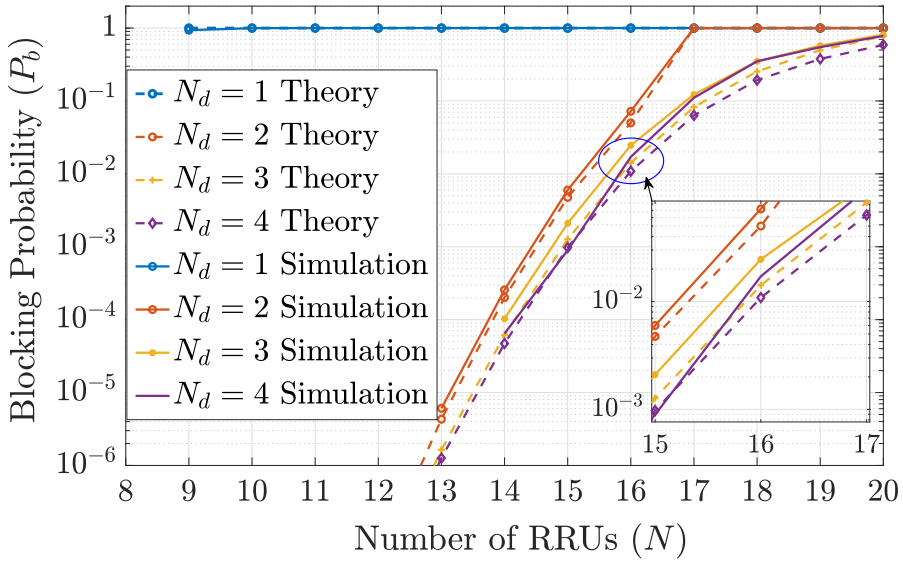

Figure 4.10: Blocking probability (P_b) *vs.* number of RRUs for $a = 0.3$

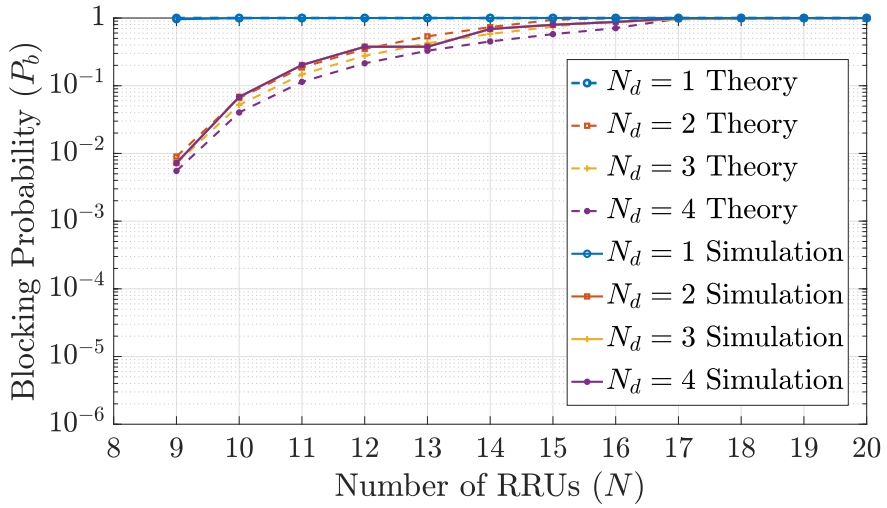

Figure 4.11: Blocking probability (P_b) *vs.* number of RRUs for $a = 0.5$

Secondly, it is clear from the figures that our proposed variable rate fronthaul scheme achieves noticeable lower blocking probability compared to the static CPRI case. Indeed for values of $a = 0.2$, we can see that a blocking probability below 10^{-4} can be achieved even when 15 cells are aggregated (i.e., almost the double of using traditional fronthaul). The number of cells can be increased to 18 if four different rates are employed. However, it should also be noticed from Figure. 4.10 that we obtain almost the same performance for $N_d = 3$ and $N_d = 4$ for $a = 0.3$ (the difference

is highlighted in the small square). This is due to the fact that in our system the different CPRI rates are considered in descending order and, for example, the $N_d = 4$ configuration only adds the low CPRI 153.6 Mbps data rate with respect to $N_d = 3$, which has a small impact on the aggregated rate. Furthermore, the maximum number of end users supported at the CPRI 153.6 Mbps rate is only 6, meaning that if more than six calls arrive at the RRU, this will move to a higher data rate. Thus, the probability for the RRU to remain in this lowest rate is significantly small. The difference can only be noticed when either the call arrival rate is very low or the size of the RRU cluster is large. As we increase the traffic intensity, for $a = 0.5$ (and for rates above $a = 0.5$, which are not shown here), there is very little difference in performance for $N_d = 2$, $N_d = 3$ and $N_d = 4$. This is illustrated in Figure. 4.11 and happens because at this high call arrival rate, higher rate configurations (614 Mbps and 1228.8 Mbps CPRI rates) are adopted most of the time.

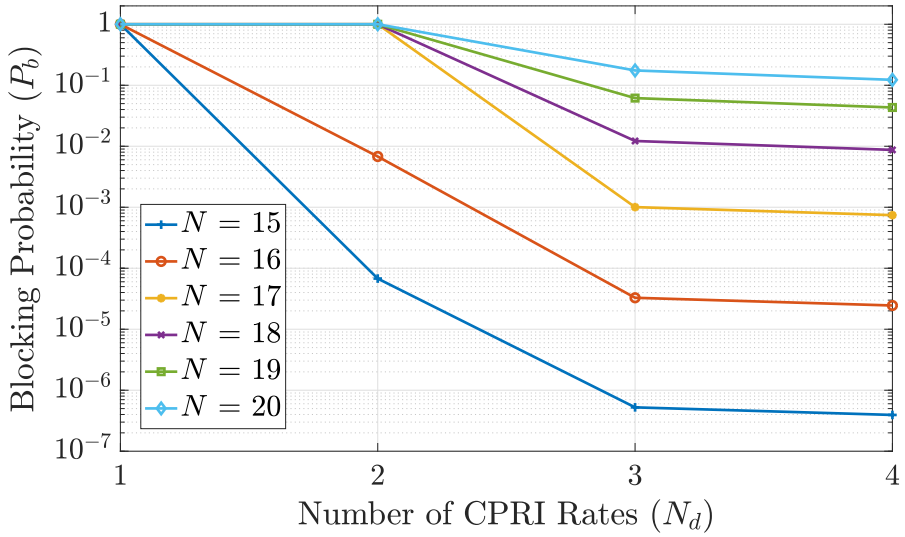

Figure 4.12: Blocking probability (P_b) *vs.* number of data rates for $a = 0.25$

In Figure. 4.12, we use our results to determine the maximum cluster size for which a given blocking probability can be achieved, for $a = 0.25$. For example, a blocking probability of 10^{-3} can be attained with a cluster size of up to 17 RRUs, if we use a VRF configuration with three CPRI rates. However, if we only use two CPRI rates, then the maximum cluster size is reduced to 15.

The results shown up to this point have considered a difference between the forward

(F_l) and reverse threshold (R_l) of 1. Figure. 4.13 shows how the blocking probability increases when we increase the difference between these thresholds. This result considers normalized traffic load as $a = 0.2$, where each RRU employs three different CPRI rates ($N_d = 3$). Indeed, as we increase such difference ($F_l - R_l$), a given CPRI state is retained for a longer period, till a higher number of call departs from the system (i.e., with respect to the case where $F_l - R_l = 1$). This result is important when considering the finite amount of time required for a mobile system to transition between different CPRI rates.

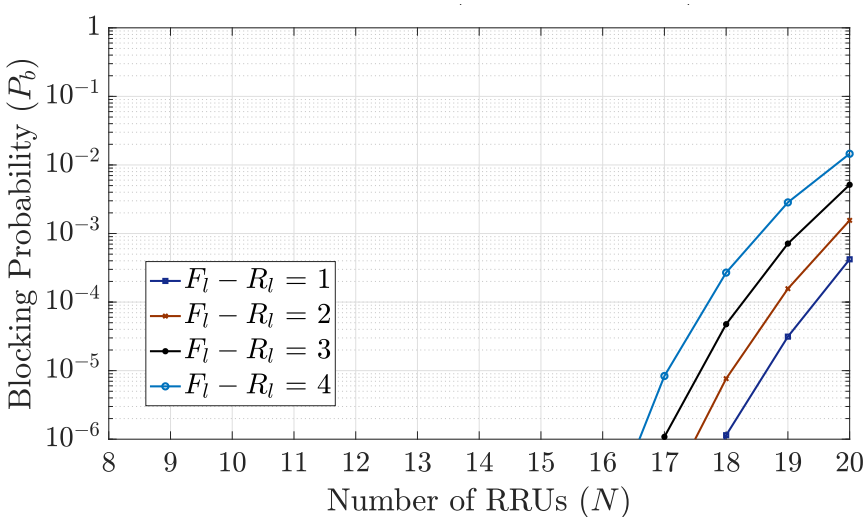

Figure 4.13: Blocking probability (P_b) *vs.* number of RRUs for various differences between forward and reverse thresholds ($F_l - R_l$ =1,2,3,4), for $a = 0.2$ and $N_d = 3$

The difference between the forward and reverse threshold helps to prevent hysteresis i.e., to prevent looping between two CPRI configurations. Furthermore, it also takes care of the latency encountered by the SDN controller to configure between different CPRI rates. For example, if the latency for reconfiguration between different CPRI rates is somewhere between 500ms (this is a value we experimented in our dynamic-bandwidth SDR testbed implementation [96]) to 5seconds (this is the average value taken to completely reboot the SDR BBU) and if we consider normalized traffic load (for example $a = 0.2$) per RRU then we can write,

$a = \lambda/\mathcal{K}\mu = 0.2$

$\implies \lambda = 0.2 \times \mathcal{K} \times \mu = 10\mu$ ($\mathcal{K} = 50 \rightarrow$ maximum number of users supported)

Now, if we consider a moderate traffic scenario, with $\lambda = 10$ calls/min (e.g., in terms of number of users joining the cell) and $\mu = 1$ calls/min, this means that one call, two calls and three calls are expected to arrive, within the reconfiguration window (0.5 seconds), respectively, with probabilities $(p) = 0.0767, 0.0032$ and 8.8739×10^{-5}. Therefore, if we choose $F_l - R_l < 2$ (say for example $F_l - R_l = 1$), then while the SDN controller is triggering a transition to a lower CPRI rate, it is highly probable that one or more calls might arrive. This immediately triggers a transition to a higher CPRI rate which makes the system unstable. Therefore $F_l - R_l \geq 2$ is a good choice to efficiently address the hysteresis (keeps it within 1 percent) and SDN controller configuration timings. However, if instead of implementing a dynamic reconfiguration mechanism, the BBU needs a reboot, with a reconfiguration timing window of 5 seconds, then using the same argument we see that for $a = 0.2, \lambda = 10$ and $\mu = 1$, the probability of arrival of two, three and four call requests is $0.1509, 0.0419$ and 0.0087 respectively. Therefore a choice of $F_l - R_l \geq 3$ efficiently keeps the hysteresis within 1 percent for this case.

It should be noted that the choice of the parameter $F_l - R_l$ only depends on the arrival and departure rates considered in the system. Furthermore, given an average traffic load with moderate intensity, if we let the reconfiguration window to be higher, then the system will be more prone to stay in the higher CPRI rate and the lower rates will be used fewer times, thereby reducing the advantage of using more CPRI configurations (N_d) in VRF. However, as the traffic load increases, the lower rates would be used less frequently in any case (as shown in Figure. 4.9 - Figure. 4.11), therefore reducing the effect of the reconfiguration window on the system performance.

Finally, while in the future we expect Cloud-RAN software-based BBUs to all have similar reconfiguration times to the values above, today some hardware-based BBU might have notably longer reconfiguration times, in which case the rate should be modified less frequently, based on predicted average traffic.

Figure. 4.14 illustrates a typical deployment scenario that could be obtained summarizing the results from Figure. 4.9 to Figure. 4.13. This result provides the number

of RRUs that can be aggregated for different VRF configurations ($N_d = 1, 2, 3, 4$) under the requirement of a certain Grade of Service (GoS) (or blocking probability ($P_b = 10^{-3}, 10^{-5}$)) while considering a given normalized traffic load ($a = 0.25$) and a choice of the forward and reverse threshold difference ($F_l - R_l = 1, 2$). For $N_d = 1$, the system operates over traditional CPRI, so different values of $F_l - R_l$ don't make any difference. For $N_d = 2$ however, there is a difference between $F_l - R_l = 1$ and $F_l - R_l = 2$ for $P_b = 10^{-5}$.

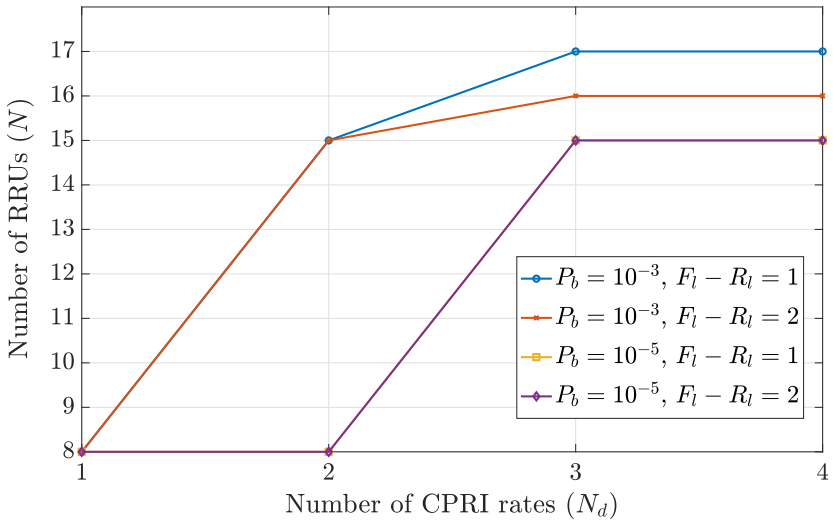

Figure 4.14: Maximum number of RRUs that can be aggregated for different VRF configuration under certain Grade of Service ($P_b = 10^{-3}, 10^{-5}$) requirement, normalized traffic load ($a = 0.25$), and different choice of forward and reverse threshold difference ($F_l - R_l = 1, 2$).

In our study, the Poisson arrival is linked to the increase of PRBs as more users join the network, which can be described as a Poisson process. However, it is true that additional PRBs could be allocated to the same user, when its requested capacity increases thus making the traffic non-Poisson in nature. The theoretical analysis for these additional cases could not be carried out because for non-Poisson distribution the system cannot be decoupled via partitions. In this work of the dissertation, we carried out simulation using two non-Poisson traffic namely, Weibull arrival process with shape factor (k)= 0.9 and 1.5 respectively. A summary of this distribution is provided in APPENDIX A.

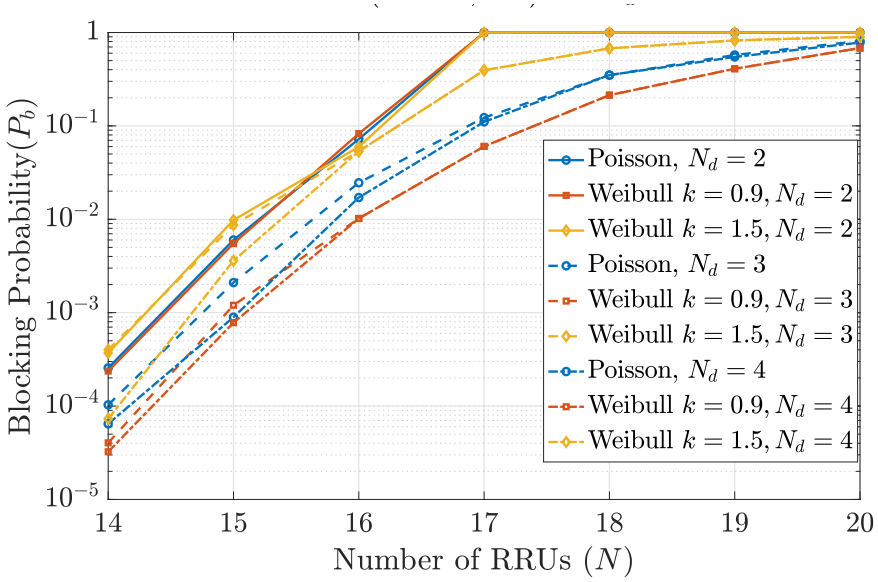

Figure 4.15: Performance comparison between Poisson and Weibull Distribution for $a = 0.3$, $N_d = 2, 3, 4$, $F_l - R_l = 1$

Figure. 4.15 shows comparative result of Weibull arrival process with two shape factors ($k = 0.9$ and $k = 1.5$) against already discussed results using Poisson arrivals under the normalized traffic load $a = 0.3$. We notice that our results with Poisson arrival process lies between Weibull process with $k = 1.5$ and $k = 0.9$. The reason for this is that the intensity of the arrival process changes according to the shape factor. For example, if the average inter-arrival time for our original exponential distribution is $1/\lambda$ (or the arrival rate for our original Poisson distribution is λ), then the inter-arrival time for Weibull arrival process is $\Gamma(1+1/k)/\lambda = 1.0522/\lambda$ for the shape factor (k)=0.9. Therefore, we see that the inter arrival time gets larger which implies a decrease in arrival rate. Thus a Weibull distribution with $k = 0.9$ yields a lower blocking probability compared to our Poisson arrival case. The same argument can be used to explain the reason for $k = 1.5$ having a higher blocking probability when compared to our Poisson arrival case. In addition, as the average rate of arrival increases for higher values of a (which is not shown here), the difference becomes less pronounced as higher capacity will move the RRU states towards the higher bandwidth, thus masking the difference in distribution.

4.6 Conclusions

In this chapter, we have introduced the concept of variable rate fronthaul for cloud-RANs. After providing the description of the network architecture, we have formulated a mathematical description of our model. We have used Queuing Theory with a two-phase approach to solve the model and obtain an analytical form for blocking probability at the fronthaul aggregator. We have then performed simulation of the model using Matlab's discrete event simulator, and the results were compared with those obtained from our analytical model.

Besides showing a close match between analytical and simulation models, our results prove that by dynamically varying the cell's bandwidth, according to the actual end user demand, we can achieve a more efficient fronthaul transport of a group of Cloud-RAN cells, without increasing the complexity, cost and energy consumption of the RRUs. This is especially relevant for next generation of high-density cell deployment, as multiplexing several cells, can sensibly lower fronthaul costs.

5 PON virtualization with EAST-WEST Communication for Ultra-low latency converged MEC

PON Virtualisation with EAST-WEST Communications for Low-Latency Converged Multi-Access Edge Computing (MEC)

Ultra-low latency end-to-end communication with high reliability is one of the most important requirements in 5G networks to support latency-critical applications. A recent approach towards this target is to deploy edge computing nodes with networking capabilities, known as Multi-access edge computing (MEC), which can greatly reduce the service end-to-end latency. However, the use of MEC nodes poses radical changes to the access network architecture. This requires to move from the classical point-to-multipoint (or point-to-point) structure, used to deliver residential broadband and Cloud-RAN services, to a mesh architecture that can fully embed the MEC nodes with all other end points (i.e., mobile cells, fixed residential and businesses, etc.).

In this work, we propose a novel PON based Mobile Fronthaul (MFH) transport architecture based on PON virtualisation, that allows EAST-WEST communication along with traditional NORTH-SOUTH communication. The architecture enables the endpoints of a PON tree, where usually ONUs are located, to also host MEC nodes by deploying an edge OLT capable of communicating directly with adjacent ONUs, by reflecting wavelength signals from the splitter nodes. We experimentally show that signal backscattering due to the reflection at the splitter does not affect the system performance. In addition, using protocol level simulations, we show how this architecture can maintain low-latency ($\approx 100\mu s$) in varying mobile traffic condi-

tions by offloading ONUs (i.e., where remote units of Cloud-RAN cells are located) to other edge OLTs through dynamic formation of virtual PON (vPON) slices. Furthermore, our results show how an efficient migration strategy for ONUs can be chosen depending on the traffic load, different functional split configurations, and the PON capacity.

5.1 Introduction

As the deployment of 5G networks picks up pace, telecommunication industries are continuously challenged by the need to support ever-increasing data traffic demand, massive connectivity and highly diverse quality of service. Some of the key network requirements in modern 5G networks [104] include high throughput, ultra-low end-to-end latency and deterministic Quality of Service (QoS). Specifically, ultra-low end-to-end latency (<1ms) has been a critical requirement in 5G networks to support various latency-critical 5G applications such as tactile internet, logistics, mission-critical control and traffic and road safety [105].

