Resource Allocation Enhancements for 5G New Radio Architecture

Ogechi Akudo Nwogu

Thesis

Submitted in fulfillment of the requirements for the degree of Doctor of Philosophy of

University Sorbonne Paris Nord

Specialization : "Computer Engineering"

presented and defended by

Ogechi Akudo NWOGU

15 December 2021

Resource Allocation Enhancements for 5G New Radio Architecture

Supervisor : Mrs. Gladys Diaz

Co-Supervisor : Mr. Marwen Abdennebi

JURY

Rami Langar	Professor, Université Gustave Eiffel	Reviewer
Yassine Hadjadj-Aoul	HDR, Université de Rennes 1 INRIA	Reviewer
Zoubir Mammeri	Professor, Université Paul Sabatier	Examiner
Gladys Diaz	HDR, Université Sorbonne Paris Nord	Supervisor, USPN
Marwen Abdennebi	MdC, Université Sorbonne Paris Nord	Co-supervisor, USPN
Elizabeth N. Onwuka	Professor, Fed. Uni. of Tech. Minna, Nigeria	Examiner
Kinda Khawam	HDR, Université de Versailles	Examiner
Nawel Zangar	MdC, Université Marne la Valle	Examiner
Christopher Cérin	Professor, Université Sorbonne Paris Nord	Examiner

THÈSE

Pour obtenir le grade de

Docteur Université Sorbonne Paris Nord

Discipline : "Ingénierie informatique"

présenté et soutenue publiquement par

Ogechi Akudo NWOGU

le 15 Décembre 2021

Améliorations de l'allocation des Ressources pour la nouvelle Architecture Radio 5G

Directeur de thèse : Mrs. Gladys Diaz

Co-Directeur de thèse: Mr. Marwen Abdennebi

JURY

Rami Langar	Professeur, Université Gustave Eiffel	Rapporteur
Yassine Hadjadj-Aoul	HDR, Université de Rennes 1 INRIA	Rapporteur
Zoubir Mammeri	Professeur, Université Paul Sabatier	Examinateur
Gladys Diaz	HDR, Université Sorbonne Paris Nord	Directeur des thèse
Marwen Abdennebi	MdC, Université Sorbonne Paris Nord	Co-Directeur de thèse
Elizabeth N. Onwuka	Professeur, Fed. Uni. de Tech. Minna, Nigéria	Examinateur
Kinda Khawam	HDR, Université de Versailles	Examinateur
Nawel Zangar	MdC, Université Marne la Valle	Examinateur
Christopher Cérin	Professeur, Université Sorbonne Paris Nord	Examinateur

Declaration of Authorship

I, Ogechi Akudo Nwogu, declare that this thesis titled, “Resource Allocation Enhancements for 5G New Radio Architecture” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

Abstract

With the fast integration of the Fifth Generation (5G) networks into the heterogeneous mobile networking Eco-system together with its wide range of service specifications along with the present advance into the next Sixth Generation (6G) standard would require high levels of dynamic system control to handle the new challenges associated with increased data volumes and latency stringent service requirements.This can be achieved through the enhancement of standard networking techniques, distributed cell architectures to improve cell coverage and through the intelligent integration of various promising mechanisms such as SDN (Software Defined Network), NFV (Network Function Virtualization), massive MIMO (Multiple-Input Multiple-Output) systems adoption, Time Sensitive Networking (TSN) technique application and a host of other schemes.

To this end, in this thesis we proffer low-latency scheduling and resource usage enhancements to various segments of the standard LTE/LTE-A network majorly to encourage URLLC services in the application of 5G/6G broad use-cases such as network slicing and RAN functions cloudification which require real-time guarantees to execute.

Our main contributions are threefold, to begin we proposed a dynamic/fixed resource reservation and usage mechanism in a bid to control congestion in heterogeneous cellular networks as a result of densely populated mixed traffic scenarios associated with the 5G networks. We also propose a new random access (RA) resource usage algorithm to efficiently schedule different priority class traffic.

In our second contribution we focus on implementing a dynamic load balancing scheme among heterogeneous cell deployments within a 5G context. Our contribution lends to the effective usage of mobile network heterogeneous cells, URLLC scheduling and power savings. Additionally, we proposed a game-based C-RAN computational resource usage scheme.

Our final contribution addressed the problem of time-sensitive networking along 5G fronthaul networks.We make use of the IEEE Time Sensitive Networking (TSN) Burst Limiting Scheduler (BLS) algorithm by implementing our contribution to this algorithm through a Dynamic Reserved capacity (DRC) scheduling scheme. With our strategy, higher traffic output rates were recorded through simulation results compared to the WRR strict priority scheduler for both high and low priority traffic classes in different traffic load conditions.

Résumé

L'intégration rapide des réseaux de cinquième génération (5G) dans l'écosystème des réseaux mobiles hétérogènes, avec son large éventail de services définis, ainsi que l'avancée actuelle vers un standard de sixième génération (6G), exigent un contrôle dynamique performant des systèmes afin de relever les nouveaux défis liés à l'augmentation des volumes de données et aux exigences en terme de latence pour les services temps-réel. Cet objectif ne peut être atteint qu'en améliorant les techniques usuelles de mise en réseau, les architectures de déploiement distribuées qui optimisent la couverture cellulaire et par l'intégration intelligente de divers mécanismes innovants, tels que SDN (Software Defined Network), NFV (Network function virtualization). D'autres améliorations sont déjà réalisées via le déploiement de systèmes MIMO (Multilple-Input Multiple-Output) massifs, l'application de la technique TSN (Time Sensitive Networking), en plus d'une multitude d'autres mécanismes.

Dans cette optique, nous proposons dans cette thèse des améliorations pour un ordonnancement à faible latence et une meilleure utilisation des ressources sur divers segments du réseau 5G/6G, l'objectif étant principalement d'assurer un meilleur support des services de type URLLC dans divers contextes d'application tels que le découpage du réseau et la cloudification des fonctions RAN qui nécessitent des garanties en temps réel pour pouvoir être exécutés.

Nos principales contributions sont au nombre de trois.Pour commencer, nous avons proposé un mécanisme dynamique/fixe de réservation et d'utilisation des ressources dans le but de contrôler la congestion dans les réseaux cellulaires hétérogènes en raison des divers scénarios de trafic mixte à forte densité d'usagers associés aux réseaux 5G. Nous proposons également un nouvel algorithme d'utilisation des ressources à accès aléatoire (RA) pour une allocation de ressources efficace pour les trafics de différentes classes de priorité.

Dans notre deuxième contribution, nous nous concentrons sur la mise en œuvre d'un schéma d'équilibrage dynamique de la charge dans le cadre d'un déploiement radio cellulaire hétérogène dans un contexte 5G. Notre contribution permet d'utiliser efficacement les cellules hétérogènes des réseaux mobiles, de planifier les URLLC et de réduire l'énergie consommée. En outre, le schéma d'utilisation des ressources de calcul C-RAN basé sur la méthode de la théorie des jeux que nous avons proposée optimise les performances de l'algorithme d'allocation.

Enfin, notre dernière contribution porte sur le problème des délais d'accès dans

les réseaux fronthaul de la 5G. Nous proposons de mettre en œuvre une part dynamique e de la capacité réservée (Dynamic Reserved capacity) afin d'améliorer l'algorithme BLS (Burst Limiting Scheduler) de l'IEEE Time Sensitive Networking (TSN) en terme de performances de l'accès sur le fronthaul. Avec cette stratégie, des amélioration en terme de délai et de débit de trafic sont été mesurés par simulation en comparaison avec l'ordonnanceur à priorité stricte WRR et ce aussi bien pour les classes de trafic à haute et basse priorité, dans différentes conditions de charge de trafic.

Acknowledgements

This Thesis would not have been possible without the support of many people whom i sincerely appreciate.

I extend my profound gratitude to my advisors, Mrs. Gladys Diaz and Mr Marwen Abdennebi, who took a chance on me to begin with, and who also motivated, guided and the took time to share their wealth of experience and knowledge with me, i could not have asked for better supervision.

I would like to thank all members of laboratory L2TI, my PhD colleagues, and in particular Prof. Anissa Mokraoui our lab director for being kind to me.

I would like to thank the Tertiary Education Trust Fund (TETFUND) Nigeria for providing me with the financial means to complete this thesis, Alex Ekueme Federal University Ndufu-Alike Ikwo for granting me this opportunity and Campus France for making my time as a PhD student in France rewarding in so many ways.

Finally, my deep gratitude goes to my loving parents Prof. Kevin Ngozi Nwogu and Mrs Emilia Elege Nwogu who have been my source of constant support and love through this long process. Also to my dear siblings Chigozie, Ihuoma, Ndidiamaka and Onyekachi, i love you guys.

Above all, i thank Abba for being awesome as always.

Contents

List of Figures

List of Tables

Dedication

I wholeheartedly dedicate this Thesis to
God Almighty of whom there are no words to describe.

Chapter 1

Introduction

1.1 Evolution toward Networks of The Future: A 5G and beyond Context

Fifth generation (5G) cellular networks are expected to support a projected estimate of 100 billion connections at near zero latency with end-to-end data rates of up to 10-20Gbps [122], [7] with diverse use-cases which are majorly categorized under (i) ultra-reliable low latency communications (uRLLC), (ii) enhanced Mobile broadband (eMBB) and (iii) massive machine type communication (mMTC) [122].

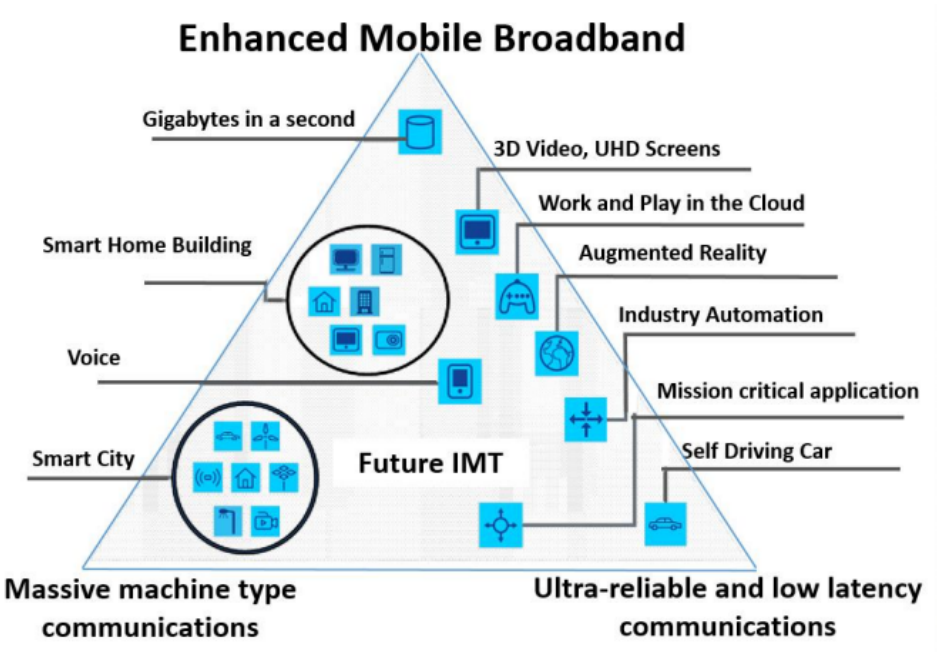

FIGURE 1.1: 5G services and use cases. Source: ITU, 2018

The latency requirements for a number of 5G applications to be hosted on next generation wireless communication networks require seamless connections owing to stringent reliability and latency targets as depicted in Figure 1.1, represented under uRLLCs. The most challenging end-to-end (E2E) design requirements are created by

uRLLC which serves as the key enabler for numerous latency sensitive applications to be hosted across diverse vertical industries. Examples of such delay sensitive applications include healthcare applications (remote robotic surgeries), intelligent transport systems, industry control/automation, and some classes of multi-media services. mMTC on the other hand, are representative of a large class of low complexity, narrow-bandwidth devices which transmit and receive small volumes of data. These devices may require coverage range expansion functionality, and rely majorly on battery power supply. Major mMTC use cases include wireless sensors, smart meters, actuators, trackers, and wearable devices. eMBB deals mostly with human-centric use cases such as media delivery and mobile telephony which require high data rates across wide coverage areas, thereby, enabling large volumes of data transfers.

In both academia and the industrial sector, solution proposals for a host of ultra low latency 5G use cases have been extensively explored and as a result, various new radio network architecture prototypes and frameworks have been developed with the aim for vendors to successfully slice the network and deliver service based isolation throughout all operational layers to customers in next generation networks.

The third generation partnership projects (3GPP) objective is to realise such uRLLC applications with low latencys of 1ms and 10ms for user plane and control plane, respectively, with ultra-high reliability of up to 99.999% in terms of packet delivery. One of the most critical sources of latency in advanced LTE radio access network (RAN), results from the initial link connection using random access channel (RACH) procedure which can take several tens of milliseconds [34]. This link establishment latency will potentially result in problems for factories-of-the-future (FoF) and internet of things (IoT) applications contending for a fixed amount of resources along with other applications. This results in severe congestion at the LTE/5G medium access control (MAC) layer.

In light of the challenges described above associated with efficient 5G use case function, present day architectures are unable to guarantee the end-to-end delay requirements of around 60-100ms and the 1ms round-trip latency in addition to required jitter figures necessary to sustain latency-sensitive applications in 5G networks. To this end, end-to-end mobile network solutions are required to keep up with next generation user QoS and QoE projections.

For the above described reasons, we focus throughout this thesis on proposals geared towards the end-to-end optimization of network segments related to latency reduction, load balancing and resource usage in order to meet the QoS and QoE targets required for efficient utilization of 5G and beyond (5GB) network application use-cases.

Subsequently, we highlight the limitations of existing solutions and define our research challenges and questions.

1.2 Problem Definition and General Approach

Problem Definition The use of SDN and NFV technologies in 5G networks to decouple control plain functions from data plain functions is lauded to be able to facilitate flexible and scalable network functions deployment in cloud environments geared to specific service requirements. The cloudification in RAN domain remains challenging due to the stringent time bounded critical functions of which if not satisfied will adversely affect the QoS for a host of 5G use case applications. Also, capacity concerns surrounding the the expected rapid growth in connected devices in addition to the former described constraints clearly require advanced latency reduction techniques in addition to smart resource allocation techniques and load balancing of heterogeneous multi-cell deployments.

Therefore, within the context of these defined limitations, this thesis focuses on providing solutions along the lines of these identified problems as illustrated in figure 1.2.

General Approach Following the defined problem context, the general approach we reflect in this thesis is made up of our proposed schemes in the domains of (i) Random Access optimization in preparedness for massive traffic access analogous to 5G IoT applications (ii) Load balancing in heterogeneous type cellular cell deployments (iii) resource management and (iv) Fronthaul (FH) QoS scheduling. All proposed optimization schemes are developed to promote latency reduction and energy conservation in future cellular networks. Fig. 1.2 further describes the wider domains under which our target proposals fall under of which we briefly introduce.

1.2.1 Real-time Applications Support

5G networks are expected to deliver high-speeds, ultra-low latency, scalable and reliable mobile broadband in its preliminary deployment [1]. A mix of mobile edge computing (MEC) technology and 5G will support time-stringent applications such as latency-critical IoT manufacturing processes, autonomous vehicular technology, video streaming, augmented and virtual reality (AR/VR). 5G and beyond networks are expected to provide real-time services to new applications with more demanding performance requirements such as

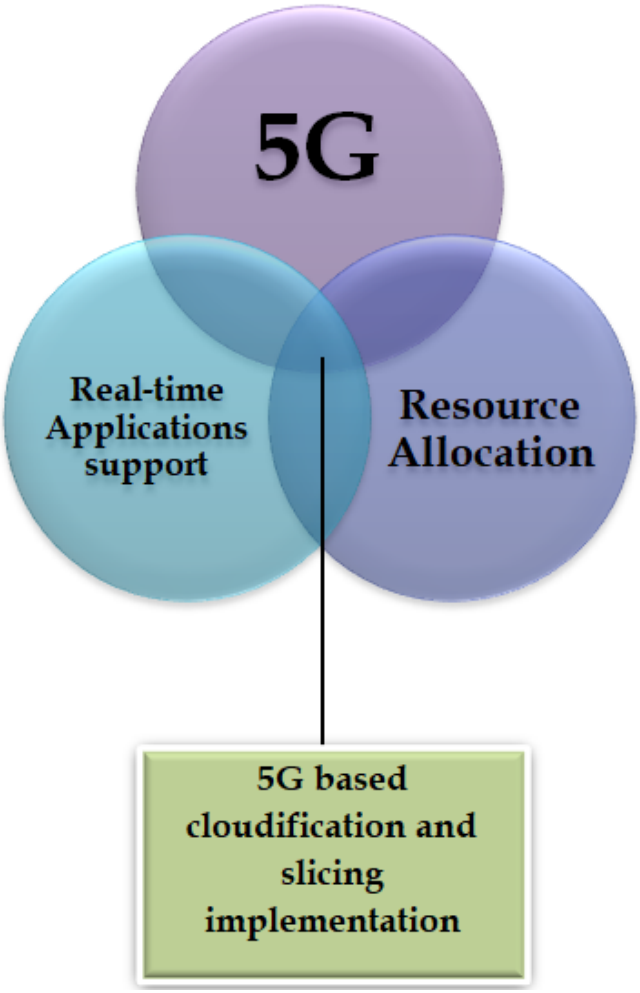

FIGURE 1.2: Proposed solutions

- **Ultra-high definition 4K display video provision:** These applications will be very sensitive and demanding in terms of reliability and latency and would require wireless cellular network availability, ease and speed of deployment and largely improved bandwidth

- **Real-time support to formerly non-cellular network use-cases:** applications like remote surgery support (telemedicine) and self- determination medicine fall into this category. therefore, the data transmission speeds required to support such applications should be up to 100x more than more wide-spread LTE network data speeds.

- **Scalability challenge in terms of load**

- **Interactive applications QoE requirement:** Interactive applications such as video conferencing and video Telephony require extreme reliability to provide crystal clear video streams with no lags and dropped connections, more so than most standard real-time applications.

In order to effectively support these critical systems and applications which are the hallmark in terms of the objectives of 5G networks, low latency connectivity should be guaranteed.

In this thesis therefore, we proffer solutions to optimize the mobile network via contributions to latency reduction in the random access (RA) procedure, through a novel load balancing scheme and by way of a QoS based FH network solution employing dynamic scheduling.

1.2.2 Resource Allocation

5G cellular networks will come with heterogeneous multi-tiered architectures consisting of a host cells with different operational specifications existing within the same coverage zones to meet the different QoS and QoE requirements for active mobile and stationary devices. Radio resource allocation and interference management as a result of this multi-cell architecture becomes a problem facing future networks.

In this thesis, we present a novel technique utilizing the concept of Almost Blank Subframes (ABS) to balance traffic load among cells which leads to optimized cell and resource usage. In addition, our first proposed scheme also contributes to optimized Random Access preamble resource usage.

1.2.3 C-RAN

Cloud radio access network (C-RAN) a candidate for next generation access network technology,also referred to as C-RAN or Cloud-RAN, is advantageous due to its potential for handling increased mobile traffic, and to enhance energy efficiencies [99]. Compared to the traditional distributed RAN, C-RAN is capable of instantiating baseband units (BBUs) and allocating baseband resources to Remote Radio Heads (RRHs) via Software Defined Networking (SDN) [116] and Network Functions Virtualization (NFV) technology [108]. Compared to previous generations, 5G provides an environment where a coalition of several radio access technologies, from pico to macro cells heterogeneously co-existing within the same Network topology and contributing to spectral efficiency, cell splitting gains, as well as optimized power consumption as a direct result. This is achievable through the adoption of SDN, NFV and cloud computing technologies [56, 138].

C-RAN architecture and benefits

C-RAN as described earlier, consists a cellular network architecture where the baseband signal processing and network functions of a radio access network are performed in a cloud environment (data center) . The C-RAN architecture comprises of three major components, i.e. the BBUs (located in the cloud), RRHs (acting as remote antenna elements) and high-bandwidth, low latency (mostly optical) front-haul links that interconnect the BBUs and the RRHs. This architecture fully enables virtualisation, and is also known as virtual RAN (V-RAN) [57]. In C-RAN, traditional Based Stations (BSs) are separated into two parts, the distributed RRHs and BBUs clustered into a pool. The pool is located at a centralized site (cloud) housing a set of BBUs. The instantiation of baseband functionalities over software without the need for dedicated hardware, enables the implementation of the RAN as a Service (RANaaS) feature. This arrangement allows the radio resources of various BBUs to be shared to meet dynamic user demands.

The C-RAN architecture facilitates several advantages compared to legacy wireless architectures, such as [58] :

- Cost reduction in CAPEX and OPEX: The efficient utilization of BBU resources is achieved through the pooling of BBUs in the cloud, and the centralization of computational processing. This in effect, reduces the combined number of BBUs needed to meet end-user demands. Furthermore, resulting in lower overall power consumption and consequent cost reductions in both Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) [35].

- Energy efficiency and green deployment: Network energy efficiency can be optimized through C-RAN by decreasing the number of active cell sites when needed, thereby, decreasing the total system power consumption of the network. Moreover, during low traffic periods, underutilized BBUs can be turned off and their traffic can be redistributed to the active BBUs.

- Improved Spectral efficiency and decreased inter-channel interference: The centralized BBU can share the resources dynamically and cooperatively among heterogeneous cells. Therefore, resources can be utilized per service demand. The inter-channel interference can also be greatly reduced as a result of joint scheduling and processing.

- Capacity and coverage range improvement: The C-RAN architecture is pegged to support a range of 10s to 1000s of cell sites. This capability will lend to improved network capacity through cell range expansion, and serving much more

user devices. As a direct result, the Quality of Service (QoS) for end-user is also improved.

- Flexible BBU-RRH associations: Additionally, the C-RAN architecture allows for flexible BBU-RRH associations, dynamically adjusting to varying network load conditions. For instance, when network load conditions are low, several RRHs can be dynamically handled by one or a few BBUs, reducing the number of running BBUs and also promoting energy efficiency

- Inter-cell interference management and reduction:the co-location of BBUs in a single place encourages easier adaptation of coordination and interference management techniques such as Coordinated Multi-Point (CoMP) [29] and enhanced Inter-cell Interference Coordination (eICIC).

- New business model adoption: The evolved C-RAN concept will generate more business models, such as the BBU pool resource rental system and cellular system as a service (network slicing) [118] .

C-RAN Challenges

Despite the several advantages posed by the implementation of C-RAN architecture, a number of challenges need to be addressed which can limit the facilitation and deployment of C-RAN. In this section, we focus on some C-RAN challenges in particular which will be the focus of this work.

- Infrastructure Limitations: The Fronthaul (FH) network design which connects the RRHs to the BBU pool, plays a crucial role in the overall system performance of a virtualized wireless network. The FH network is required to carry a large amount of data traffic in real time with high bandwidth and ultra low latency requirements so as to meet Hybrid Automatic Repeat Request (HARQ) requirements. Furthermore, as a rule of thumb, the end-to-end latency between the RRHs and the BBUs should be an order of magnitude lower than the latency requirement of any service or system algorithm [118].

- BBU Functional Split placement: The functional split concept consists of locating some baseband or protocol processing back from the centralized BBU to the distributed RRH according to the applied functional split between them [128] . In this way, the RRH will cease from being just a passive distributed antenna to an active one.Therefore, there is a need to find the optimal placement of RAN baseband functions (at centralized cloud or RRU) in order to effectively meet the strict timing transport requirements and at the same time also benefit from centralization.

- Cloudification and Virtualization of RAN Latency concerns: cloudification, virtualization and network slicing are especially challenging in the RAN domain due to the potential virtual network functions (VNFs) run-time in the cloud. When executing RAN lower and upper layer functions in commodity hardware, the runtime of lower physical layer functions such as, channel coding and large matrix inversions which require on-demand resource provisioning, can result in problems due to the bounded run-time requirements of these functions. In current LTE/LTE-A networks, the coding of Physical (PHY) layer functions make use of FPGA architectures to meet these delay requirements. We address this challenge in our work in Chapter 5 by proposing a differential traffic based scheduling solution to aid tackle latency concerns along the additional Fronthaul network added by the C-RAN architecture.

- Resource management: In standard LTE/LTE-A distributed RAN architecture, various optimization algorithms are utilized for Radio Resource Management (RRM), such as admission control, scheduling, interference mitigation and power control. Cloudification of RAN with its new architecture adds additional challenges in RRM, requiring new optimization mechanisms and algorithms to be designed. Such algorithms not only focus on spectral efficiency enhancement or interference mitigation, but also allow optimized BBU-RRH associations.The adequate design for BBU-RRH associations can be exigent and require careful implementation so as to dynamically adapt to traffic changes and also deliver an acceptable level of QoS for serviced devices.

On the basis of the above observations, this thesis focuses on latency reduction techniques and efficient resource allocation schemes over the RAN domain in order to ensure the effective implementation of cloudification and slicing techniques analogous to future heterogeneous mobile networks.

1.3 Thesis Contributions

This section highlights the main contributions of this thesis. Our contributions are three fold as displayed in Figure 1.3, to start with, we proposed a dynamic resource reservation and usage approach to deal with congestion in heterogeneous cellular networks as a result of densely populated mixed traffic scenarios analogous with the Fifth Generation (5G) networks. Furthermore, within the same context, we propose a new random access (RA) resource usage algorithm to efficiently assign RA preambles to 3 distinct traffic types with different latency requirements. In our second contribution, we propose and implement a dynamic load balancing scheme among

heterogeneous cells deployment within a 5G context. Our proposed scheme lends to the effective usage of mobile network cells, low latency scheduling and power savings. We also proposed a game-based C-RAN computational resource usage scheme. In our third proposal, we focused on developing and implementing a service differential based scheduling scheme suitable for latency reduction in C-RAN based fronthaul networks.

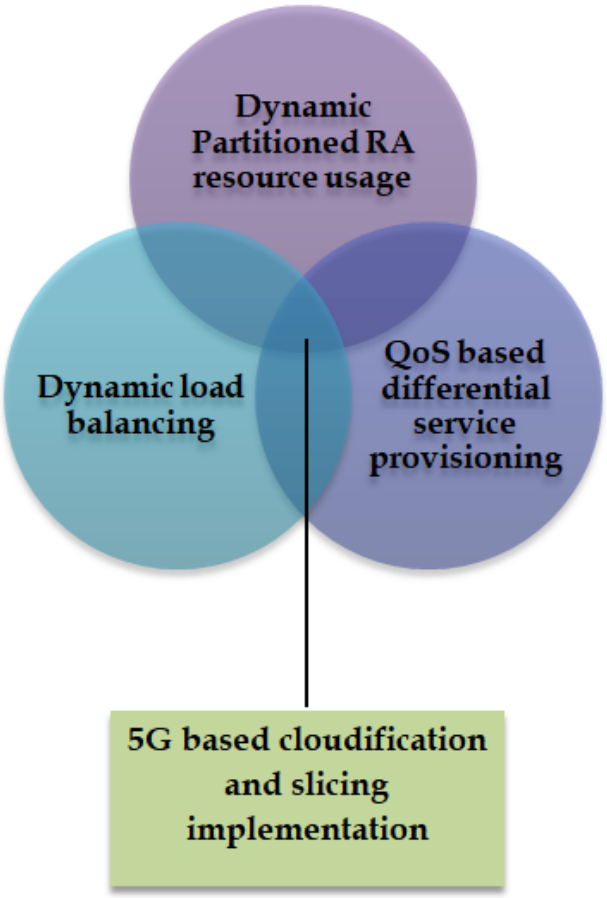

FIGURE 1.3: Proposed enablers for Cloudification and Slicing

1.3.1 Partitioned Resource Usage Approach for Random Access (RA) in 5G Cellular Networks

The ultra reliable and low latency connections required by mission critical applications to be hosted in future networks require fast data delivery services which can guarantee latency's of 10ms and under in the control plane and 1ms in the user plane [7] . In order to achieve estimated delay figures for mission critical applications in addition to overall system performance, advanced resource allocation schemes and enhancements to the network architecture, in particular the Random Access (RA)

and core protocols, need to be applied [10] . A first contribution to our work relies on RA preamble reservation for 3 distinct traffic classes (High, Medium and Low), where the reserved resources for the High priority class remains fixed, while for the medium and low class, a conditional dynamic resource sharing mechanism is introduced. Access delay figures are then compared with standard LTE/LTE-A RA figures and results show that our partitioned resource reservation mechanism significantly reduces the access delay figures in the control plane for both priority and non-priority traffic, meeting the 5G ultra Reliable and Low Latency Communications (uRLLCs) requirement. In addition, we obtain through analytical means, the optimum preamble partition for the reserved case which will also meet service requirements for low-priority traffic.

1.3.2 An Optimized Approach to Load Balancing and Resource Usage in 5G Multi-tiered Cellular Networks

Further more, in the context of end-to-end (E2E) latency reduction, we explore new techniques for optimized device to cell associations as well as efficient load balancing among heterogeneous cells within multi-tiered networks expected to be common place in 5G and future cellular network architectures. Our proposal in this part primarily consists of a major contribution to the standard LTE/LTE-A hand-over (HO) procedure by dynamically adjusting the Almost Blank Subframe (ABS) ratio (ABS to non-ABS) of large macrocells within a multi-tiered network space, to the amount of HO requests received from nearby user equipment's (UEs). Our proposal aims at encouraging attachment of devices to smaller powered nodes with less domineering reference signal received power (RSRP) as compared to larger powered nodes within the same coverage area. Properly balanced traffic load invariably leads to lower scheduling latency's and power savings, which in turn is an enabling factor for a scalable and virtual RAN. In addition to the load balancing scheme, a cooperative gamified approach to RRH-BBU resource sharing was also proposed. The validation of our proposal was done through Matlab simulations.

1.3.3 QoS based differential service provisioning for 5G New radio (NR) Fronthaul networks

With the planned data rate growth for 5G New Radio (NR), very high-rate fronthaul traffic patterns with different QoS requirements are to be expected which require cost-efficient transport solutions given the size and economic impact of the access network in the telecommunication business [123] . Therefore, we propose in this part

of our work a scheduler based optimization of the fronthaul network for transporting 5G NR signals. We implement a queue based prioritization of mobile traffic to cope with the latency requirements of inelastic ultra Reliable and Low Latency communication (URLLC) flows and non-URLLC flows associated with Networks of the Future. We make use of the IEEE Time Sensitive Networking (TSN) Burst Limiting Scheduler (BLS) algorithm by implementing our contribution to this algorithm by way of a Dynamic Reserved capacity (DRC) scheduling scheme. We apply our proposed algorithm and switching architecture to differential flows in a packet-switched fronthaul aggregator node, in the bid to mitigate the waiting time for mission critical data and thus maximize throughput. The simulation results show that our scheduling scheme is able to mitigate delay and optimize throughput for delay sensitive traffic.

1.4 Thesis Organisation

The organisation of this thesis is outlined in Figure 1.4. Each chapter comprises of the literature review, contributions and major results.

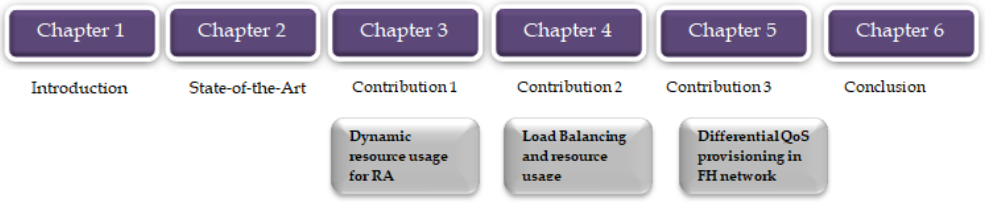

FIGURE 1.4: Thesis Organisation

Chapter 1 presents the introduction into the thesis, research challenges which are addressed, objectives and proposals. Chapter 2 presents a comprehensive State-of-the Art into 5G technology, from the background to standardization and its limitations. We also present our position about these limitations together with our proposed solutions.

Chapter 3 introduces our first contribution by way of a partitioned RA resource usage mechanism, geared towards the optimization of RA preamble resource and to meet up with acceptable 5G control plane latency figures.

Chapter 4 presents a joint optimisation algorithm for multi-tiered heterogeneous 5G networks consisting of two stages, the first being a novel contribution to standard handover (HO) decision control mechanism in a bid to effectively balance traffic

load in hierarchical or mixed cell structures analogous to networks of the future. We also present a gamified approach to RRH-BBU association for C-RAN with optimal convergence and utility guarantees.

Chapter 5 defines a novel FH network differential traffic scheduling architecture to carry out optimized scheduling making use of a credit based shaper mechanism. Our proposal is aimed at improving the overall throughput and latency figures for each differential traffic class from the time constrained to best-effort traffic.

Chapter 6 presents a summarized version of the major findings of this thesis and gives key directions into future work and open research questions.

Chapter 2

State-of-the-art in 5G networks

This chapter provides an outline of 5G Networks and its features, recent updates together with their constrains, technical aspects related to ultra-low latency provision, optimal resource usage and end-to-end Quality of service provisioning. In addition, this chapter provides an overview of the open issues related to future networks.

2.1 Introduction

As stated in the previous chapter, 5G cellular heterogeneous networks of which initial roll-out and deployment began in 2020, are expected to support a very large number of connections, from IoT devices to massive machine type communications (mMTC) in addition to standard mobile devices. 5G networks are designed to support wide host of applications with an extensive variety of requirements, including high user data rates, lower latency, enhanced indoor coverage, improved resource usage in addition to energy efficiency [63], [48], [162].

Performance targets and requirements for 5G networks have been addressed during the standardization process, therefore making 5G networks more structured compared to previous standards such as LTE/LTE-A and so on. These 5G specific performance targets include ultra reliable and low latency response to meet up with the end user QoS requirements, network densification through heterogeneous multi-tiered cell structures, enhanced energy efficiency and capacity. Table 2.1 summarizes the 5G network performance targets.

In order for these service requirements to be fulfilled, 5G networks will need to embrace multi-tiered heterogeneous architectures made up of cell nodes with diverse characteristics and capacities, from the conventional macrocell to smaller nodes such as femto, pico cells as well as device-to-device (D2D) UEs. As a result of the need for

proper coordination and management among these cells, and in light of the large expected growth in traffic and connected devices associated with future networks, network nodes will require smart self-organization capabilities to perform tasks such as autonomous load balancing, enhanced inter-cell interference coordination (eICIC), network resource allocation, power consumption minimization etc. [162], [88] .

Some of the more promising attributes associated with 5G networks is its ability to be more responsive in terms of lower latency in addition to its ability to connect more devices at once through advanced resource management schemes. In order to effectively deliver projected latency and connection figures, novel algorithms as well as improvements to RA and multiple access (MA) mechanisms and procedures need to be designed in a way that would meet user QoS and QoE needs as well as system performance requirements.

In light of all the described features, this chapter therefore will provide the *State-of-the-art* of 5G networks with major focus on these features. We will begin with the advancements to modulation technologies and solutions associated with the 5G network standard in addition to the open issues confronting the standard and current solutions. Subsequently, we will proceed to 5G radio resource allocation as it relates to random access, cell deployments and differential service provisioning along the extended FH network.

2.2 5G RAN : modulation technologies and solutions

The innovations 5G is meant to support have been categorized under three major groups of applications by 3GPP known as enhanced mobile broad band (eMBB), massive machine type communication (mMTC) an ultra reliable and low latency communications (uRLLCs) [63], [136]. These service classes are considered as the key focus under which the 5G network is designed of which we provide a brief overview.

- **eMBB:** eMBB is a communication service that is representative of an evolution from long term evolution (LTE) which already provides impressive mobile broadband speeds in the gigabit range in limited markets. However, the need for much higher bandwidths, ultra-low latency for mission critical applications and greater connectivity exceeds the capacity of the more widely adopted

5G Performance targets	Trends/Proposals
Capacity and throughput improvement, high data rate (∼1000x of throughput improvement over 4G, cell data rate ∼10 Gb/s, signaling loads less than 1∼100%)	Spectrum reuse and use of different band (e.g., mm-wave communication using 28∼GHz and 38∼GHz bands), multi-tier network, D2D communication, C-RAN, massive-MIMO
Reduced latency (2∼5 milliseconds end-to-end latencies)	Full-duplex communication, C-RAN, D2D communication
Network densification (∼1000x higher mobile data per unit area, 100∼10000x higher number of connecting devices)	Heterogeneous and multi-tier networks
Advanced services and applications (e.g., smart city, service-oriented communication)	C-RAN, network virtualization, M2M communication
Improved energy efficiency (∼10x prolonged battery life)	Wireless charging, energy harvesting
Autonomous applications and network management, Internet of Things	M2M communication, self-organizing and cognitive network

TABLE 2.1: Performance targets of 5G networks

LTE, specifically with the number of connected devices worldwide on a high increase.

To satisfy the requirements for new wireless networks, the other key use-cases - uRLLC and mMTC work together with eMBB. One way 5G will be deployed to effectively cover vast areas would be via fixed wireless access (FWA) using spectrum bands in the mm range which were not utilized in 4G networks. The number of eMBB enabled devices is projected to increase to 1 billion in the year 2024 as compared to 15 million in 2019.

- **mMTC:** Machine type communication (MTC) technology is founded on the idea that the growth value of machines are proportional to the number of network units [85], leading to the Internet-of-everything (IoE) and IoT concepts [70] and the prospect of developing a fully networked smart city society by expanding the number of networked machines [138], [30], [24]. The annual growth rate of networked machines is 25% with current numbers up to hundreds of millions. In 5G networks, mMTC is expected to pave the way for an increase in new innovative services and applications through the IoE concept [56], [88].

- **Ultra reliable and Low Latency Communications (URLLC)** URLLCs are

among the key service class cathegories in 5G networks expected to contribute to heightened user experience through emerging services and applications which have both delay constrained and reliability requirements. Among the 3 major groups of eMBB, mMTC and URLLC, the design and implementation in 5G networks of URLLCs is the most demanding due to its requirement to jointly meet the design aspects of *Ultra-high reliability* and *low latency* [68] . Improving reliability through added resources for re-transmission,redundancy, signaling and parity results in a negative trade-off for the latency [139]. The 3GPP aims at providing URLLC with user plane latencies of $\sim$ 1ms and reliability of $1-10^{-5}$ (99.999%). A symmetric latency budget of $\sim$0.5ms for both UL and DL is required to ensure the aforementioned latency target [126], [86].

A more recent inclusion is enhanced Vehicle to everything communications (eV2X) [31] . These applications require flexible connectivity with improved spectral efficiency and system throughput, subjecting the architecture of general 5G networks to significant design challenges. traditional orthogonal Frequency Division Multiplexing (OFDM) is unable to meet many new demands of 5G networks due to its requirement for high synchronization to avoid out of band leakage (OOB) among adjacent bands, whereas, in the case of technologies like mMTC, sensor nodes transmit different data types which are mostly asynchronous in narrow bands [164] .

To combat the challenges associated with future wireless communications a number of modulation and waveform candidates have been considered for their prospective advantage for 5G network use-cases. In this chapter, we provide an overview of some potential modulation and waveforms prototypes with promising interest in 5G, due to their specific advantages. We highlight the major advantages of each modulation scheme and waveform in accordance to identified design requirements. Various modulation techniques have been proposed, some of which include;

- Filtering
- Pulse Shaping
- precoding
- Guard Interval (GI) shortening

These modulation schemes have the capacity to effectively reduce OOB leakage of OFDM with the filtering technique being the most straightforward to reduce OOB leakage over the stop-band [31] . The above mentioned modulation schemes can be used with Orthogonal Multiple Access (OMA) in future networks. OMA is essential to all previous and currently used wireless networks.

2.2.1 New Orthogonal Multiple Access Modulation Schemes

OFDM is a structured multi-carrier modulation format developed for RF systems used in 5G and most modern wireless communication systems. Its capability to greatly increase the data rate in channels with bandwidth constraints allowing MIMO integration and high spectral efficiency is one of its key advantages [160].

On the other hand, a major disadvantage associated with OFDM would be its susceptibility to OOB emissions. As a result of this shortcoming, new modulation techniques for 5G networks are being proposed. However, due to its extensive use in current wireless systems as a base standard and to encourage backward compatibility in future networks, we will first introduce it [31].

A. Traditional OFDM

d_k, for $k = 0,1,N-1$, is the transmit complex symbol.The baseband OFDM signal,

$$s(t) = \sum_{k=0}^{N-1} d_k e^{j} 2\pi f_k t \tag{2.1}$$

for $0 \leq t \leq T_s$, where $f_k = k\Delta f$, Δf is the subcarrier bandwidth and T_s is the symbol duration

Therefore,

$$d_k = \frac{1}{T_s} \int_0^{T_s} s(t) e^{j2\pi f_k t} dt \tag{2.2}$$

To address delay spread in wireless channels a cyclix prefix (CP) is used in OFDM if the length of the CP is larger than the delay span, the OFDM demodulated signal is given as

$$\overline{d_k} = H_k d_k + n_k \tag{2.3}$$

Where, H_k is the frequency response of the wireless channel at $f_k = k\Delta f$ and n_k is the impact of additive channel noise.

5G networks have to support a large host of different user types with different demands. Traditional OFDM cannot hope to satisfy these lofty requirements, hence

novel modulation techniques with much lower OOB leakage will be required. Modulation techniques for OMA include,phase shaping, subband filtering, precoding design, guard interval (GI) shortening etc.

B. Modulation based on Pulse Shaping

Filter Bank Multi Carrier (FBMC)

Future mobile communications beyond 2020 will require highly diverse and optimized architecture to fully accommodate a wide range of applications. The air interface of 5G systems will provide more flexibility compared to present systems. Advanced OMA schemes like FBMC and GFDM will be considered as key modulation techniques for New Radio (NR). In light of this, these techniques are explored in depth.

Presently OFDM is widely adopted due majorly to the schemes robustness against multipath channels and easy implementation based on Fast Fourier transforms (FFT) [102]

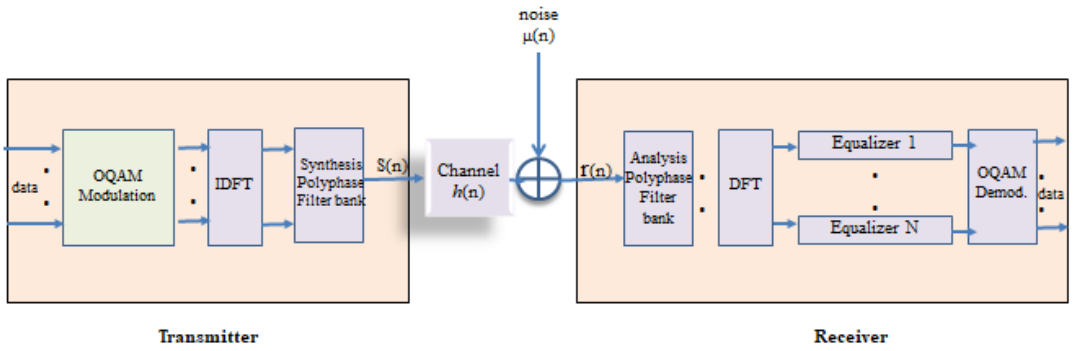

FIGURE 2.1: Filter bank multicarrier (FBMC)

As shown in figure 2.1 for FBMC to achieve the best SE, offset quadrature amplitude modulation (OQAM) is applied to make FBMC real-domain orthogonal in time and frequency domains [20],[122],[31], [51] .

FBMC consists of IDFT and DFT synthesis and analysis polyphase filter banks. The transmit signal for FBMC can be expressed as

$$s(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} d_{k,m} \Theta_{k,m} g(n - mk/2) e^{j2\pi n k_n / k} \tag{2.4}$$

Where k and M are the numbers of subcarriers and symbols respectively, $d_{k,m}$ is the transmit symbol at subcarrrier k and symbol m, and $g(n)$ is the prototype filter coefficient at the n^{th} time-domain sample. The parameter, $\Theta_{k,m}$ in (2.4) is defined as

$$\Theta_{k,m} = \begin{cases} \pm 1 & \text{if m+k is even,} \\ \pm j & \text{if m+k is odd,} \end{cases} \tag{2.5}$$

One of the main drawbacks for CP-OFDM is that it cannot provide NR required degrees of freedom, in other words, a more adaptive and efficient solution other than the fixed waveform configured to best compromise [31] . For FBMC schemes, the modulated signal on each sub-carrier is shaped by a well designed prototype filter different from traditional rectangular pulse in CP-OFDM. This is an important addition, providing a new degree of freedom.

FBMC	**CP-OFDM**
High grade prototype filter which spans over multiple symbol periods	Rectangular pulse type filter
Better spectral efficiency due to absence of CP	Additional overload due to use of CP necessary to support frequency domain channel equalization
Efficient spectrum thanks to confined power localization in frequency of the filter	Less efficient spectrum due to poor localization in the frequency domain due to sinc-shaped spectrum of pulse
Better control of out-of-band properties of wave forms	High sensitivity to frequency offsets between transmitters and receivers

TABLE 2.2: Comparison between FBMC and CP-OFDM

Disadvantages of FBMC

- Orthogonality concern: FBMC/OQAM transmission schemes inherit an intrinsic interference in time dispersive channels due to wave forms overlapping in time domain and the absence of a CP. In order to properly mitigate these effects, suitable equalization and interference cancellation schemes have to be employed.
- Packet transmission: theoretically speaking, FBMC/OQAM can achieve full time frequency efficiency through the use of OQAM. However, this is capable for symbol sequences with infinite length. Realistically, data transmission is divided into smaller time bursts. The ramp-up and ramp-down times at the edges of these intervals caused by filtering reduce the actual efficiency. This may give rise to a big problem in applications such as machine type communications, where the packets to be transmitted are expected to be short. Some solutions which include, burst truncation and packet transmission with special processing at the edge have been proposed [2]

Generalised Frequency Division Multiplexing (GFDM)

GFDM was initially proposed in 2009 by G. Fettweis *et al* as a flexible multicarrier technique for future wireless systems. More techniques for multi carrier communications are being researched for 5G network architecture and schemes like FBMC interference avoidance transmission by partitioned frequency and time domain(IA-PFT), and GFDM. All these techniques can be classified as filter bank techniques and are related to OFDM.

With GFDM, a generalization of OFDM is proposed with additional degrees of freedom in choosing system parameters. The GFDM scheme orders data in a two dimensional time-frequency block which introduces flexible pulse shaping for individual subcarriers reducing the amount of cp while still providing means for single-tap equalization in frequency domain.

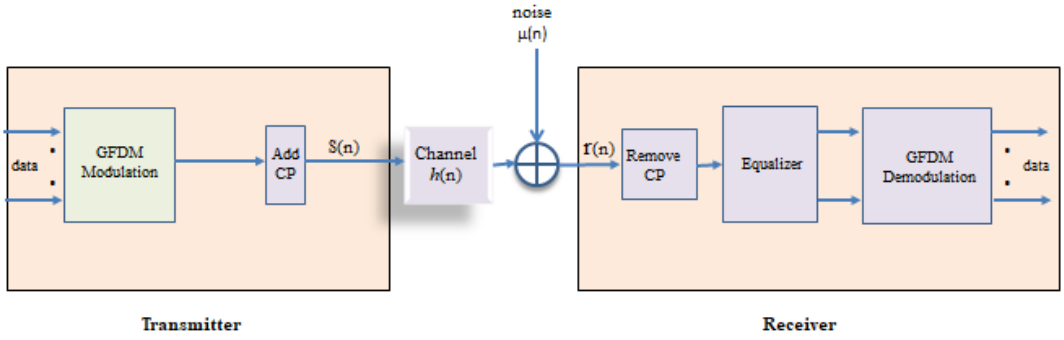

FIGURE 2.2: Generalized frequency Division Multiplexing (GFDM)

Figure 2.2 indicates the block diagram of GFDM. The difference between GFDM and FBMC is that GFDM uses circular filters instead of the linear filter used in FBMC to execute pulse shaping. By carefully choosing the filter, out-of-block leakage is minimized, M frequency samples can be easily adjusted and K time samples for a GFDM block according to application environment. The transmit signal for each GFDM block can be expressed as

$$S(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} d_{k,m} g_{k,m}(n) \tag{2.6}$$

for $0 \leq n \leq KM-1$, where $d_{k,m}$ is the transmit symbol on subcarrier k at subsymbol m and $g_{k,m}(n)$ is the circular time and frequency shift version of the prototype pulse shaping filter. From (2.6),

$$g_{k,m}(n) = g((n-mK)KM)e^{j2\pi kn/k} \tag{2.7}$$

where $(.)KM$ denotes the KM modulo operation and $g(n)$ is the prototype pulse shaping filter similar to OFDM, both modulation and demodulation can be expressed via matrix operations. However the transceiver structure for GFDM is different from OFDM. Along with FBMC and GFDM, modulators based on pulse shaping, like QAM-FBMC and pulse shaped OFDM [168], have been proposed for 5G networks.

C. Modulation based on sub-band Filtering

In addition to pulse shaping, one more technique used to reduce OOB leakage is sub band filtering. Filtered OFDM (FOFDM) [167], [5] and Universal filtered multi-carrier (UFMC) [134] are two classic modulation types built from sub band filtering. Fig.4 displays the transmitter and receiver structures of UFMC. The bandwidth of the filter in UFMC is much wider than that of the modulations derived from pulse shaping; the length in time domain is much shorter. f-OFDM uses a CP and regularly allows residual inter-symbol interference (ISI) [5]. In addition, the filter in f-OFDM can be longer than that in UFMC and has better attenuation outside the band. One more difference from UFDM is that the subcarrier spacing and the CP length do not have to be the same for the different users in f-OFDM.

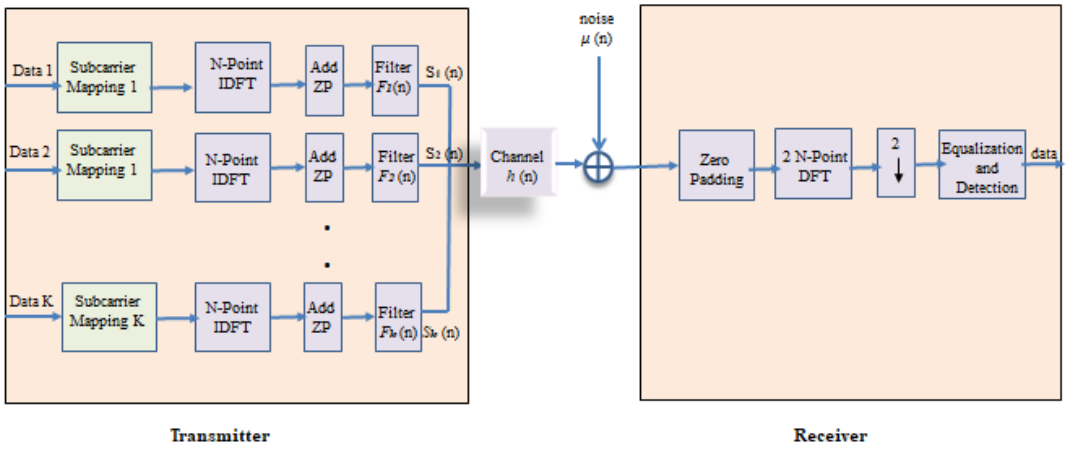

FIGURE 2.3: Universal Filtered Multi Carrier (UFMC)

Assuming N subcarriers are divided into K subbands, each with $L = N/K$ consecutive subcarriers, the transmit signal in UFMC can be expressed as

$$S(n) = \sum_{k=0}^{K-1} S_k(n) * f_k(n), \tag{2.8}$$

when $f_k(n)$ is the filter coefficient of subband k,and $S_k(n)$ is the OFDM modulated signal over subband k that can be expressed as

$$\mathrm{S}_(n) = \sum_{m=0}^{M-1} S_k(n - m(N+N_g)) = \begin{cases} S_{k,m}(n - m(N+N_g)), & m(N+N_g) \le (n) \le m(N+N_g)+N-1 \\ 0, & otherwise \end{cases} \tag{2.9}$$

with N_g denoting the length of the zero-padding (ZP) [110], M denoting the number of symbol blocks and $S_{k,m}(n)$ denoting the signal at subcarrier k and symbol m. In (9), $S_{k,m}(n)$ can be expressed as

$$S_{k,m}(n) = \sum_{l=(K-1)L}^{kL-1} d_{l,m} e^{j\frac{2\pi ln}{N}}, \qquad 0 \le n \le N-1 \tag{2.10}$$

Where $d_{l,m}$ is the l-th transmit symbol at the m-th symbol block. At the receiver, the signal at symbol interval has the length $N + N_g$ and is zero-padded to have a length of 2N so that a 2N-point FFT can be performed

TABLE 2.3: comparisons of advanced multiple access and waveform based solutions

KPI	OFDM[31]	GFDM[134][106]	FBMC [122]	UFMC[134]
Processing latency	Low	high	Low	Low
Computation complexity	Low	high	Moderate	Low
Flexibility	High	High	High	High
Spectral efficiency	Low	very high	Low	high

2.2.2 Adaptive Modulation and Coding (AMC) Implementation in 5G

Compared to LTE/LTE-A networks, in 5G communication network systems, the evolved NodeB will perform adaptive modulation and coding (AMC) in conjunction with the corresponding values of the channel state information which is fed back by the user equipment (UE). This process helps to enhance spectral efficiency as well as the throughput of the OFDM system through the selection of different code rates and modulation methods.

The principle behind AMC as opposed to fixed modulation and coding, lies with the ability to dynamically adjust the modulation and coding order, symbol coding rate and coding method such as to maximize total transmission link throughput and maintain system stability [14]. Therefore, when channel conditions are optimal, higher transmission rates are dynamically adopted for data transmission and in poor conditions, lower rates are utilized in order to ensure correct signal reception at the receiver end.

The 5G NR traffic channels (UL and DL) majorly adopt the Low Density Parity Check (LDPC) coding, while control channels are encoded using Polar codes. Modulation methods used include Quadrature Phase Shift Keying (QPSK), 16QAM (Quadrature Amplitude Modulation), 64QAM and 256QAM.

Table 2.4 shows the Channel Quality Indicator (CQI) figures defined by 3GPP Ts38.214 specification for Physical layer data procedures.

TABLE 2.4: 4-BIT CQI

CQI index	**modulation**	**code rate *1024**	**efficiency**
0	Out of range		
1	QPSK	78	0.1523
2	QPSK	193	0.3770
3	QPSK	449	0.8770
4	16QAM	378	1.4766
5	16QAM	490	1.9141
6	16QAM	616	2.4063
7	64QAM	466	2.7305
8	64QAM	567	3.3223
9	64QAM	666	3.9023

10	64QAM	772	4.5234
11	64QAM	873	5.1152
12	256QAM	711	5.5547
13	256QAM	797	6.2266
14	256QAM	885	6.9141
15	256QAM	948	7.4063

2.2.3 Requirements relating to modulation and waveforms for 5G networks

Some of the promising and key requirements in terms of waveform formats necessary to reduce connection latencies and optimize general performance for 5G and beyond application use-cases are reviewed in this section.

Summarized in figure 2.4 are the performance requirements associated with mmWave transport over analog radio over fiber (ARoF) which are both top candidates for 5G FH network transmission as well as general requirements for 5G wireless communications.

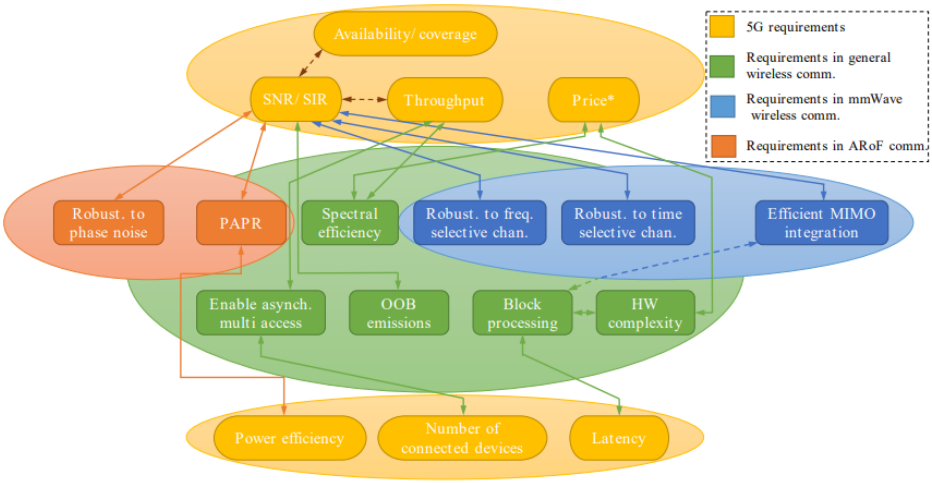

FIGURE 2.4: Relationships among requirements for candidate waveforms and 5G requirements [124]

Requirements for general 5G wireless communications

The waveform format relating to wireless communications ideal for fifth generation and future network use-cases would require the following specific features:

1. **Permitting asynchronous multiple access:** The importance of asynchronous multiple access comes with its ability to efficiently utilize resources. In both time division duplex (TDD) and frequency division duplex (FDD) systems, dynamic allocation of communication resources for varying bandwidths is of the essence in order to effectively accommodate asymmetric traffic. Therefore, it follows that waveform formats which enable asynchronous MA would utilize channel resources optimally corresponding to higher total throughput figures [135].