Cloud Radio Access Networks (Cloud-RAN), along with Functional Split processing, are regarded as the most promising 5G radio technologies that support these requirements. In the 5G New Radio (5G-NR) architecture of Cloud-RAN, the baseband processing functions are split into three parts: Central Unit (CU), Distributed Unit (DU) and Radio Unit (RU) [6]. The CU and DU processing functions of several cells can be centralised and virtualised in either at a Central Office (CO) or at a nearby cloud edge processing site, while the cell processing part is retained at the cell site and is called RU. This architecture better facilitates Radio Access Networks (RAN) virtualisation with flexible assignment of computing resources across different entities. The distribution of baseband processing functions between CU, DU and RU is identified by 8 split points in the Long Term Evolution (LTE) baseband processing chain as described in detail on Chapter 2, Section 2.1. Interfaces based on these RAN functional split options are broadly classified in two categories: High Layer Split (HLS) and Low Layer Split (LLS) interface. In the 5G-NR architecture, the CU contains

all RAN functions above the HLS interface while the DU contains all RAN functions between LLS and HLS interfaces, and the RU contains all RAN functions below the LLS interface. As the HLS interface is defined for a split option at the higher layer of the protocol stack, the fronthaul transport bandwidth and latency is relaxed for this interface. On the other hand, the LLS interface has higher transport bandwidth and stricter latency requirements.

Sharp data rate increase in 5G and stringent latency requirements in fronthaul transport makes the use of the legacy CPRI interface impractical, as the fronthaul rate is fixed and, for example, as can be seen from Table. 2.1 in Chapter 2, for a 100 MHz radio bandwidth with 32 antennas and 16 bits resolution per I/Q sample, the fronthaul bandwidth requirement would already exceed 150 Gbps ($\approx$ 157.3 Gbps) [6]. Moving to a higher layer split (such as split option-6, between MAC-PHY) would relax the fronthaul bandwidth requirement; however, less processing functions can be centralised. The choice of optimal 5G NR split points depends on specific deployment scenarios. 3GPP introduced the use of split option 2 (PDCP/high RLC) as the reference HLS split [21] (standardised as F1 Interface [22]), while it left the selection of LLS open across a range of different split options (option 6 for MAC/PHY split or option 7 for intra-PHY split) (TR 38.816 [2]). The Fx interface is a generic notion for these LLS configurations at ITU-T Gsupp-66 [6]; it is standardised as NGFI-I interface by IEEE [3] and Open-Fronthaul interface by the O-RAN alliance [23]. However, any LLS split option below the MAC layer (split-6), that uses the evolved Common Public Radio Interface (eCPRI) fronthaul transport scheme (the evolved CPRI standard for 5G [106]) at the Fx interface requires a stringent fronthaul transport latency of $\approx$100 μs. In order to meet this low transport network latency requirement, the transmission distance between the RU and the first processing site of the LLS (DU) may be shortened through the deployment of limited capacity cloud processing resources, called Multi Access Edge Computing (MEC), close to the cell-sites.

As described in Chapter 2, Section 2.3, MEC is an emerging technology to assist Cloud-RAN meeting the above 5G requirements by bringing the cloud computing unit

towards the network edge. MEC brings highly efficient cloud computing and storage capabilities at the edge and can be used by a RAN to offer low-latency and high bandwidth data processing for latency-critical applications. It can also offer content caching near to end users in order to alleviate the overall network load on data transmission through caching and forwarding contents at the edge of the network. Standardisation efforts are actively ongoing within the European Telecommunications Standards Institute (ETSI) Industry Specification Group (ISG) [107] to effectively integrate MEC in the 5G networks.

Because 5G will increasingly make mobile cells more dense, there is a danger that the use of dedicated point-to-point fibre optical transport network will make the cost of cell deployment prohibitively expensive. In addition, point-to-point solutions do not offer much flexibility, when RU connectivity must be migrated between edge cloud nodes. Passive Optical Networks (PONs) are instead recognised as a low cost alternative to dedicated point-to-point fibre to provide high capacity to end users, being the optical solution of choice for fibre to the premises services. In addition, where a PON installation is available, customers can be connected in a short time to a high-capacity fibre connection. Multi-wavelength solutions such as NG-PON2 have already been developed to further increase the capacity and flexibility of PONs, where for example a wavelength channel could be used to support a small pool of mobile cells that require high-priority [108]. In addition, PON rates are further increasing, with 50 Gb/s due to be standardised soon. For this reason, PONs are widely regarded as a competitive solutions for mobile fronthaul services (e.g., considering different functional split options). Indeed, there has been increasing interest in the use of PONs as optical Mobile Fronthaul (MFH) transport media, as they can use an already deployed Optical Distribution Network (ODN) to provide fronthaul transport for RUs along with serving residential users [6]. However, achieving low latency is a major challenge for PONs in the upstream direction, because of its Time Division Multiple Access (TDMA) nature, which requires a centralised and deterministic scheduling operation from the Optical Line Terminal (OLT). A solution was proposed in [36] and recently standardised by ITU-T [109], as Cooperative Transport Interface (CTI),

which adopts a mechanism where User Equipment (UE) scheduling information is passed directly from the DU to the OLT. This bypasses the report/grant mechanism of the classical Dynamic Bandwidth Allocation (DBA), thereby enabling low MFH transport latency.

Cooperative DBA optimises latency in PON upstream scheduling when transporting LTE (and future 5G) Cloud-RAN low split signals. However, as the load from cells increases, migration of DUs across MEC nodes might be required in order to maintain such low levels of latency. Due to statistical multiplexing of cells traffic in the PON, such migration of traffic between edge nodes is needed to even out the load, and as a consequence performance, so as to reduce application level latency.

5.2 Problem Description and Related works

To this end, researchers in the last few years, have begun to conduct research into how edge cloud nodes may be integrated seamlessly into the optical access architecture. Authors in [13] proposed a solution where additional physical links are deployed between PONs to interconnect ONUs directly. Through the deployment of an edge node at one of the ONU sites, a low-latency network can be created. It uses, however, an excessive number of wavelengths, typically, of the order of N^2 (where N is the number of PON trees in the network). In addition, the wavelength assignment in the proposed scheme is static. In [110], a similar solution was proposed, which relies on enhancing local connectivity between ONUs by deploying a star coupler and additional fibers for each ONU; however this approach does not scale as network densification increases.

In order to overcome this bottleneck, we propose a novel MFH architecture, based on PON virtualisation, which also enables EAST-WEST PON communication. Here the PON end-points can serve Broadband end-users and 5G RUs, but can also host edge nodes. This gives end-points the ability to carry out ultra-low latency communication that is direct and does not need to be routed electronically through the CO where the main OLT is located. In this way, for example, RUs can dynamically redirect

their connection from DUs/CUs located at the central offices (i.e., at the source of the PON tree) towards ones located at the edge (i.e., at the leaves of the PON tree) using dynamic reconfiguration of virtual PON (vPON) slices. This largely improves the statistical multiplexing ability of the PON to support low-latency services. While the concept of dynamic vPON was also proposed in [111], there the authors restrict the location of edge nodes at the the splitter nodes only and consider dynamic offloading only between edge nodes and CO.

Our work introduces instead the ability to create virtual PONs across a mix of COs and edge nodes, which can be located anywhere in the PON. In addition, we propose a novel CO-assisted dynamic vPON slice formation mechanism for offloading ONUs between edge OLTs, to provide ultra-low end-to-end transport latency for MFH. This is important, because our virtualisation mechanism enables the seamless creation and management of slices to support diverse traffic patterns and requirements, thus delivering a fully integrated transport mechanism for MEC.

We evaluate the proposed virtualized EAST-WEST PON mechanism in the following way: Firstly, we provide experimental proof that our envisaged mechanism of reflecting back wavelength channels does not introduce sensible impairments into the system, thus validating the architecture from a physical layer perspective. Secondly, we investigate the performance of end-to-end latency on multiple functional splits (split-8 with Variable Rate Fronthaul (VRF) and split-7.1). The results obtained using discrete event simulation show how the proposed scheme helps to determine the maximum number of ONUs in a vPON slice for a dense deployment of RUs with heterogeneous splits.

The rest of this chapter is organized as follows: Section 5.3 provides the details of our proposed PON architecture. In Section 5.4, the vPON slice formation and EAST-WEST communication mechanism are presented. Section 5.5 describes the experimental setup for the validation of the proposed physical layer architecture and discusses the testbed results. Section 5.6 provides an overview of the simulation framework, and discuss the simulation results. Finally, Section 5.7 concludes the

study.

5.3 Proposed vPON architecture for MEC support through EAST-WEST communication

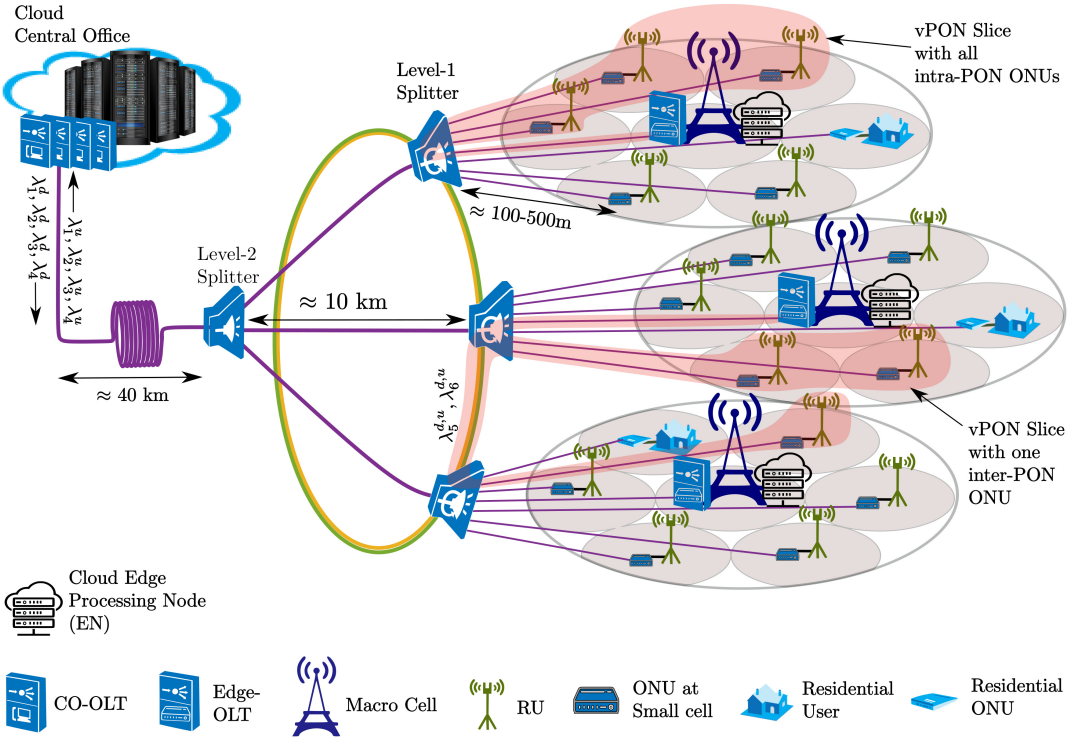

Figure 5.1: System Architecture.

Figure. 5.1 illustrates the system architecture of our proposed Cloud-RAN over PON scenario. We consider a Time-Wavelength Division Multiplexing (TWDM)-PON based mobile fronthaul network, shared with residential users, as shown in Figure. 6.1. RUs are connected with DUs through a two-stage splitter hierarchy (although more stages can be considered). While our architecture can support multiple scenarios of edge cloud convergence, in this work we consider a popular mobile cell placement strategy, where several small cells are deployed to provide offload capability to a macro cell. Further, we consider that MEC servers with limited processing capacity are deployed at the macro-cell sites in order to process delay-sensitive traffic. The level-1 splitter connects all the RUs belonging to the coverage area of each macrocell. We refer to this as level-1 PON tree. The level-2 splitter interconnects the level-1

splitters to the CO. However, unlike [111], we propose an interconnection between level-1 splitters to establish communication between level-1 PON branches. It is important to emphasise that this interconnection can be implemented either through direct cable routes between level-1 splitters or else, if existing ducts are not available, as an overlay over the existing level-1 to level-2 splitters' fibre routes (the difference in performance is shown later in Figure. 5.8), *which does not require any additional fiber ducting* (although it increases the latency by a fixed and deterministic amount of time, due to the additional propagation distance).

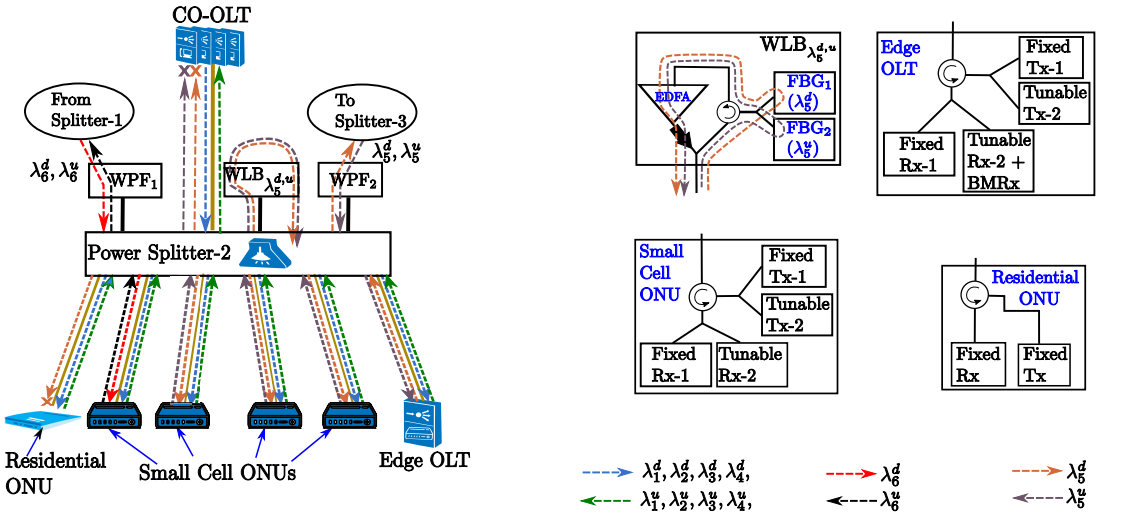

Figure 5.2: Architecture of the level-1 splitter.

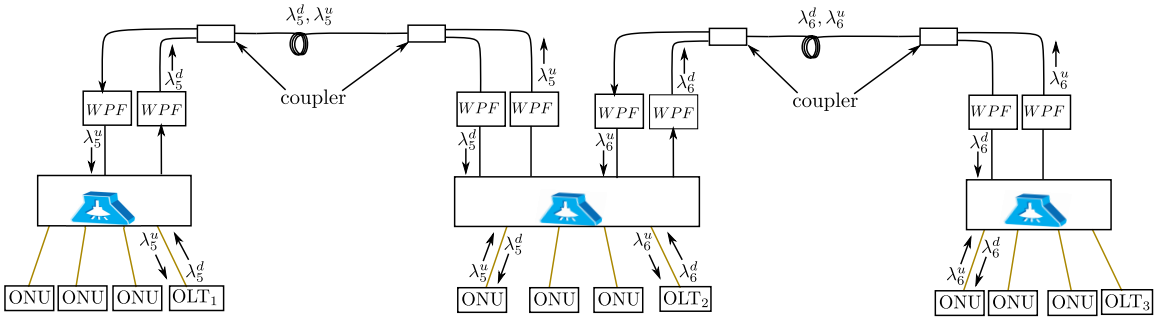

Figure 5.3: Architecture of inter-splitter communication using the proposed EAST-WEST communication over PON.

Figure. 5.2 presents the architecture of the proposed level-1 splitter. Each level-1 splitter uses three additional blocks, namely Wavelength Loop Back $(\text{WLB})_{\lambda_i^{d,u}}$, Wavelength Pass Filter $(\text{WPF})_1$ and WPF_2, where $\lambda_i^{d,u}$ is the operating wavelength pair (λ_i^d for downstream and λ_i^u for upstream) of the edge OLT. Each block connects to

the upper side of the splitter, which can have as many ports as there are on the lower side (splitter are inherently symmetrical in this sense). As shown in the right hand side of Figure. 5.2, $\text{WLB}_{\lambda_i^{d,u}}$ makes use of two reconfigurable Fibre Bragg Gratings (FBGs) connected through a coupler, a circulator and, where required (depending on the splitter loss) an Erbium-doped or semiconductor optical amplifier, reflecting back selected wavelengths towards the edge. Therefore, if the operating wavelength of the edge OLT is λ_5^d (for downstream) and λ_5^u (for upstream), $\text{WLB}_{\lambda_5^{d,u}}$ (as inFigure. 5.2), reflects back λ_5^d and λ_5^u towards the edge OLT (notice that λ_5^d and λ_5^u are different physical wavelengths and the subscript $_5$ indicates that they are both associated to the same vPON slice). This enables the edge OLT to connect to the ONUs of its own level-1 PON tree (the vPON is shown as the red shaded area inFigure. 6.1). WPF_2 lets λ_5^d and λ_5^u pass through to connect to the upper side of the splitter of the adjacent level-1 PON tree, enabling the edge OLT to connect to the ONUs of its neighbouring level-1 PON tree (also shown as red shaded area). Similarly, WPF_2 lets the operating wavelength of the edge-OLT of the adjacent level-1 PON tree (λ_6^d and λ_6^u in this case) pass through, enabling the ONUs of the current Level-1 PON tree to send and receive upstream and downstream traffic to/from the neighbouring edge-OLT. This process of inter-splitter communication is illustrated in Figure. 5.3. Two wavelength pass filters are used to filter wavelength channels in both directions towards the adjacent splitters. An additional coupler (not shown in the figure) can be inserted between the WPFs and the splitter in order to save on the number of splitter ports. As mentioned above, from a physical perspective, the fibre linking level-1 splitters can be routed directly between them if a fiber duct is available. It should also be noted that while a ring structure enables direct connectivity also between the two furthest splitters, this is not a necessity for the system to operate. In addition, if an existing fiber duct is not available along the direct path between splitters, the fibre route can re-use the existing ducts, by going forward and back through the routes linking the two Level-1 splitters, respectively, to the Level-2 splitter above. We refer to this second longer route as the logical ring architecture in the subsequent sections.

This architecture has many advantages compared to other solutions from the liter-

ature. Firstly, unlike [13], each edge OLT can communicate with the ONUs of its one-adjacent level-1 PON tree by using only one pair of wavelengths. In addition, by using one more pair of wavelengths (λ_7^d and λ_7^u), each edge OLT can connect to two of its adjacent level-1 PON trees. This also greatly simplifies the wavelength assignment and lightpath allocation for the dynamic vPON formation. Secondly, unlike [110], it does not require the deployment of additional fibres to realize intra-PON communication within the same Level-2 splitter, as the signal is looped back through the same splitter. It should be emphasized here that looping back the signal through the splitter using the proposed WLB action has the potential to introduce backscattering, as part of the signal is reflected back towards the source. However, we experimentally show (see our results provided in the experimental evaluation section) that this has a negligible effect on the system performance.

The OLT located at the central office (the CO-OLT) can employ a one-channel XGS-PON or a TWDM PON (e.g., NG-PON2) with four (or more) pair of wavelengths ($\lambda_1^i, \ldots, \lambda_4^i \mid i \in d, u$) for upstream and downstream. $\lambda_1^{d,u}$ is dedicated for exchanging control information such as wavelength reconfiguration and vPON slice information in the MFH with all small cell Optical Networking Units (ONUs) and edge OLTs. The surplus bandwidth of $\lambda_1^{d,u}$ is shared with the users for data transmission along with $\lambda_2^i, \ldots, \lambda_4^i \mid i \in d, u$, which are used for data transmission only. In order to dynamically connect to the edge OLT and CO-OLT, each small cell ONU (which can be expected to be more expensive than residential ONUs) employs one fixed (i.e., to reduce cost) and one tunable transceiver, so that a control channel to the CO is always available. Thus, the service disruption period of an ONU is significantly reduced when the virtual association is dynamically switched from one OLT to another. The OLT hosted at the MEC node also incorporates a similar pair of transceivers where the fixed transceiver is dedicated for exchanging the control channel information with CO, and the tunable transceiver is dedicated for providing the datapath for the dynamic vPON slices. At this point, it is important to emphasise that the higher cost ONUs and OLTs, with multiple wavelength channels, are only required for the small cells and MEC end points. All other residential ONUs can adopt

traditional, single wavelength XGS-PON units, thus enabling the use of low-cost end points where required. We should point out that in point-to-point fibre deployments, similar low-latency performance could in principle be achieved by deploying dedicated mesh routes to each edge-OLTs and/or CO for each small-cell ONUs. However, the network would be much more complex with additional fibre routes and transceivers, making the cost of deploying such networks prohibitive. Our proposed architecture can achieve low-latency in standard PON deployment scenarios, while keeping the fibre deployment cost at a minimum, and only using higher-cost ONUs (i.e., with respect to XGS-PON units) at the small cell and MEC sites. Although a detailed cost-benefit analysis is out-of-scope of this work, studies in literature ([79], [80]) shown that the usage of TWDM-PON in the access can lead to a quantified cost benifits by at least 10%, which can be saved even further by sharing the TWDM-PON mobile fronthaul transport with the residential broadband infrastructure running over legacy G-PON and XGS-PON. The vPON slice allocation is carried out at the CO and communicated to the edge OLTs through Physical Layer Operation and Maintenance (PLOAM) messages from the OLTs located at the CO.

5.3.1 Power Budget Analysis of the Proposed Architecture

One of the main concerns of this physical layer implementation is the signal loss due to propagating to a splitter twice. Depending on the splitter size and configuration this might require the use of amplified splitter nodes.We have carried out an analysis in Table 5.1, showing the power loss for different splitter configurations at different stages of the PON hierarchy. On the left-hand side, we show 4 different budget loss classes, with different color coding. According to Table 5.1, a first-stage split configuration up to 4x16 is possible without the use of amplifiers, when the EAST-WEST interconnection is confined to the Level-1 splitter (i.e, single stage of splitter architecture). However, the 4x16 is in the E1 loss budget class, which is currently not available in the initial 25G PON specifications [112] (although it could be made available for specific-purpose applications, such as cloud-RAN). The 4x32 splitter instead cannot be handled without the use of an optical amplifier. For the case of

logical ring architecture (going via Level-2 splitter and back), this architecture always requires amplification at the second splitter. In this case, we assume a 4x4 stage-2 splitter, giving an overall PON split ratio between 1:32 and 1:128. Here the loss can be within the range of N1 class devices, assuming amplifier gains between 17 and 30 dB. As a result, for the split ratio up to 1:64, it is possible to maintain the Level-1 splitter unamplified. The second splitter instead requires amplification for all cases. In some cases, this might be located in the central office, in others in powered street cabinets. One interesting configuration is where the 1:128 split ratio is achieved through a 4x16 1st stage split followed by a 4x8 second stage split. In this case, it is possible to leave the 1st stage splitter unamplified. The lower part of the table reports the values used for the computation of the overall system loss. Splitter losses were averaged from [113] (based on real component measurements).

Table 5.1: Loss incurred due to signal travelling twice through splitter nodes for achieving EAST-WEST communication (calculated loss is shown for different splitter configurations).

Budget class	N1	N2	E1	E2	Total split ratio	Splitter configuration	OLT (MEC) to ONU (small cells) loss via 1st stage and 2nd stage splitter LB				
							1st stage (without EDFA)	1st stage (with EDFA)	Via 2nd stage 4×4 splitter (without EDFA)	Via 2nd stage 4×4 splitter (with EDFA)	Required EDFA gain at second stage (dB)
Loss budget	29	31	33	35	32	4×8 - 4×4	24.8	9.8	45.4	28.4	17
					64	4×16 - 4×4	31.36	16.36	51.96	28.96	23
					128	4×32 - 4×4	37.96	22.96	58.56	28.56	30
									4×8 second stage	4×8 second stage (with EDFA)	
					128	4×16 - 4×8	31.36	16.36	58.86	28.86	30

Parameters:

Splitter architecture	Avg. splitter loss (one way)	Fiber loss dB/km (including splice loss)	Overall connector loss (dB)	FBG loss (dB)	Fiber loss (dB/km)	Fiber length between 1st and 2nd stage Splitter (km)	Fiber length (avg.) between ONU and 1st stage splitter (km)
4×4	7.3	0.3	1	2	0.3	10	0.5
4×8	10.75						
4×16	14.03						
4×32	17.33						

5.4 vPON Slice Allocation and EAST-WEST Communication

We consider the topology of a converged access/metro architecture [114, 115], where the main CO is located 50km away from the edge. Of this, 40 km are used by the main feeder fibre, 10 km by the distance between level-1 and level-2 splitters, while the

distance from the last splitter to the edge is up to 500 meters. Although the proposed system can support different distance distributions, this is an example of a popular converged access/metro architecture [115], currently under standardisation. In our proposed architecture, each wavelength channel follows the XGS-PON specification. The small cells implement Cloud-RAN with LLS split, as described in [1], where each RU is served by an ONU and the OLT, DU, and CU is either at the edge (MEC) or CO. The mobile core network functions are hosted at the CO regardless of the placement of the CU/DU. We consider eCPRI traffic over the fronthaul interface between RU and DU. More in details, we consider two types of split. One is a split-8 (CPRI), operating over an intelligent adaptive VRF scheme [116], whose experimental operation was demonstrated in [93],[94], which makes the line rate proportional the cell load. The other is a split-7.1 [1]. Both splits use a variable transport rate, that is proportional to the actual traffic at the cell.

The vPON slice allocation is carried out at the CO and communicated to the edge OLTs through PLOAM messages from the OLT located at the CO. Once they power up, the ONUs and the edge OLTs tune to the wavelength corresponding to the control channel of the CO (for instance, λ_1) and then complete the standard XG-PON ranging process.

After the ranging process is completed, the CO generates the required vPON slices, providing information on the edge OLT, ONUs associated to the vPON slice and the wavelength/s for the slice. The vPON slice information is sent to the edge OLTs through the PLOAM messages in the same control channel wavelength (λ_1). In the same downstream PHY frame, a wavelength tuning command is sent to the member ONUs of the corresponding vPON slice through PLOAM messages. In this way, the OLT and member ONUs corresponding to the particular vPON slice configure the wavelength channel simultaneously.

5.4.1 DBA procedure and Dynamic vPON slicing

The allocation of upstream bandwidth in a vPON slice is done by the corresponding OLT independently of other slices, once the vPON slice is configured from the CO. The DBA process in each vPON slice works as follows. Following the cooperative DBA concept [109], each OLT receives scheduling information of the UEs 4 ms prior to the transmission of data corresponding to the particular Transmit Time Interval (TTI). In this work, we consider the case where each UE connected with a RU is scheduled with one Resource Group (RG) (which is equal to two Physical Resource Block (PRB) in LTE). The eCPRI data rates corresponding to the particular TTI can then be obtained from Table. 5.2, considering that the bandwidth adaptation scheme is applied according to the number of UEs linked to a given RU, the OLT aims to schedule the entire eCPRI payload output from the RU-ONUs for the corresponding TTI within its duration. Therefore, considering the DBA cycle of 125 μs, the OLT is required to schedule the entire eCPRI payload over 8 grant cycles. The allocation algorithm for upstream bandwidth in a vPON slice, described below, follows a similar approach of three stage bandwidth allocation of XGS-PON.

Stage-1: Fixed bandwidth assignment: In this first stage, a fixed amount of upstream bandwidth (R_F^i) is allocated to each ONU (ONU_i) regardless of its traffic demand.

Stage-2: Guaranteed bandwidth assignment: After scheduling the fixed bandwidth assignment (R_F^i), the OLT carries out the guaranteed bandwidth assignment (R_G^i) by allocating bandwidth to each ONU until either their respective provisioned level (defined as assured bandwidth, R_A^i) is reached or their traffic demand (R_L^i) is satisfied, i.e., $R_G^i = \min\{R_F^i + R_A^i\,;\, \max\{R_F^i; R_L^i\}\}$. We define $R_A^i = (C - \sum_i R_F^i)/N_{\text{ONU}}^{sl}$, where C denotes the fronthaul capacity and N_{ONU}^{sl} denote the number of ONUs in the vPON slice.

Stage-3: Non-assured bandwidth assignment: The surplus bandwidth is calculated and distributed in a non-assured form ($S_{NA} = C - \sum_i R_G^i$), among the eligible ONUs whose traffic demands were not satisfied in assignment stage-2.

The OLT allocates non-assured bandwidth components to eligible ONUs until either all the ONUs reach their saturation level (whichever is smaller between their maximum provisioned bandwidth (R_M^i) and the offered load (R_L^i), i.e., $\min\{R_M^i; R_L^i\}\}$) or the surplus bandwidth pool (S_{NA}) is exhausted

If the DBA cannot schedule the entire eCPRI payload within 8 grant cycles, (for example if the aggregated upstream bandwidth is higher than the available bandwidth), the leftover segment of the eCPRI packet is queued at the ONU side, which will be scheduled for transmission in the successive grant cycles, along with the eCPRI data for the next TTI. This, leads to increased latency in the successive fronthaul packets.

Since the fronthaul rate varies depending on the actual load at the RU, statistical multiplexing could be exploited by oversubscribing the number of ONUs per edge OLT in a vPON slice. As the load per ONU increases however, the fronthaul transport latency also increases. When the latency reaches a pre-established threshold, the CO re-configures the vPON slices to dynamically offload some ONUs to a nearby edge OLT, keep the latency below threshold.

It is important to note that the tunability of the transceivers is one of the key requirements for enabling the proposed EAST-WEST PON architecture. Therefore, the tuning time of the transceivers would affect the performance especially when the vPON slices are reconfigured dynamically. This would potentially cause an increased latency at the fronthaul transport during the wavelength tuning time of the transceiver as fronthaul packets is needed to be queued while transceivers are being tuned into a another wavelength during the vPON slice reconfiguration. However, current tuning time for the NG-PON2 transceivers goes as low as <10 μ for class 1 transceivers and ranges from 10 μ to 25 ms for class 2 transceivers [117]. Therefore, with the class 1 transceivers, it would have a minimal effect on the instantaneous upstream latency. With the class-2 transceivers (with tuning time of the order of a few hundreds of microseconds to few milliseconds), the queuing latency would predominate during the vPON reconfiguration time. Therefore, the instantaneous latency may not meet the required latency threshold, however, the average latency may meet the

target low-latency threshold as the system goes into steady state after the vPON slice reconfiguration. It is worthwhile to note that typically the frequency vPON slice reconfiguration requirement is the order of a few seconds to minutes in case of low-moderate mobile traffic (for e.g., cycling, low-speed transport etc.) to static traffic (for e.g, walking, residential mobile etc.). Therefore, both class-1 and class-2 transceiver would be sufficient to meet the average latency requirement. In case of highly dynamic and moving traffic (for e.g., high-speed transport, Intelligent transport and prioritized-vehicular services etc.), the frequency vPON slice reconfiguration requirement is the order of a few hundreds of milliseconds to a few seconds. Therefore, in this case, a class-1 transceivers would be better suited to meet the latency requirement and providing Quality of Experience (QoE).

5.5 Experimental Evaluation of the proposed architecture

Figure. 5.4 illustrates the experimental setup for the proof-of-concept of the proposed architecture. In this setup, we use a Xilinx VCU-108 board to generate *burst mode traffic* at 10.3125 Gbps rate to emulate the upstream ONU traffic. The original signal (S_i^λ) is generated by a wavelength-tunable SFP+ module which is programmed to transmit the optical signal at $\lambda = 1546$ nm. The optical signal propagates through 15 km single-mode fibre (where back scattering from the reflected signal will also occur) which is then connected to one of the two downside ports of a 2x4 splitter. On the other side of the splitter, one port is connected to an FBG centered at 1546 nm wavelength. The reflection port of the FBG is fed to the Erbium-Doped Fibre Amplifier (EDFA) for amplification, which is then looped back to the second upside port of the splitter (signal S_a^λ, refers to the signal after amplification), so that the signal is reflected back towards the access side of the PON. We can see that the S_a^λ is reflected also towards the 15km fiber, where it will generate backscattering (S_b^λ), which will also be amplified by the EDFA, after going through the FBG, although delayed with respect of the original signal (S_i^λ). Finally, the output of one of the unconnected downside ports (which is the sum of S_a^λ plus the back scattered signal

S_b^λ) is detected using a photodiode, the output of which is then terminated on a real time scope, which operates burst-mode reception, to measure the BER performance. The power falling on the detector is controlled using a variable optical attenuator in order to measure the receiver performance against different received optical powers. The performance of the system is tested by measuring the BER as a function of the received optical power.

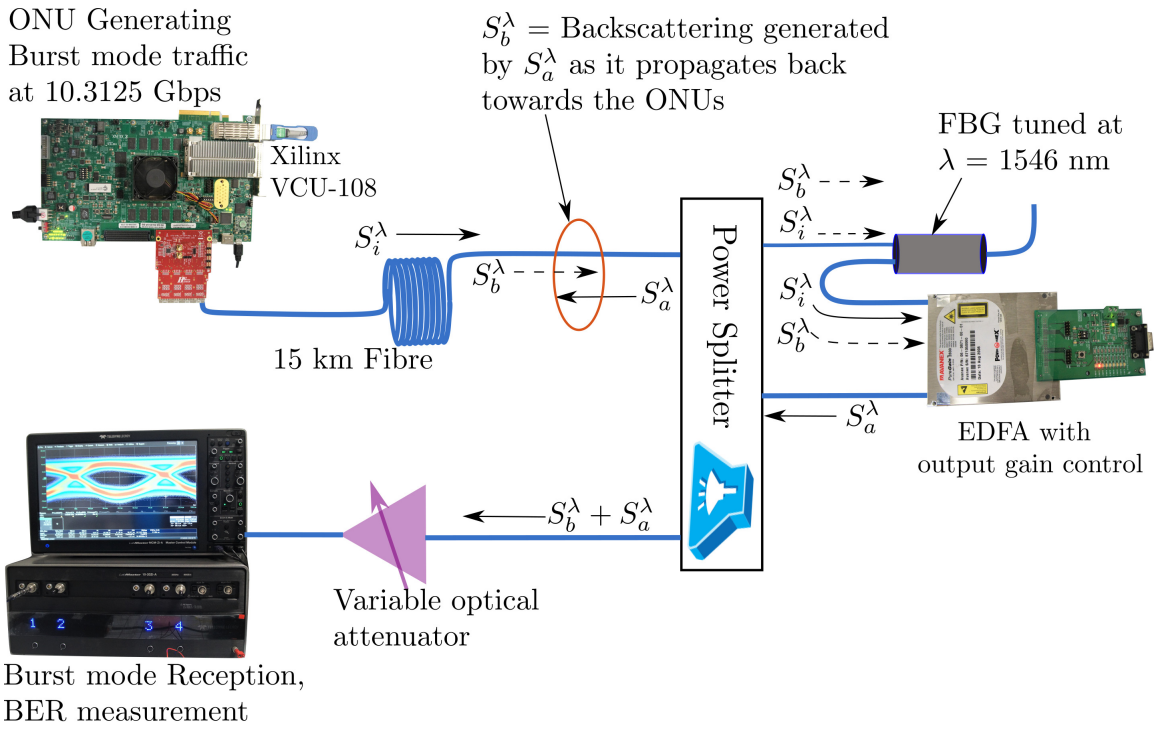

Figure 5.4: Expermental setup for the proof-concept of the proposed architecture.

We measure the Bit Error Rate (BER) against the received optical power for three configurations:

1. back-to-back (B2B) where no fibre and no splitter is inserted in the path, acting as benchmark.
2. Configuration-1, where the splitter is not connected so that there is no backscattering generated. Here the signal travels through the fibre, is reflected at the FBG, amplified at the EDFA and fed into the receiver through the variable optical attenuator.
3. Configuration-2, where the WLB action is reproduced by introducing the splitter loopback mechanism, which generates backscattering as the signal backpropag-

ates in the fibre.

The variation in the path loss due to the removal of the fibre and the splitter (in B2B and configuration-1) is compensated using fixed optical attenuators of values 2 and 13 dB, respectively. As a result, the input power to the EDFA for all cases is kept constant.

Figures 5.5 and 5.6 show the eye diagram of the configuration-1 and configuration-2 of the experiment. The BER performance measured through the eye diagram ($2.8 \times$

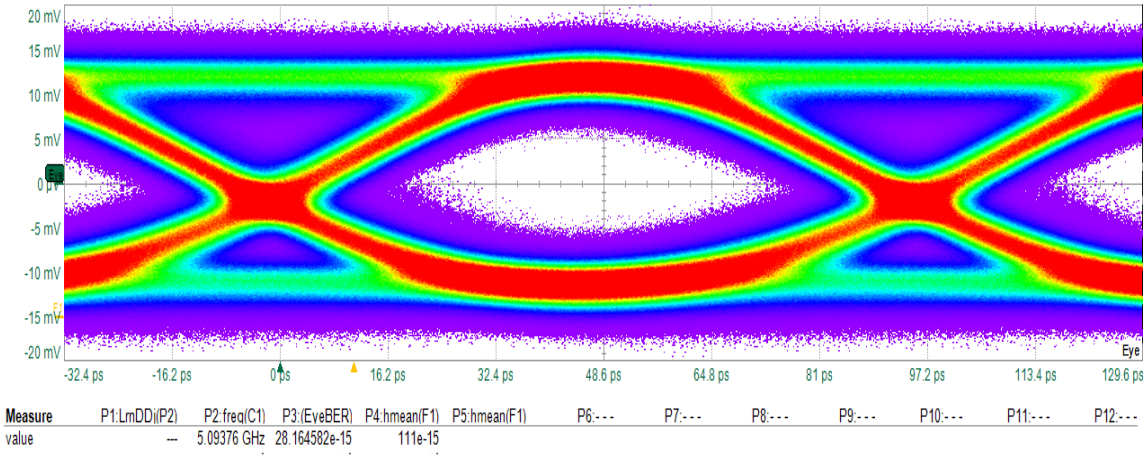

Figure 5.5: Experiment Configuration-1: with 15km fiber, EDFA, FBG no splitter loopback. Fixed loss of 13dB in the path to account for the two-way splitter loss

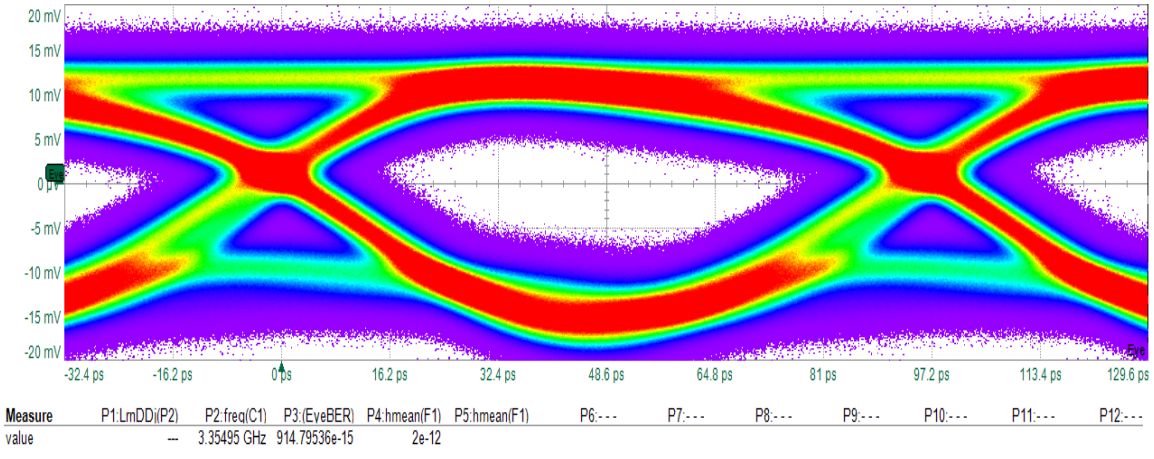

Figure 5.6: Experiment Configuration-2: with 15km fiber, EDFA, FBG and splitter loopback. Configuration of the WLB action

10^{-14} for configuration-1 and 9.1×10^{-14} for configuration-2) shows that although backscattering does introduce a distortion in the received signal, it has minimal effect

on the system performance. This is more evident from the BER performance shown in Figure. 5.7, where we compare these two configurations against the benchmark B2B configuration (shown by the yellow curve). The result shows a penalty of 2dB (at BER range of 10^{-10}) for configuration-1 with respect to the benchmark, which results from the signal broadening due to the fibre dispersion. The introduction of the WLB in the configuration-2 on the other hand incurs only an additional penalty of 0.3 dB over the configuration-1. These results prove that the backscattering introduced by the splitter loopback has negligible effect on the performance, while most impairments are simply due to the optical propagation through fibre.