2. **hardware complexity:** Due to the increase in the heterogeneity of next generation mobile networks by way of multiple access adoption, the complexity and cost of hardware present in each cell should be considered when deciding the feasibility of a modulation format in order to manage OPEX and overall system complexity.

3. **Structured MIMO integration:** MIMO systems are an appropriate technique to overcome the low attenuation associated with mmWave wireless communications which is a key propagation candidate for 5G networks. mmWave signals are only able to travel shorter distances as compared to current signals utilized in LTE/LTE-A and would require a modulation format with efficient MIMO integration [8].

4. **Out of band emissions:** In order to guarantee the high levels of spectral efficiency required to foster next level mobile communication networks which will be shared by heterogeneous user type technologies and network operators. It is important that OOB emissions be kept to a minimum so as not to interfere with simultaneous service transmissions on adjacent channels. This restriction of OOB leakage can be bounded through the use of considerable bandwidth as guard band in order to obtain adequate service frequency multiplexing.

5. **Spectral efficiency:** The increase in spectral efficiency has long been an important factor in the wireless communications sector, as it relates directly to performance by way of achieved bit rate. Deploying 5G networks with targeted service at peak spectral efficiency exceeding 30bps/Hz downlink (DL) and 15bps/Hz uplink (UL), which doubles the 15bps/Hz DL provided by LTE-Advanced is essential to meet the increased mobile data usage [159], [50], [166].

6. **Peak-to-average power ratio (PAPR):** PAPR is the relation between the maximum power of a sample in a given transmit symbol and the average power of the symbol [115]. In more simple terms, PAPR is the ratio of peak power to the average power of a signal expressed in decibel (dB).Large power fluctuations as a result of high PAPR figures lead to high power consumption which 5G networks seek to effectively minimize [135].

2.3 Radio resource management in 5G HetNets

The ultimate goal of 5G NR is to support a wide range of services with diverse QoS requirements. To arrive at this goal, an extensive range of techniques and models will have to be adopted to effectively manage the performance degradation associated with interrupted services. A host of novel algorithms and strategies, scheduling, power allocation, interference coordination and link adaptation, optimization and heuristic models, game theoretic models, AI/machine learning etc. [80],[73] which leverage on utilizing limited spectral resource optimally would be employed in 5G.

Some of the enabling techniques which will contribute to efficient radio resource management (RRM) which have been adopted in this thesis include; leveraging of ABS techniques in heterogeneous multi-tiered networks, modified RA procedures, game theoretic approaches and differential service scheduling in 5G FH networks. The adoption of these techniques have been tailored to optimization objectives of low control and user plane latencies, higher throughput and energy efficiency. Some system performance metrics that require optimization for 5G and future networks are listed below.

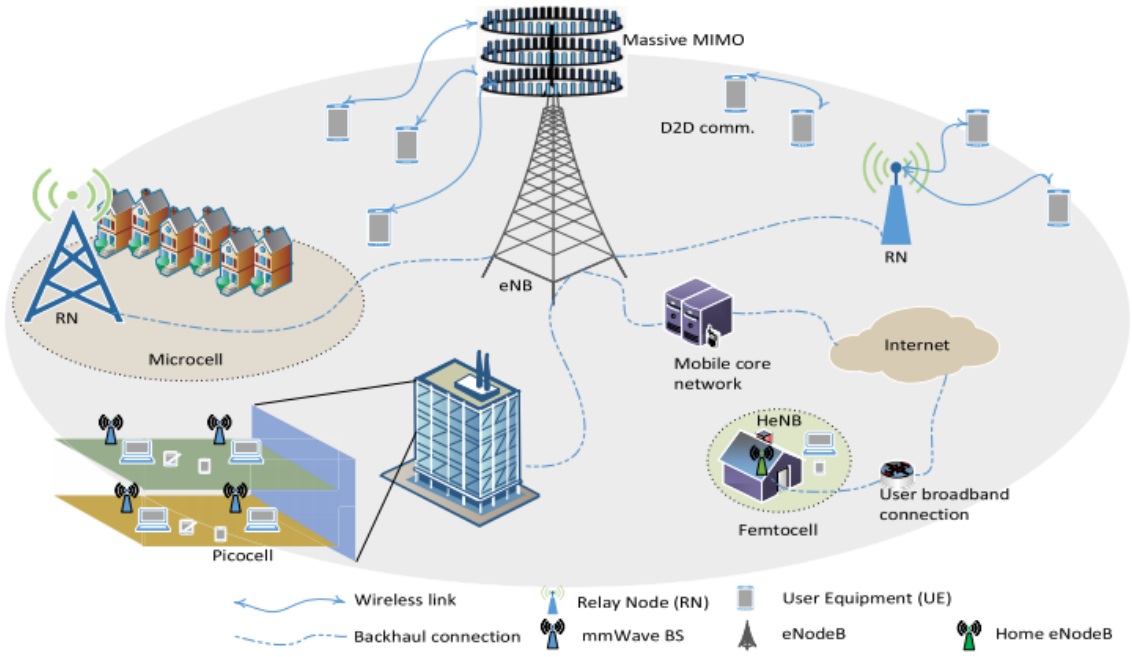

FIGURE 2.5: 5G multi-tiered HetNet Architecture [103]

QoS based Scheduling

5G and beyond networks are expected to consist of ultra-high dense multi-cell deployments necessary for optimized system capacity [15]. This upgrade to the 4G network is not without its challenges which include cell boundary ICI challenges, frequency reuse [112], and the efficient allocation of already scarce radio resource. In light of these challenges, intelligent and dynamic scheduling of radio resource is paramount to reaching system targets and satisfying QoS/QoE requirements of connected devices.

The scheduling mechanism in LTE and LTE-A [83] systems is controlled by the eNB which allocates resources to the active users in a cell every transmission time interval (TTI) which is equal to 1ms. OFDMA is utilized in downlink transmission and it divides available bandwidth into multiple sub-carriers which carry multiple data flows, the uplink radio access scheme is single carrier frequency division multiple access (SC-FDMA). The allocated resources are expressed in terms of physical resource blocks (PRBs) or RBs. The RB is defined as seven OFDM symbols

(one 0.5ms time slot) in time domain and 12 subcarriers in the frequency domain (180KHz).

The major parameters utilized by the eNB to compute scheduling metrics are as follows:

- **QoS/QoE Parameters:** When a user device runs an application, a logical channel is created between UE and eNB referred to as a radio bearer which has a defined QoS profile named QCI. In order to meet these QoS requirements, the eNB uses the QCI parameters mapped to their flows to assign scheduling priority to the flows.
- **Channel quality:** In LTE-A systems, the eNB makes use of the channel quality indicator (CQI) to obtain an estimation about a connected UEs channel quality. The CQI has a value which ranges between 0 to 15 that indicates the level of modulation and coding the UE could operate thereby allowing the eNB to prioritize traffic scheduling based off this information.
- **Resource allocation record:** System fairness is an important aspect in mobile networking whereby all connected devices require service. The records of past achieved performance of UEs are utilized by the eNB to improve UE fairness during the traffic scheduling process.

Some of the main scheduling strategies introduced in LTE/LTE-A which are still adopted in 5G and beyond networks are summarized as follows (more extensive surveys can be found in [81], [125])

1. **First in First out (FIFO):** This strategy falls under the cathegory of channel-unaware scheduling strategies. As the acronym implies, FIFO allocates resources to UEs on a first-come first-served basis.
2. **Round Robin(RR)/Weighted Round Robin (WRR):** RR scheduling technique is a starvation-free RBs algorithm which cyclically selects equal amounts of data flows (allocates a fixed amount of RBs) from an ordered queue buffer for transmission.This method ensures there is equality for all UE traffic. The WRR on the other hand is a modification to standard simple RR whereby weights are assigned to each queue or traffic class mostly in the order of priority.
3. **Largest Weighted Delay First (LWDF):** The LWDF scheduler is used to ensure Real-Time (RT) application packets are received within set worst-case

deadlines to avoid dropped packets [93]. Users with more latency-stringent requirements in relation to deadline expiration and acceptable loss rate are granted scheduling priority. This scheduler also falls among the channel-unaware strategies.

4. **Blind Equal Throughput (BET):** This scheduler attempts to provide throughput fairness among UEs. To achieve this, the scheduler algorithm allocates resources to traffic flows from UEs which have previously been served with the lowest average throughput [74]. This scheduler also falls under the cathegory of channel quality unaware schedulers.

5. **Maximum Throughput (MT):** The MT scheduler belongs to the channel-aware cathegory of schedulers which take into cognisance candidate UE channel quality. The scheduler prioritizes UEs with the best channel conditions which will in turn achieve maximum throughput [129]. One of the major drawbacks with this scheduler is its lack of consideration to the fairness criterion among UEs.

6. **Proportional Fair (PF):** This channel-aware scheduler was designed primarily for downlink transmission and for Non-Real Time applications (NRT). the PF scheduling algorithm offers a balance between achieved throughput and fairness by taking into cognisance both resource allocation history and CQI [72]. Therefore, the UE with an average past throughput and worst CQI ratio are ensured to be served within a specific time frame.

From a 5G FH network scheduling point of view, during the transportation of data packets between RRH and BBU, a considerable number of switches within the FH network are utilized as aggregation and multiplexing points where the packets are scheduled, de-queued and routed [52]. The fairness and latency of these packets and connected RRHs are greatly affected by the packet scheduling policies applied at these aggregate switch nodes [39], [40].

Therefore, optimized packet scheduling is required to better apply the trade-off between fairness, QoS provisioning per traffic class and C-RAN multiplexing gains.

System throughput

This parameter is measured in bits per second (bps) and is referred to as the sum of successful data rate being sent over all network devices and nodes [95]. Network balancing affects overall system throughput and thus in hierarchical networks of the future, careful planing and technique application should be carried out in order to optimally benefit from the multi-cell architectures being adopted in 5G and beyond

networks. In our work we have proposed a load-balancing based technique to achieve optimized throughput and power optimization figures.

Other techniques adopted to achieve improved system throughput include, non-cooperative and cooperative repeated games [27], markovian chain approximations, linear programming based solutions, optimized power allocation [148], among a host of other techniques.

Spectral efficiency

Optimized spectral efficiency (SE) falls among one of the key performance metrics associated with resource allocation in 5G and future networks whereby the 5G standard targets 10x more spectral gains than the previous LTE standard. This would be achieved partly through the adoption of more cell types compared to previous standards.

SE represents the maximum amount of services to be obtained from a measure of spectrum and is measured in bits per second per hertz (b/s/Hz). A few methods in academia and industry that have adopted SE as one of the performance metrics for optimal resource allocation include clustering game algorithms [157], [155], range expansion and load balancing schemes [153], [117], stochastic optimization schemes [153].

User association

In 5G and future network HetNets, users are associated with any available network within their range. In order to select the most favourable network, a user association mechanism is applied based on channel quality - channel quality indicator (CQI) and proximity to the evolved node B (eNB). User association is also a key requirement in load balancing optimization, energy efficiency (EE) and spectral efficiency (SE) [92].

Load balancing

A key component of 5G multi-tier HetNets is the load balancing of user traffic between network cells. As part of our work, we proposed a dynamic load-aware balancing mechanism utilizing 3GPPs ABS mechanism [47], [117].

2.4 5G: Towards a more flexible RAN

In order to effectively manage and control the integration of heterogeneous network components and technologies hosted in networks of the future (NoF), 5G networks have adopted the well established architectures and protocols associated with the cloud-computing space, where virtualization software-centric techniques are combined to abstract functionalities and services from the previous majorly hardware-centric mobile network platforms [23].

Therefore, in 5G RAN and 5G core (EPC), openness and scalability through NFV and SDN [13] has emerged as viable approaches to increase network flexibility in order to reduce time to market deployment for new services etc. The sharing of data centres and the use of open application interfaces will enable a large numbers of applications and services to be provisioned cost-effectively over fixed mobile broad band networks [133].

A major part of the cost of mobile networks lie in the large number of distributed base stations and antenna sites. Future cloud and Open-RAN (O-RAN) architectures [57],[13],[114] will compose of a combination of Virtualisation, Centralization and coordination techniques all interacting with each other in the network.

2.4.1 5G RAN Architectures

A review of available C-RAN architectures being implemented and proposed for NR, ranging from traditional distributed and centralized RAN architectures to new options enabled by SDN and NFV are presented.

A. Distributed RAN

A fully distributed baseband deployment consists of an interface between RAN and core network at the radio site. Majority of current LTE networks make use of the distributed baseband model. Thanks to collaboration between base stations over the X_2 interface, LTE hand-overs are majorly seamless. X_2 coordination is evolved to support carrier aggregation as well as coordinated multi-point (CoMP) reception across layers.

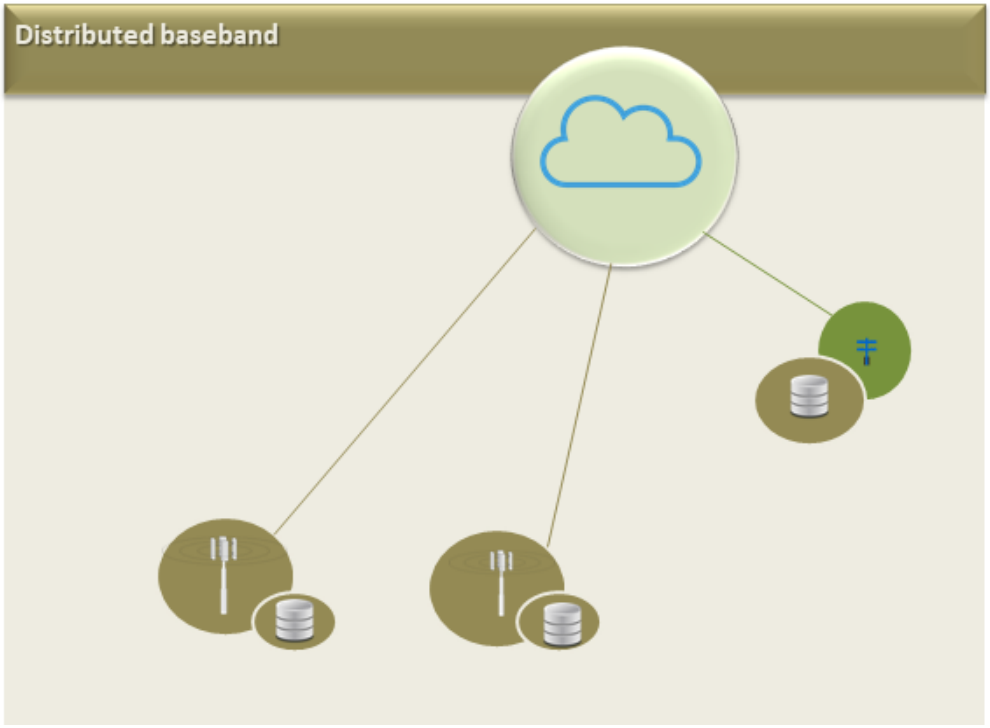

FIGURE 2.6: Distributed baseband deployment

B. Centralized RAN

Centralized RAN deployments have become increasingly adopted by operators to boost performance in traffic hot spot such as offices, city squares, malls, stadiums etc. A fully centralized baseband deployment model consists of all baseband processing (RAN L_1,L_2 and L_3 protocol layers) located at central locations, serving multiple distributed radio sites as shown in figure 2.7

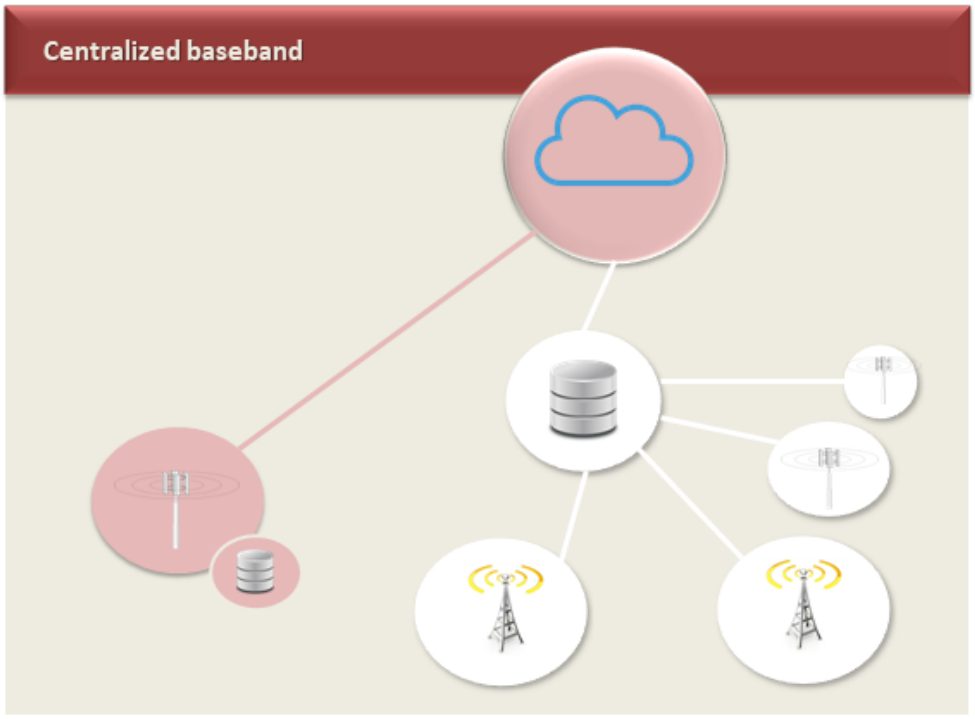

FIGURE 2.7: centralized BB deployment complementing a distributed BB deployment

Transmission between BBU's and RRH's use CPRI front-haul over dedicated fiber or microwave links (DROF) [49]. In many situations, CPRI connectivity requirements will be too strict for centralized RAN architectures to be affordable.

C. Virtualized RAN

When introducing high bandwidth layers with partial coverage for 5G, the currently deployed distributed/centralized architecture for 4G networks will not be efficient. Virtualized RAN addresses the issues which will arise due to vastly different throughput capabilities and limited coverage exhibited by new spectrum [130].

Virtualized RAN architectures will exploit NFV techniques and data center processing abilities and enables both coordination and centralization in mobile networks, as highlighted in figure 2.8.

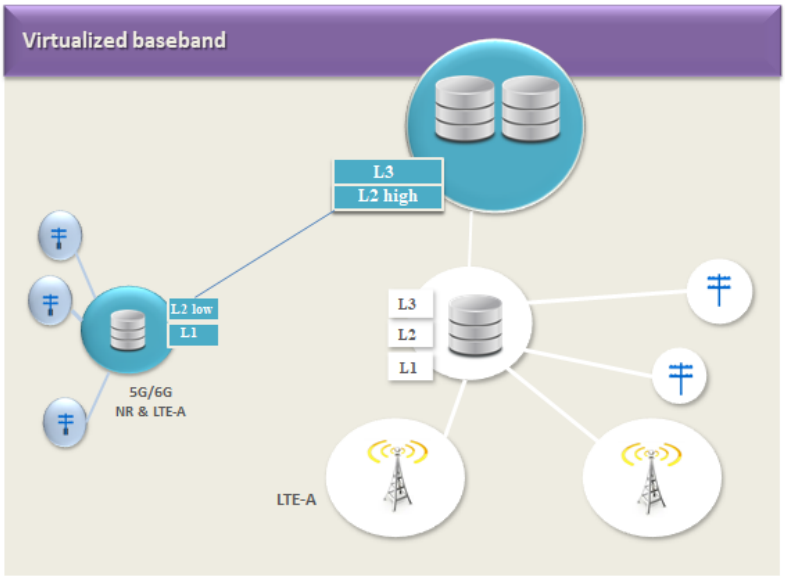

FIGURE 2.8: Virtualized RAN with some BB functions in separate environment

An important aspect of virtualized RAN is the fact that certain benefits from the split separation of the higher asynchronous layers of the radio protocol stack can be achieved, see figure 2.9.

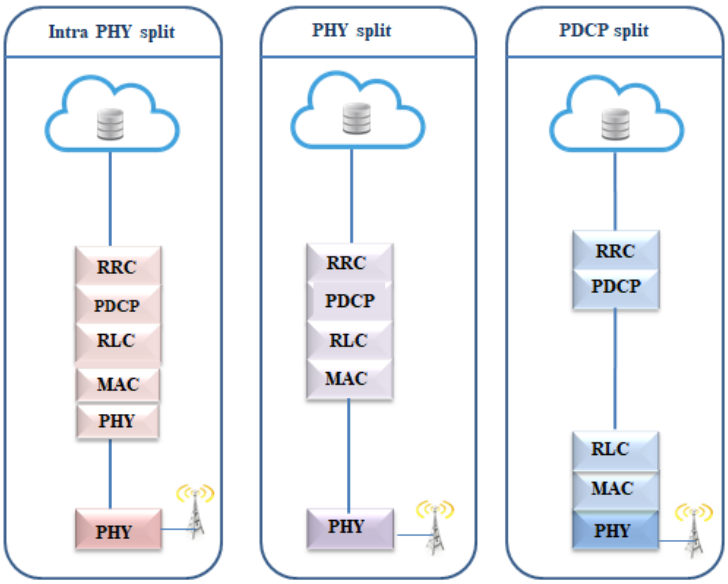

FIGURE 2.9: RAN functional splits: Intra-PHY split; PHY-MAC split; PDCP split

In order to reduce the bandwidth requirements on the fronthaul link between BBU and RRH (for base-band signal transmission ease), a host of diverse functional split options have been explored for use by major standardization bodies in 5G NR.

The 3^{rd} Generation Partnership Project (3GPP) have come up with eight main functional split options see (TS 38.401, TS 38.806, TR 38.816), in addition to some sub-options [84]. Enhanced Common Public Radio Interface (eCPRI) [66] propose split options A to E as depicted in figure 2.10, while the Next Generation Fronthaul Interface (NGFI) came up with split options 1 to 5 [39]. NGFI (xHaul) supports two major functional split standards, the IEEE 1914.1 standard for packetized fronthaul transport networks and the IEEE 1914.3 standard for Radio over Ethernet (RoE) encapsulations and mappings [84]. The split options 1 through to 4 are termed Higher Layer Splits (HLS), while lower (from 4 below) are Lower Layer Splits (LLS).

The different downlink (DL) split options proposed by 3GPP, NGFI and eCPRI can be seen in figure 2.10

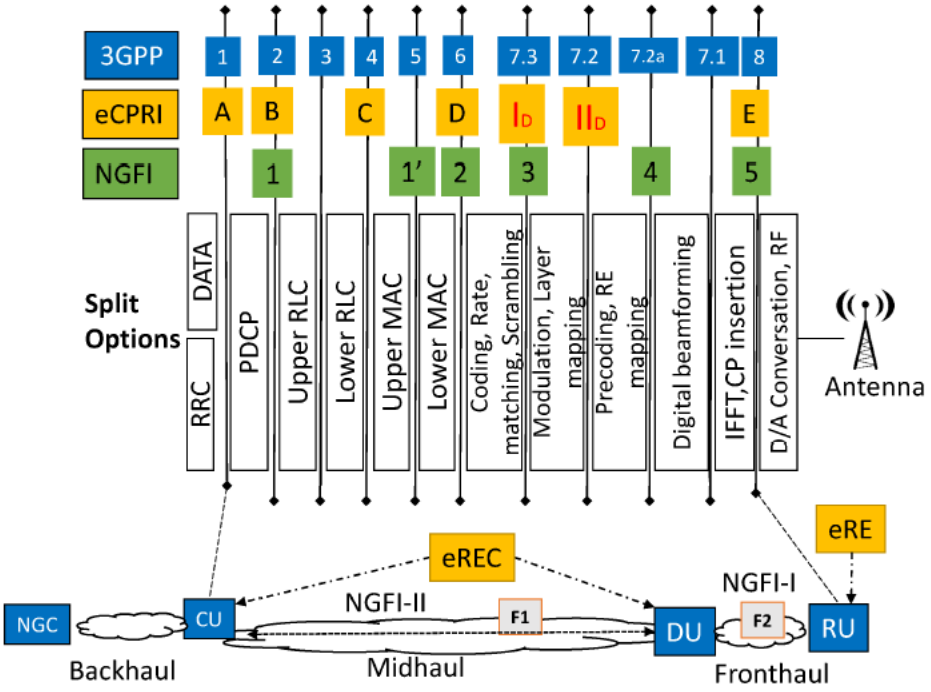

FIGURE 2.10: Functional split options proposed by 3GPP, NGFI and eCPRI [22]

A more detailed look into the LTE/LTE-A protocol stack layers with the location of functional split partitions adopted by 3GPP [84] is described in figure 2.11.

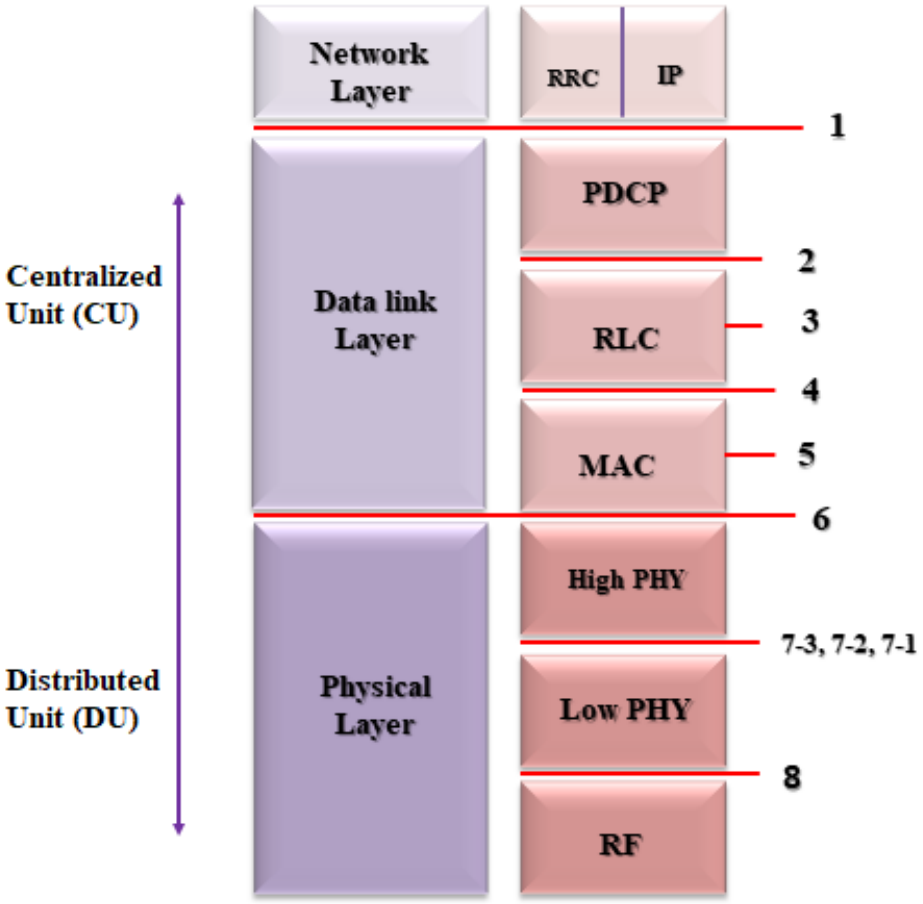

FIGURE 2.11: Functional split options proposed by 3GPP

In figure 2.11, the red partition lines represent the different functional split options where the functions below the lines represent function implementation in the distributed unit (DU) while those above the red lines will be performed at the central unit (CU). The functions located in the DU are situated close to the UEs in remote radio systems (RRS) while those located in the CU pool will be exposed to the benefits of higher processing power, coordinated multi-point (CoMP) gains among cells and processing centralization. With more baseband functionality located at the DU, the less need for higher bandwidth on the fronthaul link.

It therefore sounds feasible to have all programmable OSI layers, but there is a strong consensus among network owners about the danger it poses to allow low level network programming by third parties due to network security and management issues [31]. Also, advanced features such as CoMP and cooperative processing for mMIMO will not be possible if keeping, for instance, layer 1 in dedicated hardware. Communication between layers 1 and 2 will also become complex.

2.4.2 SDN for Cloud-RAN

The major focus of SDN technology is towards decoupling of the software-based control plane from the hardware-based data plane of networking and switching pieces of equipment [49]. SDN brings with it unprecedented ease in innovation, openness, optimum resource utilization, support for virtualization, etc.,. However, many experts have stated that 5G may just see some point solutions with SDN in some places [161], and the focus will more likely remain on densification, mm-wave technology, and distributed antenna systems (DAS) to achieve 1000x capacity, 100x data rate, and 100x active connections from the already deployed LTE network.