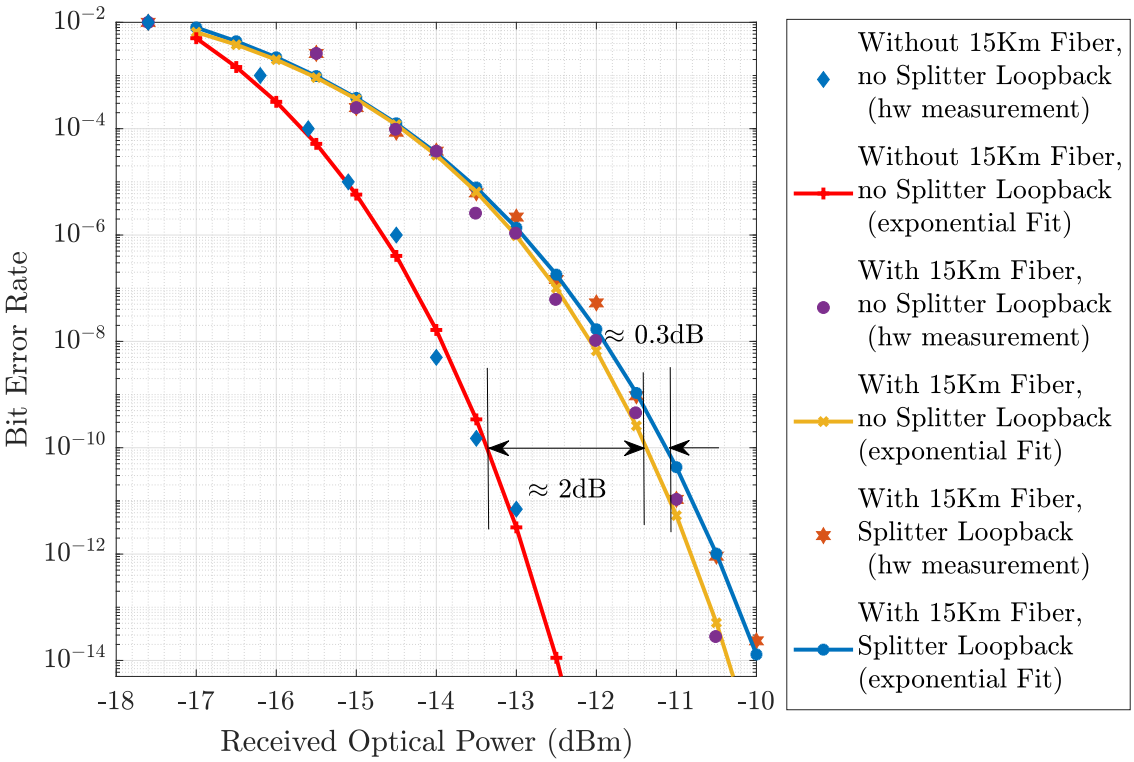

Figure 5.7: BER performance against the received optical power at the OLT with burst mode receiver

5.6 Performance evaluation of the proposed architecture through Discrete Event simulation

5.6.1 Simulation Overview

We simulate the protocol-level performance of the proposed architecture with the OMNET++ discrete event simulation. The topology described in Section 5.4 is created using OMNET's network descriptor. The simulation framework closely follows the ITU-T XG-PON specification [4], where the communication protocol over the PON follows the XGTC layer specification. We consider LLS split-8, and split 7.1 for the simulation. Traditional CPRI is a fixed rate traffic. However, we have experimentally demonstrated a VRF scheme which provides variable rate over CPRI split by dynamically adjusting the radio bandwidth (and thus the LTE sampling rate) depending on the cell load with the help of an SDN controller [93]. For split option 7.1, the processing of FFT/IFFT and removal of unused subcarriers is carried out at the RU. Therefore, by dynamically adjusting cell bandwidth depending on the cell load, as previously described, the fronthaul rate for split option 7.1 also becomes variable. We use [87] to derive the fronthaul rates for the VRF split-8 (the cell bandwidth varies from 1.4 to 20MHz) and split-7.1. We then further extend this procedure to derive the variable fronthaul rates for the 5G-NR scenario. The equations to derive the fronthaul rates for the CPRI split-8 (R_{CPRI}) and split-7.1 ($R_{7.1}$) are given in Eq. (5.1) and Eq. (5.2) respectively.

$$R_{CPRI} = 2N_{ant}R_sN_{res,CPRI}N_{ovhd}N_{8B10B} \tag{5.1}$$

$$R_{7.1} = 2N_{MIMO_L}\Big(N_{res,traffic}\frac{N_{scrr}}{T_{OFDMsymb}} + N_{bins}N_{res,PRACH}\frac{1}{T_{PRACH}}\Big) \tag{5.2}$$

In the equations above, N_{ant} is the number of antennas and N_{MIMO_L} is the number MIMO layers at the RU. R_s is the LTE sampling rate, $N_{res,CPRI}$ is the resolution in terms of number of bits, N_{ovhd} and N_{8B10B} are the CPRI specific overheads for control and line coding. N_{scrr} is the number of usable subcarriers per multi-carrier OFDM symbol, which scales linearly with the bandwidth. $T_{OFDMsymb}$ is the duration of a single multi-carrier symbol including the cyclic prefix (we assume the normal cyclic prefix mode here). $N_{res,traffic}$ is the resolution bits for the traffic channel and

$N_{res,PRACH}$ is the resolution bits for the Physical Random Access Channel (PRACH). T_{PRACH} is the periodicity of PRACH (which is 10 ms) and N_{bins} is the number of bins per PRACH allocation. Table. 5.2 lists the fronthaul rates derived form the equations, using the parameter values provided in Table. 5.3. The fronthaul rate for split-8 goes from 153 to 2,457 Mb/s, while the split-7.1 from 43 to 675 Mb/s. We use Eq. (5.1) and Eq. (5.2) along with the parameters listed in the Table. 5.3 to extend the computation of the fronthaul rates to the 5G scenario. Table. 5.4 provides the fronthaul rates used in the simulation corresponding to 5G-NR scenario where the sample bandwidth configuration and parameters are collected from 3GPP recommendation for 5G-NR [118]. However, in this work we only consider 2 antennas and 2 MIMO layers per RU. For a 100MHz bandwidth, 32 antennas and 8 MIMO layers per RU, with 16 bits of resolution per I/Q, the split-8 reaches 157.6 Gbps while split-7.1 reaches a fronthaul rate of 22.06 Gbps.

LTE BW Config	Sampling Rate	N_{scrr}	eCPRI rate (Mbps)	
			Split-8	Split-7.1
1.4	1.92	72	153.6	43.694
3	3.84	180	307.2	104.2
5	7.68	300	614.4	171.43
10	15.36	600	1228.8	339.50
15	23.04	900	1843.2	507.58
20	30.72	1200	2457.6	675.65

Table 5.2: eCPRI rates corresponding to split-8 with VRF and split-7.1. $N_{ant} = N_{MIMO_L} = 2$, $T_{OFDMsymb} = 71.4\mu s$

$N_{res,CPRI}$	N_{ovhd}	N_{8B10B}	$N_{res,traffic}$	$N_{res,PRACH}$	N_{bins}
16	16/15	10/8	10	10	839

Table 5.3: Parameters for calculating the CPRI and eCPRI rates for LTE and NR

We consider a Poisson distributed end user traffic (i.e., measured at the UE) with exponential inter-arrival time. The mapping of user traffic to fronthaul traffic follows the same process described in [116] and summarized as follows. Let us consider the users arrive at the RU following a Poisson distribution with intensity γ (arrivals/unit time) and submit a connection request, which we refer to as the call request, following the terminology used in queuing theory. Upon arrival, if accepted in the system, a

5G-NR Bandwidth Config (MHz)	Sampling Rate (MHz)	N_{scrr}	eCPRI rate (Mbps)	
			Split-8	Split-7.1
20	30.72	612	2457.6	689.104
30	46.08	936	3254.4	1052.14
50	61.44	1596	4915.2	1791.68
70	92.16	2268	7372.8	2544.66
100	122.88	3276	9830.4	3674.13

Table 5.4: Fronthaul rates corresponding to split-8 (CPRI) with VRF and split-7.1 (eCPRI) for 5G-NR scenario. $N_{ant} = N_{MIMO_L} = 2$, $T_{OFDMsymb} = 71.4\mu s$, Subcarrier Spacing (SCS) = 30KHz

service session is initiated and one RG resource (which is equal to two PRBs for LTE) is allocated (we refer to this as server) which is occupied for the duration of the accepted call (we refer to this as the holding time). Therefore the number of servers (κ) can be determined by the number of RGs in the highest bandwidth configuration of the cell. Following this, we can calculate the maximum number of users that can be served for each bandwidth configuration of the RU from Table. 5.2. In order to explain how the variable rate fronthaul system operates, let us consider a scenario where an RU is serving already the maximum number of users for the current cell bandwidth configuration. If a new user arrives, which cannot be handled within the current bandwidth, the SDN controller triggers a request to increase the cell bandwidth. As a consequence, the fronthaul rate (for both CPRI and split 7.1) also increases to support the higher bandwidth configuration (as listed in Table. 5.2). Similarly, when a call departs and the number of remaining users can be supported by the next lower bandwidth configuration, both wireless spectrum and fronthaul rate are decreased accordingly. If the average holding time of the call is τ time units (or the call departure rate is $\mu = \frac{1}{\tau}$ calls/unit time), then the traffic load (Erlang) is given by $\rho = \gamma/\mu$. From the RUs perspective, the system maintains a steady state if $\frac{\gamma}{\kappa\mu} < 1$. As the RU requires some local processing time for the LLS functional split processing and the encapsulation of the eCPRI traffic, we model this through a uniform distribution with an upper limit of 125 μs. We measure the latency as the time between the packet arrival at the RU corresponding to a particular TTI, and its reception at the DU.

5.6.2 Results

Figure. 5.8 shows a latency reduction of over 10 times between RU and DU, obtained by edge vPON slicing w.r.t. the use of OLTs located at the CO. The figure also shows the difference in latency when the fibre routes are overlaid on top of current PON routes to interconnect the level-1 splitters (i.e., logical ring, shown in red curve), versus the physical ring architecture where direct fiber routes were used to interconnect them (shown in blue curve). We observe that end-to-end transport latency is

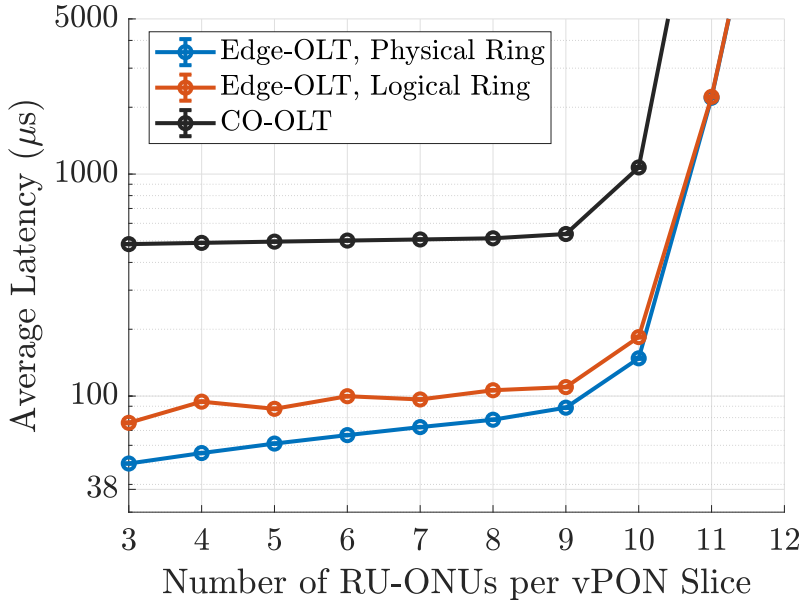

Figure 5.8: Comparison of MFH transport Latency (μs) w.r.t vPON slice size (number of ONUs per vPON slice) for traffic intensity of 12.5 Erlang and split-8 (VRF).

somewhat higher for the logical ring case due to longer fiber propagation distance, however still around 100 μs, thus compatible with our selected threshold. In this configuration, half of the ONUs per vPON slice is from the adjacent level-1 PON tree (50 % inter-PON load). In the case when all the ONUs in a vPON slice are from the adjacent level-1 PON tree or 100 % inter-PON load (as shown in Figure. 5.9), the latency for the logical ring increases slightly ($\approx$ 110 μs) due to increase in average propagation distance, while the latency for the physical ring still remains below the 100 μs threshold. This suggests that while offloading ONUs to a nearby edge-OLT, the CO should optimally reconfigure the vPON slices so that the overall latency

remains below the target threshold level.

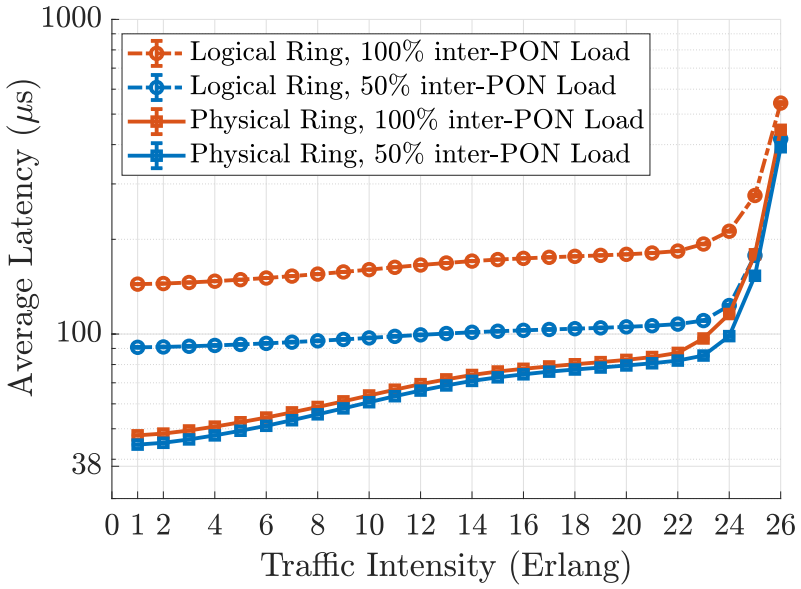

Figure 5.9: Comparison of MFH transport Latency (μs) w.r.t traffic intensity on logical ring vs. physical ring for 50% and 100% inter-PON ONUs per vPON slice and split-8 (VRF).

Figure. 5.10, illustrates how our proposal can be exploited to considerably improve statistical multiplexing of cells through MEC migration of DUs, by dynamically re-configuring vPON slices, depending on the traffic intensity reports from the DU. The architecture we use in the following results are obtained for the case of physical ring architecture. We consider two edge OLTs and 24 ONUs, where half of them are residential and served by the CO. Initially, at low traffic volumes, the edge OLT1 starts with all 12 RU-ONUs (these are the ONUs attached to RUs, e.g., providing the mobile fronthaul service) and we can see that the latency increases as the traffic at each RU increases. At 10 Erlang traffic per RU, the latency reaches our threshold, set at 100 μs (this value can be set to the most appropriate value required by the service). The CO thus activates the edge OLT2 and reconfigures the vPON slices, offloading one ONU to the vPON slice served by OLT2. This causes a sharp reduction in uplink transport latency at OLT1. As the traffic from the RUs further increase, the process is repeated as soon as the latency grows close to the threshold level. Another possible approach to load balancing is instead to offload 6 of the 12 RU-ONUs to the vPON slice corresponding to the OLT2 at once, when the latency approaches the

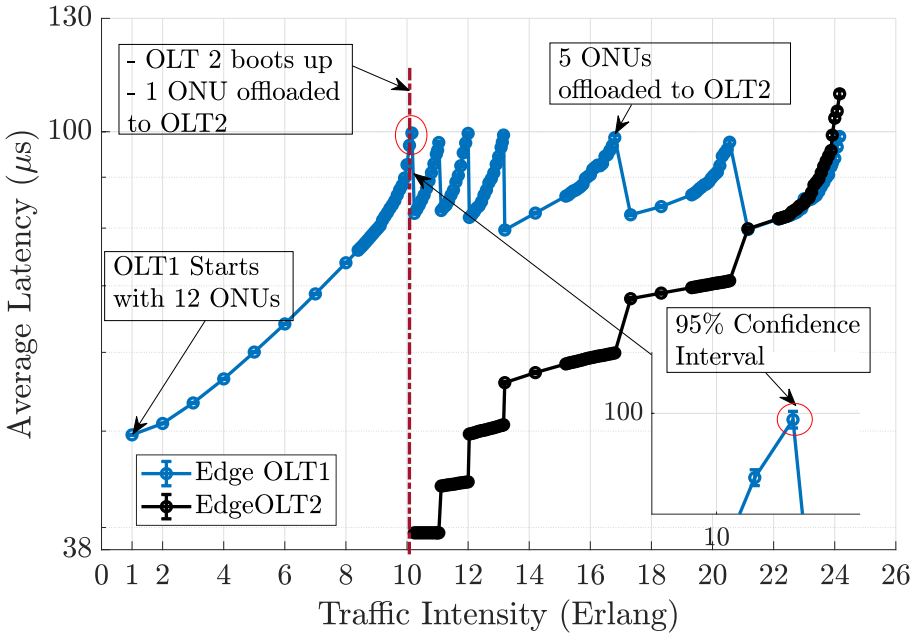

Figure 5.10: Illustration of MFH transport Latency w.r.t RU traffic intensity for unbalanced migration of RU-ONUs across edge OLTs using the proposed dynamic vPON slicing technique. All RU-ONUs are using split-8 (VRF).

threshold. The performance results from applying this second method, which reduces the frequency of offloading events, are reported in Figure. 5.11.

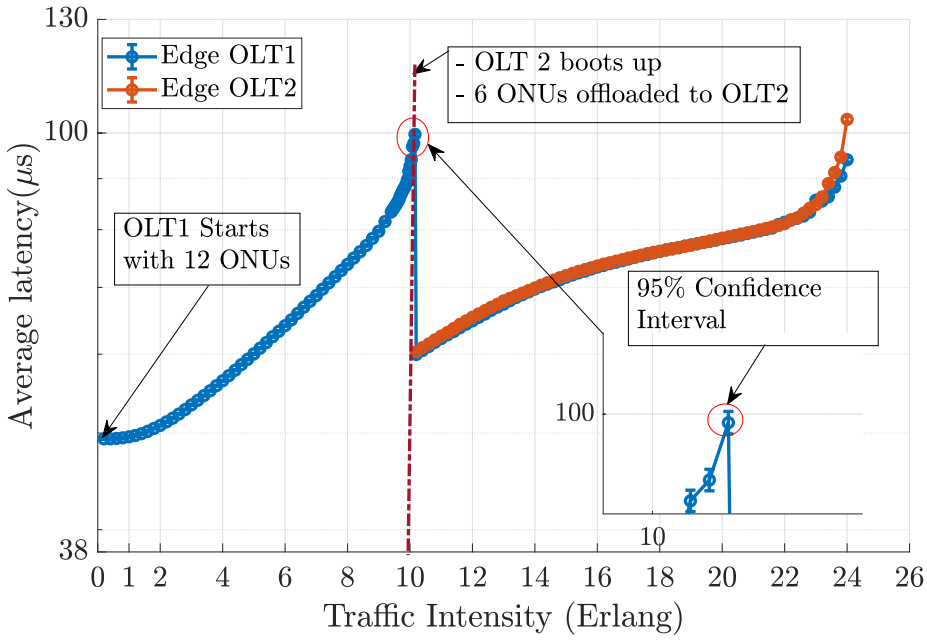

Figure 5.11: Illustration of MFH transport Latency w.r.t RU traffic intensity for balanced migration of RU-ONUs across edge OLTs using the proposed dynamic vPON slicing technique. All RU-ONUs are using split-8 (VRF).

All the results described above use split-8, which remains the most bandwidth hungry across all the possible functional splits. Therefore, as can be seen from Figure. 5.12, the queuing latency raises quickly as cell traffic increases in a given vPON slice (at 10 erlang for 12 ONUs per vPON slice). On the other hand, if all RUs in the vPON slice use split-7.1, the queuing latency at the ONU is negligible for 6 ONUs per vPON slice. If we increase this to 12 ONUs per vPON slice, the queuing latency becomes higher but still without suffering a sharp increase. This figure also shows the case when half of the RUs in a vPON slice uses split-8 with VRF and the other half adopts split-7.1 (labelled as 50% split-7.1 in the figure), which present intermediate performance results.

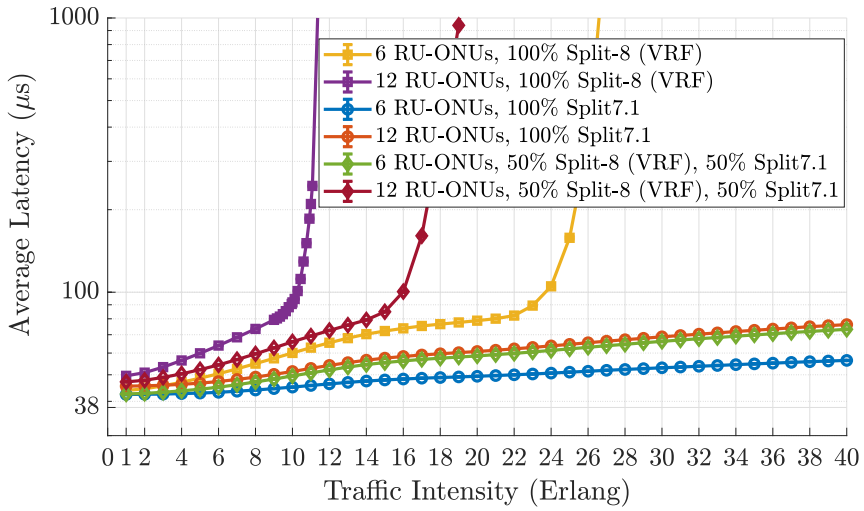

Figure 5.12: Comparison of MFH transport Latency w.r.t traffic intensity over physical ring for different functional split configurations (split-8 (VRF) and split-7.1)

Figure. 5.13 plots the latency performance against the number of ONUs per vPON slice, for the two split points for a moderate traffic intensity of 12.5 Erlang. Here we can observe that we can accommodate as many as 20 ONUs per vPON slice if all RUs adopt split-7.1 (compared to 9 for split-8 VRF) while keeping the latency below the chosen threshold of 100 μs. Finally, Figure. 5.14 shows the maximum number of ONUs per vPON slice depending on the average traffic intensity that still allows to keep latency below 100μs, for the same three different split configuration. Therefore, given a required QoS in terms of fronthaul transport latency, and average traffic load

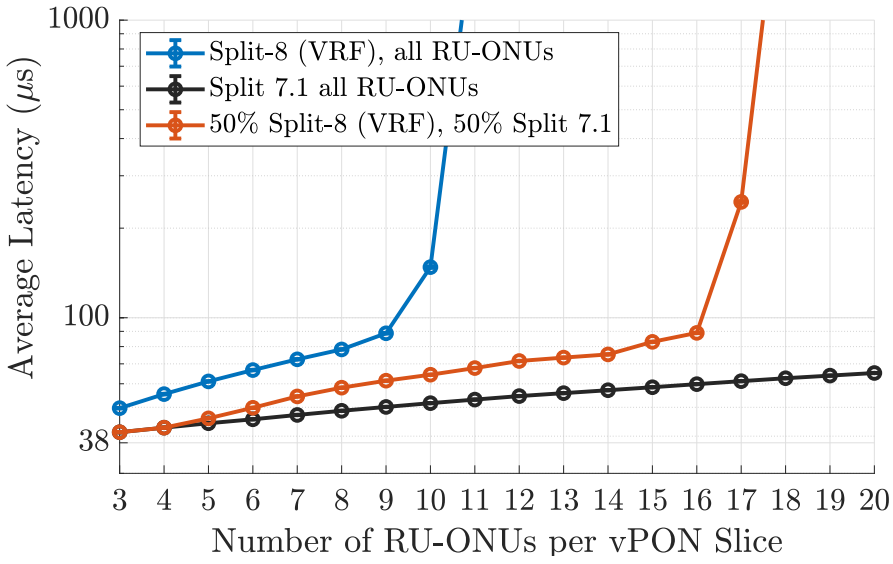

Figure 5.13: Comparison of MFH transport Latency w.r.t number of ONUs per vPON slice for different functional split configurations (split-8 (VRF), split-7.1 and mixed split deployments), physical ring, traffic intensity = 12.5 Erlang.

over the network, this result helps in determining the maximum number of ONUs per vPON slice depending on the split configuration of the deployed RUs.

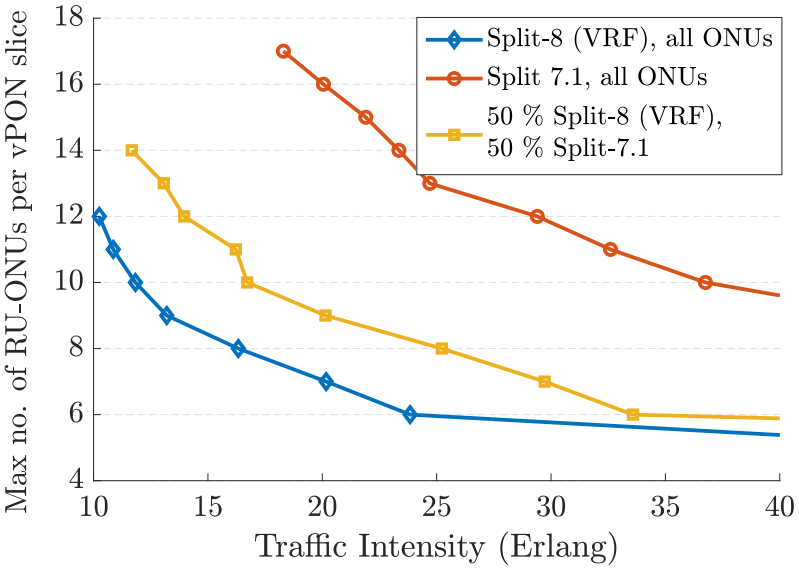

Figure 5.14: Performance comparison showing maximum number of ONUs per vPON slices vs. the average traffic intensity, for different functional split configurations to achieve below 100μs MFH transport latency.

It is also worth to consider how our proposed system performs in a 5G-NR scenario. For 100 MHz transmission bandwidth, with only two antenna configuration (or MIMO

layers as we are considering them equal in this work of the dissertation), each RU starts to push Fronthaul data at 9.83 Gbps rate with split-8 and 3.68 Gbps with split-7.1. Therefore, a 10G PON (such as XGS-PON) as considered in the previous results, is not suitable. This is evident from Figure. 5.15 as we can see the red solid curve corresponding to split-8 increases steeply even at very low traffic, whereas with split-7.1, a vPON slice configuration with 4 RU-ONUs per slice can still be used for a moderate traffic intensity (till 25 Erlang). Therefore, a high bandwidth PON such as 50G PON should be used to overcome the queuing latency/ONU buffering (dotted blue and red curve in Figure. 5.15). Finally, Figure. 5.16 provides a more conclusive result by showing the maximum number of RU-ONUs per vPON slice configuration depending on the traffic condition when a particular PON (10G-PON or 50G-PON) carries fronthaul data over a 5G-NR Fx interface. We have not shown the results for higher MIMO layers, as the data rate can easily reach values above 150Gb/s for split-8 and above 20Gb/s for split-7.1, which would be difficult to manage even in a 50G PON. These would indeed require PON channels rates of 100 Gb/s and above, with more than one wavelength channel. Although not shown here, as they might be considered speculative, such results could be easily extrapolated from the current results.

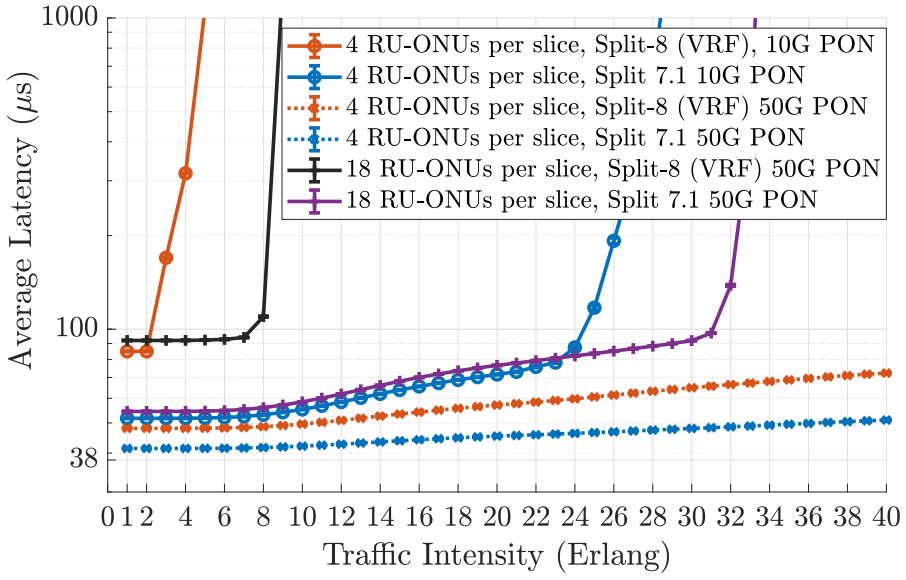

Figure 5.15: Performance comparison showing latency performance over 5G-NR fronthaul over 10G PON and 50G PON.

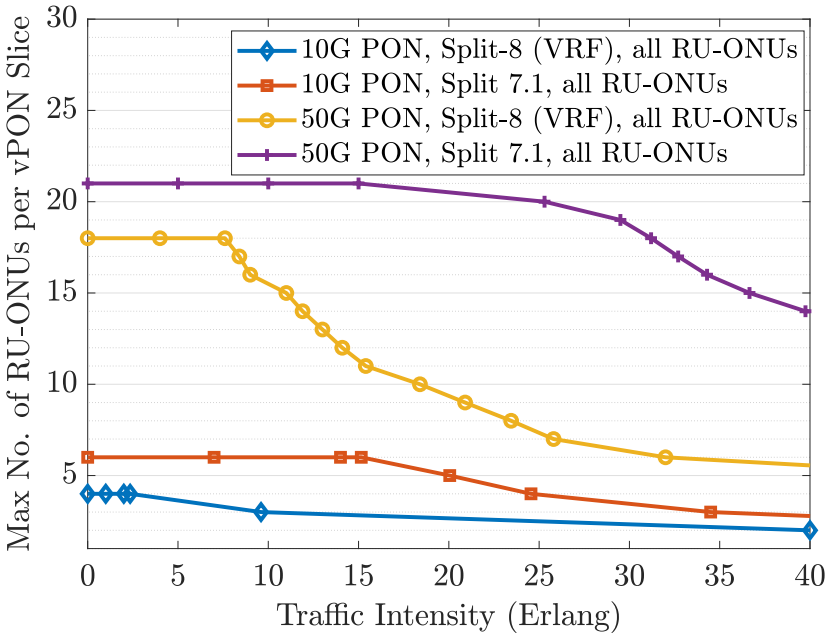

Figure 5.16: Performance comparison showing maximum number of ONUs per vPON slices vs. the average traffic intensity, for different functional split configurations on 5G-NR to achieve below 100μs MFH transport latency over 10G and 50G PON.

5.7 Conclusion

In this Chapter, we have introduced a novel PON architecture which enables both NORTH-SOUTH and EAST-WEST communication, giving the ability to set up multiple virtual mesh topologies with low latency. The architecture enables end points to host both mobile cell sites and MEC nodes, with the positioning of additional OLTs at the user end points and using reflective filters at the splitter locations to back-propagate signals towards other ONUs. We experimentally show that back scattering due to this reflective action has negligible effect on the system BER performance. Through protocol-level simulations, we then show how cooperative DBA can be exploited to maintain low fronthaul transport latency under varying mobile traffic conditions, while at the same time achieving statistical multiplexing of RU-ONUs employing heterogeneous functional splits. Our results show that even if a direct physical fiber deployment is not possible between level-1 splitters, existing ducts through level-2 splitters can be used to realise EAST-WEST communication for low-latency fronthaul transport. We further show that under highly dynamic

traffic scenarios, by dynamically offloading functional split computation across edge nodes using dynamic vPON slicing technique, our EAST-WEST PON architecture can maintain the system latency below a given threshold (set to 100 μs in our work). This is achieved through appropriate MEC migration strategies, so that as traffic in the RU-ONUs increases, their computation can migrate towards other local OLTs hosting MEC nodes. Following this, we give insights on how these vPON slices can be formed dynamically, depending on the cell load, processing capacity at the MEC nodes, and functional split option employed at the RU-ONUs, so that migration of DU processing across MEC nodes meets the target latency threshold. We finally show how our proposed architecture and the corresponding results can in general be applied for a scaled up 5G-NR system with high bandwidth configuration and supported by next-generation high bandwidth PON based fronthaul. In conclusion, we show how our PON architecture enables the convergence of mobile and MEC nodes, delivering deterministic low-latency performance under highly dynamic traffic scenarios. The proposed virtualized EAST-WEST PON mechanism and the corresponding results discussed in this chapter have been published in [119] and [120].

6 Optimal virtual PON slicing supporting mesh traffic pattern in MEC-based Cloud-RAN

Optimal virtual PON slicing to support ultra-low latency mesh traffic pattern in MEC-based Cloud-RAN

As progressive densification of cells, deployment of Cloud-RAN and Multi Access Edge Computing (MEC) are coming into reality to support the ultra low latency with high reliability in 5G and beyond, it generates mesh traffic pattern across fronthaul network. This requires the evolution of the mesh PON architecture in order to support such mesh traffic pattern. In the previous chapter (Chapter 5), we have shown how a novel EAST-WEST PON architecture based on PON virtualization, has the potential to transport such mesh traffic with ultra low latency using dynamic virtual PON slicing. However, dynamic optimal allocation of virtual PON slices over such mesh-PON based fronthaul transport is a challenging task. Moreover, with the progressive cell densification in 5G, such MESH-PON based fronthaul transport architecture grows too complex to find optimal virtual PON slices within real-time or near-real time. In this chapter therefore, we propose and discuss a mixed analytical-iterative model to compute optimal virtual PON slice allocation, providing mesh access connectivity with ultra-low end-to-end latency in next generation MEC-based Cloud-RAN. Our proposed method can compute optimal virtual PON slice allocation in timescales compatible with real-time or near real-time operations.

6.1 Introduction

In the previous chapter (Chapter 5), we have shown how an EAST-WEST PON architecture along with PON virtualization can support ultra-low end-to-end latency in next generation converged networks. Besides the traditional end point to central office (i.e., NORTH-SOUTH) communication, these architectures enable direct (i.e., EAST-WEST) communication between PON end points, which becomes a key feature to support high capacity and low latency interconnection of MEC nodes in next generation access network. The ONF AETHER [121] is a prominent example of this distributed access architecture, requiring a high-capacity, low latency mesh access network.

Although, dynamic virtual PON slicing over such mesh PON architecture can achieve low-latency offloading of Optical Networking Units (ONUs), migration of Distributed Unit (DU) between MECs, the formation of such slices dynamically is not an easy task. It requires a careful consideration of the network layout, functional split between DU and Radio Unit (RU), and traffic load at each RU site while offloading RU-ONUs across different edge OLTs for load balancing scenarios. For example, as can be seen from Figure. 5.9 in Chapter 5, for a particular traffic load over a given network layout, offloading a large number of inter-PON RU-ONUs to a nearby edge OLT may surpass the end-to-end latency from a target value. Therefore, dynamic formation of such vPON slices optimally over a given fronthaul network layout, heterogeneous functional splits across RUs and traffic load per RUs, requires solving a network optimization problem in real-time or near-real time.

As 5G is expected to experience progressive cell densification, consequently the complexity of virtual PON slice allocation in MESH-PON based fronthaul network would grow exponentially complex. Therefore, the optimal vPON slice allocation over such complex network in real-time or near-real time would become immensely challenging.

Therefore, in this chapter, we address the dynamic virtual PON slice allocation problem of an access network where a mesh PON topology enables dynamic interconnec-

tion of RUs, MEC nodes and central offices. The virtualized mesh topology can be created according to the architecture mentioned in the previous chapter (Chapter 5), where MEC nodes can be placed at the PON endpoints and EAST-WEST communication between PON end nodes can be established using Wavelength Loop Back (WLB) technique with reflective splitters. Here we are able to create virtual PON slices (whose capacity is allocated through dynamic use of wavelength channels) to enable direct communications between RUs and MEC nodes (hosting DUs and possibly CUs) to support operation of Cloud Radio Access Networks (Cloud-RAN) instances.

This work proposes a method for optimal formation of virtualised PON (vPON) slices under dynamic traffic scenarios. Given a number of RUs supporting a mix of 7.1 and 7.2 functional split, with varying traffic load and pattern, we determine the optimal set of small cells, macro cells and MEC nodes (our virtual group of end points), that can support the required traffic while maintaining latency below a target threshold. Once the slices are created, our approach is also used to maintain the latency target, in real time, below that threshold. As changes in traffic load and patterns produce latency increase above threshold, we re-configure the virtual topology (i.e., MEC node migration) to reduce latency. A key achievement of this work is the development of an analytical model for PON latency, which significantly reduces the slice computation time, down to few tens of second (depending on load and number of iterations), which makes this algorithm suitable for real time network optimisation.

6.2 System Model

Figure. 6.1, presents the system architecture and use case of the MESH-PON scenario where a Macro Cell with embedded MEC computation hosts an OLT, which enables direct communication to the nearby small cells (connected through ONUs). In this work we assume such direct connectivity is achieved through Fibre Bragg Grating reflectors located at splitter locations (as discussed in Chapter 5), although our solution is transparent to the specific physical layer implementation.

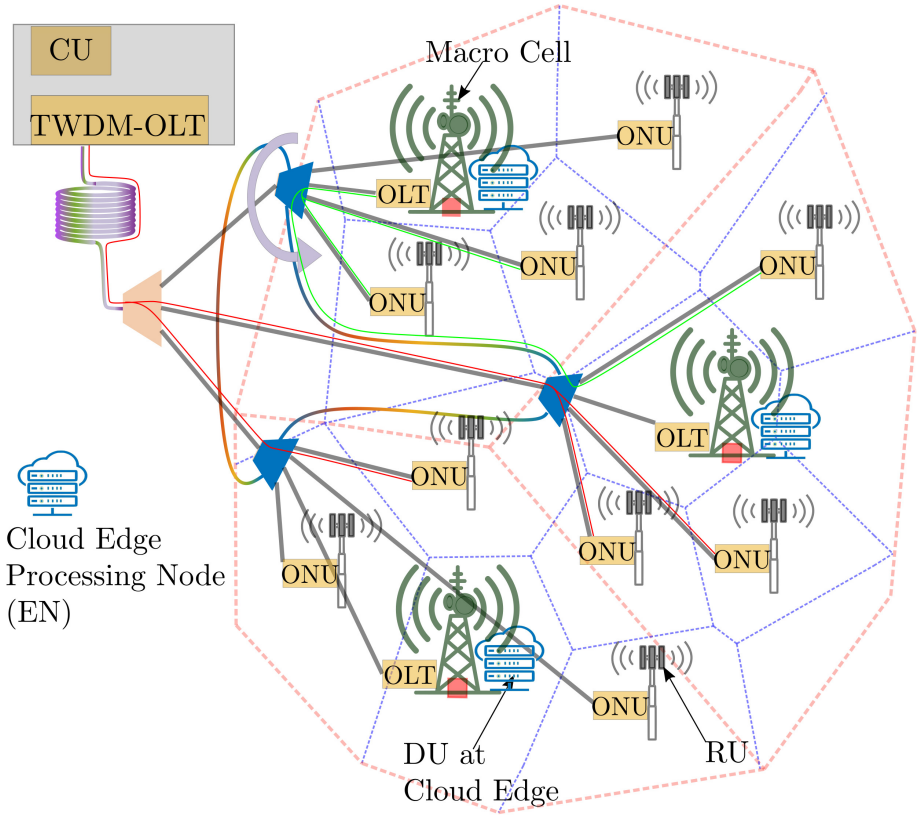

Figure 6.1: Virtualised Mesh-PON architecture supporting MEC-based Cloud-RAN showing EAST-WEST links (green) along with traditional NORTH-SOUTH links (red).

Figure. 6.2 shows a sample solution returned by our virtual PON allocation model, with minimum number of MEC nodes to guarantee latency below 100 μs threshold. Here, the macrocell and small cell coverage areas are modeled using the polybound-vornoi diagrams. The small cells within the boundary of the corresponding macrocell (red color borders) are connected by a level-1 PON tree with a possible MEC node (with OLT) deployed in the macrocell site. RUs at the small cells (blue dot) implement Cloud-RAN with functional split 7.1 or 7.2 which is served by an ONU. Their OLT is located at the computing node that implements the corresponding DU (and possibly CU): this can either be an MEC node or a central office (depending on latency requirements). In addition, we use the common assumptions that MEC nodes are physically co-located with macro cell sites. The core network functions are hosted at the CO regardless of the placement of DU/CU. The sample solution shows the optimal no. of MEC nodes and their respective location to be deployed, or if already deployed, should be enabled (as represented by yellow circle) or to be

put into hibernation (as represented by grey circle) to save energy. The green lines show a snapshot of the optimal virtual PON slices illustrating the dynamic connection of RU-ONUs with edge OLTs and consequently the MEC nodes via low-latency EAST-WEST links.