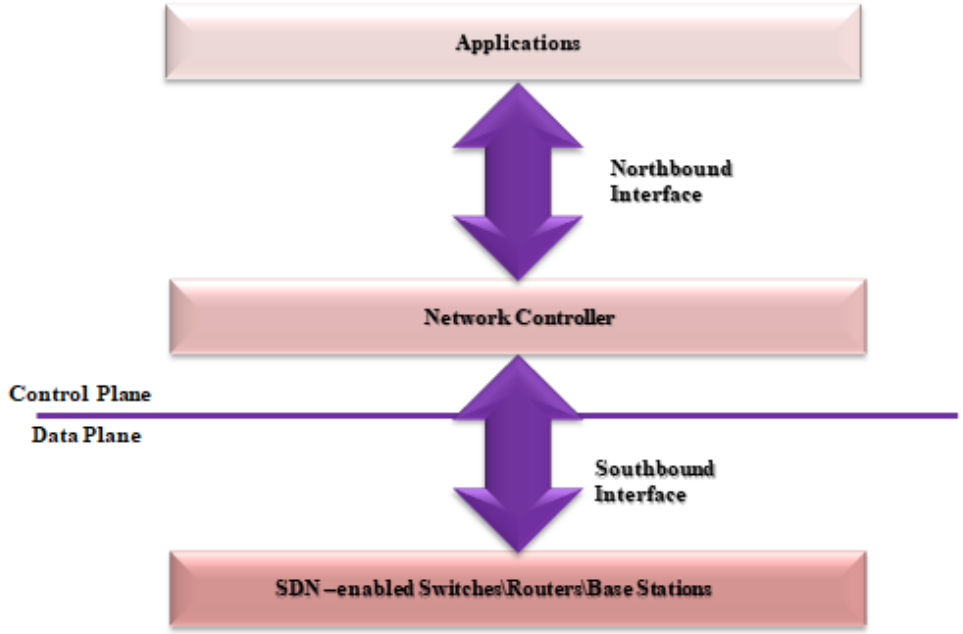

FIGURE 2.12: A generic SDN model of three layers

With 87% of the total Internet users now in possession of smart-phones, mobility is very important for global telecommunication networks. LTE Evolved Packet Core (EPC) has done a remarkable job in simplifying the core and separating control and user planes to a reasonable extent. 3GPP's 5G NGC has further improved this separation. The base station eNodeB, however, still contains both planes. A few major issues hindering the progression of SDN technology in mobile cellular networks particularly include;

- fronthaul

- latency of general purpose platforms (majorly for lower layer RAN)
- disruptive deployment
- SDN specific security issues
- compelling business case

Fronthaul solutions

China mobile presented Next Generation Fronthaul Interface (NGFI) [64],[40] as a NR FH solution which decouples the dependency of CPRI on a number of antenna elements by putting all antenna related functionality such as, downlink antenna mapping, FFT/IFFT, channel estimation and equalization in the RRU. Currently NGFI is being standardized under IEEE 1914.1 project.

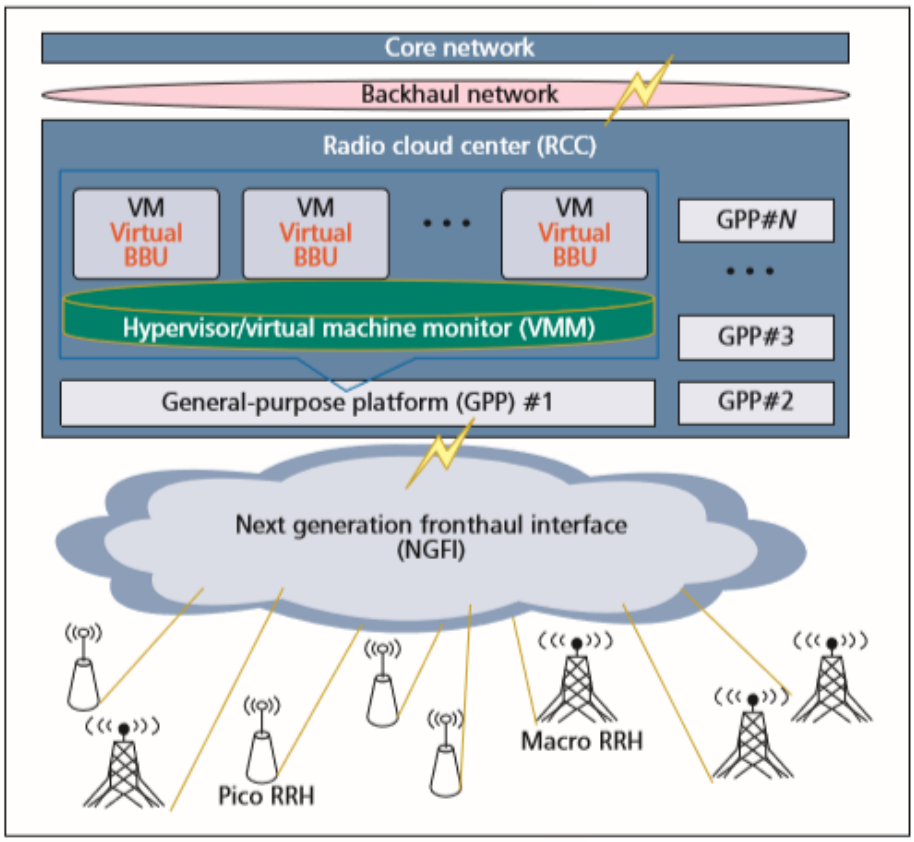

FIGURE 2.13: C-RAN Network architecture with NGFI [49]

As presented in figure 2.13, NGFI was proposed to connect Radio Cloud Centre (RCC) and new RRHs. The new RRHs contain not only radio processing but also partial baseband processing.

NGFI design principles

- Redesign of functional split between BBU and RRH
- Decoupling FH bandwidth from number of antennas
- Decoupling cell/ UE Processing
- Focusing on high performance gain collaborative technologies
- Reliable synchronization on packetized networks(SYNC-E)/1588v2 hybrid network

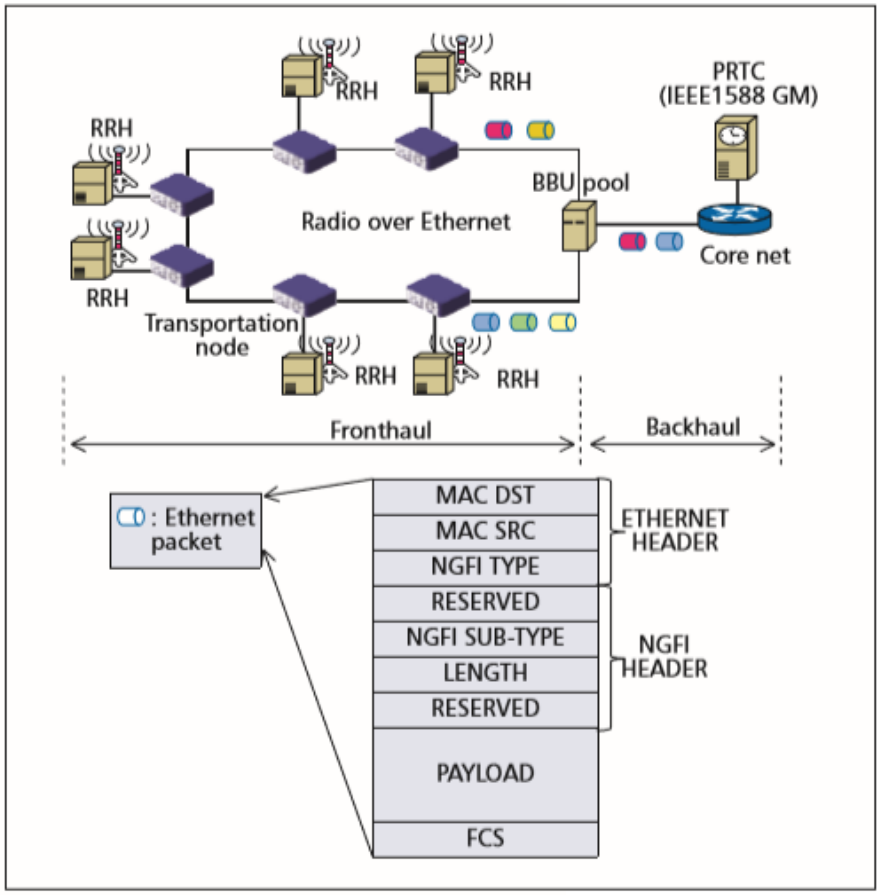

FIGURE 2.14: Fronthaul Ethernet packet format to support NGFI [49]

Limitations for NGFI

- Careful design of SYNC-E and 1588v2 to support 5G technologies like CoMP

- Jitter and latency are key difficulties to overcome to finally realise NGFI

- Limitations in the support of novel 5G RAN architectures such as the air interface C/D decoupled RAN architectures

2.4.3 SDN architectures

A few proposed SDN based architectures for cloud-based wireless mobile networks in literature are presented in this section.

CONCERT

In their work, Jingehu liu *et al* [94] propose a CONvergence of Cloud and cEllulaR sysTems (CONCERT) a converged edge infrastructure for cellular networking and mobile cloud computing.They introduce new design for physical resource placement and task scheduling, so that CONCERT can overcome the drawbacks of existing baseband-up centralization approach.

Virtualisation in this architecture is achieved through a control/data (C/D) decoupling mechanism by which a logical control plane entity dynamically coordinates data plane physical hardware. They claim to overcome drawbacks of baseband-up centralization by allowing flexible combination of distributed and centralized strategies in allocating data-plane computational resources for signal processing functions.

Architectural Design

The infrastructure is divided into

- heterogeneous physical data-plane resources and
- decoupled control plane entity called conductor which is responsible for virtualizing data-plane resources

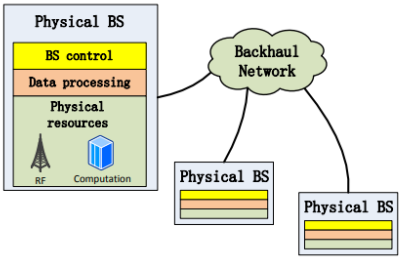

FIGURE 2.15: conventional cellular network [94]

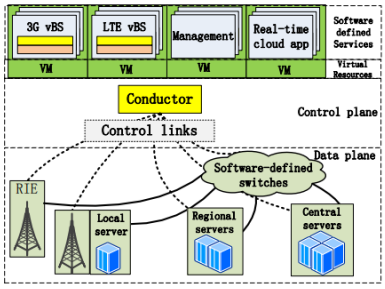

FIGURE 2.16: CONCERT architecture [94]

Software-Defined Services

On top of the virtual resources, various services can be easily deployed in a software defined way. Service developers should be able to remotely manage the virtual resource, thus greatly simplifying services deployment and construction.

CONCERT is said to support virtual base stations (vBSs), the operator will have to initiate an instance of vBS software, and the conductor will provision the required resources so the network functions of the physical BS can be delivered. CONCERT is said to support traditional mobile communication services on top of their vBSs. CONCERT is said to also supports the deployment of centralized RAN, the conductor provisioning radio resources for RIEs to form virtual antennas; virtualize a portion of the underlay SDN as the fronthaul network; and virtualize some computational resources in a data center as the baseband processing pool.

SoftAir

I.F. Akyildie *et al* [11], in their work proposed a new software defined architecture called softAir for 5G wireless systems. They proposed the utilization of SDN concept for next generation 5G wireless networks. In their softAir architecture, the control plane consists of network management and optimization tools and is implemented on network servers. The data plane consists of software defined base stations (SD-BSs) in the RAN, and software defined switches (SD-switches) in the core network. Their control logic are implemented in software on general purpose hardware and remote data centers.

In softAir, the service providers are provided with the ability to control, optimize, and customize the underlying infrastructure without owning it and without interfering with the operations and performance of other service providers, leading to more cost efficient operations and enhanced QoS. Thanks to the programmable data plane, the network resources, e.g spectrum, can be dynamically shared among the service providers.

SoftAir Architecture design

Consists of forwarding plane and control plane. The forwarding plane is an open, programmable, and virtualizable network forwarding infrastructure which consists of software defined core network (SD-CN) and software defined radio-access network (SD-RAN). The control plane consists of two primary components

1. Network management tools

2. customized applications for service providers or virtual network operations

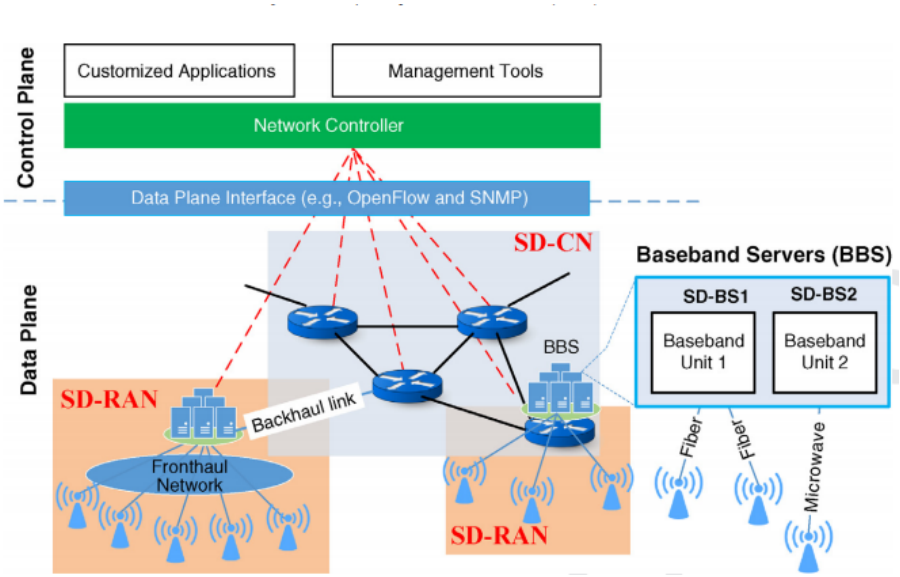

FIGURE 2.17: SoftAir architecture [11]

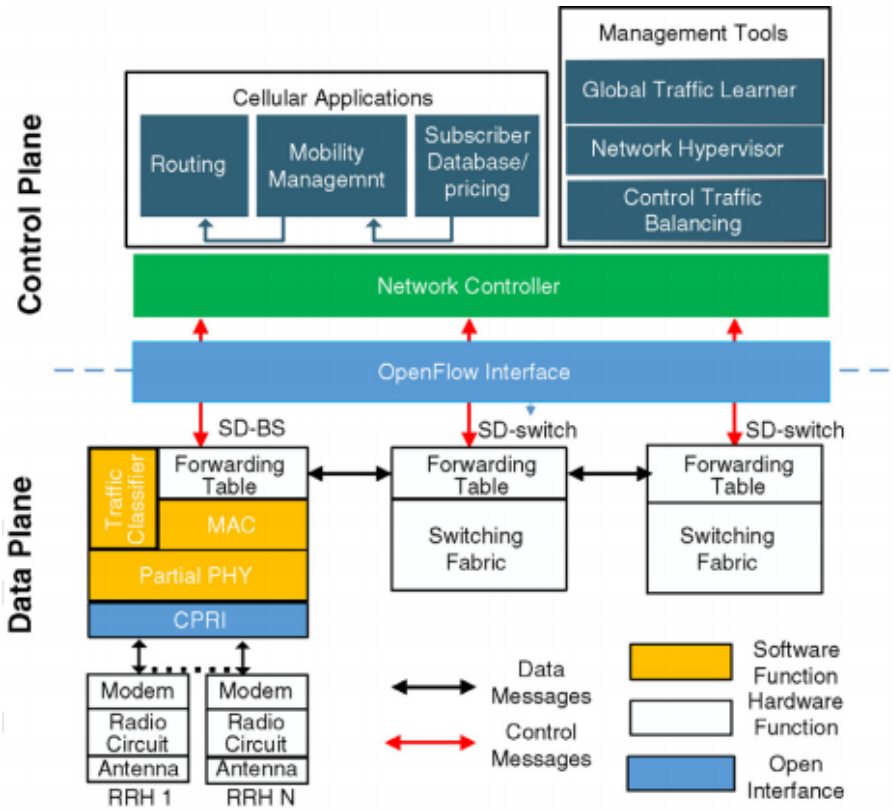

FIGURE 2.18: SoftAir architecture [11]

scalable Network function cloudification

1. Scalable SD-CN design: Their proposed softAir design adopts SD switches to provide cellular core network with high flexibility as illustrated in figures 2.17 and 2.18. They propose to minimise the control message forwarding delay between controllers and switches by their proposed control traffic balancing scheme which employs emerging parallel optimisation theories, e.g the Alternating Direction Method of Multipliers (ADMM) to achieve fast and reliable control message forwarding.

2. Scalable SD-RAN: The proposed SD-RAN follows a distributed RAN architecture as shown in figure 2.17. Here each SD-Bs is split into hardware only radio heads and software implemented BBUs on base band servers (BBS) through fronthaul network using standardized interfaces. Centralizing base band processing at data centers facilitates network wide cooperative processing among different base stations. Existing distributed RAN architectures such as cloud-RAN mainly focus on the high-performance computing of BB processing functions at remote servers. This system is limited in that;

    - They cannot achieve scalable PHY/MAC layer cloudification. The SoftAir proposed SD-RAN offers significantly enhanced scalability and cooperativeness via fine-grained base station decomposition. Through the approach of a completely centralized BBU processing, digital I-Q samples must be transport between BBS and RRHs which demands high data rates on fronthaul networks. With the introduction of massive MIMO, full-duplex transceivers, mm- waves and Terahertz band, even more higher data rates will have to be supported.To address this challenge, SoftAir adopts a new base station decomposition architecture by leaving partial baseband processing at the RRH(e.g modem), while implementing the remaining BB functions at the BBS as shown in figure 2.18 In addition, SoftAir decomposition still preserves sufficient flexibility offered by the distributed RAN architecture.

Table 2.5 gives a summary of some cloud architecture based solutions in literature as the technologies and frame works implemented in them.

TABLE 2.5: Cloud architecture based solutions

Reference	centralized	NFV	SDN	MEC/Fog
[165]	✓			
[149],[11]	✓	✓	✓	
[121]-[62]			✓	
[55],[105]		✓		✓
[49]		✓	✓	✓

2.5 Conclusion

The need for smart networks capable of servicing the volumes of data predicted to increase exponentially in the near future has given birth to the 5G heterogeneous network architecture coupled with NR upgrades.

As we have extensively shown in this chapter, the migration from the previous generation of mobile networks into the 5th generation and beyond of mobile networks is very essential to maintaining adequate QoS and QoE for the end user. As a way to fulfil these high demands, 5G networks have introduced a wide range of mechanisms aimed at increasing data volume by $\sim$1000 times, decrease latency by $\sim$5 times and increased device-to-device (D2D) connectivity [107].

Therefore, in this thesis we have proposed effective strategies and schemes which lends towards the optimization of URLL in various network segments, in addition to energy reduction through unnecessary signaling control through our ABS adaptive load balancing scheme. We have also contributed towards an optimized FH network packet scheduling and resource allocation.

In the following chapters, we will further elaborate on our proposed strategies developed to improve network function associated with 5G and networks of the future.

Chapter 3

Resource allocation for Random Access in 5G Cellular Networks

This chapter presents a major contribution developed to optimize the Random Access (RA) procedure in LTE/LTE-A networks in readiness for 5G and beyond 5G (B5G) mobile wireless network use-cases. This contribution consists of a dynamic preamble selection scheme to be applied to the grant-based RA adopted in current 5G wireless networks.

The performance of the proposed scheme has been evaluated through simulations and compared to the base LTE scheme. The results show that our developed scheme outperforms the standard LTE/LTE-A RA mechanism.

3.1 Introduction

From previous chapters, it has been established that the fifth generation of mobile wireless communication networks are expected to support a wide range of new services in addition to the data and voice services. Some of these emerging services include IoT and IoE which can be applied to a broad array of fields such as traffic, medicine, factories and agriculture [163],[9],[45],[67]. With an estimated 75.4 billion connections world-wide by the year 2025 [140], [54], advanced wireless access technology capable of delivering communication reliability is a top priority in order to effectively benefit from novel services such as M2M, IoE and IoT [36].

To this end, we address in this chapter, the problem associated with massive random access for 5G and B5G networks through our contribution to the grant-based random access procedure by way of an optimized preamble resource usage scheme which lends to reduction in connection latency for both URLLC and non-URLLC traffic.

3.2 A Dynamic resource allocation scheme for 5G Random Access

3.2.1 Context

In order to meet the QoS requirements for 5G applications, end-to-end enhancements to current mobile networks have to be applied. In both Long Term Evolution Advance (LTE-A) and 5G networks, before data transmission can be carried out by devices, air interface connections are required to be established between these devices and the network. The Random Access Channel (RACH) in particular is the transport layer channel responsible for the management of this initial connection of uncoordinated channel access requests from various user devices within the network.

Due to the nature of emerging services in future networks such as IoT with unique requirements characterized by very large amounts of transmitted data with high rates and also, MTC [132], with smaller transmitted data,low data rates, low computational capabilities and an overall lower power consumption. As a result of the characteristics of these services, significant changes need to be made to the air interface and connection mechanism through enhancements to the RA channel preamble distribution among traffic classes in order to reduce connection latencies in readiness for new services operation quality requirements.

As a result, we propose a partitioned dynamic/reserved Random Access preamble resource allocation scheme geared towards reducing the latency incurred during the RA process and meeting the 5G latency target of under 10ms for control plane transmissions.

3.2.2 Related Work

To reduce network latency in LTE/LTE-A networks, several techniques have been proposed for RA. In [37] a reserved Random Access Channel (RACH) resource for enhanced mobile broadband (eMBB) and uRLLC traffic was proposed which are both next generation network use cases. Their method was termed prioritized resource reservation, results showed a reduction in control plane access delay of below 10ms for 95% of uRLLC devices using their resource reservation method. Although this scheme is very insightful, it does not specify the amount of preambles to be reserved for the URLL traffic class which will result in optimized performance in lower priority class.

In their paper [146] proposed a RA enhancements for 5G latency sensitive traffic from a Factory of the Future (FoF) perspective via the combination of a three-fold solution which includes parallel preamble transmission, dynamic reserved preambles and enhanced back-off. Simulation results reduced the access delay by 90% compared to standard LTE solution. However, their strategy was focused on optimization for high priority traffic alone.

Results from [142] show that the standard static Access Class Baring (ACB) cannot support next generation congestion control and latency. They propose a dynamic ACB which shows efficient congestion control aiding the coexistence of both human-to-human (H2H) and machine-to-machine (M2M) traffic.

Among all these works, no one considered joint optimization for the lower priority traffic classes in a bid to foster fairness and a more robust QoS solution across all access traffic. We therefore in our proposed resource scheduling proposal take into consideration access latency reduction for both URLLC and non-URLLC traffic.

3.2.3 Random Access Procedure

In a grant based LTE-A RACH procedure a preamble is selected out of 54 available orthogonal sequences. The RA procedure from start to finish consists of a four-way handshake of exchanged messages between the device and next generation node B (gNB), this is used to establish uplink (UL) synchronization between device and gNB, to further schedule a connection request, re-connect to a radio link in the event of connection failure and carry out handovers (HOs) from one gNB to another of which 10 preambles are reserved for HOs.The grant based RA occurs over pre-defined time slots known as Access Grant time Interval (AGTI) or simply RA opportunity.

The physical Random Access Channel (PRACH) is the time-frequency resource through which the RA procedure is performed and has a minimum time based scheduling duration of 1ms. The available set of PRACH resources for transmitting preambles is broadcasted by the gNB periodically. The grant-based four-way RA procedure is summarized bellow;

Message 1: Preamble transmission

The user device seeking a network connection randomly chooses an orthogonal Frequency Division Multiplexing (OFDM) based preamble randomly from a predetermined set of 64 preambles. In LTE, 10 of these 64 preambles created by the Zadoff-Chu sequence are reserved for contention-free access for events like HOs. This is carried out through the PRACH and thus defining the RA-radio network temporary identifier (RA-RNTI).

Message 2: Random Access Response (RAR)

Preamble detection by the gNB is carried out through cross-correlating received signals with all preambles from the set at each time epoch. Once there is a correlation match, a Channel Impulse Response (CIR) occurs. Next, the transmitted access request will be decoded by the gNB and if more than one device selects the same preamble, it will be interpreted by the gNB as a severely degraded signal which is interpreted by the gNB as a collision and therefore discarded. Successful user devices are sent a Random Access Response (RAR) over the Physical Downlink Shared Channel (PDSCH) with an identification (ID) indicating the slot where the preamble was sent, timing advance (TA), uplink scheduled resource grant following Message (msg) 1 success, Physical Uplink Shared Channel (PUSCH) assigned resource and Cell Radio Network Temporary Identifier (C-RNTI).

Message 3: Connection Request

Msg 3 is sent by the device requesting a Radio Resource Control (RRC) connection through the PUSCH, including a random initial device identity after receiving the time-frequency RB from msg 2. If the user device does not receive a RAR containing its preamble and within a predefined time interval from the gNB, it flags its access attempt as failed.

Message 4: Contention resolution

Finally, contention resolution is sent as message 4 by the gNB to the devices using the Downlink shared channel (DL-SCH) with IDs of successfully decoded devices which can proceed to transmit uplink data. For cases of collision, the device retries the whole procedure after a back-off interval. In the event of a defined number of failed access attempts, the network is deemed unavailable to the device and the issue is subsequently raised to the upper layers for resolution.

In current LTE and LTE-A systems, an increase in traffic load leads to an increase in RA collisions, therefore standard LTE and LTE advanced systems may not be able to effectively handle the load associated with massive Machine-Type communications (mMTC) which is a key use case associated with 5G and beyond wireless networks [36].

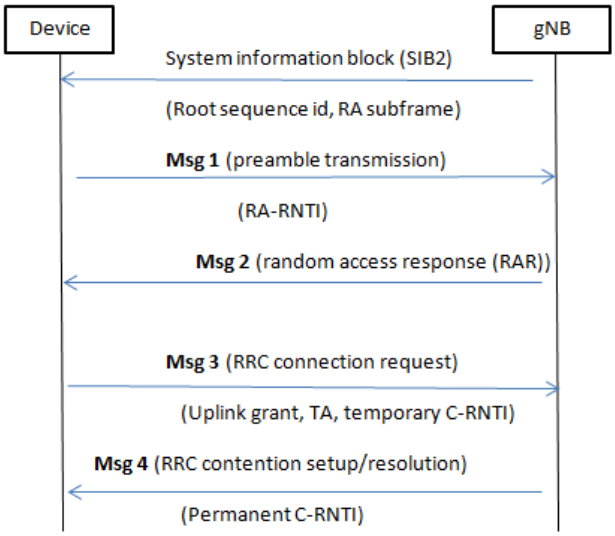

FIGURE 3.1: Standard LTE/LTE-A RACH procedure

Figure 3.1 describes the four-way, grant based, LTE RA procedure between user device and gNB, while figure 3.2 describes what occurs when two or more devices select the same preamble and collision occurs.

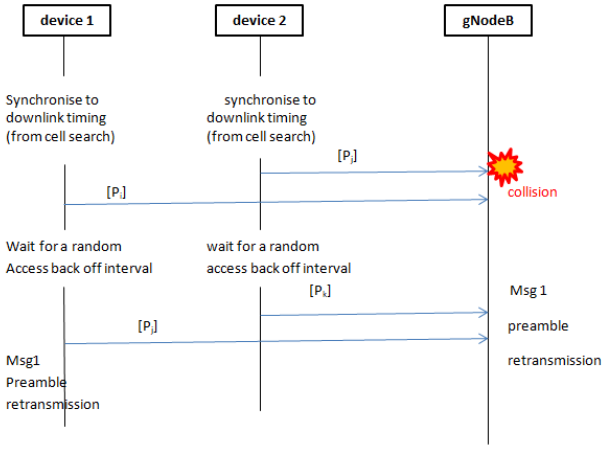

FIGURE 3.2: Collision event in Msg 1

3.2.4 Random Access Channel Congestion Control

Many solutions have been proposed in both industry and academia to tackle the issue of congestion in LTE-A networks in readiness for 5GB networks massive traffic expectations [86],[90],[6].

We therefore look at a few standardized solutions proposed to optimize RACH load balancing for access channel congestion control.

Dynamic Resource Allocation

In the LTE RA system, an increase in the number of RA opportunities (slots) leads to a decrease in the resources available for data transmission. Therefore in order to make way for massive access for 5G and beyond traffic types, dynamic resource allocation schemes through self-optimizing algorithms [42],[152], [96] and other novel mechanisms are required. On this note, our proposed scheme falls within this category of access channel congestion control

Access barring procedure

The idea behind this access control mechanism also known as Access Class Baring (ACB) is that the gNB broadcasts an access probability which is mapped to the access load on the system per time epoch and before every access attempt, the user device draws a random number usually between 0 and 1 [164]. If this selected number exceeds the access probability, the devices access attempt gets blocked, if it is below this probability, access is granted.