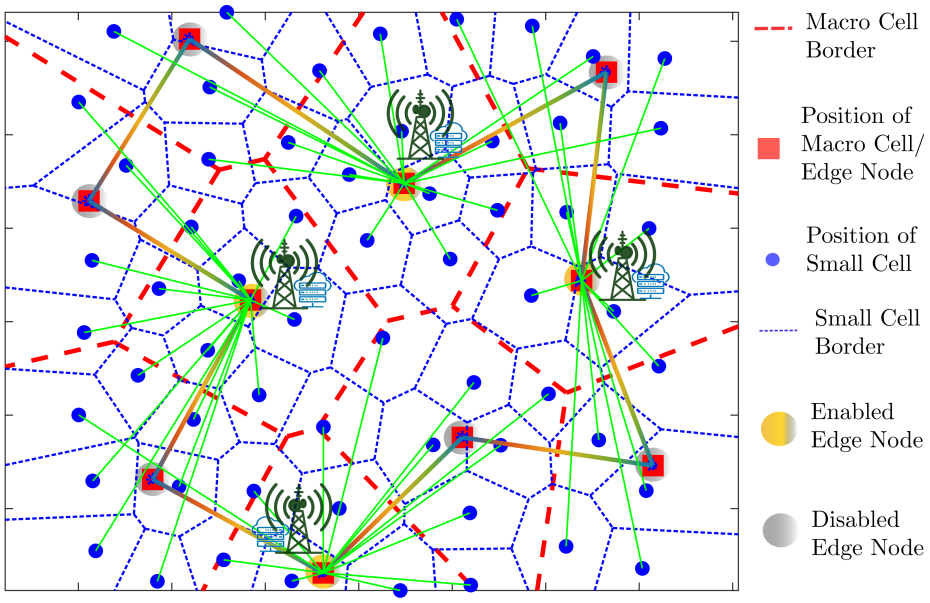

Figure 6.2: Sample of network layout optimal solution computed by our mixed analytical-Iterative model. Only EAST-WEST links (green) are present due to the 100 μs latency constraint.

6.3 Optimal formation of vPON slice to provide ultra-low latency in uplink CoMP clusters

In order to find a solution for the virtual PON slice allocation problem, we first need to find an analytical expression for upstream latency per vPON slice based on parameters such as number of RUs, traffic at RUs and functional split (i.e., 7.1 or 7.2). To achieve this, we first find the packet queuing latency per vPON slice as a function of these parameters and then add the propagation latency, according to the slice configuration to obtain the end-to-end latency per vPON slice. We use the Kingsman heavy traffic approximation method for G/G/1 system for finding the mean packet queuing time in ONU queue as described in Eq. (6.1) below.

$$T_W \leq \frac{1}{\mu C} + \frac{\lambda(\sigma_a^2 + \sigma_b^2)}{2(1-\rho)} \tag{6.1}$$

In Eq. (6.1), $1/\lambda$ is the average packet inter-arrival times, σ_a^2 is the variance of inter arrival times, σ_b^2 is variance of the service times, $\rho = \lambda/\mu$ is the utilisation factor and $1/\mu$ is the average service time per packet.

Let X be the random variable defining the service time of a fronthaul packet in uplink considering coordinated DBA for uplink packet scheduling. Let D be the set of possible fronthaul packet sizes arriving per Grant Cycle (GC) in a particularr vPON slice, then the mean and the variance of the service times can be calculated as:

$$E\{X\} = E\{k \times \Pr\{D_j = k\}\} \times D_{cap}^{UL} \tag{6.2}$$

$$1/\mu = E\{X\}, \quad \sigma_b^2 = E\{X^2\} - E\{X\})^2 \tag{6.3}$$

$$X = \{X_k \,|\, X_k = k \times (1/D_{cap}),\ k \in D\}$$

The state probability $Pr\{D = k\}$ (or $P(D)$) in Eq. (6.1), where random variable $D = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ denotes the set of possible aggregated fronthaul packet size ($\underline{x}_i$) per uplink GC in a vPON slice can be found using successive convolution method [122] as given Eq. (6.4) and Eq. (6.5). Here, $P_j = \{p_j(d_i) \,|\, d_i \in \{(d_1, d_2, \ldots d_{K_{7.1}}) | (d_1^{'}, d_2^{'}, \ldots d_{K_{7.2}}^{'})\}\}$ are the state probabilities of fronthaul packet size for j^{th} RU considering two possible split configurations per RU, Split-7.1 or 7.2 and d_s is the eCPRI packet segment size.

$$P(\underline{x}) = P_1 \circledast P_2 \circledast \ldots P_N, \quad D = \sum_{j=1}^{N} x_j.d_s \leq d_{cap} \tag{6.4}$$

$$\text{Where,} \quad P_i \circledast P_j = \{p_i(0).p_j(0), \sum_{x=0}^{1} p_i(x).p_j(1-x), \sum_{x=0}^{2} p_i(x).p_j(2-x), \ldots, \sum_{x=0}^{u} p_i(x).p_j(u-x)\} \tag{6.5}$$

$p_j(d_i)$ can be found by considering an M/M/m/m system (queuing theory) at the RU as follows.If γ and ν be the user call arrival and depart rate at each RU, then the

probability that at the steady state, k-user is connected with RU is given by

$$\begin{aligned} p_k &= p_0 \left(\frac{\gamma}{\nu}\right)^k \frac{1}{k!} && k \leq m \\ &= 0 && k > m \end{aligned} \tag{6.6}$$

$$\text{where} \qquad p_0 = \left[\sum_{k=0}^{m} \left(\frac{\gamma}{\nu}\right)^k \frac{1}{k!}\right]^{-1}, \quad \text{and} \tag{6.7}$$

m is the maximum number of users that can be supported at the RU. If We consider thresholds $F_1 < F_2 < \cdots < F_n$ as the number of active users per RU to jump between different eCPRI rates, then we can calculate the probability $p_j(d_i)$ that the fronthaul rate for RU-j is d_i as

$$\begin{aligned} p_j(d_i) &= \{F_{i-1} \leq \text{number of users in the RU} \leq F_i\} \\ \text{or} \qquad p_j(d_i) &= \frac{\sum_{k=F_{i-1}}^{F_i} \left(\frac{\gamma}{\nu}\right)^k \frac{1}{k!}}{\sum_{k=0}^{m} \left(\frac{\gamma}{\nu}\right)^k \frac{1}{k!}} \end{aligned} \tag{6.8}$$

The next step is to define the model for optimal vPON slicing that satisfies a ultra-low latency threshold while minimising total number of MEC nodes to be deployed for a given traffic intensity per RU. The decision variable $\alpha_i \in \{0, 1\}$ determines whether a MEC-node$_i$ is to be deployed at the i^{th} level-1 PON tree or not. The objective function for the slice optimisation model is given in Eq. (6.9), and the constraints are described in Eq. (6.10)-Eq. (6.15). The optimal level-1 ring to minimize the uplink latency over EAST-WEST PON can be realised by getting a Hamiltonian tour for which the travelling distance is minimized, which is a classical Travelling salesman optimisation problem and can be formulated using Eq. (6.16) and Eq. (6.17),

Table 6.1: Notations for mathematical symbols

Symbol	Description
$\mathcal{W}$	Set of wavelengths for EAST-WEST PON
r	RU-ID $r \in \{1, 2, \ldots, \lvert R\rvert\}, (R := \text{No. of RUs})$
v	Denotes the vPON ID $v \in \{1, 2, \ldots, \lvert V\rvert\}$
α_i	Binary variable. 1 if an MEC-node, is to be deployed at the i^{th} level-1 PON tree.
Sl_v	Set of RUs belonging to vPON slice v i.e., $\{r \mid r \in R, \text{ and } X_{r,v} = 1\}$
$\mathcal{T}_{Sl_v,a}$	Theoretical value of uplink latency for vPON slice Sl_v and traffic load a (% load)
$\mathcal{T}_v^{max}$	Maximum value of uplink latency for vPON slice. Latency threshold (100 μs)
$\mathcal{L}_i$	Describes i^{th} physical Level-1 PON tree cluster. i.e., $\{r \mid r \in R \text{ and } X_{r,i} = 1\}$
$X_{r,v}$	Binary decision variable $\in \{0, 1\}$. 1 if RU-r is assigned to vPON-v $(r \in R, v \in V)$
$\mathcal{K}_i^{\mathcal{W}}$	Set of $\mathcal{W}$-nearest neighbours of the i^{th} L1 PON tree.

$$\text{minimize:} \quad \sum_{i=1}^{N_{L_1}^{PON}} \alpha_i C_i^{CAP} + \sum_{i=1}^{N_{L_1}^{PON}} \alpha_i \mathcal{W} C_i^{OLT} \tag{6.9}$$

subject to (constraints):

$$\sum_{v \in V} X_{r,v} = 1 \quad \forall\ r \in R \tag{6.10}$$

$$\sum_{r \in R} X_{r,v} \leq \alpha_v |R| \quad \forall\ v \in V \tag{6.11}$$

$$\sum_{r \in R} \sum_{v \in V} X_{r,v} = |R| \tag{6.12}$$

$$\mathcal{T}_{Sl_v,a} \leq \mathcal{T}_v^{max} \quad \forall\ v \in V \tag{6.13}$$

$$\sum_{v \in K_v^{\mathcal{W}}} X_{r,v} = 1 \quad \forall\ r \in R \tag{6.14}$$

$$\sum_{i} \alpha_i = |V| \tag{6.15}$$

$$\text{minimize:} \sum_{i \in 0,\ldots,n-1} dist(e_i, e_{i+1}) + \ dist(e_n, e_0) \tag{6.16}$$

$$\text{subject to:} \qquad e_i \text{ are a permutation of } N \tag{6.17}$$

$$e_i := \text{edge Node Location}(< x_i, y_i > \ \ i \in \{1, 2 \ldots N\})$$

The latency constraint given by Eq. (6.13) is a nonlinear function that can be solved with known nonlinear discrete optimisation solvers. However, the exhaustive search with such non-linear solvers is extremely slow due to large search space and time-intensive non-linear constraint evaluation. Therefore, we propose an iterative optimisation method (Algorithm 1)(governed by parameter “Max iterations") along with integer-linear programming to take care of the non-linear constraint and speed-up the optimisation significantly.

In a nutshell, this iterative method first evaluates the optimal no. of MEC nodes by evaluating the integer-linear programming model without the non-linear constraint ($\mathcal{O}_{\mathcal{L}}$). The obtained optimal no. of MEC nodes from the linear model is used as a lower bound (N_{MEC}^{lb}) for further exploration. Non-linear latency constraint is evaluated on solution obtained for each vPON slices (corresponding to MEC nodes) and checked for constraint violation. A slice configuration constraint is added for each of the latency violating slice of the linear-model solution, and the integer-linear model is run again. This process iterates until the

Algorithm 1: Iterative algorithm to solve the nonlinear discrete slice optimization model within bounded time

Set up and Initialize integer linear model $\mathcal{O}_{\mathcal{L}}$;
Solve $\mathcal{O}_{\mathcal{L}}$;
N^{lb}_{MEC} = optimal no. of MEC nodes (output of $\mathcal{O}_{\mathcal{L}}$);
Evaluate non-linear constraint Eq. (6.13) for each vPON slice;
Find $\mathcal{N}$:= set of vPONs where non-linear constraint Eq. (6.13) is not satisfied ;
iteration_ID = 0;
while $\mathcal{N}$ *is not empty* **do**
    if *max no. of iterations passed* **then**
        increase lower bound: $N^{lb}_{MEC} = N^{lb}_{MEC} + 1$;
        iteration_ID=0;
    else
        iteration_ID=iteration_ID+1;
    end
    for *each vPON slice* $v \in \mathcal{N}$ **do**
        $X^{sol}_{r,v} \leftarrow$ current vPON config solution $X_{r,v}$;
        add constraints to $\mathcal{O}_{\mathcal{L}}$:
            $\sum_{\substack{r\in R,\\ X^{sol}_{r,v}=1}} X_{r,v} \leq \alpha_v(\sum_{r\in R} X^{sol}_{r,v} - 1)$;
    end
    solve updated $\mathcal{O}_{\mathcal{L}}$;
    if *feasible/optimal solution obtained* **then**
        Evaluate non-linear constraints Eq. (6.13);
        Find updated $\mathcal{N}$;
    else
        Result: No feasible solution. Exit
        ;
    end
end
Result: return the optimization output (slice config)

non-linear latency constraint is passed for all the obtained slices. We define maximum number of iterations to attempt to satisfy the non-linear constraint so that the algorithm is not stuck in the iteration for indefinite time. Once the maximum number of iterations has passed for the current lower-bound of the number of active MEC nodes, we increase the lower bound and start the process again until an optimal solution is found. Therefore, max-number of iterations creates a trade off on the quality of the optimal solution vs the speed of the optimization.

6.4 Performance Evaluation and Results

Our first step is the validation of the analytical model with simulations carried out in OMNET++. We consider a multi-wavelength architecture (i.e. following NG-PON2), although we assume a next-generation rate of 50Gbps per channel. The traffic from RU to DU is modeled as evolved Common Public Radio Interface (eCPRI) traffic. We consider split 7.1 and 7.2, both providing variable rate depending on the actual traffic at the cell. The corresponding fronthaul rates are derived from [87] and scaled to 5G configuration of 100 MHz cell bandwidth. For an RU having four antennas and 4-MIMO layers, the fronthaul rate for split-7.1 goes from 1.378 to 7.384 Gb/s, while the split-7.2 it goes from 273.98 Mbps to 2.92 Gbps. Our first result, in Figure. 6.3, reports the feasibility region, showing the optimal mix of small cells using 7.1 and 7.2 split, that satisfies a given latency threshold (100 μs in this case). In Figure. 6.3, different curves refer to different load at RUs, expressed as percentage of average cell load. This results show how our analytical results based on queuing theory are in close agreement with simulations.

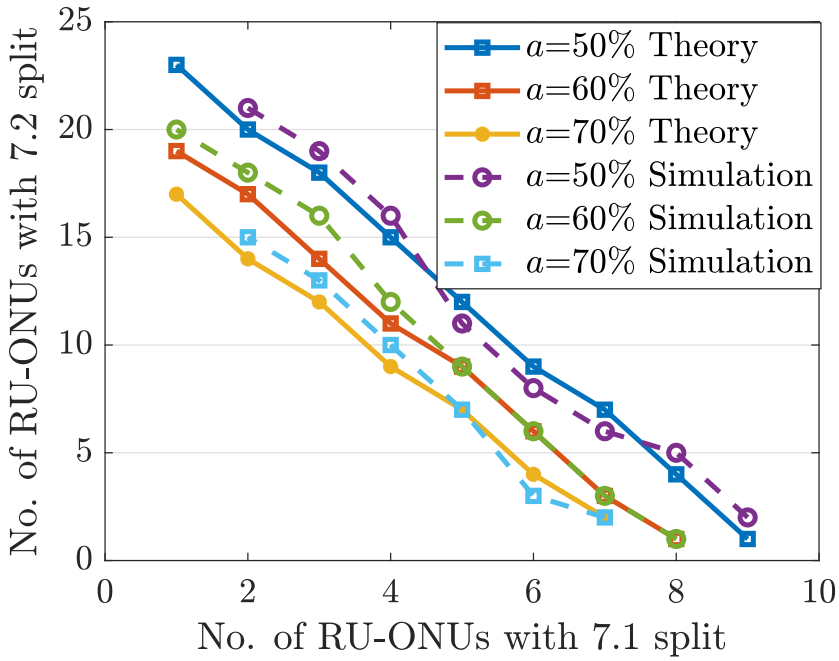

Figure 6.3: Feasible vPON slice config. region: a is the RU percentage load.

This is important because it means we can use the analytical model for calculating upstream latency to quickly find the optimal vPON slices through the proposed optimisation model

instead of going through the extensive simulation for all possible vPON slice configurations for the considered network layout. It should be noticed that the solution to our optimisation problem also returns the specific MEC node location and virtual PON configuration (in variable $X_{r,v}$), a snapshot of which is shown in Figure. 6.2 ($X_{r,v}$ corresponds to the EAST-WEST green colored links).

In Figure. 6.4 and Figure. 6.5, we report the algorithm performance as a function of load and algorithm iterations. Figure. 6.4 shows how a higher number of iterations can improve the solution, returning a configuration with smaller number of MEC nodes, as the solution is explored over a larger search space. We can also see that the computation time increases with the increase in traffic load. This is because at high traffic load, it is more difficult to find a solution that satisfies the latency constraint, therefore the algorithm spends more time in iteration to find the optimal values. Figure. 6.5 reports the exact computation time as a function of load and iterations.

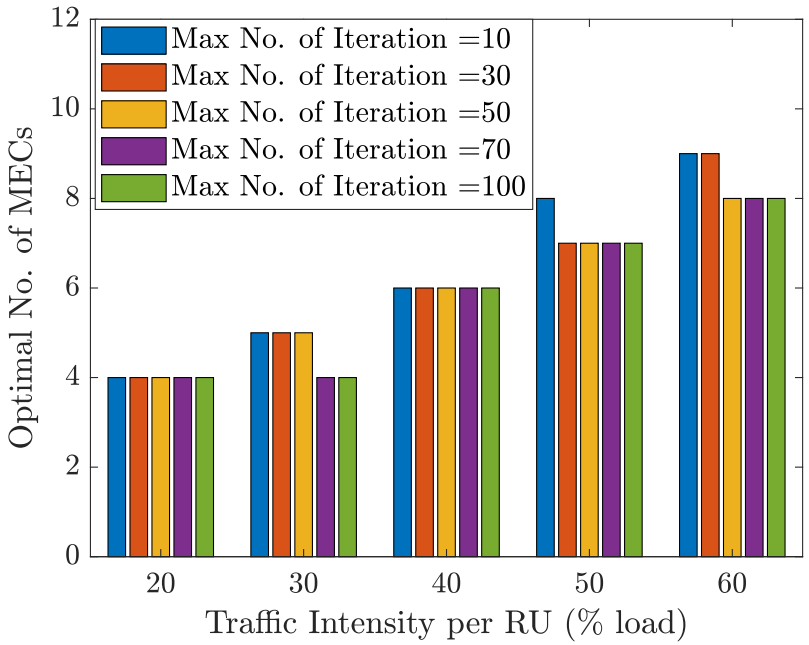

Figure 6.4: Performance of solution for different algorithm iterations.

From Figure. 6.4 and Figure. 6.5 we can conclude that our analytical-iterative model can quickly (i.e. within 10 iterations) find a solution suitable for real time optimisation (i.e., following burst increase in RU load), which is close to optimal (form the figures, we see just one more MEC node compared to the higher iteration ones). At the same time, even

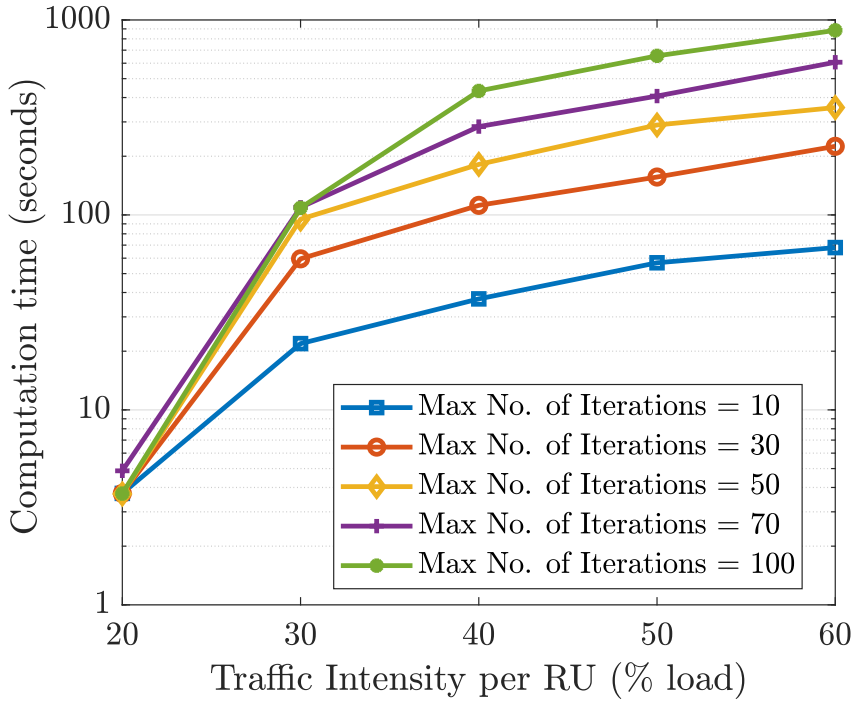

Figure 6.5: Algorithm computation time vs load and iterations.

the best solution can be calculated in times ranging from few seconds to few minutes. For comparison, we run the case of 30% load and 70 iteration in simulation without making use of the analytical model. While our model could return an optimal solution in about 100 seconds, the same result required 3 hours and 9 minutes without it (all computations were carried out on Intel i7-6600 mobile processor and simulations parallelised over 4 threads).

6.5 Conclusion

In this Chapter, we have proposed a mixed-analytical iterative optimization method that computes optimal virtual PON slice configuration in a mesh-PON type fronthaul network to support ultra-low latency under dynamic traffic scenarios. To achieve this, we first derived an analytical form for uplink latency in a virtual PON slice under varying traffic load, number of RUs in the vPON slice and RU-split configurations (7.1 or 7.2). With the help of discrete event simulation in OMNET++, we validated the analytical model. Using this analytical form of uplink latency, we then formed a non-linear discrete optimization framework to compute optimal virtual PON slices. We further proposed an iterative algorithm to solve the optimization model that can quickly find the optimal virtual PON slices with significantly

reduced computation time. Our results show that using the proposed mix-analytical iterative optimization method, optimal virtual PON slices can be computed in as quickly as seconds or tens of seconds (based on traffic load and max iterations). Thus making it suitable for real-time or near-real time network optimization.

7 Summary and Open Challenges

Summary and Open Challenges

The core objective of the thesis was to explore enhanced Passive Optical Network (PON) architectures for converged access networks for 5G and beyond. First, we assessed the motivation for using next-generation PON based fronthaul architectures for converged access. We then explored the potential challenges and bottlenecks associated with it. Finally, we addressed how these challenges and bottlenecks can be addressed with architectural enhancements of next-generation PON systems.

7.1 Summary

In Chapter 2, we provided a technological overview of the underlying concepts and reviewed the state-of-the-art to set the background of this dissertation. Subsequently, we identified the motivations for exploring next-generation PON based fronthaul architectures to support converged access. These include the potential to enable diversified services such as residential, mobile, IoT, etc over a common physically deployed transport access network, while supporting diversified Quality of Service (QoS) and network requirements such as ultra-high bandwidth and low latency. As a consequence, we identified three major technical implications first, an architectural enhancement of optical transport networks to support such diversified QoS over shared fronthaul transport. Second, a low-cost, highly scalable optical transport technology (e.g WDM or TWDM-PON) to support progressive cell densification, and third, a need of supporting various use cases and service scenarios with end-to-end network slicing. These implications then became the central theme of the thesis producing Research Question (RQ) 1-3.

In Chapter 3, we provided an overview of the software and hardware tools that we have used to evaluate the performance of the proposed schemes in the technical contributory works

of this dissertation. We covered a brief overview of modelling and discrete event simulation using MATLAB's discrete event toolbox Simevents and C++ oriented discrete event simulation framework OMNET++. We then discussed MATLAB's optimization toolbox for finding the optimal solution to a given optimization problem. Finally, we discussed the hardware testbed development framework that we used to evaluate the performance of the proposed schemes in the corresponding contributory works.

In Chapter 4, we have provided an answer to the first Research Question: "How to improve the statistical multiplexing of cells over Centralized Radio Access Network (C-RAN) to transport multiple fronthaul links effectively over a shared Time-Wavelength Division Multiplexing (TWDM)-PON fronthaul in order to meet the high bandwidth requirements of converged access in 5G and beyond?". We proposed a variable rate fronthaul scheme that leverages on the Software Defined Networking (SDN) control to enable statistical multiplexing over shared PON based fronthaul by dynamically adjusting the cell wireless bandwidth (and consequently the fronthaul transport rate) depending upon the cell load. Using theoretical analysis and discrete event simulation, we have demonstrated how the proposed variable rate fronthaul scheme can achieve a more efficient fronthaul transport of a group of Cloud-RAN cells without increasing the complexity, cost and energy consumption of the RUs. This work therefore provides an answer to the high transport bandwidth requirement in next-generation converged networks with high-density cell deployment, as statistical multiplexing of several cells can sensibly lower fronthaul costs while maintaining a certain target QoS.

In Chapter 5, we have provided an answer to the second Research Question: "How the ultra-low end-to-end latency requirements for 5G and beyond can be satisfied through PON architectural enhancements in transport network for MEC based Cloud-RAN? What architectural modifications are required to support them?" We proposed a virtualized EAST-WEST PON architecture supporting direct communication between PON endpoints to enable ultra-low latency fronthaul transport for Cloud-RAN cells with functional split processing. We have experimentally demonstrated the feasibility of the proposed EAST-WEST PON scheme and proved that the potential backscattering introduced by the wavelength loopback action towards the PON endpoints does not affect the overall system performance. Further,

using discrete event simulation with OMNET++, we have shown how our EAST-WEST PON architecture can maintain the system latency below a given threshold by dynamically offloading functional split computation across MEC nodes using dynamic virtual PON slicing technique. Therefore, the work in this chapter enables the convergence of mobile and MEC nodes, delivering deterministic low-latency performance under highly dynamic traffic scenarios.

The virtualized EAST-WEST PON architecture proposed in Chapter 5 can enable EAST-WEST communication in a physical PON along with the traditional NORTH-SOUTH communication, therefore enabling the support for the transport of mesh traffic pattern in converged MEC architecture. However, optimal formation of such virtualized PON slices to maintain low latency while maximizing the statistical multiplexing of Cloud-RAN cells is a challenging task. In Chapter 6, we answered the third Research Question that arose from work in Chapter 5: "How can we use transport network virtualization to support multiple Cloud-RAN, MEC-based services, such as to meet the high fronthaul transport bandwidth required in Cloud-RAN based converged access by exploiting the statistical multiplexing over a shared fronthaul while simultaneously meeting the target ultra-low end-to-end latency requirement?" For this, in Chapter 6, we have proposed a mixed-analytical iterative optimization method that computes optimal virtual PON slice configuration in mesh-PON type fronthaul network to support ultra-low latency under dynamic traffic scenarios. To achieve this, we first derived an analytical form for uplink latency in a virtual PON slice under varying traffic load, number of RUs in the vPON slice and RU-split configurations (7.1 or 7.2). With the help of discrete event simulation in OMNET++, we validated the analytical model. Using this analytical form of uplink latency, we then formed a nonlinear discrete optimization framework to compute optimal virtual PON slices. However, due to nonlinear latency constraint, the exhaustive search to find optimal solution would take up long times. We, therefore, proposed an iterative algorithm to solve the optimization model to quickly find the optimal virtual PON slices with significantly reduced computation time. Our results show that with the proposed mix-analytical iterative optimization method, optimal virtual PON slices can be computed in timescales suitable for real-time or near-real-time network optimization.

Therefore, in summary, our first contributory work answers the ultra-high bandwidth requirements of next-generation converged access in 5G and beyond by enabling statistical multiplexing over shared PON fronthaul. The second contributory work answers the ultra-low latency requirements in next-generation converged MEC based access by proposing a PON architectural enhancement based on PON virtualization and EAST-WEST communication (thus enabling mesh traffic over PON fronthaul). The third contributory work merges these two together by looking at how we can exploit such enhanced PON architecture and PON virtualization to provide ultra-low latency transport across the access network while simultaneously addressing the high transport bandwidth requirement by maximizing the statistical multiplexing of Cloud-RAN cells over a shared PON fronthaul.

7.2 Open Issues and Future Work

In this dissertation, we have looked at how we can address the ultra-high bandwidth and ultra-low latency requirement for converged access in 5G and beyond. In the following, we provide a number of open challenges, including possible extensions of the works from this dissertation and potential research ideas inspired by the topics addressed in this dissertation. Some of these future works are already underway by the author.

7.2.1 Mobility aware ultra-low Latency CoMP enabled by virtualized MESH PON

In Chapter 5, we have discussed how PON virtualization along with MESH PON architecture enables ultra-low end-to-end latency in converged MEC based access networks. A procedure for optimal virtual PON slicing mechanism depending on the traffic load, functional split and the network layout is presented in Chapter 6. All this work implicitly considers the user traffic to be static. In fact, the mobility of the user traffic has been assumed as the connection detachment (or traffic departure) from one cell and connection attachment (traffic arrival) to another cell. This procedure, therefore, enforces the handover when the mobility of the user traffic is considered. This potentially introduces a DU migration across MECs and consequently connection disruption. For the case of emergency transport, such as remote treatment while on the move (e.g in the ambulance), which requires uninterrupted

high-reliable service with low latency, the number of handovers and consequently the DU migrations should be reduced to which the current solution fails to address. However, by forming dynamic virtual PON slices, that consider the mobility pattern of the user traffic along with other parameters (e.g., network layout, traffic load per cell and functional split), this method can potentially reduce the number of DU migration and handovers to provide uninterrupted service with ultra-low latency.

7.2.2 Low Latency Service Migration using PON virtualization and EAST-WEST Communication

Migration of DU processing across MECs in Cloud-RAN is often required for load balancing, maintaining low-latency transport between RU and DU etc. In traditional PON based fronthaul with NORTH-SOUTH communication, this migration of DU processing happens via the Central Office (CO) which requires an extra Optical-Electrical-Optical (OEO) conversion process at CO, which incurs a significant latency for the service migration. Furthermore, transporting the DU processing instance (which is a snapshot of the virtual container, and generally refered to as the migration traffic) to the other MEC node over the same physical TWDM-PON fronthaul would require sharing of the uplink/downlink bandwidth with the fronthaul traffic [123] which causes delays in the actual fronthaul traffic.

Our proposed EAST-WEST PON architecture can be exploited for low-latency transport of such migration traffic while not interrupting any fronthaul traffic. This can be achieved by re-using the control channel on edge-OLT in the EAST-WEST PON architecture (described in Chapter 5) opportunistically when it is not used to communicate with CO for receiving virtual PON slice information. However, there are certain challenges associated with the realization of such scheme in realistic scenarios. First, how to re-use the control channel of the edge OLT by coordinating it with the CO is one of the main challenges, secondly the corresponding wavelength (whether fixed or dynamic) and capacity allocation during the service migration and how these can be facilitated is another important issue to look at.

Appendix

APPENDIX A

The additional simulations as discussed in Chapter 4 are based on the use of a Weibull distribution, which generalizes a large class of distributions (exponential, Rayleigh, chi-squared etc.) depending on the shape factor k. For example, for $k = 1$ we obtain an exponential distribution (Poisson arrival process = exponentially distributed inter-arrival times). Equation-(7.1) provides the details of an Weibull distribution with mean = $\gamma\Gamma(1 + 1/k)$ and variance = $\gamma^2 \left[\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2\right]$. Figure 7.1 plots the distribution under different shape factor (k).

$$f(x, \gamma, k) = \begin{cases} \frac{k}{\gamma} \left(\frac{x}{\gamma}\right)^{k-1} e^{-(x/\gamma)^k} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{7.1}$$

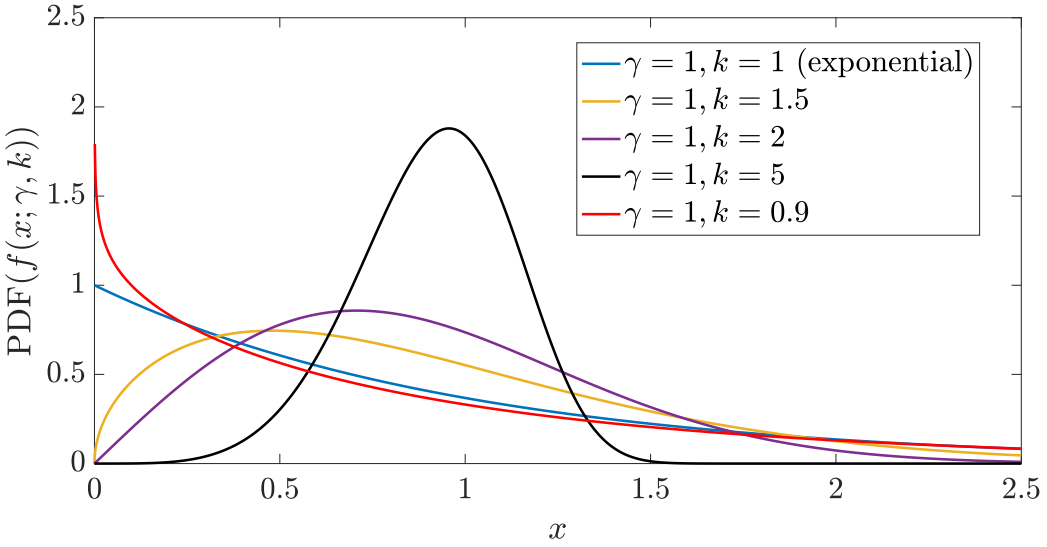

Figure 7.1: PDF of Weibull distribution with different shape factors

List of Acronyms

3GPP Third Generation partnership Project

4G Fourth Generation

5G Fifth Generation

5G-NR 5G New Radio

AxC Antenna Carrier

B2B back-to-back

BBU Baseband Unit

BCDR Burst mode Clock and Data Recovery

BER Bit Error Rate

BGP Border Gateway Protocol

BS Base Station

CapEx Capital Expenditure

CDR Clock and Data Recovery

CFS Customer Facing Service

Cloud-RAN Cloud Radio Access Networks

CO Central Office

CO-DBA Coordinated DBA

COMAC Converged Multi-Access and Core

CoMP Coordinated Multipoint

CORD Central Office Re-architected as Datacenter

CPRI Common Public Radio Interface

C-RAN Centralized Radio Access Network

CTI Cooperative Transport Interface

CU Central Unit

DBA Dynamic Bandwidth Allocation

DC Data Center

DU Distributed Unit

D-RAN Distributed RAN

DWDM Dense Wavelength Division Multiplexing

E2E End-to-End

eCPRI evolved Common Public Radio Interface

EDFA Erbium-Doped Fibre Amplifier

EFM Ethernet in the First Mile

EPON Ethernet Passive Optical Network

ETSI European Telecommunications Standards Institute

FBG Fibre Bragg Grating

FFT Fast Fourier Transform

FTTH Fiber to the Home

GC Grant Cycle

GIANT GigaPON Access Network

GPON Gigabit Passive Optical Network

gRB grouped Resource Block

HLS High Layer Split

ICIC Inter Cell Interference Coordination

IEEE Institute of Electrical and Electronics Engineers

IoT Internet of Things

IP Internet Protocol

IPACT Interleaved Polling with Adaptive Cycle Time

ISG Industry Specification Group

ITS Intelligent Transport Systems

ITU International Telecommunication Union

LCM Lifecycle Management

LLS Low Layer Split

LTE Long Term Evolution

LXC Linux Container

M-CORD Mobile CORD

MEC Multi Access Edge Computing

MFH Mobile Fronthaul

MIMO Multiple Input Multiple Output

MMC Mobile Micro Cloud

NFV Network Function Virtualization

NGFI Next Generation Fronthaul Interface

NG-PON2 Next Generation Passive Optical Networks 2

Ng-RAN Next Generation RAN

NSR-DBA Non-Status Report based DBA

ODN Optical Distribution Network

OLT Optical Line Terminal

ONF Open Networking Foundation

ONOS Open Network Operating System

ONU Optical Networking Unit

OpEX Operational Expenditure

OSS Operation Support System

OTDN Open and Disaggregated Transport Network

OTN Optical Transport Network

P2MP Point to Multi Point

PDCP Packet Data Convergence Protocol

PLOAM Physical Layer Operation and Maintenance

PON Passive Optical Network

PRACH Physical Random Access Channel

PRB Physical Resource Block

QoE Quality of Experience

QoS Quality of Service

RAN Radio Access Networks

R-CORD Residential CORD

RG Resource Group

RQ Research Question

RRC Radio Resource Control

RRU Remote Radio Unit

RU Radio Unit

RV Random Variable

SCC Small Cell Cloud

SDN Software Defined Networking

SDR Software Defined Radio

SLA Service Level Aggrement

SNMP Simple Network Management Protocol

SR Status Report

SR-DBA Status Report based DBA

TC Transmission Convergence

T-CONT Transmission Containers

TDMA Time Division Multiple Access

TM-DBA Traffic Monitoring based DBA

ToR Top-of-Rack

TTI Transmit Time Interval

TWDM Time-Wavelength Division Multiplexing

UE User Equipment

URLLC Ultra Reliable Low-Latency Communication

VM Virtual Machine

VNF Virtual Network Function

VOLTHA Virtual OLT Hardware Abstraction

VON Virtual Optical Network

vPON virtual PON

VRF Variable Rate Fronthaul

WDM Wavelength Division Multiplexing

WLB Wavelength Loop Back

WPF Wavelength Pass Filter

XGEM XG-PON Encapsulation Method

XG-PON 10-Gigabit-capable Passive Optical Network

XGTC XG-PON Transmission Convergence

Bibliography

[1] N. Alliance, "NGMN Overview on 5G RAN Functional Decomposition," Feb. 2018.

[2] "Study on CU-DU lower layer split for NR; (Release 15)," 3GPP, Standard TR 38.816, V15.0.0, Jan 2018.

[3] "IEEE Standard for Packet-based Fronthaul Transport Networks," IEEE, New York, USA, Standard 1914.1-2019, 2020, oCLC: 8585976188.

[4] ITU, "10-Gigabit-capable passive optical networks (XG-PON): Transmission convergence (TC) layer specification," Recommendation ITU-T G.987.3, Jan 2014.

[5] ETSI, "Mobile-Edge Computing (MEC); Framework and Reference Architecture," Group Speciication, March 2016.

[6] "5G wireless fronthaul requirements in a passive optical network context," ITU-T, ITU Standard: Series G- Supplement 66, July 2019.

[7] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016–2021," 2016. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-whitepaper-c11-481360.html

[8] P. Loskot, M. A. M. Hassanien, F. Farjady, M. Ruffini, and D. Payne, "Long-term drivers of broadband traffic in next-generation networks," *annals of telecommunications - annales des télécommunications*, vol. 70, no. 1, pp. 1–10, Feb 2015. [Online]. Available: https://doi.org/10.1007/s12243-014-0424-9

[9] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks -A Technology Overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[10] NEC Corp. and Nortel Networks SA and Siemens Networks GmbH & Co. KG and Ericsson AB and Huawei Technologies Co Ltd., "V6.1 Common Public Radio Interface (CPRI); Interface Specification," *in Specification CPRI*, July 2014.

[11] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G Backhaul Challenges and Emerging Research Directions: A Survey," *IEEE Access*, vol. 4, pp. 1743–1766, April 2016.

[12] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 152–159, February 2016.

[13] J. Li and J. Chen, "Passive Optical Network Based Mobile Backhaul Enabling Ultra-Low Latency for Communications Among Base Stations," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 10, pp. 855–863, Oct 2017.

[14] B. Mukherjee, I. Tomkos, M. Tornatore, P. Winzer, and Y. Zhao, Eds., *Springer Handbook of Optical Networks: Chapter-29: PON Architecture Enhancements (by Thomas Pfeiffer*, ser. Springer Handbooks. Cham: Springer International Publishing, 2020. [Online]. Available: http://link.springer.com/10.1007/978-3-030-16250-4

[15] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. WÃŒbben, "Cloud technologies for flexible 5G radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, May 2014.

[16] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible centralization through Cloud-RAN," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov 2014.

[17] China Mobile Research Institute, "C-RAN: The Road Towards Green RAN," *White paper, Version 2.5*, Oct 2011.

[18] K. F. Nieman and B. L. Evans, "Time-domain compression of complex-baseband lte signals for cloud radio access networks," in *2013 IEEE Global Conference on Signal and Information Processing*, 2013, pp. 1198–1201.

[19] B. Guo, W. Cao, A. Tao, and D. Samardzija, "LTE/LTE-A signal compression on the CPRI interface," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 117–133, 2013.

[20] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed Transport of Baseband Signals in Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3216–3225, 2012.

[21] "Study on new radio access technology: Radio access architecture and interfaces (Release 14)," Standard TR 38.801, V14.0.0, Mar 2017.

[22] "NG-RAN; F1 general aspects and principles (Release 16)," 3GPP, Standard TS 38.470, V16.1.0, Mar 2020.