ACB is efficient in its ability to reduce collisions but also leads to much longer access delays which will be destructive to mission critical type traffic that require stringent delay figures.

Back-Off Schemes

As mentioned above, UEs with failed access attempts can refrain from re-attempting access in the next available RA opportunity. By delaying access through random backoff times, the access attempts are dispersed over time and the number of collisions is reduced. It follows that the random back-off time increases with an increase in network load.

Clustering

User device grouping or clustering as a way to monitor and control massive traffic access can be achieved by considering several cluster specifications such as the average distance between devices and gNB, Channel State Information (CSI), type of application, and so on. Each cluster or meshed area could be allocated a cluster head charged with exchanging information with the gNB on behalf of every device within the cluster, thus restricting the amount of access request and access failure rate as a direct consequence.

3.2.5 Random Access with Priority Class Partitioning

In systems with high collision rates and subsequent re-transmissions, RA delay is destructively affected. Therefore, the Msg 1 collision probability is the probability that a preamble collides with another preamble sequence, and is dependent on the number of available preamble sequences and terminal density in a cell.

We therefore analyse the successful access probability for the three traffic class partitions of which is used to evaluate random access delay. According to [38], the expected number of successful RA requests per PRACH slots is

$$P(N) = (1 - \frac{1}{P})^{N-1} \tag{3.1}$$

Where N are the number of devices attempting to access the gNB and P is the number of available preambles.

When the number of competing devices and available preambles are equal at the start of a Random Access Opportunity (RAO), maximum successful access probability is achieved. In order to meet delay requirements on the RA front, we propose to have 3 traffic classes with different preamble partition reservation policies which are:

- High priority traffic (T_H): This traffic class is dedicated to mission critical traffic or latency critical data, with stringent delay requirements such as remote surgeries, serious gaming applications, tactile internet apps etc. There will be reserved preambles for this class and contention will be within high priority traffic alone.
- Medium priority traffic (T_M): This traffic class is dedicated to medium priority level traffic which do not require as much precision in terms of delay as the T_H

class, e.g non-serious gaming, multimedia streaming, browsing, file transfer etc.

- Low priority traffic (T_L) This traffic class hosts the least priority level traffic

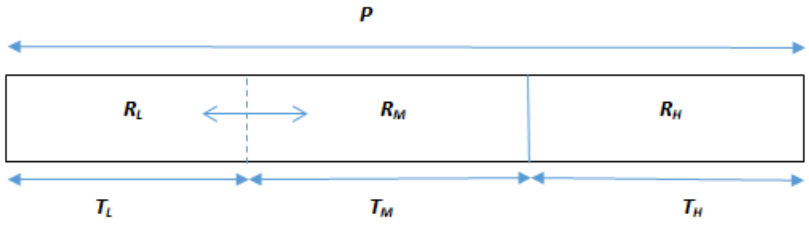

FIGURE 3.3: Traffic Based Preamble Resource Partition

Figure 3.3 illustrates the different traffic class preamble resource partitions where,

R_H: High priority preamble resource partition

R_M: Medium priority preamble resource partition

R_L: Low priority preamble resource partition

T_H: High Priority device traffic

T_M: Medium priority device traffic

T_L: Low priority device traffic and

P : Total number of RA preambles

We propose that the medium priority and low priority devices have dynamic access to their respective preamble partitions, given that some load balancing conditions are met.

Case Where T_L Traffic Can Share R_M Resources

In this condition, some T_L traffic (starting with the re-transmitted T_L traffic class preambles) share in the T_M class resources once the average number (Av_2 in this

case) of accessing T_M devices per Random Access Opportunity (RAO) is below a certain threshold th_2. Figure 3.4 below illustrates this relation

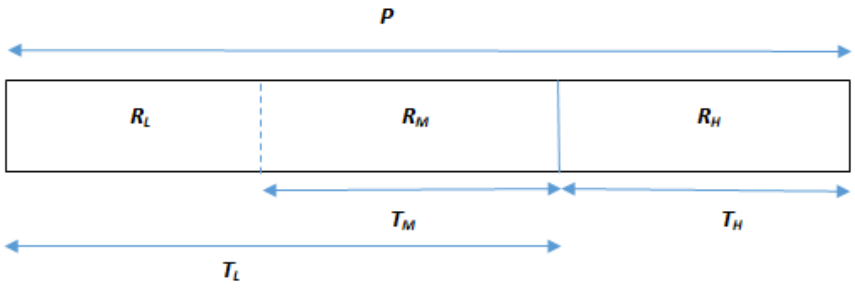

FIGURE 3.4: resource partition with condition 1

Given the mixed traffic case between medium and some low priority traffic, the access success probability for the T_L with reference to equation 3.1 and with respect to the reserved preambles for medium and low traffic classes can be written as;

Total successful access probability = successful access probability of low priority nodes within its preamble partition + successful access probability of some low priority nodes within medium priority partition

$$P_s^{T_L} = (1-\tfrac{1}{P_L})^{f_{T_L}-1}(\tfrac{R_L}{P_L}) + (1-\tfrac{1}{P_L})^{f_{T_L}-1}(\tfrac{R_M}{P_M}(1-\tfrac{1}{R_M})^{f_{T_M}})$$

$$= (1-\tfrac{1}{P_L})^{f_{T_L}-1}[\tfrac{R_M}{P_M}(1-\tfrac{1}{R_M})^{f_{T_M}} + \tfrac{R_L}{P_L}]$$

where f_{T_L} represents low priority traffic flows and $P_L = P-(R_H+R_M)$

$$P_s^{T_L} = (1-\frac{1}{P_L})^{f_{T_L}-1}[\frac{R_M}{P_M}(1-\frac{1}{R_M})^{f_{T_M}} + \frac{R_L}{P_L}] \tag{3.2}$$

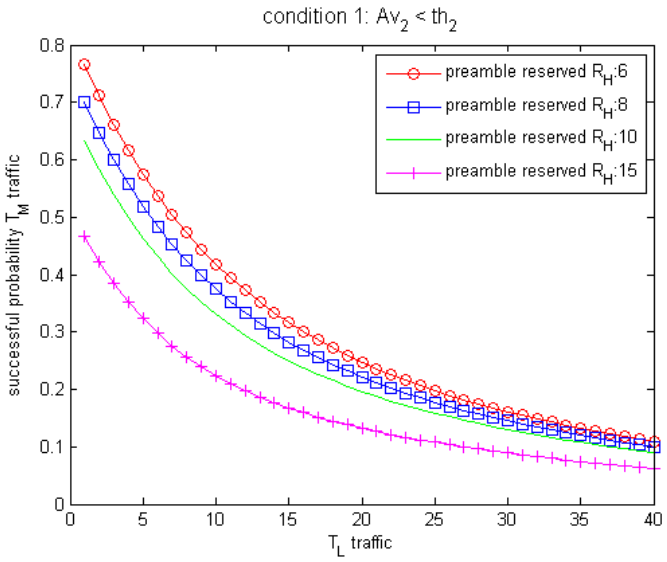

FIGURE 3.5: Successful access probability of T_L traffic within R_M partition for different R_H reserved preambles

Figure 3.5 shows the successful access probability performance of T_L traffic when sharing in the medium priority resource (R_M) for different high priority reserved preambles, analytically represented by equation 3.2. As is to be expected, T_L traffic access probability reduces with increase in R_H reserved preambles and vice-versa.

Case Where T_M Traffic Can Share R_L Resources

In this condition, some T_M traffic (starting with re-transmitted T_M traffic class preambles) share in the R_L class resources once the average number of accessing T_L devices per RAO is below a certain threshold th_1. Figure 3.6 below illustrates this relation

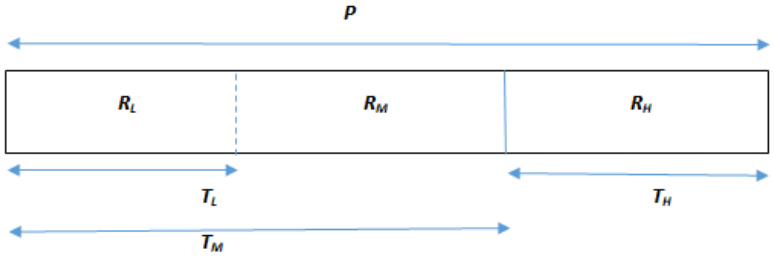

FIGURE 3.6: resource partition with condition 2

The successful access probability of T_M traffic within R_L partition is given as;

$$P_s^{T_M} = (1 - \frac{1}{P_a})^{f_{T_M}-1}[\frac{R_L}{P_a}(1 - \frac{1}{R_L})^{f_{T_L}} + \frac{R_M}{P_a}] \tag{3.3}$$

where f_{T_M} represents medium priority flows.

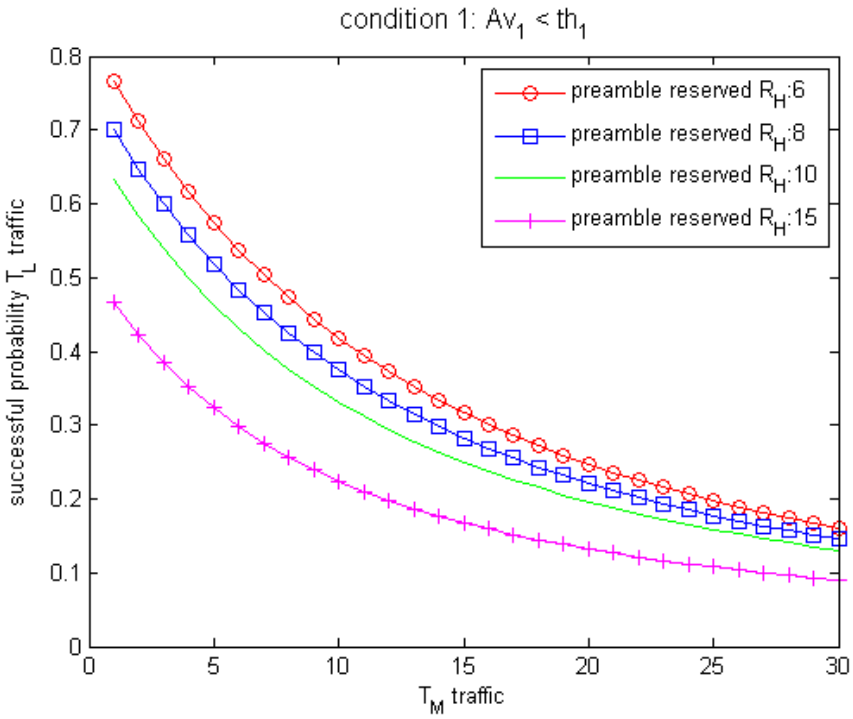

FIGURE 3.7: Successful access probability of T_L traffic within R_M partition for different R_H reserved preambles

Figure 3.7 shows the successful access probability performance of T_M traffic when sharing in low priority resource (R_L). Similar to figure 3.5, the access probability reduces with increase in high priority reserved preamble.

In both cases, the high priority devices will have dedicated resources R_H and hence will not compete with the T_M and T_L for resources. The successful access probability for T_H traffic can be written as

$$P^{T_H} = (1 - \frac{1}{R_H})^{f_{T_H}-1} \tag{3.4}$$

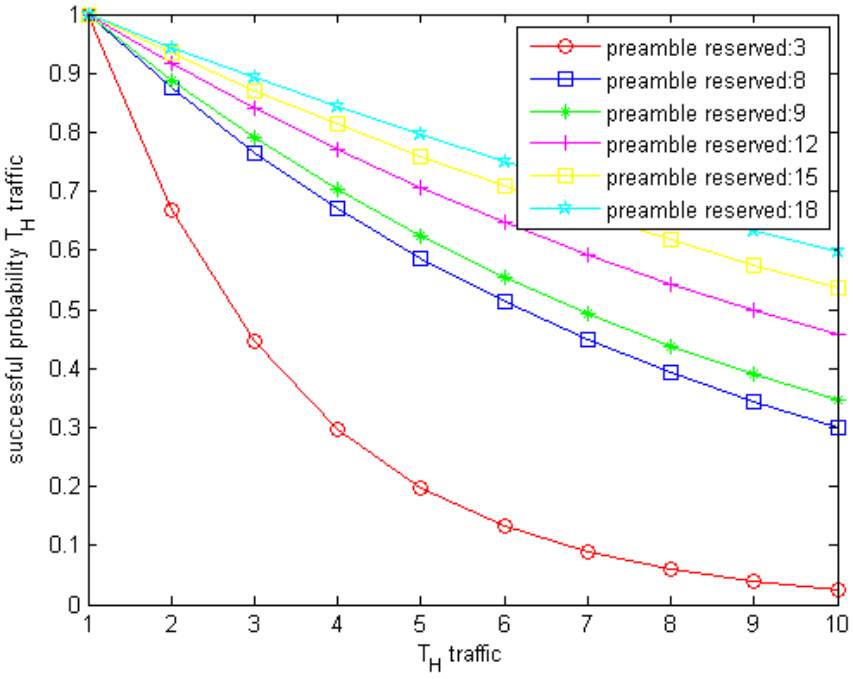

FIGURE 3.8: Successful access probability of Hp traffic with separate resource

Figure 3.8 shows the successful access probability performance of the high priority traffic (T_H) within its reserved preamble resource hence, the T_H requests will not compete with T_M and T_L traffic when a RAO appears.

Success probability can be seen to increase with increased reserved resource, with overall better performance compared to the T_M and T_L success probabilities.

Dynamic resource usage algorithm

The dynamic resource usage mechanism between the medium and low priority traffic priority effectively balances the load between these two traffic classes, optimizes the usage of preamble resource and serves as a fairness metric.

Our resource usage method is implemented in the following steps as depicted in Algorithm 1.

- **Step 1:** Initialize all parameters which include the new mid and low priority devices (N_{mid-p}, N_{low-p}) and the mid and low priority devices which previously failed the RACH procedure (F_{mid-p}, F_{low-p}).
- **Step 2:** For every System Information Block2 (SIB2) period (80ms) and Random Access Opportunity (RAO) period (5ms), compute the moving average

(AV1) of the number of devices in the mid-priority list and compare to dynamic schedulability threshold Th1.

- **Step 3:** Based on results from step 2, if AV1 is less than the dynamic schedulability threshold Th1, **then** schedule new mid-priority devices, RACH failed mid-priority devices and Msg 1 failed low-priority devices in that order.
- **Step 4: else**, schedule new and failed mid-priority devices.
- **Step 5: else-if** the moving average (AV2) of low-priority devices in the low priority list for each SIB2 period is less than the dynamic threshold Th2 for low-priority devices, then schedule new low-priority devices, RACH failed low-priority devices and Msg 1 failed mid-priority devices, in that order.
- **Step 6: else** , schedule new and failed low-priority devices. **end**

Algorithm 1: Dynamic preamble usage algorithm

Input:
$N_{mid-p}, N_{low-p}, F_{mid-p}, F_{low-p}$
Output: output the updated traffic profile N
while *SIB2 period* T_{SIB2}=*80 ms* **do**
  while *RAO = 5ms* **do**
    if $Av1 \leq Th1 == $ *'true'* **then**
      $N = N_{mid-p}, F_{mid-p}, F_{low-p}$;
    else
      $N = N_{mid-p}, F_{mid-p}$;
      else if $AV2 \leq Th2 == $ *'true'* **then**
        $N = N_{low-p}, F_{low-p}, F_{mid-p}$;
      else
        $N = N_{low-p}, F_{low-p}$;
      end
    end
  end
end

Numerical Results and Analysis

In this section, we evaluate the performance of our three-way resource partition scheme. We first determine the parameters threshold 1 (th_1) from which the access

of T_M traffic can use the R_L reserved preambles and threshold 2 (th_2) from which the T_L traffic can use the R_M preambles.

These values directly depend on PRACH collision values, the more collision, the more re-attempts and access delay. Similarly, the drop probability can be analyzed for re-transmission attempts with the maximum limitation N_{max} [164].

The collision probability P_c for combined T_M and T_L traffic is given as

$P_c = 1 - P_{acc}$

where P_{acc} is the access probability

$$P_c = 1 - (1 - \frac{1}{P - R_H})^{N_{eq}-1} \tag{3.5}$$

Where N_{eq} is the equivalent number of flows in the system, N_{TM} and N_{TL} are the flows for medium and low priority devices:

$N_{eq} = th_1 N_{T_M} + th_2 N_{T_L}$

The average delay T_m depends on both the PRACH duration T_{PRACH} and re-transmit duration in case of collision T_c, both values approximated by [3] for early data transmission (EDT) to be 10ms and 9ms.

The average delay T_m can therefore be computed via the the expected value E(t).

$$T_m = E(t) = \sum_{k=0}^{N_{max}} T_{PRACH} + kT_c(1 - P_c)P_c^k \tag{3.6}$$

Where, N_{max} is the maximum number of reattempts.

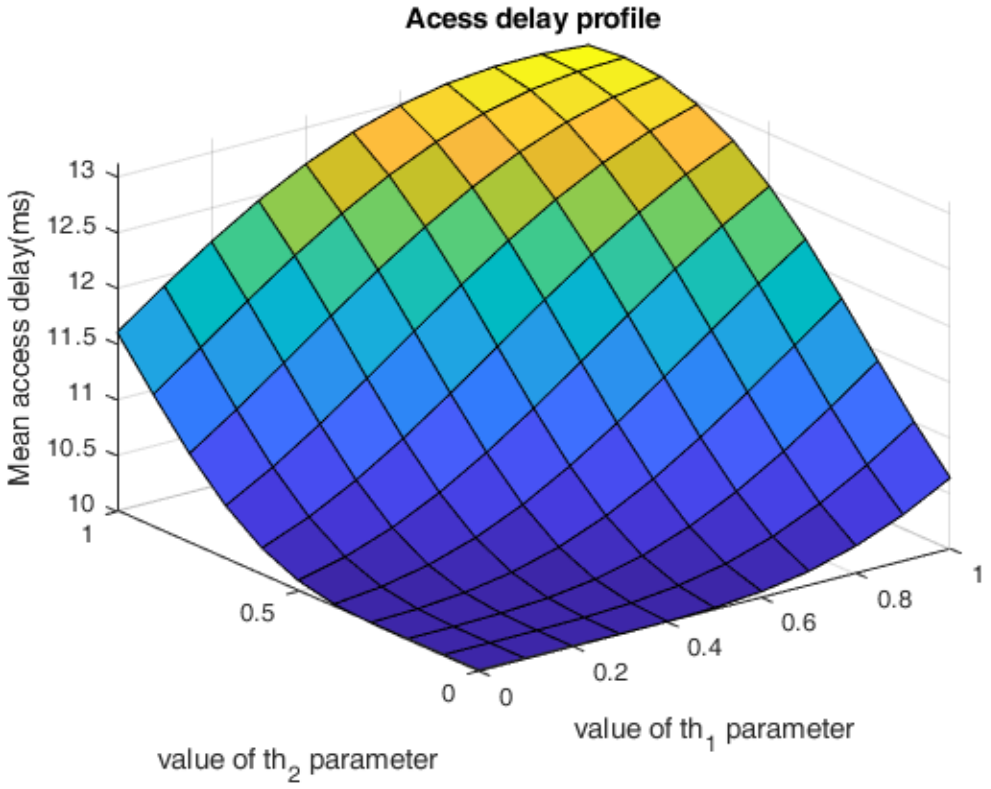

FIGURE 3.9: Average access delay

Packet drop rate is dependent on the maximum number of re-transmissions N_{max} as well as P_c and is given as

$$P_d = (P_c)^{N_{max}+1} \tag{3.7}$$

The th_1 and th_2 dynamic access adjustments can be obtained from figures 3.9 and 3.10 for the access delay and drop rate.

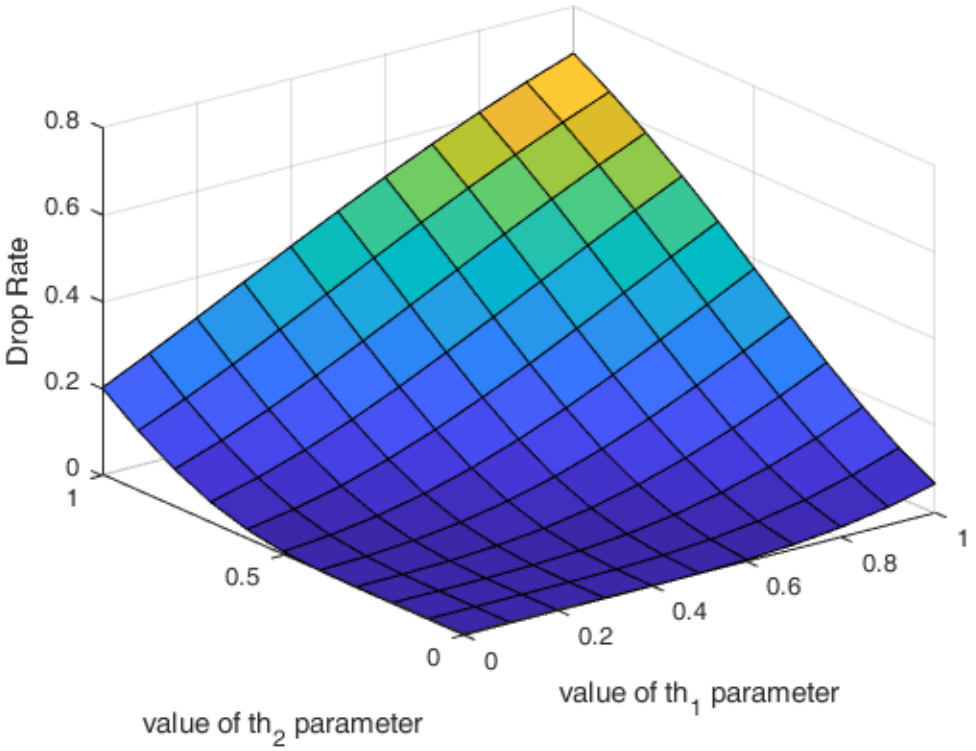

FIGURE 3.10: Drop Rate

From figures 3.9 and 3.10 we are able to appropriate values for th_1 and th_2 which determines at what condition the T_M traffic can access R_L preamble resource and vice-versa for th_2. It can also be seen from figures 3.9 and 3.10, that the values of th_1 and th_2 increase with increment in access delay and drop rate.

The conditions for dynamism between both traffic class resource partitions depend on the considered Quality of Service (QoS) threshold for each traffic class.

Considering a threshold of 10ms for average access delay and $\leq 1\%$ for drop rate, the values for th_1 and th_2 are displayed in table 3.1.

TABLE 3.1: Parameters th_1 and th_2

QoS threshold	T_M	T_L	th_1	th_2
T_m	10ms		0.3	
P_d		1%		0.5

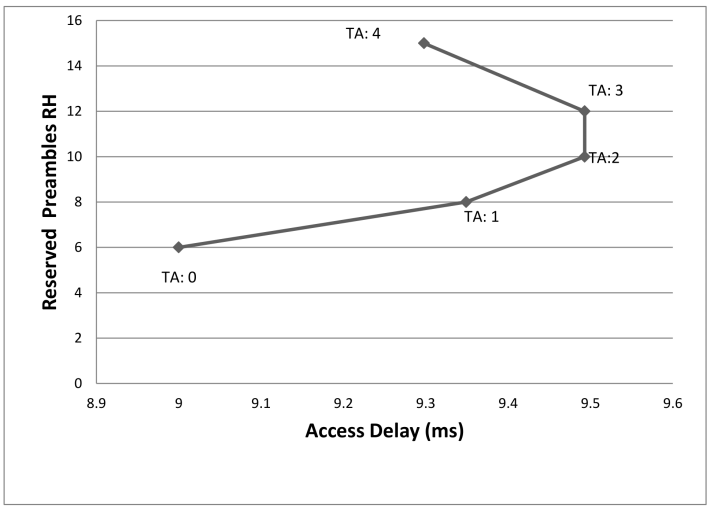

FIGURE 3.11: Impact of increased reserved preambles R_H and transmission attempts (TA) on access delay for T_M and T_L

Figures 3.11 and 3.12 show the access delay figures using our scheme for high, medium and low priority traffic with respect to various high priority preamble reservations. We make the assumption that both T_M and T_L traffic have the same amount of preambles for RA, and the total amount of preambles P, is 30.

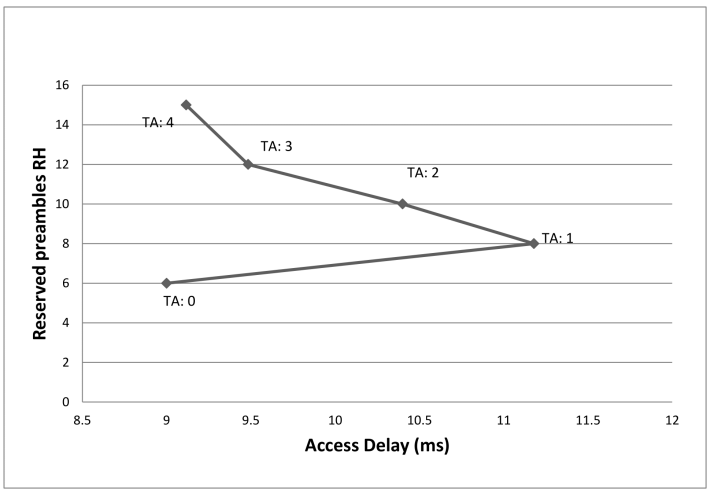

FIGURE 3.12: Impact of increased reserved preambles R_H and transmission attempts on access delay for T_H

Figure 3.13 shows the traffic access delay for high, medium and low priority classes in terms of R_H reserved preambles and number of re-transmissions.

The results show that the access delay performance for reserved preambles $R_H = 14$ performs the best across board, given that the number of transmission attempts are the highest, with the delay for T_H, T_M and T_L falling within the uRLLC delay requirement of 10ms in the control plane.

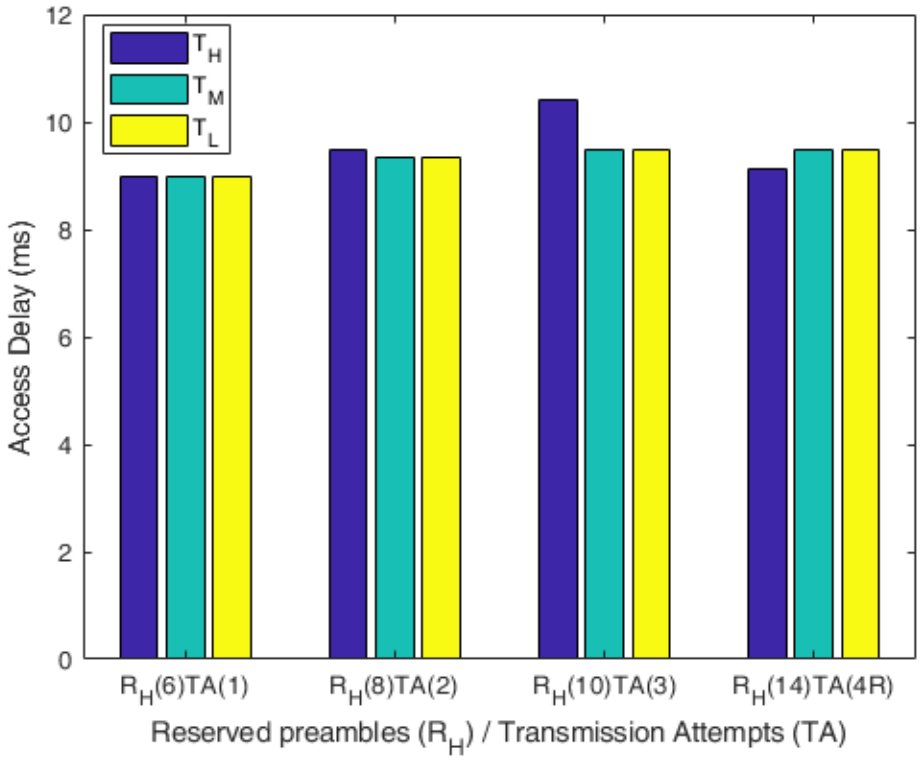

FIGURE 3.13: Number of reserved high priority preambles vs access delay for priority classes

In summary, for a mixed traffic case, with our proposed resource partitioning scheme, given 10% high priority uRLLC devices and 90% mid and low priority devices, figure 3.13 shows that by reserving 46% of preambles for high priority RA procedure, a control plane access delay profile of 9ms for high priority traffic and average of also 9ms for mid and lower priority traffic can be realised as compared to an average of 30ms in standard LTE.

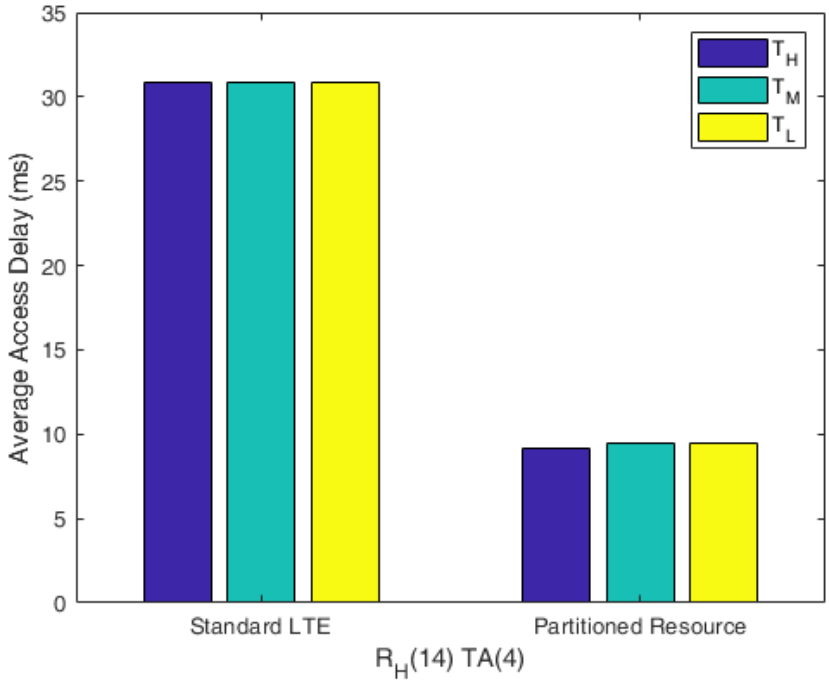

FIGURE 3.14: Access delay for proposed approach vs Standard LTE

3.3 CONCLUSIONS

With the ushering in of next generation wireless networks, research is focused on improving the latency figures with greater emphasis on URLLC applications via some kind of preamble reservation for this high priority class in RA.

In this chapter, we derive through analytical means,the optimum preamble partition for the reserved case which will also meet service requirements (average to high performance) by reducing the access delay for the non-URLLC traffic classes which is of particular interest for FoF environments.

We also propose a dynamic resource sharing mechanism which can exist between the medium and lower priority traffic, further reducing the delay for non-URLLC traffic. Our results show that the proposed scheme can meet the 5G URLLC latency requirements.

In the next chapter, we will present our novel scheme related to load balancing and RRH-BBU association.

Chapter 4

User association, Load balancing and Resource Usage in 5G Networks

In this Chapter our main contributions designed to improve User-RRH association, load balancing in heterogeneous cell clusters and RRH-BBU resource usage. Different from past works, in particular, our load balancing proposition has the originality to take into consideration the use of Almost Blank Subframes (ABS), an enhanced Inter-cell Interference Coordination (eICIC) technique proposed in LTE release 10 as a means to effectively balance device to cell association and hence traffic load among heterogeneous cells within similar clusters. The objective is to optimize user offloading among different cells, reduce transmission latencies fostering 5G uRLLCs and enhance network power consumption which can be achieved through reduced control signalling. We have also proposed a gamified solution to RRH-BBU resource usage. The performance of our proposed joint optimization scheme is evaluated through extensive simulations using MATLAB.

4.1 Introduction

Heterogeneous networks have been considered as a viable solution to maximize spectral efficiency through improved spectral re-use [118],[47],[87],[59] and also improve network performance by way of reduced connection latencies and improved data rates. In heterogeneous clusters, the macrocell provides basic coverage while the smaller cells are underlayed and serve as complimentary cells. This results in improved user off-load latencys as compared to legacy single cell deployments.

One of the major concerns for successful implementation of heterogeneous networks is the inter-cell interference by macrocells on User Equipment's (UEs) connected to smaller cells especially users at the cell edge also known as cross-tier interference, due to the large transmitting power difference [98]. Therefore, this Chapter elaborates on our work to address the problem of improving user offloading which is

an important aspect of multi-tiered heterogeneous networks [17], in addition to the issue of optimized resource usage in cloud based RAN architectures.

4.2 User Association and Load Balancing in HetNets

4.2.1 Context

As mentioned in previous chapters, heterogeneous networks have been considered as a viable solution to maximize spectral efficiency through improved spectral re-use [118], [46], [59] and also improve network performance by way of reduced connection latencies and improved data rates.

In order to choose the best network for a user device, a user association scheme is implemented. The association of the user device is carried out according to their service demands, channel quality and their distance from the eNodeBs. User association is important to improve the spectral efficiency (SE), energy efficiency (EE) and the load balancing in network sectors [92],[103].

One of the major concerns for successful implementation of heterogeneous networks is the sever interference by macro-cells on User Equipment's (UEs) connected to smaller cells especially users at the cell edge, due to the large transmitting power difference thus hindering efficient cell-UE load balancing and consequently the gains to be derived from dynamic UE offloading.

Due to the need for solutions to this problem, we propose an advancement to the standard handover (HO) decision control mechanism, adopting the use of an enhanced inter-cell interference coordination technique (eICIC) [43], as a means to effectively balance traffic load between RRH and UEs for a more efficient 5G network QoE.

4.2.2 Related Work

The efficient UE association, load balancing and resource allocation problem is one that is being addressed in both academia and industry. In this section we focus on current 5G optimisation techniques associated with these issues that have been proposed and implemented.

Authors in [59] propose a sequentially distributed coalition formation game to maximize the throughput in C-RAN architectures. Their proposed non-cooperative model incorporates a handover or power control interference scheme. The authors conclude that their proposed model produces better throughput and average interference per RRH in C-RAN systems compared to two other tested models.

A scheme for RRH clustering was proposed in the work of [28]. They tackle the BBU-RRH re-association which in a centralised C-RAN architecture with a fronthaul (FH) link constraint, can lead to costly HO's resulting in signalling overheads which further interrupts data flow, consequently reducing UE throughput. They proffer a solution by means of a joint centralisation and gamification approach. Their results demonstrated a good trade-off between power savings and re-association rate. In contrast to [28] and [101], our work tackled specifically the air interface excessive HO problem which would arise between high and low powered nodes in a multi-tiered network scenario.

The authors in [101] define a Linear Integer Program which selects virtual BBUs from different cloud services according to a price and failure probability metric. Authors in [16] modeled the optimal resource allocation between RRH and BBU in addition to a heuristic solution with less power consumption.

The joint optimization of almost blank sub-frame and user association was studied in [158] and [69], where each macrocell was assumed to have the same blank sub-frames, but the joint optimization in their work is combinatorial and users can only associate with one BS at a time. However, non of these works proposed a game-based resource allocation algorithm which is uniquely guaranteed to global optimally converge. Thus we propose in this chapter a new load balancing and user association scheme in addition to a BBU-RRH resource usage approach. Our solutions contribute to optimized average user throughput as well as energy conservation due to the reduction of large signalling overheads.

4.2.3 Enhanced Inter-Cell Interference Coordination for HetNets

In multi-tiered heterogeneous cellular network deployments with macrocells, femtocells or picocells users may suffer from significant co-channel cross-tier interference as seen in figure 4.1.

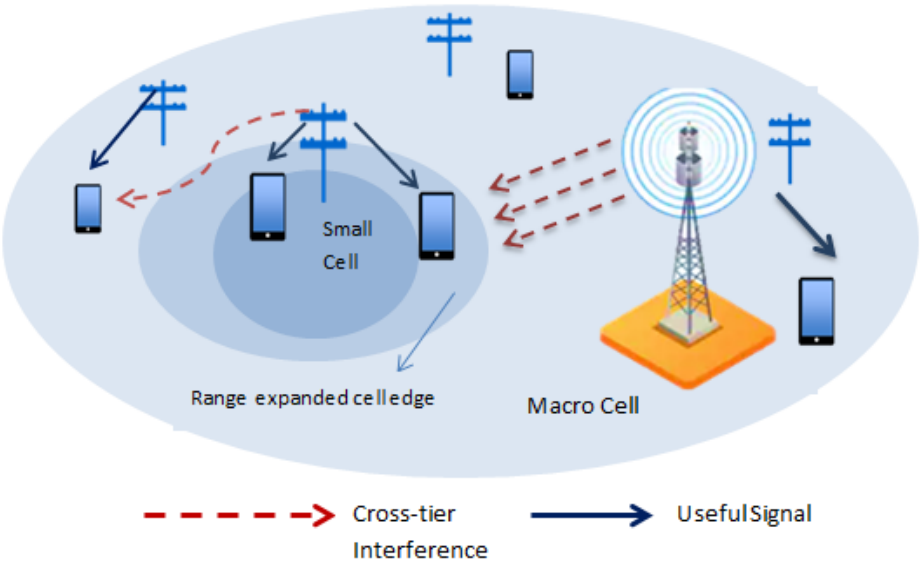

FIGURE 4.1: Cross-tier Interference in HetNet scenario

To manage this interference several proposals have been made in both industry and in research [32], [78], [71], [65], [154].

Almost Blank Subframes in Heterogeneous Cellular Networks

In a bid to address the need for increased capacity, particularly in more crowded network environments and at cell edges which experience significant network performance degradation, mobile operators are adding a lot of smaller cells together with macrocells to effectively spread traffic load among heterogeneous cells, maintain network quality of service and optimize network efficiency.

Small cells are majorly deployed to enhance network capacity in locations with high traffic activity and to fill in blind zones not covered by macrocells. They improve network service quality by encouraging traffic offloading from large high powered macrocells and therefore increase per unit area bit rates (see: Fig. 4.1)

In networks with frequency reuse of one (1), the user devices would normally attach to the cell with the strongest DL reference signal received power (RSRP) or strongest received DL signal (SSDL). An issue of major concern in heterogeneous network planning comes from the ability to ensure that smaller powered nodes serve an adequate number of user devices.

Therefore, in order to obtain more offloading to smaller powered nodes, reducing the scheduling load on macro-cells, and to improve connection latencies/performance gain for attached devices in addition to mitigating interference among cells in HetNets, 3GPP proposed the implementation of almost blank subframes (ABS), an enhanced inter-cell interference coordination (eICIC) technique [43] where only reference signals are transmitted through blanked subframes from the interferer tier and at-risk users get a chance to be scheduled in ABSs as a means to reduce cross-tier or cross-cell interference [120], [61], [117], [4].

Figure 4.2 describes a situation where 50% of DownLink (DL) macro cell subframes are silenced for the ABS. In this described case, macro cells do not schedule any UEs in even subframes, only in odd subframes, while UEs connected to small cells, more particularly UEs at the cell edge are allocated downlink (DL) resources in both [118], [61]. In addition to ABS, cell range expansion (CRE) techniques are used to improve cell coverage for UEs connected to smaller cells. In CRE, a positive bias is added to the measured signal strength of a small cell within range of a macrocell in order to shield UEs connected to the smaller cells from interference from high powered macro cells as shown in figure 4.1 [97].

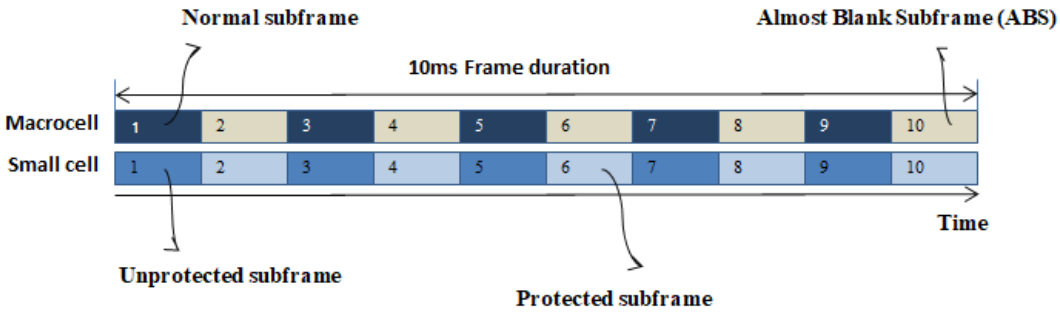

FIGURE 4.2: Almost Blank Subframes for range expansion in heterogeneous networks

4.3 Dynamic Load Balancing Scheme for 5G HetNets

User association schemes in HetNets highly depend on the load balancing which is an important part of multi-tiered HetNets [17] in addition to optimization metrics, applied traffic models and network distribution.

Given a multi-tiered network which adopts the ABS mechanism to encourage balanced user offloading and discourage inter-cell interference, if the ratio of ABS to non-ABS is low, the average throughput of offloaded users might be lower due to heavily loaded users within ABSs and vice-versa. As a result, mechanisms to dynamically adapt this ABS ratio needs to be developed to cope with the massive traffic load associated with networks of the future.

In a bid to address this issue, we propose an advancement to the standard handover (HO) decision control mechanism based on our dynamic ABS ratio scheme which is a function of the control signalling overhead component and aggregated HO rates for each macro cell. The cell selected is chosen among those which give better rate which largely depends on Signal to Interference and Noise Ratio (SINR) values which we compute below.

4.3.1 System Model

Consider the downlink data transmission of a multi-tiered network, the coverage area is hexagonally meshed into many discrete zones with each zone consisting of both small and macro-cells. Therefore a zone is represented by a number of UEs $u \in U$ and high to average levels of radio conditions. A zone z, is associated to a mix of macro-cells denoted as, m $\in M$ and small cells s $\in S$. The $SINR_{c,s,u}$ of UE u, associated to small cell s, which is mapped to serving BBU cluster $c \in C$ (where c is an index of BBU cluster C) can be represented as:

$$SINR_{c,s,u} = \frac{G_{s,u}.P_s}{N_u + \sum P_m.G_{m,u}.\varphi(n)} \tag{4.1}$$

Where,

P_s : The power emitted by the serving small cell s

$G_{s,u}$: The channel gain from small cell s to UE u

N_u : Noise power of UE u

P_m : The power emitted by macrocell m

$G_{m,u}$ The channel gain from macrocell m to UE u

$$\varphi(n) = \begin{cases} 0, & if\ subframe\ n\ is\ an\ ABS \\ 1, & otherwise \end{cases} \tag{4.2}$$

Similarly, when UE u selects macrocell m, mapped to BBU cluster c, its $SINR_{c,m,u}$ would be expressed as:

$$SINR_{c,m,u} = \frac{G_{m,u}.P_m}{N_u + \sum P_s.G_{s,u}} \tag{4.3}$$

The expected peak throughput when the UE $u \in U$ selects a macrocell $m \in M$ mapped to BBU c, can be expressed as:

$$E\{R_u^m\} = BW.(1-\alpha)log_2(1+SINR_{m,u}) \tag{4.4}$$

Where α is the ABS ratio of the macrocell and BW is the total bandwidth of the BBU mapped to the sector or sector cluster which contains macrocell m. Similarly, the expected throughput when the UE $u \in U$ selects a smallcell $s \in S$ mapped to BBU c, can be expressed as:

$$E\{R_u^s\} = BW.(1-\alpha)log_2(1+SINR_{s,u^{non-ABS}}) + \alpha log_2(1+SINR_{s,u^{ABS}}) \tag{4.5}$$

The effective SINR is dynamically mapped to the corresponding Modulation and Coding Scheme (MCS) in table 2.4 (referenced in Chapter 2) which affects the resulting expected bit-rate.

After comparing expected data rate of each cell, the cell of closest proximity to the UE and which gives higher expected throughput will be chosen by the UE as defined in [118].

$$i^* = \arg\max_{i \in M \cup S} E\{R_s^i\} \tag{4.6}$$

Area decomposition

Our system model hinges on the classical meshing method [145]. The covered area is meshed into 7 hexagonal zones with maximum cell radius of 250 meters (m) as shown in Fig. 4.3. Devices with different service requirements are uniformly distributed within the cell with 6 small cells and 20 UEs per macrocell. We adopt Cost Hata 231 to model the pathloss between cells/RRHs and UEs which covers a more robust range of frequencies for urban environments associated with 5G networks.

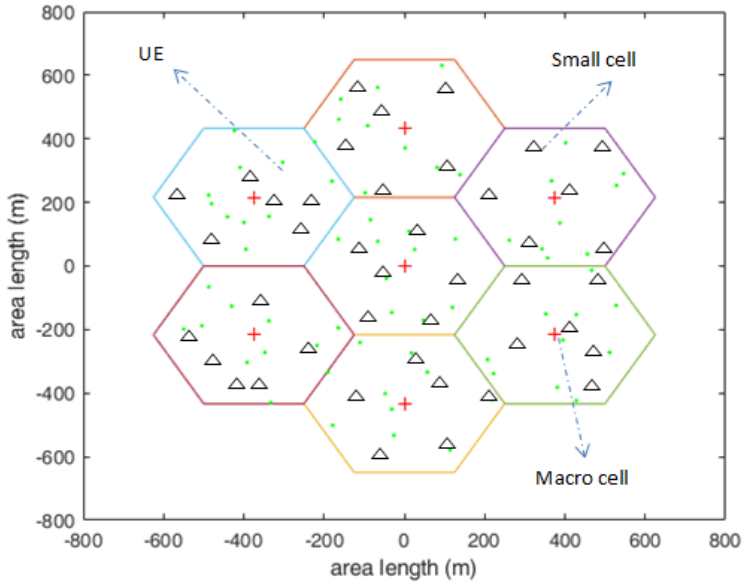

FIGURE 4.3: Meshed heterogeneous cell layout

4.3.2 Dynamic ABS ratio

A large number of smaller cells within a sector containing macrocells will create a large number of outbound handovers (HO's) majorly for UEs at the edge of small cells due to the stronger Reference Signal Received Power (RSRP) transmitted by the larger macro cells, in addition to destructive massive signalling flows and macro cell congestion as a direct consequence.

In coalition with the ABS approach, we propose to balance the UE load per cell and meet the QoS requirements related to user throughput, reduced signalling overhead due to excessive/unnecessary HO's and optimised network power consumption

by dynamically adjusting the ABS duration based off the number of HO requests received by a macro cell from UE's attached to nearby smaller cells every subframe duration.This mechanism is expected to achieve a good balance between traffic load balancing, traffic scheduling latencies and throughput maximization.

We therefore make the assumption that for every heterogeneous system $S \cup M \in r$, each sector contains an average number of UEs with average radio conditions. Therefore, the ABS ratio α in equations 4.4, 4.5 and 4.6 increases or reduces subject to a HO threshold θ where if it is surpassed the ABS duration will increase, consequently balancing any excessive offloading of UEs to macrocells from smaller nodes in addition to reduced HO air-interface signaling overhead and power consumption.

As a result, $i^* = \arg\max_{i \in M \cup S} E\{R_s^i\}$

is maximized subject to (s.t.):

$$\sum_{m=1}^{M} \theta_m^{HO} \geq R_h, \quad \forall m \in M \tag{4.7}$$

where R_h denotes the rate of handover UEs from s to m.This procedure is described in Algorithm 2.

Dynamic Load balancing algorithm

Algorithm 2 describes our dynamic load balancing algorithm which we derive using the ABS mechanism

- **Step 1:** Initialize total number of UEs, small cells ad macrocells. Set the value for handover threshold θ.
- **Step 2:** For every subframe period of 1ms and for every connected user device to both macro and small cells solve equation 4.6.
- **Step 3:** For every UE connecting to a macrocell within cell layout, compute equation 4.7.
- **Step 4: if** the hand-over (HO) request rate to macrocells (excluding mobility cases) from UEs connected to smaller surrounding nodes is greater than HO threshold (θ) then increase current almost blank subframe (abs) to non-almost blank subframe (non-abs) ratio by abs adjustment factor B.

- **Step 5: else-if** θ is greater than UE HO rate (R_h), decrease current abs to non-abs ratio by adjustment factor B.
- **Step 6: else** , current abs to non-abs ratio remains fixed.
- **Step 7:** Output abs to non-abs ratio. **end**

Algorithm 2: Dynamic ABS ratio for Handover Decision Control, Load Balancing and Power savings

Input:
u = { 1, 2, ..., U }: Total number of UEs
s = { 1, 2, ..., S }: Total number of small cells
m = { 1, 2, ..., M }: Total number of macro-cells
R_h, θ, B(almost blank subframe)
Output: i^*: The cell expected to give higher expected data rate.
While Subframe period $T_s = 1ms$ **do**
 for $s \cup m \in r$ **do**
  Solve equation (4.6)
  Calculate i^* from equation (4.6)
 end for
 for $m \in M$ **do**
  Solve equation (4.7)
  if $\theta R_h ==$ 'true' **then**
   $abs = abs + B$;
  else if $\theta R_h ==$ 'true' **then**
   $abs = abs - B$
  else
   $abs = abs$;
   Output abs
  end if
 end for
end While

4.4 A Gamified RRH-BBU Association Scheme

In this section, we formulate a cooperative game to model the function of a cloud based centralised controller in the cooperation between individual BBU clusters and their associated cells in the network side, connected through FH links, this description can be seen in figure 4.4. The BBU-RRH mapping as described in figure 4.4 can be one-to-one or one-to-many mapping.

Some BBU-RRH association optimization strategies in existence are categorized as follows;

Graph based schemes

in this BBU-RRH mapping approach, RRHs are represented as clusters (nodes) which can be associated to a single BBU or a collection of BBU's. In [111] a clustering algorithm was developed to reduce the HO rate of user devices between cells.

Results from graph based formulations proposed in [141] have RRH's belonging to the same clusters linked to each other through edges associated to similar BBU's. This is repeated till the primary BBU's resources are completely depleted.

In [150], the authors propose a graph-based scheme which optimize energy efficiency as compared to Fractional Frequency Reuse (FFR) [156] and reuse-1 schemes.

Knapsack approach

Not many works develop a BB-RRH association problem as a knapsack problem [104]. This graph based approach category consists of combining the most important objects into a set of fixed capacity bins. Therefore objects are mapped by value to their respective weights.

A few works which adopt this principle to different degrees include [100], [99] and [147].

Bin Packing method

Arguably one of the more widely adopted formulations in current research [79], the bin packing method takes into consideration two finite sets of bins and objects. The bins in this case are a fixed capacity while the objects are of variable volumes. The goal is to optimize the object-bin utilization in a way to decrease bin usage.

Relating this to BBU-RRH association, the BBU's are modeled as fixed capacity bins, while the RRHs are the objects [109], [26], [148], [75]. Works in [33],[127], [137], [12] and [109] utilize this approach in their formulations.

Evolutionary methods

This comprises of providing sufficient enough results to optimization problems by sampling a subset of solutions and applying iterative searches to them until a feasible optimal solution is reached. works utilizing this meta heuristics based approach can be found in [76], [77] and [44].

4.4.1 BBU-RRH Mapping and resource usage approach

The interaction between cloud based BBU clusters (players) that assign baseband processing resources is modeled. The major components of the C-RAN architecture are displayed in Fig. 4.4 with RRHs of different color schemes mapped to BBU clusters of the same coloring showing their cell coverage domain.

Each physical machine is representative of a centralized BBU pool containing individual BBUs of different capacities. The structural mapping between cells and BBUs can be one BBU pool to one RRH/cell zone or one BBU pool to many RRHs/cell zones.

Each BBU entity possesses computing capacities with a finite number of heterogeneous cells serving user devices. The BBU pool processing capacity is the total computing resource required by the total cells in the associated/mapped network system.

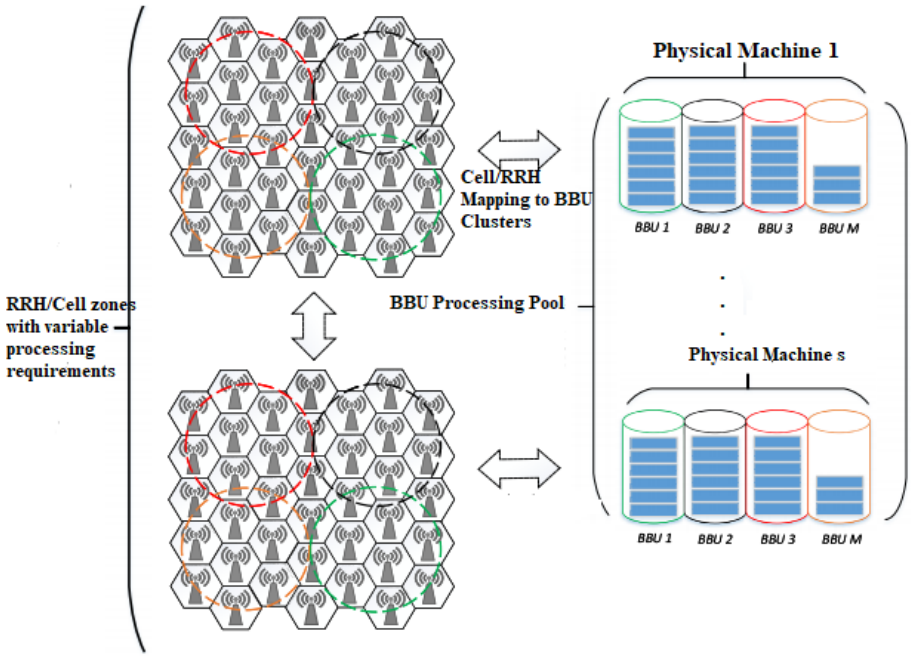

FIGURE 4.4: BBU-RRH/cell mapping

Our game $\supset = [N, \mathrm{S}, \{U_i\}_{i \in N}]$ comprises of the following components:

- The set of $N = \{1, ..., N\}$ players which are BBU clusters in the cloud tasked at processing network cell traffic flows
- The Strategy profile space S where a strategy profile S, will consist of a centralised controller optimally assigning processing resources to scheduled network flows from linked cells/RRHs in a way to ensure QoS metrics are maintained.
- The cooperatives game Utility function U_i

We assume K available multi-channel fronthaul (FH) links mapped to BBU clusters N at any given time epoch t.

let $\mathbf{A} \in 0, 1^{N \times K}$ be the cloud processing resource assignment matrix whose element $a_{i,k,z} \in \{0, 1\}$ equals 1 when flows from cell/RRH zone z, in FH link k mapped to BBU cluster i is scheduled, and 0 otherwise.

Player i's strategy is given by $S_i = a_i^T$, the $1 \times K$ *ith* row vector extracted from **A**. Therefore, every cell/RRH zone mapped to a BBU cluster is represented by a vector. Where,

$$a_i^T = \begin{bmatrix} a_{i1} & a_{i2} & .... & a_{iK} \end{bmatrix}$$

The games strategy space S is equal to $\{0,1\}^{N \times K}$

Utility Function

The management and control system in the game seeks to maximize this QoS performance metric given by

$$Q_{i,k,z} = \frac{\alpha_{i,k,z}}{\beta_{i,k,z}} \sum_{k=1}^{K} a_{i,k,z} \frac{R}{l_{i,k}}, \ \forall i \in N \tag{4.8}$$

Where

- $\alpha_{i,k,z}$ is the binary connectivity decision coefficient.If the controller maps cell zone z to BBU cluster i, $\alpha_{i,k,z} = 1$ otherwise $\alpha_{i,k,z} = 0$
- $\beta_{i,k,z}$ the hop count along FH link from cell zone z, to BBU processing cluster i.
- R is the FH link data rate, defined by the chosen optimal modulation and coding scheme (MCS) (refer to Table 2.4) and also the link type.
- $l_{i,k}$ counts the number of interfering cells sharing link k with cell zone z, mapped to BBU cluster i (player).

As the players are cooperative, a common network objective (Utility function) is defined which is jointly maximized among all players and given as:

$$U_i(S) = P(S) = \sum_{1=1}^{N} Q_i \ \ \forall i \in N \tag{4.9}$$

Where

$U_i(S)$ and $P(S)$ are the utility and potential functions of our cooperative game, $\Im$.

Our strategic resulting game is a game of identical interest and an exact potential game with potential function P(s) equal to utility function $U_i(S)$. Such games are guaranteed to converge to a Nash Equilibrium (NE) using well known best and better response dynamics [82].

The Nash Equilibrium concept

A Nash equilibrium (NE) is an key notion in predicting a game's outcome. By definition, it represents a strategy profile such that if opponent strategies remain unchanged, no player would gain from moving away from their current strategy. Nash Equilibrium is widely considered as the most important concept in game theory application.

From a cellular systems context, when a NE is reached, no player would gain from deviating from its current position and can be viewed as a "stable operating point" [82]. Therefore, in an optimized resource allocation game, attaining Nash Equilibrium is the ideal system goal and has an overall optimized effect on system bit rate.

Best Response Optimization

Similar to the algorithm described in [28], the RRH chooses the BBU that provides the highest utility (4.9).

A local optimum response is when all BBU clusters choose the same strategies as in the previous round. The best response is described in algorithm 3.