[23] "O-RAN Architecture Description," O-RAN alliance, standard TS, v01.00.00, Feb 2020.

[24] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen, "Evaluating C-RAN fronthaul functional splits in terms of network level energy and cost savings," *Journal of Communications and Networks*, vol. 18, no. 2, pp. 162–172, April 2016.

[25] "IEEE Standard for Information Technology– Local And Metropolitan Area Networks-Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks," IEEE, New York, USA, Standard Std 802.3ah-2004, Sep 2004.

[26] "IEEE Standard for Ethernet – Amendment 1: Physical Layer Specifications and Management Parameters for 10Gb/s Passive Optical Networks," IEEE, Standard Std 802.3av-2009, Oct 2009.

[27] "IEEE Standard for Ethernet - Amendment 9," IEEE, Standard Std 802.3ca-2020, July 2020.

[28] ITU, "Gigabit-capable passive optical networks (XG-PON): Transmission convergence (TC) layer specification," Recommendation ITU-T G.984.3, Feb 2004.

[29] ——, "40-Gigabit-capable passive optical networks (NG-PON2): Transmission convergence layer specification: Ammendment-3," Recommendation ITU-T G989.3 Amendment 3, Mar 2020.

[30] G. Kramer, B. Mukherjee, and G. Pesavento, "Ipact a dynamic protocol for an ethernet pon (epon)," *IEEE Communications Magazine*, vol. 40, no. 2, pp. 74–80, 2002.

[31] H.-C. Leligou, C. Linardakis, K. Kanonakis, J. D. Angelopoulos, and T. Orphanoudakis, "Efficient medium arbitration of FSAN-compliant GPONs," *international journal of communication systems*, vol. 19, no. 5, pp. 603–617, 2006.

[32] Z. Qi-yu, L. Bin, and W. Run-ze, "A Dynamic Bandwidth Allocation Scheme for GPON Based on Traffic Prediction," in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, May 2012, pp. 2043–2046.

[33] P. Sarigiannidis, D. Pliatsios, T. Zygiridis, and N. Kantartzis, "Dama: A data mining forecasting dba scheme for xg-pons," in *2016 5th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, May 2016, pp. 1–4.

[34] Y. A. Mahmud, N. A. M. Radzi, F. Abdullah, and N. M. Din, "Fuzzy-logic based NSR DBA for upstream GPON," in *2015 IEEE 12th Malaysia International Conference on Communications (MICC)*, Nov 2015, pp. 169–174.

[35] A. G. Sarigiannidis, M. Iloridou, P. Nicopolitidis, G. Papadimitriou, F. Pavlidou, P. G. Sarigiannidis, M. D. Louta, and V. Vitsas, "Architectures and Bandwidth Allocation Schemes for Hybrid Wireless-Optical Networks," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 427–468, 2015.

[36] T. Tashiro, S. Kuwano, J. Terada, T. Kawamura, N. Tanaka, S. Shigematsu, and N. Yoshimoto, "A novel DBA scheme for TDM-PON based mobile fronthaul," in *Optical Fiber Conference and Exhibition (OFC) 2014*.

[37] Nomura, Hiroko and Ou, Hiroshi and Shimada, Tatsuya and Kobayashi, Takayuki and Hisano, Daisuke and Uzawa, Hiroyuki and Terada, Jun and Otaka, Akihiro, "First demonstration of optical-mobile cooperation interface for mobile fronthaul with tdm-pon," *IEICE Communications Express*, vol. 6, no. 6, pp. 375–380, 2017.

[38] H. Uzawa, H. Nomura, T. Shimada, D. Hisano, K. Miyamoto, Y. Nakayama, K. Takahashi, J. Terada, and A. Otaka, "Practical Mobile-DBA Scheme Considering Data Arrival Period for 5G Mobile Fronthaul with TDM-PON," in *2017 European Conference on Optical Communication (ECOC)*, 2017, pp. 1–3.

[39] D. Hisano and Y. Nakayama, "Two-stage optimization of uplink forwarding order with cooperative DBA to accommodate a TDM-PON-based fronthaul link," *Journal of Optical Communications and Networking*, vol. 12, no. 5, May 2020.

[40] M. Chiang and T. Zhang, "Fog and iot: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, 2016.

[41] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[42] ETSI, "Mobile-Edge Computing (MEC); Service Scenarios," Group Speciication ETSI GS MEC-IEG 004, Nov 2015.

[43] N. Takahashi, H. Tanaka, and R. Kawamura, "Analysis of Process Assignment in Multi-tier mobile Cloud Computing and Application to Edge Accelerated Web Browsing," in *2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, 2015, pp. 233–234.

[44] Yuan Zhang, Hao Liu, Lei Jiao, and Xiaoming Fu, "To offload or not to offload: An efficient code partition algorithm for mobile cloud computing," in *2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET)*, 2012, pp. 80–86.

[45] J. Dolezal, Z. Becvar, and T. Zeman, "Performance evaluation of computation offloading from mobile device to the edge of mobile network," in *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2016, pp. 1–7.

[46] X. Sun and N. Ansari, "EdgeIoT: Mobile Edge Computing for the Internet of Things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, 2016.

[47] G. Yan and Q. Qin, "The application of edge computing technology in the collaborative optimization of intelligent transportation system based on information physical fusion," *IEEE Access*, vol. 8, pp. 153 264–153 272, 2020.

[48] A. Martin, R. Viola, M. Zorrilla, J. Flórez, P. Angueira, and J. Montalbán, "Mec for fair, reliable and efficient media streaming in mobile networks," *IEEE Transactions on Broadcasting*, vol. 66, no. 2, pp. 264–278, 2020.

[49] A. Ndikumana, N. H. Tran, T. M. Ho, Z. Han, W. Saad, D. Niyato, and C. S. Hong, "Joint communication, computation, caching, and control in big data multi-access edge computing," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1359–1374, 2020.

[50] H. E. Project, "Small cEllS coordinAtion for Multi-tenancy and edge services (SESAM)," 2015. [Online]. Available: http://www.sesame-h2020-5g-ppp.eu/

[51] S. Wang, G.-H. Tu, R. Ganti, T. He, K. Leung, H. Tripp, K. Warr, and M. Zafer, "Mobile micro-cloud: Application classification, mapping, and deployment," in *Proceedings of Annual Fall Meeting of ITA (AMITA)*, 2013.

[52] K. Wang, M. Shen, J. Cho, A. Banerjee, J. Van der Merwe, and K. Webb, "Mobiscud: A fast moving personal cloud in the mobile network," in *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, 2015, pp. 19–24.

[53] T. Taleb and A. Ksentini, "Follow me cloud: interworking federated clouds and distributed mobile networks," *IEEE Network*, vol. 27, no. 5, pp. 12–19, 2013.

[54] J. Liu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "Concert: a cloud-based architecture for next-generation cellular systems," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 14–22, 2014.

[55] ETSI, "Mobile-Edge Computing (MEC); Service Scenarios," Group Speciication, March 2016.

[56] D. Levin, A. Wundsam, B. Heller, N. Handigol, and A. Feldmann, "Logically Centralized? State Distribution Trade-Offs in Software Defined Networks," in *Proceedings of the First Workshop on Hot Topics in Software Defined Networks*, ser. HotSDN '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 1–6. [Online]. Available: https://doi.org/10.1145/2342441.2342443

[57] T. D. Nadeau and K. Gray, *SDN: Software Defined Networks: an authoritative review of network programmability technologies*. O'Reilly Media, Inc., 2013.

[58] T. Ulversoy, "Software defined radio: Challenges and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 4, pp. 531–550, 2010.

[59] A. S. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein, and W. Kellerer, "Software defined optical networks (sdons): A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2738–2786, 2016.

[60] D. F. Macedo, D. Guedes, L. F. M. Vieira, M. A. M. Vieira, and M. Nogueira, "Programmable networks—from software-defined radio to software-defined networking," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 1102–1125, 2015.

[61] H.-H. Cho, C.-F. Lai, T. K. Shih, and H.-C. Chao, "Integration of sdr and sdn for 5g," *IEEE Access*, vol. 2, pp. 1196–1204, 2014.

[62] A. Lara, A. Kolasani, and B. Ramamurthy, "Network innovation using openflow: A survey," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 493–512, 2014.

[63] E. R. Enns, M. Bjorklund, J. Schoenwaelder, and A. Bierman, "Network configuration protocol (netconf)," June 2011. [Online]. Available: http://www.rfc-editor.org/rfc/rfc6241.txt

[64] F. Hu, Q. Hao, and K. Bao, "A survey on software-defined network and openflow: From concept to implementation," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 2181–2206, 2014.

[65] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[66] R. Munoz, R. Vilalta, R. Casellas, R. Martinez, T. Szyrkowiec, A. Autenrieth, V. Lopez, and D. Lopez, "Integrated sdn/nfv management and orchestration architecture for dynamic deployment of virtual sdn control instances for virtual tenant networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 11, pp. B62–B70, 2015.

[67] R. Vilalta, R. Muñoz, A. Mayoral, R. Casellas, R. Martínez, V. López, and D. López, "Transport network function virtualization," *Journal of Lightwave Technology*, vol. 33, no. 8, pp. 1557–1564, 2015.

[68] "SD-RAN, Open Networking Foundation," (accessed on 24/04/2021). [Online]. Available: https://opennetworking.org/sd-ran/

[69] L. Peterson, A. Al-Shabibi, T. Anshutz, S. Baker, A. Bavier, S. Das, J. Hart, G. Palukar, and W. Snow, "Central office re-architected as a data center," *IEEE Communications Magazine*, vol. 54, no. 10, pp. 96–101, 2016.

[70] "XOS- Open Networking Foundation," 2021, (accessed on 13/04/2021). [Online]. Available: https://opennetworking.org/xos/

[71] "Open Source Cloud Computing Infrastructure- OpenStack," 2021, (accessed on 13/04/2021). [Online]. Available: https://www.openstack.org/

[72] P. Berde, M. Gerola, J. Hart, Y. Higuchi, M. Kobayashi, T. Koide, B. Lantz, B. O'Connor, P. Radoslavov, W. Snow, and G. Parulkar, "Onos: Towards an open, distributed sdn os," in *Proceedings of the Third Workshop on Hot Topics in Software Defined Networking*, ser. HotSDN '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1–6. [Online]. Available: https://doi.org/10.1145/2620728.2620744

[73] "CORD Platform | Open Networking Foundation (ONF)," 2021, (accessed on 13/04/2021). [Online]. Available: https://opennetworking.org/cord/

[74] A. Campanella, H. Okui, A. Mayoral, D. Kashiwa, O. G. de Dios, D. Verchere, Q. Pham Van, A. Giorgetti, R. Casellas, R. Morro, and L. Ong, "Odtn: Open disaggregated transport network. discovery and control of a disaggregated optical network through open source software and open apis," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*, 2019, pp. 1–3.

[75] N. Nadarajah, E. Wong, and A. Nirmalathas, "Implementation of multiple secure virtual private networks over passive optical networks using electronic cdma," *IEEE Photonics Technology Letters*, vol. 18, no. 3, pp. 484–486, 2006.

[76] X. Wang, L. Wang, C. Cavdar, M. Tornatore, G. B. Figueiredo, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Handover Reduction in Virtualized Cloud Radio Access Networks Using TWDM-PON Fronthaul," *Journal of Optical Communications and Networking*, vol. 8, no. 12, p. B124, Dec. 2016. [Online]. Available: https://www.osapublishing.org/abstract.cfm?URI=jocn-8-12-B124

[77] R. I. Tinini, D. M. Batista, and G. B. Figueiredo, "Energy-Efficient VPON Formation and Wavelength Dimensioning in Cloud-Fog RAN over TWDM-PON," in *2018 IEEE Symposium on Computers and Communications (ISCC)*. Natal: IEEE, Jun. 2018, pp. 521–526. [Online]. Available: https://ieeexplore.ieee.org/document/8538610/

[78] X. Wang, 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," in *2016 IEEE Global Communications Conference (GLOBECOM)*. Washington, DC, USA: IEEE, Dec. 2016, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7841923/

[79] B. Skubic, M. Fiorani, S. Tombaz, A. Furuskär, J. Mårtensson, and P. Monti, "Optical Transport Solutions for 5G Fixed Wireless Access [Invited]," *Journal of Optical Communications and Networking*, vol. 9, no. 9, p. D10, Sep. 2017. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=jocn-9-9-D10

[80] A. D. Giglio and A. Pagano, "Scenarios and Economic Analysis of Fronthaul in 5G Optical Networks," *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 585–591, Jan. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8526294/

[81] "SimEvents-MATLAB and Simulink: Model and simulate message communication and discrete-event systems," 2021, (accessed on 24/04/2021). [Online]. Available: https://www.mathworks.com/products/simevents.html

[82] "OMNeT++ Discrete Event Simulator," 2021, (accessed on 24/04/2021). [Online]. Available: https://omnetpp.org/

[83] "OMNEST: High Performance Simulation for All Kind of Networks," 2021, (accessed on 24/04/2021). [Online]. Available: https://omnest.com/

[84] "Optimization Toolbox - MATLAB: Solve linear, quadratic, conic, integer, and nonlinear optimization problems," 2021, (accessed on 24/04/2021). [Online]. Available: https://www.mathworks.com/products/optimization.html

[85] "Global Optimization Toolbox - MATLAB: Solve multiple maxima, multiple minima, and nonsmooth optimization problems," 2021, (accessed on 24/04/2021). [Online]. Available: https://www.mathworks.com/products/global-optimization.html

[86] B. Guo, W. Cao, A. Tao, and D. Samardzija, "LTE/LTE-A signal compression on the CPRI interface," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 117–133, Sept 2013.

[87] U. Dötsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, 2013.

[88] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "On the statistical multiplexing gain of virtual base station pools," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 2283–2288.

[89] L. Wang and S. Zhou, "On the fronthaul statistical multiplexing gain," *IEEE Communications Letters*, vol. 21, no. 5, pp. 1099–1102, May 2017.

[90] S. Bidkar, J. Galaro, and T. Pfeiffer, "First Demonstration of an Ultra-Low-Latency Fronthaul Transport Over a Commercial TDM-PON Platform," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[91] R. Nadiv and T. Naveh, "Wireless backhaul topologies: Analyzing backhaul topology strategies," *Ceragon White Paper*, pp. 1–15, 2010.

[92] M. Ruffini, "Multidimensional Convergence in Future 5G Networks," *Journal of Lightwave Technology*, vol. 35, no. 3, pp. 535–549, Feb 2017.

[93] P. Alvarez, F. Slyne, C. Bluemm, J. M. Marquez-Barja, L. A. DaSilva, and M. Ruffini, "Experimental Demonstration of SDN-controlled Variable-rate Fronthaul for Converged LTE-over-PON," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[94] F. Slyne *et al.*, "Coordinated fibre and wireless spectrum allocation in SDN-controlled wireless-optical-cloud converged architecture," in *45th European Conference on Optical Communication (ECOC)*, Sept 2019, pp. 1–3.

[95] M. Ettus, "USRP User's and Developer's Guide," *Ettus Research LLC*, 2005. [Online]. Available: https://www.ettus.com/

[96] P. Alvarez, F. Slyne, C. Bluemm, J. M. Marquez-Barja, L. A. DaSilva, and M. Ruffini, "Experimental Demonstration of SDN-controlled Variable-rate Fronthaul for Converged LTE-over-PON," in *Optical Fiber Communication Conference*. Optical Society of America, March 2018, p. Th2A.49.

[97] L. Golubchik and J. C. S. Lui, "Bounding of performance measures for a threshold-based queueing system with hysteresis," *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, vol. 25, no. 1, pp. 147–157, Jun. 1997. [Online]. Available: http://doi.acm.org/10.1145/258623.258684

[98] P. J. Courtois and P. Semal, "Computable bounds for conditional steady-state probabilities in large markov chains and queueing models," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 926–937, Sep 1986.

[99] P.-J. Courtois, *Decomposability:Queuing and Computer System Applications*. ACM monograph series, Jan 1997.

[100] C. D. Meyer, "Stochastic Complementation, Uncoupling Markov Chains, and the Theory of nearly Reducible Systems," *SIAM Review*, vol. 31, no. 2, pp. 240–272, 1989.

[101] W. Grassmann, "Transient Solutions in Markovian Queueing Systems." *Computers and Operations Research*, vol. 4, pp. 47 – 53, 1977.

[102] J. Kaufman, "Blocking in a Shared Resource Environment," *IEEE Transactions on Communications*, vol. 29, no. 10, pp. 1474–1481, Oct. 1981. [Online]. Available: http://ieeexplore.ieee.org/document/1094894/

[103] H. Akimaru, *Teletraffic Theory and Applications*. Springer London Ltd, 2012, oCLC: 933620017.

[104] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5g: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[105] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3098–3130, Fourthquarter 2018.

[106] "Common Public Radio Interface: eCPRI Interface Specification," Tech. Rep., Oct. 2019.

[107] ETSI Industry Specification Group (ISG), "Multi-access Edge Computing (MEC); (ETSI GS MEC 001 - 029)," https://www.etsi.org/deliver/etsi_gs/MEC/001_099/, Group Specification, oct 2019.

[108] D. Nesset, "PON Roadmap [Invited]," *Journal of Optical Communication and Networking (JOCN)*, vol. 9, no. 1, pp. A71–A76, Jan 2017. [Online]. Available: http://jocn.osa.org/abstract.cfm?URI=jocn-9-1-A71

[109] "40-gigabit-capable passive optical networks: TC layer specification amd. 1," ITU-T, Standard, Nov. 2016.

[110] C. Ranaweera, E. Wong, C. Lim, and A. Nirmalathas, "Next generation optical-wireless converged network architectures," *IEEE Network*, vol. 26, no. 2, pp. 22–27, March 2012.

[111] R. I. Tinini, D. M. Batista, G. B. Figueiredo, M. Tornatore, and B. Mukherjee, "Low-latency and energy-efficient BBU placement and VPON formation in virtualized cloud-fog RAN," *IEEE/OSA Journal of Optical Communications and Networking*, April 2019.

[112] MSA specification for 25GS-PON, "25 Gigabit Symmetric Passive Optical Network, Version 2.0," Aug 2021. [Online]. Available: https://www.25gspon-msa.org/wp-content/uploads/2021/09/25GS-PON-Specification-V2.0.pdf.

[113] M. Hajduczenia and S. Pato, "Channel insertion loss for 1x64 and 1x128 split EPONs," IEEE802.3 Plenary Meeting, Dallas, TX, November 14-16 2006. [Online]. Available: https://www.ieee802.org/3/av/public/2006_11/3av_0611_hajduczenia_1.pdf

[114] M. Ruffini *et al.*, "Access and metro network convergence for flexible end-to-end network design," *IEEE/OSA Journal of Optical Communications and Networking*, June 2017.

[115] "IEEE p802.3cs increased-reach ethernet optical subscriber access (super-pon) task force," Nov 2018, http://www.ieee802.org/3/minutes/nov18/1118_spon_close_report.pdf.

[116] S. Das and M. Ruffini, "A Variable Rate Fronthaul Scheme for Cloud Radio Access Networks," *Journal of Lightwave Technology*, July 2019.

[117] D. Nesset, "Ng-pon2 technology and standards," *Journal of Lightwave Technology*, vol. 33, no. 5, pp. 1136–1143, 2015.

[118] "NR; Base Station (BS) radio transmission and reception (Release 16)," 3GPP, standard TS 38.104, V16.3.0, Mar. 2020.

[119] S. Das and M. Ruffini, "PON Virtualisation with EAST-WEST Communications for Low-Latency Converged Multi-Access Edge Computing (MEC)," *Optical Fiber Conference and Exhibition (OFC)*, March 2020.

[120] S. Das, F. Slyne, A. Kaszubowska, and M. Ruffini, "Virtualized east–west pon architecture supporting low-latency communication for mobile functional split based on multiaccess edge computing," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 12, no. 10, pp. D109–D119, 2020.

[121] "Aether: Enabling a New Era of Smart Enterprises," ONF, whitepaper, Dec 2020.

[122] V. B. Iversen, *Teletraffic engineering and Network planning.pdf*, jan 2007.

[123] J. Li, X. Shen, L. Chen, J. Ou, L. Wosinska, and J. Chen, "Delay-Aware Bandwidth Slicing for Service Migration in Mobile Backhaul Networks," *Journal of Optical Communications and Networking*, vol. 11, no. 4, p. B1, Apr. 2019. [Online]. Available: https://www.osapublishing.org/abstract.cfm?URI=jocn-11-4-B1