Algorithm 3: Stage 2: Best Response

1 Initialize all vectors $S_{i\in N(0)}$ of pure strategies;
2 repeat
3 Each active BBU cluster is made to selects a strategy Q^* that respects $Q^* = \arg\max_{I\in N} U_i(eqn.4.8)$;
4 Actions $a_{i\in N(i)}$ of each instantiated BBU cluster is updated;
5 Until Attaining NE

A centralized control host manager located where the BBU resides in the cloud, provides a suitable configuration for BBU activation and deactivation according to realized average user throughput per cluster.

A positive user throughput is easily a function of effectively balanced UE association to individual small and macro cells as provided in our solution in (4.6). If a minimum average throughput and user connection latency per cluster is not realized, the algorithm will activate more computational resources by way of additional processing resource, and reverts back to the start of algorithm 2 in order to set new strategies.

4.4.2 Performance analysis

To verify the performance of our proposed load balancing/cell selection and BBU-RRH association scheme, we use MATLAB. The simulation results aim to compare our novel joint optimization scheme to the base maximum Reference Signal Received Power (RSRP) based scheme. The results obtained reveal the ability of our proposed scheme to reduce scheduling latencies and improve average throughput values.

The simulation parameters are shown in table 4.1.

TABLE 4.1: SIMULATION PARAMETERS

Parameter	**Value**
Scheduling Scheme	Proportional fairness
Cell Layout	7 macro cells and 6 small cells per MC
UE number	20 (uniformly distributed)
propagation model	Cost Hata 231
Shadowing Standard Deviation	6dB (Macro cell) 4dB (small cell)
Thermal Noise Power	-174dBm/Hz
Meshing Step(a)	100 m
Cell Radius	200 m (Macro cell)
Transmit Power of eNB/sc	46dBM/30dBm

We evaluate the performance of the standard Max RSRP centralized cell selection in comparison to our adaptive load balancing scheme. The results are obtained for the aggregate signalling rate which is a function of the HO rate.

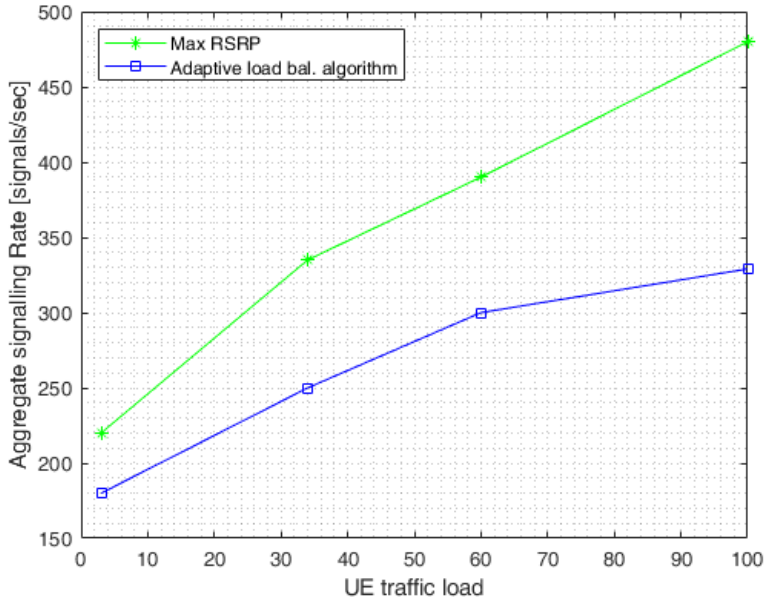

FIGURE 4.5: HO rate Cost: Impact on standard Max RSRP vs Adaptive load bal. Algorithm

It can be seen from fig. 4.5 that our scheme shows improvement due to the threshold our algorithm placed on excessive HOs, thereby reducing uplink (UL) interference caused by high cell border crossing from small to macro cells.

Figure 4.6 shows the packet loss for different traffic load volumes for the Max RSRP and our Adaptive load balancing and scheduling algorithm. Our algorithm performs similarly to Max RSRP for UE traffic loads from 0 to 30 due to the under load conditions. When the traffic load increases from 30 onwards, our scheme outperforms the standard Max RSRP due to our methods robust user offloading onto smaller cells. It is found that there is an approximately 2% improvement from our scheme compared to standard Max RSRP.

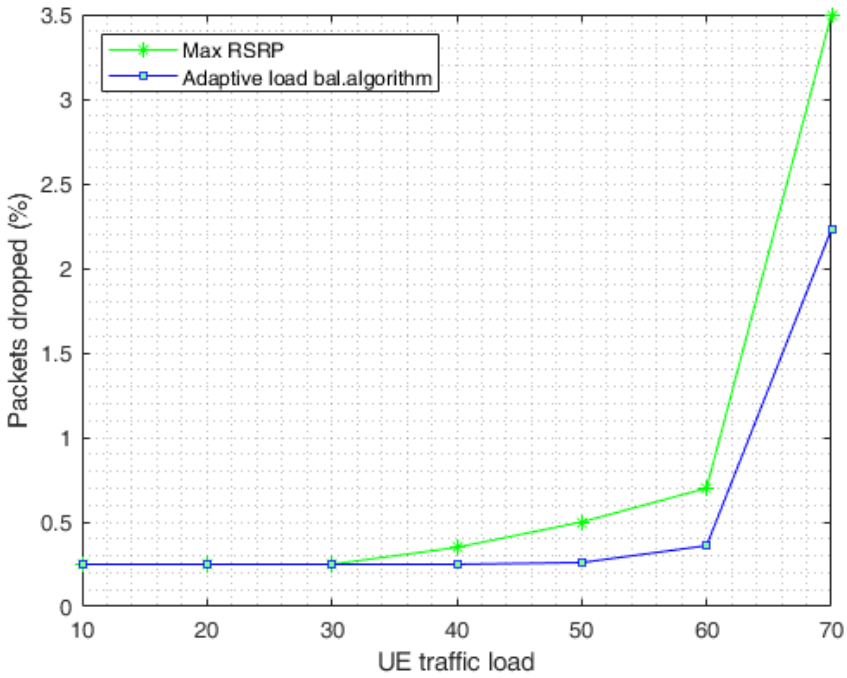

FIGURE 4.6: packet loss rate per traffic load for standard Max RSRP vs Adaptive load bal. Algorithm

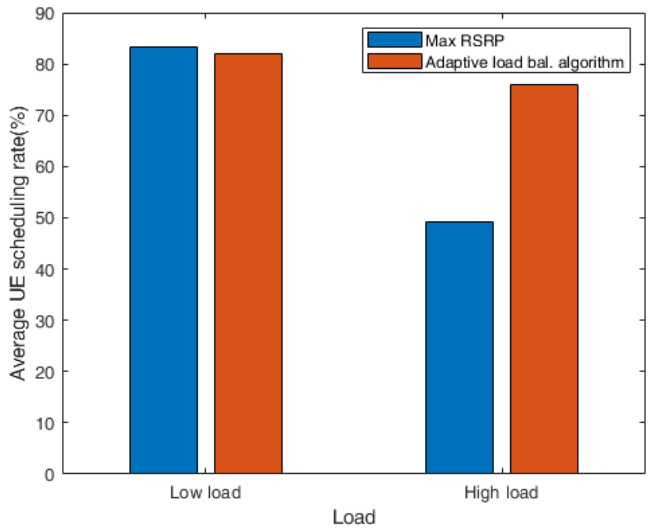

FIGURE 4.7: schedule response rate: Existing Max RSRP vs. Adaptive cell association with load balancing

Fig.4.7 shows the average user scheduling rate per cell cluster. This is computed for low and high user load conditions. Our proposition is denoted as Adaptive load balancing algorithm.It can be seen that our adaptive algorithm and the existing Max RSRP realized similar scheduling rates for low load conditions.

The adaptive scheme however, shows a much better response for higher load

conditions, this can be attributed to the fact that each cell per sector is effectively utilized as a result of the load balancing component which discourages unnecessary HOs from smaller cells to macro cells.

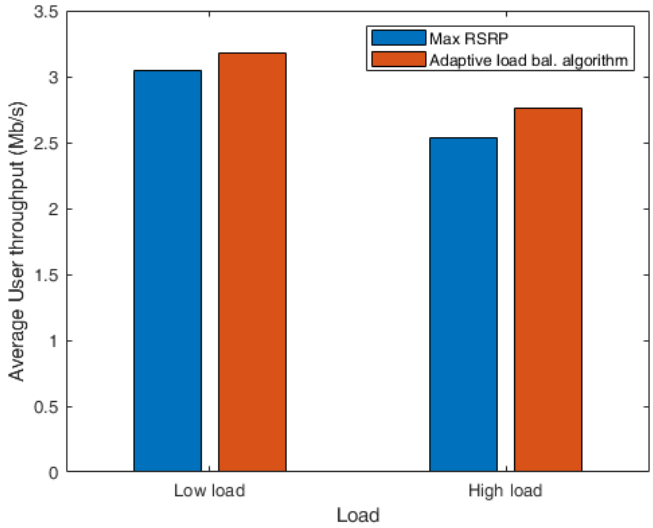

FIGURE 4.8: Average user throughput per UE: Existing Max RSRP vs. Adaptive cell association with load balancing

Figure 4.8 displays the results of average user throughput for different user load conditions. The proposed scheme outperforms the existing scheme due to the adaptive ABS ratio and load balancing component in both low load and high load conditions.

4.5 Conclusions

In this Chapter, an optimized scheme for load balancing and resource usage in 5G networks is proposed. The approach stems from an adaptive mechanism which optimizes user throughput and simultaneously balances user load among cells in the network side, in addition to a centralized application which manages BBU computational resources as a function of balanced load.

The results are compared to the standard existing baseline approach and shows that the performance of the adaptive scheme significantly improves network performance in terms of the number of offloaded users and average user rate.

In the next chapter, we present our novel strategy related to efficient differential traffic scheduling along the C-RAN based fronthaul network in 5G ans beyond heterogeneous networks.

Chapter 5

Time-Sensitive Traffic Scheduling for 5G and beyond Fronthaul Networks

In this chapter, our major contribution is towards the design of a novel switch node architecture to optimize scheduling for differential traffic flows with diverse QoS requirements along the additional Fronthaul Network introduced by the 5G network C-RAN. Our contribution utilizes the IEEE 802.1 Time Sensitive Networking Technology (TSN) in particular, the 802.1 cm Burst Limiting Shaper (BLS) scheduler mechanism of which we optimize in order to account for more robust data rates and latency figures which would enable the better adoption of novel mechanisms such as Network Slicing, and will also aid to meet the stringent delay requirements of applications for heightened QoE for the end user. We evaluate our contributions ability to provide guaranteed scheduling performance for both high and low priority traffic in the region of ultra-low latency and achievable rates for the fronthaul network.

5.1 Introduction

To meet the throughput, jitter and latency stringent requirements of 5G and beyond fronthaul networks, quite a number of solutions have been explored which revolve around deterministic communications as well as Ethernet-based packet switched solutions [22].

Some of the more promising 5G transport layer solutions have been proposed by well known standardization bodies such as IEEE, ITU and 3GPP. Our main focus in this chapter however, is on the IEEE 802.1 working groups TSN technology, and in particular their event-triggered Burst Limiting Shaper mechanism [143], [60], [144].

5.2 A novel Dynamic Reserved Capacity/BLS algorithm for improved Fronthaul Traffic Scheduling

5.2.1 Context

The link between user devices (UEs) and radio equipment controllers is known as the fronthaul (FH) in 5G Centralized Radio Access Networks (C-RAN) architecture has very stringent requirements which are quite challenging to meet. As previously mentioned in the introduction to this chapter, various research activities bordering around Time Sensitive Networking (TSN) technology to combat these challenges have been proposed in literature [18], [89], [91], [119]. One of particular interest to us is the credit based shaping IEEE 802.1Qav mechanism [60]. In addition to the connection latency requirement of traffic flows along the FH network that have to be met, there is the problem of allocation of limited radio resource in the system in an optimized fashion for all traffic flows. These scenarios present a challenge for existing scheduling algorithms and solutions suitable for the C-RAN fronthaul context in particular are to be developed.

Therefore, in this chapter we present the final contribution of this thesis by way of a new scheduling algorithm for the fronthaul of 5G and beyond networks which takes into consideration the efficient use of scarce radio resource and improves traffic scheduling for both TSN high priority flows as well as lower priority flows in traffic underload and overload scenarios. We implement our proposal as a scheduling modification to the TSN/BLS mechanism on top of a non-preemptive strict priority WRR scheduler together with our analysis.

5.2.2 Related Work

Full-fledged solutions have been proposed in both academia and the industry to tackle the problem of meeting 5G and beyond transport layer requirements. However, only a few of these contributions have addressed the issue of improving C-RAN fronthaul network scheduling for mixed criticality traffic flows.

The authors in [19] propose a core router traffic scheduling architecture used to differentiate traffic flows with diverse QoS demands. They employ a rate scheduler (e.g WFQ) between the default (DE) and assured forwarding (AF) traffic classes to share the unused capacity of the highest priority class. The downside to this method is highlighted in the absence of resource isolation for the lower priority traffic flows in the event of expedited forwarding (EF) traffic fluctuations.

In [52], the authors utilize the credit-based BLS to improve core network scheduling for assured forwarding traffic with respect to expedited forwarding class of traffic. They demonstrated that their scheme could isolate the AF class output rate in spite of traffic variations in the EF class. However, this strategy did not address the impact of limited resource for best- effort traffic in the presence of expedited forwarding traffic overload scenarios.

The authors in [41] study the delay and jitter figures associated with employing Common Public Radio Interface (CPRI) on Ethernet FH links for traffic scheduling. Through simulations they were able to show that the delay associated with high priority delay stringent traffic can be significantly reduced. However, their focus was solely on optimization for critical high priority traffic without ensuring fairness in resource distribution. In comparison to the mentioned works, we provide a unique means of addressing the trade-off between meeting the latency requirements of high-priority fronthaul traffic through resource reservation and the resource usage efficiency.

5.2.3 IEEE Time Sensitive Networking

Table 5.1 below shows the IEEE TSN standards in their various cathegories. Flow management aspects of the standard are covered in 802.1Qcp , 802.1Qat, 802.Qcc and 802.1CS. TSN flow synchronization is covered in IEEE 802.1AS and 802.1AS-Rev. Flow control and delay guarantees are addressed in TSN mechanisms like Credit Based Shaping (CBS) IEEE 802.1Qav [60], frame scheduler IEEE 802.1Qbv, frame preemption IEEE 802.1Qbu and cyclic queuing IEEE 802.1Qch [113].

These industry standardized mechanisms define how frames belonging to different traffic classes with different priority tags are processed in TSN enabled networks.

5.2.4 TSN BLS for 5G Fronthaul

The TSN Burst Limiting Shaper (BLS) is presented in this section, then our proposed Dynamic Reserved Capacity (DRC) scheduling algorithm which is an advancement to the BLS algorithm is also presented in section 5.3.2.

Some key notations used in this section are summarized in table 5.2.

TABLE 5.1: Classification of IEEE TSN Standards

Category	Standards
Latency & Jitter Traffic class categorization, transmitting & frame queuing according to traffic classes	IEEE 802.1Qav (Credit Based Shaping) IEEE 802.1Qbv (Scheduled Traffic) IEEE 802.3br & IEEE 802.1Qbu (Frame Preemption) IEEE 802.1Qch (Cyclic Queuing) IEEE 802.1Qcr (Asynchronous Traffic Shaping
Flow Management Resource allocation management & registration, network configuration, network discovery and monitoring.	IEEE 802.1Qcp (YANG Models) IEEE 802.1CS (Link-Local Reservation) IEEE 802.1Qat & IEEE 802.1Qcc (Stream Reservation)
Time Synchronization Provides network wide stringent synchronization of all network element clocks at Layer 2.	IEEE 802.1AS & IEEE 802.1AS-Rev (Network Timing & Synchronization)
Flow Integrity Maintaining system wide integrity by ensuring flows deliver their frames in spite of dynamic network conditions.	IEEE 802.1CB (Frame Replication & Elimination) IEEE 802.1Qca (Path Control & Reservation) IEEE 802.1Qci (Per-Stream Filtering)

Key design aspects

The BLS is part of the credit-based shapers class [21]. Each shaped queue is mapped to a specified traffic class and is defined by an upper threshold *M_L*, a lower threshold *R_L* and a reserved bandwidth *BwF* which is a fraction of the total link capacity C of the system. The priority of an arbitrary class k shaped by the BLS, p(k) is designed to vary from high priority $p_1(k)$ to low priority $p_n(k)$ (where priority value p = 1, 2, ...,n in descending order of priority) as the BLS credit in bits of the shaped traffic goes from high to low.

In our fronthaul network design context, the BLS spaces out high priority time-sensitive traffic packets T_{URLLC} to reduce bursting and bunching effects thereby protecting lower priority flows $T_{non-URLLC}$ as the maximum high priority packets burst is bounded. Fig. 5.1 below describes the scheduling architecture for our proposed DRCS to be implemented on a FH network Ethernet switch.The total capacity is shared between the URLLC high priority traffic T_{URLLC} and the lower priority traffic $T_{non-URLLC}$. The BLS manages the T_{URLLC} traffic flows which is mapped to a priority

TABLE 5.2: Notations

C	Link Rate
I_{slope}	BLS idle slope
S_{slope}	BLS send slope
BwF	DRCS/BLS reserved capacity
q_{ULLC}	shaped high priority queue
$q_{non-ULLC}$	low priority queues
T_{URLLC}	Ultra Reliable Low latency Traffic
$T_{non-URLLC}$	Non-Ultra Reliable Low Latency traffic
M_L	maximum credit Level
R_L	resume credit level
MFS_k	Maximum Frame Size of flow k
$\triangle_j^i$	defined DRCS/BLS windows with $j \in \{send, idle\}$ and i $\in \{max, min\}$

scheduler (PS). The Priority of T_{URLLC} changes between one (1) and three (3) depending credit counter and traffic conditions in its queue. Therefore, the $T_{non-URLLC}$ is of higher priority at some instances.

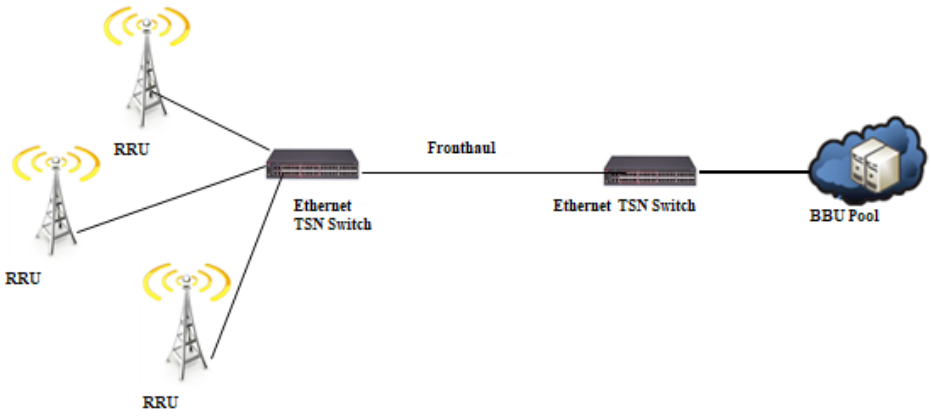

FIGURE 5.1: 2 hop fronthaul network topology example

The credit-based burst-limiting shaper implementing our DRCS algorithm separates a queue into two traffic classes as illustrated in Fig. 5.2, High Priority class (tighter delay bound) and Low Priority class (looser delay bound). When no frame is available in either queue, the credit for the queue is set to zero. A queue is eligible for transmission if the credit is non-negative (i.e $R\_L \geq 0$).

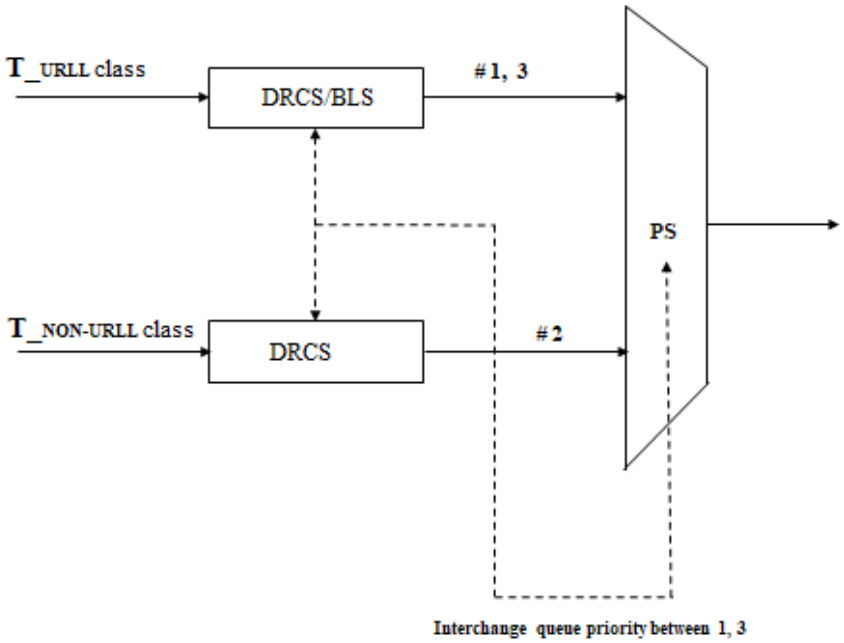

FIGURE 5.2: Proposed output scheduling architecture for TSN switch with DRCS on top of Priority Scheduler

In our DRCS coupled with a Priority Scheduler at the output, packets are scheduled with respect to the following conditions:

1. In order of priority, highest priority with non-empty queue first
2. The credit of the T_{URLLC} class (assuming it has highest priority) increases linearly with rate I_{slope} when there is at least a single packet in the queue q_{URLLC}.
3. The credit of the URLLC class decreases linearly with rate S_{slope} when a packet is transmitted.
4. Packets from $T_{non-URLLC}$ queues q_{URLLC} can be scheduled during T_{URLLC} Transmission Period (TP) as long the DRCS condition is fulfilled where T_{URLLC} has right to preemption in the event of sudden traffic bursts.
5. Priority change occurs for T_{URLLC} queue from #1 to #3 when its credit falls to 0 from M_L at the rate I_{slope} and vice-versa when the shaped T_{URLLC} traffic begins to rises from 0 (R_L) at the rate S_{slope}. This behaviour is illustrated in Figs. 5.2 and 5.3.
6. The BLS throttles the shaped T_{URLLC} class not to exceed their preconfigured bandwidth limits.

The behaviour of the DRCS/BLS is displayed in Fig.5.3. As described, the credit interchanges between 0 (resume level R_L) and M_L (Maximum Level). The DRCS/BLS shaper credit rates are defined as follows:

- The decreasing rate (Idle slope):

$$I_{slope} = C.BwF \tag{5.1}$$

Where C is the fronthaul link speed and BwF is the BLS high priority shaped traffics reserved link capacity.

- The increasing rate (send slope):

$$S_{slope} = C - I_{slope} \tag{5.2}$$

Fig. 5.3 depicts the operation of the BLS for the two traffic classes T_{URLLC} and $T_{non-URLLC}$,. The first transmission window denoted as $\triangle_1$, describes the case where bursty high priority packets are sent and reaches the max. level (M_L) when its priority is highest. The next transmission window $\triangle_2$, after M_L has been reached by the T_{URLLC} is the $T_{non-URLLC}$ packets transmission window with send rate I_{slope} and priority 2. The T_{URLLC} class priority is 3 (the lowest) at this transmission time window.

5.2.5 DRCS/BLS Output Scheduling algorithm

In order to guarantee low latency characteristics for high priority burst limited traffic, the BLS mechanism reserves a percentage of link bandwidth or capacity (BwF in % of total BW in Mbps)[60].

With any case of reserved (dedicated) resource, there always exists a trade-off between guaranteeing sufficient bandwidth for high priority traffic and scarce resource wastage which could be optimally utilized for lower priority traffic classes. In our dynamic reserved capacity (DRC) enhancement, the reserved capacity for the BLS traffic can dynamically accommodate non-URLLC traffic in instances where the reserved capacity is underutilized (after long periods of inactivity) by T_{URLLC} packets

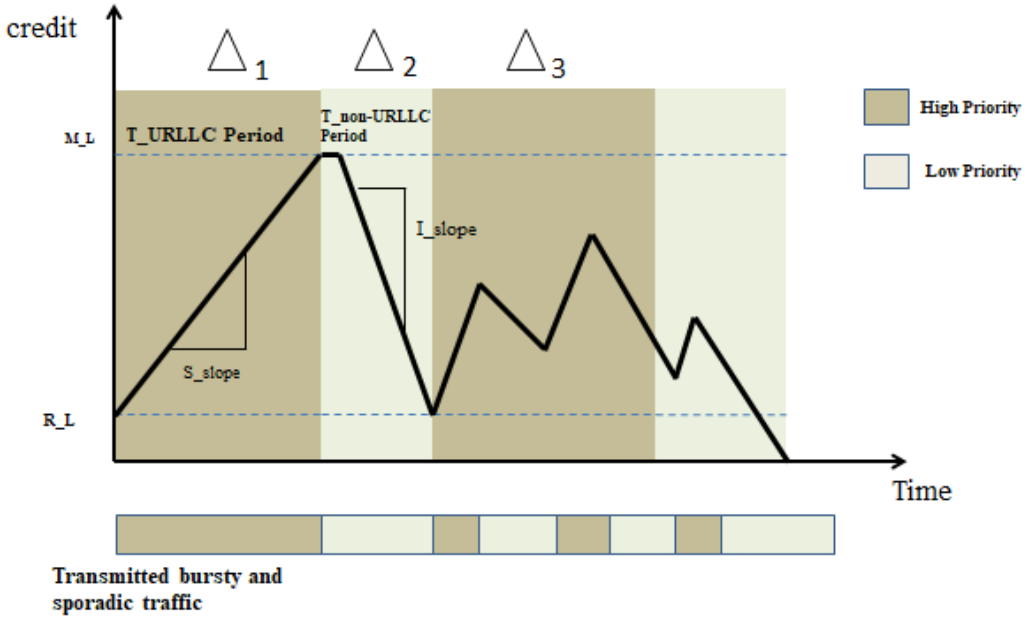

FIGURE 5.3: Operation of URLLC traffic shaping algorithm in FH switch output port

in every T_{URLLC} transmission window $\triangle_j^i$. Therefore, the dynamic component of our scheme is mapped to the periodic high priority class traffic load.

Therefore, our DRCS/BLS algorithm corresponds to a modification of the standard BLS algorithm and strict priority scheduler with our DRCS schedulability condition being the main modification. The new output scheduling algorithm is presented in Algorithm 4.

The Resume Level credit is initialized to zero (line 9). If the priority of the high priority traffic queue is low, it changes to high (line 10). Then with the updated priority, the strict priority scheduler performs based on the update. If the high priority traffic is not empty and the DRCS schedulability condition is not met (the number of packets are above a certain threshold) then packets are transmitted from the high priority queue by the WRR priority scheduler (lines 16 and 17). Else if the High priority queue is empty and the DRCS schedulability condition is met, then some low priority packets can be scheduled.

Algorithm 4: Dynamic BLS output Scheduling

Input:
credit ← credit in bits
M_L ← Maximum credit Level
R_L ← Resume credit Level
q_{URLL} ←Burst Limited (URLLC) queue
$q_{non-URLL}$ ← non-URLLC queues
P_{high}, P_{low} ← high and low priority queues
K ← Updated priority scheduled queue
DRCS ← Dynamic reserved capacity scheduled condition
R_L ← 0
for credit ≤ R_L **do**
  if $q_{URLL} = P_{low}$ **then**
    $q_{URLL} = P_{high}$
  end if
end for
for $q_{URLL} = P_{high}$ **do**
  if $q_{URLL} \neq$ NULL and DRCS == 'false' **then**
    K = q_{URLL}: transmit packets from q_{URLL} using WRR
  else if $q_{URLL} \neq$ NULL and DRCS == 'true' **then**
    K = $q_{URLL}, q_{non-URLL}$: transmit packets from
    $q_{URLL}, q_{non-URLL}$ using WRR
  end if
end for
for credit ≥ M_L **do**
  if $q_{URLL} = P_{high}$ **then**
    $q_{URLL} = P_{low}$
  end if
end for

5.2.6 Load Aware Scheduling component

In order to implement the DRCS the traffic in each URLLC queue has to be below a certain threshold. Therefore, to monitor the traffic variations of each URLLC traffic queue to ascertain when to schedule non-URLLC traffic we implement a dynamic load computation.

Suppose at any given time t, the high priority traffic in queue $i \in \{T_{URLL}\}$ at a FH network switching element/node $n \in \{sw\}$ is given by $load_{i,t}$, where $1 \leq i \leq N$ for a class with N total queues. The queue variance at time $t(\alpha_t)$ is computed as in [131].

$$\alpha_t = \frac{1}{N}\sum_{i=1}^{N}(load_{i,t} - (load_t))^2 \tag{5.3}$$

the root mean square error (rms) at time t is given as:

$$\beta_t = \sqrt{\alpha_t} \tag{5.4}$$

The dynamic load coefficient of queue i at time t ($l_{i,t}$)

$$l_{i,t} = \frac{load_{i,t}}{\beta_t + 1} \tag{5.5}$$

The dynamic load in queue i at time t which we used as a sufficient condition to implement DRCS is given by:

$$L_{i,t} = l_{i,t}P_i(t) \tag{5.6}$$

Where $P_i(t)$ is the number of packets in queue i at time t.

The DRCS condition is fulfilled if equation (5.6) falls below a load threshold θ and thus is implemented subject to:

$$\sum_{i=1}^{N} L_{i,t} \leq \theta \quad \forall i \in T_{URLLC} \tag{5.7}$$

DRCS Schedulability condition

To further elaborate on the load-aware scheduling component described in the previous section, we define more concisely our dynamic schedulability condition.

In order to optimally take advantage of the reserved capacity for the shaped high-priority traffic, and to minimize the trade-off between reserved resource and resource available to lower-priority traffic in service differential FH network systems, we introduce our opportunistic scheduling scheme.

We first define a **sufficient low-priority flow schedulability condition**. This consists of monitoring the arrival rates of the high priority shaped flows i at any crossed switching node n along the FH transport link which has to be lower than a defined arrival rate threshold θ to ensure system stability and service quality for T_{URLLC} flows. This constant θ is the **arrival threshold**:

$$\forall node\ n \in FH\ network,\ \sum \theta_i \leq R_n \tag{5.8}$$

Where,

R_n is the input cumulative function of shaped flows at any node n along the FH network.

5.3 DRCS Response Time analysis

We present in this section the timing analysis utilizing Network Calculus (NC) theory as adopted in [53] and applying this to our New Radio (NR) FH network use-case. We compute the network switching node worst case delay bound for both traffic classes and model the traffic flows/Ethernet switching element present at each hop of the FH network.

In order to infer real-time guarantees on both URLLC and non-URLLC traffic classes based on our proposed DRCS/BLS scheduler, a sufficient schedulability condition is applied to verify that the delay bounds of each traffic class conforms to set deadlines.

The expression for end-to-end delay of a flow k in the class $1 \in \{T_{URLLC}, T_{non-URLLC}\}$, $EED_{i,k}$ along its path $Path_k$ is given by [53]:

$$EED_{i,k} = d_j^{es} + d_{prop} + \sum_{i \in path_k} d_{i,k}^{sw,j} \tag{5.9}$$

Where d_j^{es} is the end-system delay (es) contributed by the aggregate traffic of class i, d_{prop} represents the propagation latency along the path and $d_{i,k}^{sw,j}$ is the delay within the switching element along the flow path.

For simplicity we are only concerned with the delay associated with the switching elements (D_{sw}) along the FH flow path which houses the DRCS/BLS coupled with the strict priority scheduler (PS). Therefore, equation (5.8) is reduced to

$$D_{sw_{i,k}} = \sum_{i \in path_k} d_{i,k}^{sw,j} \tag{5.10}$$

5.3.1 Fronthaul Ethernet Switch Modeling

In order to compute the response time associated with our DRCS/BLS on top of the strict priority scheduler (PS) located in the FH switch component (see Fig. 5.2) we model both traffic class flows using Network Calculus (NC) parameters [25] to obtain the maximum arrival and minimum service curves at the crossed switch.

Network Calculus Model

As earlier described, we utilize Network Calculus theory timing analysis to compute the delay bounds on arrival and service traffic defined in NC as both arrival curve α and minimum service curve β. The definitions of these curves are elaborated as follows.

Definition 1 (Arrival Curve). [25] A function $\alpha(t)$ is used to denote an arrival curve for a data flow with input cumulative function given by, R(t), which represents the number of received bits until time epoch t, *iff*:

$$\forall t, R(t) \leq R \otimes^1 \alpha(t)$$

Definition 2 (Strict Minimum Service Curve). [25] The function β is used to represent the minimum strict service curve for a data flow with output cumulative function R^*, if for any backlogged period

$$]s,t]^2, R^*(t) - R^* \geq \beta(t-s)$$

Definition 3 (Maximum Service Curve).[25] The function $\gamma(t)$ represents the maximum service curve for a data flow with input cumulative function $R(t)$ and output cumulative function $R^*(t)$

$$\forall t, R^* \leq R \otimes \gamma(t)$$

According to [53], the arrival curve of the aggregate traffic at any time t made up of each flow k in class i at the input of a jth node $n \in \{sw\}$ or component $n \in \{DRCS, PS\}$ is:

$$\alpha_i^{n,j}(t) = \sum_{k \in i} \alpha_{i,k}^{n,j}(t) \tag{5.11}$$

For the Ethernet-based switches along the FH network between RRH and BBUs, we model the impact of our DRCS/BLS on top of the strict priority scheduler (PS) on the T_{URLLC} and $T_{non-URLLC}$ traffic classes.

- T_{URLLC} Class: As detailed in figure 5.2, the high priority T_{URLLC} traffic adopts two credit dependent positions; (I) the case where the priority is low at (3). The guaranteed minimum service curve within the Ethernet switch at this instant is based primarily on the strict priority scheduler (PS) denoted as $\beta_{URLL_3}^{PS}$ at the output. (II) the second case where the priority switches to the high position (1). Thus, the minimum service curve guaranteed within the switch is as a result of the combination of service curves within the DRCS/BLS component β_{URLL}^{DRCS} and the PS component $\beta_{URLL_1}^{PS}$.

Therefore, the expression for the service curves guaranteed within the switch for the high priority traffic class T_{URLL} is given as:

$$\beta^{sw}_{URLLC}(t) = max(\beta^{PS}_{URLLC(3)}, \beta^{PS}_{URLL(1)} \otimes \beta^{DRCS}_{URLLC(1)}(t)(1 + \frac{\beta^{DRCS}_{non-URLLC(2)}}{\beta^{DRCS}_{URLLC(1)}})) \tag{5.12}$$

Where the last term in Eq. 5.12 accounts for the DRCS dynamic schedulability condition.

- $T_{non-URLLC}$ Class: For this traffic class we model minimum service curve guaranteed within both the DRCS/BLS (when the schedulability condition accepts some $T_{non-URLLC}$ flows) and within the strict priority scheduler using cor.1 in [53] when considering the max output arrival curve of T_{URLLC} and some $T_{non-URLLC}$ from the DRCS/BLS component.

The expression for the minimum service curve guaranteed for the $T_{non-URLLC}$ traffic at the strict Priority scheduler (PS) and some $T_{non-URLLC}$ within the DRCS/BLS when taking into consideration the maximum output arrival curve of T_{URLLC} traffic from the DRCS/BLS component $\alpha^{*,DRCS}_{URLLC}$:

$$\beta^{sw}_{non-URLLC}(t) = [C.t - \alpha^{*,DRCS}_{URLLC}(t)(1 + \frac{\alpha^{DRCS}_{non-URLLC}}{\alpha^{DRCS}_{URLLC}}) - \underset{k \in i, p(i) \geq p(non-URLLC)}{\text{Max}} MFS_k] \tag{5.13}$$

To compute the delay associated with the switching component in each node along a FH network flow path from (5.9), the unknown service curves from equations (5.12) and (5.13) relating to the DRCS/BLS scheduling component need to be determined.

URLLC Traffic Service Curves

To compute the worst case cumulative traffic for T_{URLLC} flows from the DRCS/BLS, the strict minimum service curve β^{DRCS}_{URLLC} which maximizes the latency within the DRCS burst limiting scheduler and serves as an upper constraint for T_{URLLC} flows.

To compute the min. service curve parameter we combine both minimum sending window $\triangle^{min}_{send}$ and maximum idle window $\triangle^{max}_{idle}$ durations which combine to make up the upper constraint/max. delay bound on cumulative T_{URLLC} flows. $\triangle^{min}_{send}$ describes the time in the DRCS/BLS scheduler for the consumed credit to go from

resume credit level R_L to maximum credit level M_L with increasing rate S_{send} as illustrated in fig. 5.3.

$$\triangle_{send}^{min} = \frac{M\_L - R\_L}{S_{send}} \tag{5.14}$$

On the other hand, the max idle window duration, $\triangle_{idle}^{max}$ is the time for the credit to go from M_LtoR_L with a decreasing rate I_{idle} together with the transmission duration of the maximum frame of the $T_{non-URLLC}$ traffic which occurs when the $T_{non-URLLC}$ frame starts its transmission just before the credit reaches the lower threshold R_L.

It is worth noting that this non-preemption feature does not extend to some $T_{non-URLLC}$ traffic scheduled during the DRCS scheduling window for T_{URLLC} traffic. The max idle window is thus given as:

$$\triangle_{idle}^{max} = \frac{M\_L - R\_L}{I_{idle}} + \frac{MFS_{non-URLLC}}{C} \tag{5.15}$$

Therefore, the strict minimum service curve representative of the lower cumulative service response time for T_{URLLC} traffic in DRCS/BLS crossing a FH network node with link capacity C is given as [53] :

$$\beta_{URLLC}^{DRCS}(t) = \frac{\triangle_{send}^{min}}{\triangle_{send}^{min} + \triangle_{idle}^{max}}.C.(t - \triangle_{idle}^{max})^{+} \tag{5.16}$$

where $[y]^{+}$ is the maximum between y and 0.

The maximum service curve of T_{URLLC}, γ_{URLLC}^{DRCS} offers the lower constraint and best service to the T_{URLLC} traffic within the DRCS/BLS compared to β_{URLLC}^{DRCS}.

To compute γ_{URLLC}^{DRCS}, a combination of the maximum send window $\triangle_{send}^{max}$ and the minimum idle window $\triangle_{idle}^{min}$ is carried out:

$$\triangle_{send}^{max} = \frac{M\_L - R\_L}{S_{send}} + \frac{MFS_{URLLC}}{C} + min\left(\frac{MFS_{non-URLLC}}{C}.\frac{I_{idle}}{S_{send}}, \frac{R\_L}{S_{send}}\right) \tag{5.17}$$

The minimum idle window $\triangle_{idle}^{min}$, represents the duration for the consumed credit to drop from M_L to R_L with a slope I_{idle} given as:

$$\triangle_{idle}^{min} = \frac{M\_L - R\_L}{I_{idle}} \tag{5.18}$$

The maximum service curve guaranteed to T_{URLLC} traffic, γ_{URLLC}^{DRCS} is therefore given as described in [53] as:

$$\gamma_{URLLC}^{DRCS} = \begin{cases} C.t: \ in\ the\ presence\ of\ no\ T_{non-URLLC\ flows} \\ otherwise: \\ \frac{\triangle_{send}^{max}}{\triangle_{\gamma_{URLLC}^{nom}}}.C.t + \triangle_{send,0}^{max}.C.\frac{\triangle_{idle}^{min}}{\triangle_{\gamma_{URLLC}^{nom}}} \end{cases} \tag{5.19}$$

non-URLLC Traffic Service Curve

To compute the minimum service curve for the low priority flows $\beta_{non-URLLC}^{sw}$ as defined in Eq. 10, the maximum output arrival curve of T_{URLLC} from the DRCS/BLS, $\alpha_{URLLC}^{*,DRCS}$ from Eq. 10 is to be computed and is given as:

$$\alpha_{URLLC}^{*,DRCS}(t) = min(\gamma_{URLLC}^{DRCS}(t), \alpha_{URLLC}^{DRCS} \oslash \beta_{URLLC}^{DRCS}(t)(1 + \frac{\beta_{non-URLLC}^{DRCS}}{\beta_{URLLC}^{DRCS}})) \tag{5.20}$$

Where the last term in Eq. 5.20 accounts for the DRCS dynamic schedulability condition.

Therefore, Eq. 5.20 is imputed into Eq. 5.13 for the minimum strict service curve $\beta_{non-URLLC}^{sw}$ guaranteed to $T_{non-URLLC}$ traffic with some crossing the DRCS/BLS and majority flows crossing the SP scheduler at the switch output.

5.3.2 Results and Analysis

In this section, we implement our proposed switch output architecture (see Figs. 5.1 and 5.2) and compare the performance of our DRCS/BLS scheduling algorithm coupled with a WRR strict priority scheduler in *ns-3* [151]. We consider two broad traffic flows with source nodes S_1 and S_2 which generate packets T_{URLLC} and $T_{non-URLLC}$ classes as described previously.

For both traffic flows,we consider packet lengths of 1500 Bytes. We generate 100 UDP/T_{URLLC} flows which in an experimental set-up for a TSN network would be annotated with 802.1Q VLAN tags to differentiate traffic flows and QoS requirements. We consider traffic overload and underload scenarios with traffic of approximately 100% of the link-rate for overload, and just under 55% for under.

Fig. 5.4 shows that with increase in the reserved capacity BwF for the high priority burst limited T_{URLLC} flows, and increase in the relative weight w_{URLLC} with respect to the share allocated to $T_{non-URLLC}$ flows, the output rate R^*_{URLLC} increases and vice-versa for both our DRCS/BLS and WRR. Therefore, both schemes exhibit similar results under average load conditions for high priority traffic.

The theoretical output rate [52] for T_{URLLC} class is given by:

$$R^{*,th}_{WRR/T_{URLLC}} = w_{URLLC}.C \tag{5.21}$$

where,

w_{URLLC} is the relative weight of the URLLC traffic class to the non-URLLC class, L_i^{avg} the average length of packets of both classes and W_i the weight of class $i \in \{URLLC, non-URLLC\}$.

$$w_{URLLC} = \frac{W_{URLLC}.L^{avg}_{URLLC}}{W_{URLLC}.L^{avg}_{URLLC} + W_{non-URLLC}.L^{avg}_{non-URLLC}} \tag{5.22}$$

Therefore, from this behaviour we can deduce that:

$R^{*,sim}_{DRCS/URLLC} = R^{*,sim}_{WRR/URLLC} = R^{*,th}_{WRR/URLLC}$

Fig. 5.5 describes the output rate $R^*_{non-URLLC}$ for $T_{non-URLLC}$ flows whereby the residual link capacity left by the T_{URLLC} is allocated for low priority traffic scheduling by both DRCS/BLS and WRR schemes. The theoretical output rate is given by:

$$R^{*,th}_{DRCS,WRR/non-URLLC} = (1 - w_{URLLC}).C \tag{5.23}$$

It can also be seen from Fig. 5.5 that the output rate for DRCS/BLS scheduler for $T_{non-URLLC}$ flows slightly outperforms the WRR. This is expected because in this

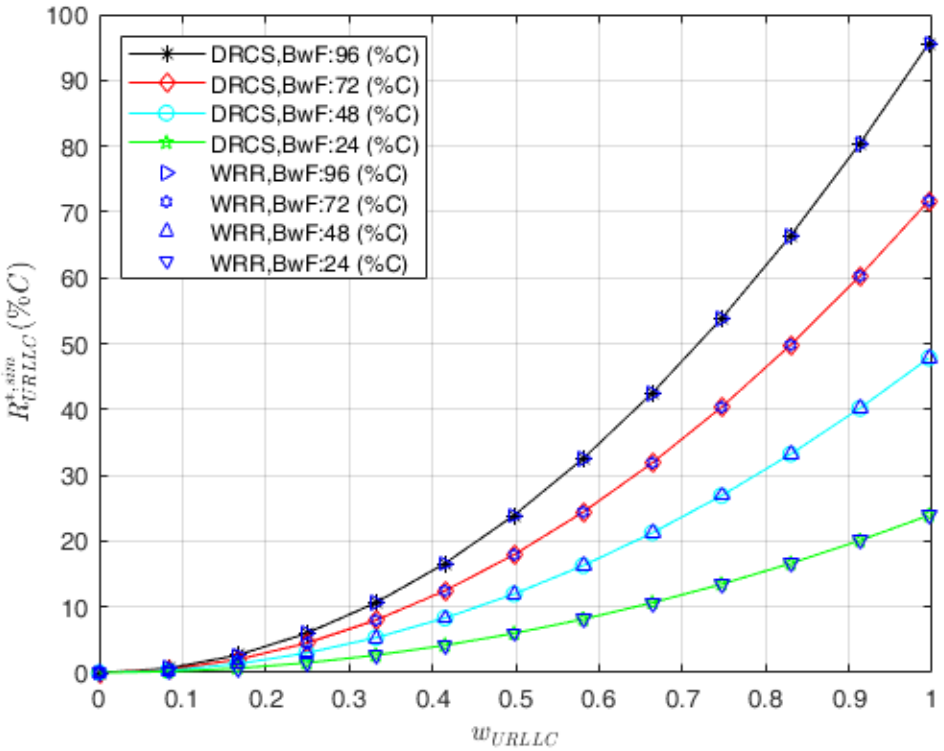

FIGURE 5.4: URLLC traffic output rate

scenario, the DRCS schedulability condition accounts for more packet scheduling for low priority flows.

Fig. 5.6 shows that for DRCS/BLS, the average latency for T_{URLLC} flows decreases as the Max. credit level (M_L) which is a function of the burst limited traffics reserved capacity, BwF. It outperforms the SP WRR scheduling mechanism in both load conditions at just under 40% and above of burst-limiting max. credit level and under performs WRR at below this level.

The latter is expected because when the high-priority packet arrives at an idle time (I_{slope}) which is when its priority has switched from high to the lowest level, it has to wait for its priority change to high before it can send packets. Therefore, this result demonstrates that the DRCS/BLS reserved capacity parameter can be optimally tuneable.

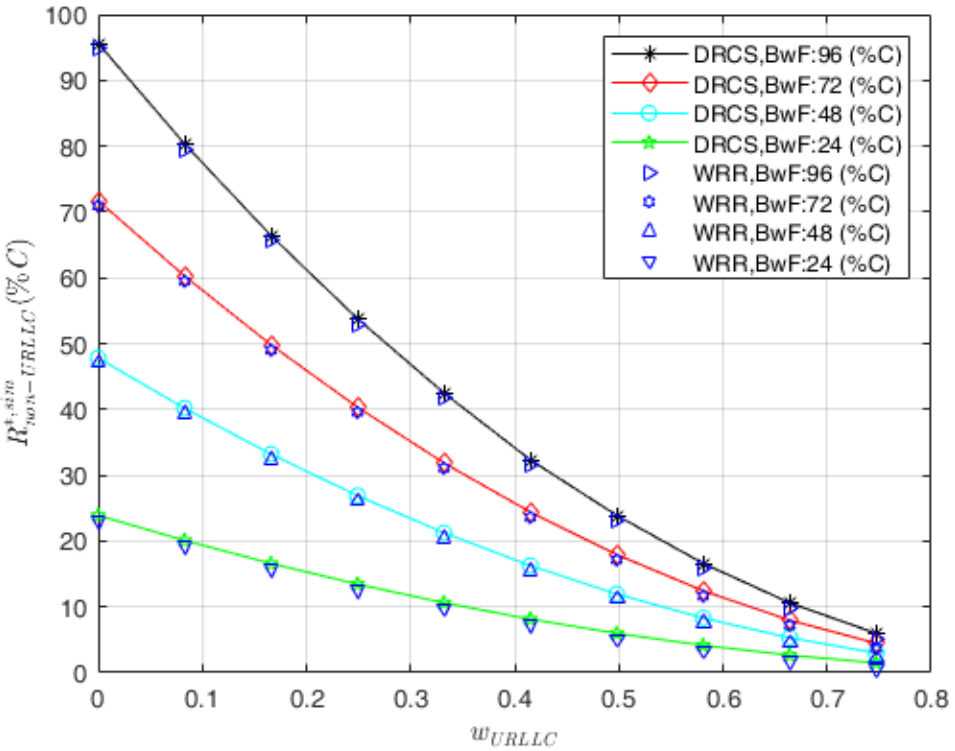

FIGURE 5.5: non-URLLC traffic output rate

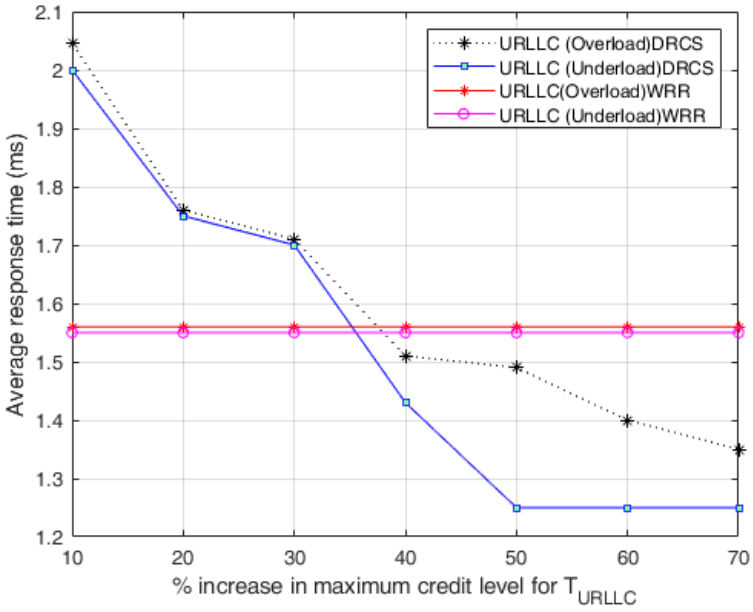

FIGURE 5.6: URLLC traffic Average Response time

Fig. 5.7 shows that the low priority traffic is also a function of the DRCS/BLS reserved capacity. This is as a result of the opportunistic scheduling of low priority flows with high priority reserved BLS capacity during periods of mostly long high

priority traffic inactivity.

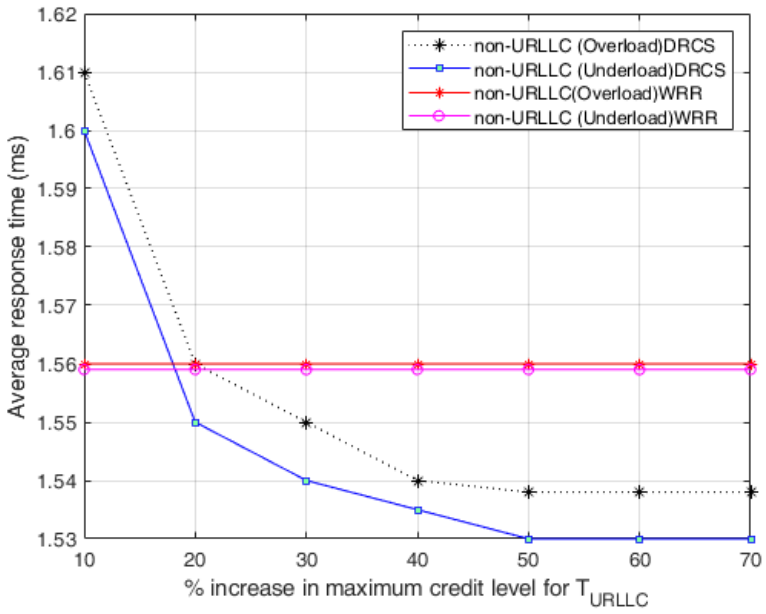

FIGURE 5.7: non-URLLC traffic Average Response time

Fig. 5.8 shows the average response latency for both WRR and DRCS scheduled traffic in relation to different packet sizes. For high- priority traffic, we adjusted the max-credit level to account for 60% link capacity $(C = 1Gbps)$ and the low-priority traffic is scheduled using 40% of the residual capacity. The packet sizes are varied from 100 to 1500Bytes which is the maximum frame size for both traffic class flows..

Fig 5.9 shows the average response time performance for both traffic class flows against the WRR mechanism. From our simulated experiment, we observe the average latency results for the low- priority traffic performs significantly better than the WRR. This is as a result of our opportunistic scheduling contribution to the standard burst limiting shaper for TSN, thereby allowing for more robust scheduling of low-priority traffic while effectively isolating high-priority flows.

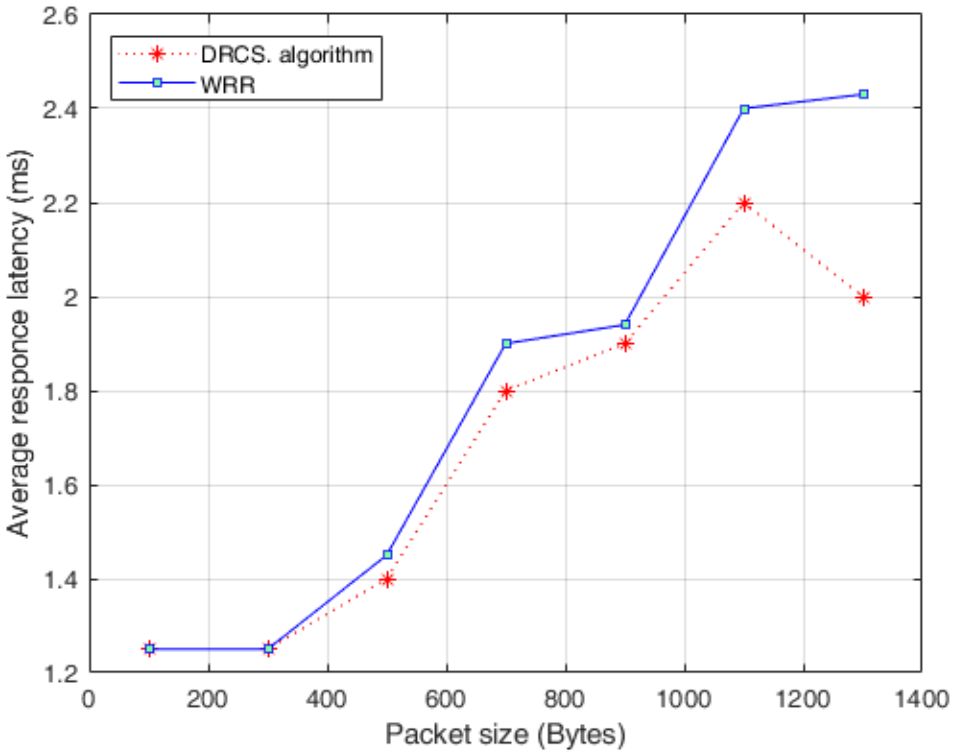

FIGURE 5.8: Average Response latency for different packet sizes

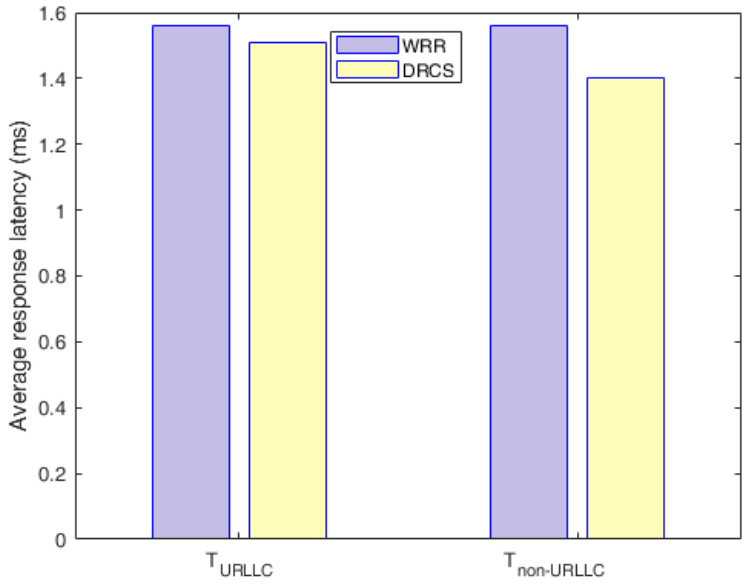

FIGURE 5.9: Average Response latency for priority class flows using WRR and DRCS

5.4 Conclusions

We have presented in this chapter, a new scheduling algorithm which is a contribution to the scheduling and dequeing process of the IEEE 802.1Q Burst Limiting Shaper algorithm. We apply our alternative architecture and algorithm in addition to Network calculus based timing analysis to achieve latency and resource usage optimized scheduling for service differential flows in 5G and onward to 6G Fronthaul Network traffic associated with Cloud-RAN New Radio topologies.

We benchmark our proposal against the popular WRR scheme and utilize the BLS reserved capacity for dynamic scheduling of low-priority traffic in order to increase overall traffic throughput and reduce response latency. Simulation experiments were carried out to evaluate our proposals performance and the results revealed that our DRCS/BLS scheme outperforms WRR in terms of average response times for both traffic classes.

Chapter 6

Conclusions and Future Perspectives

In the following two sections we present our conclusions related to this work and the substantial additional research required relating to the general scope of our work.

6.1 Conclusions

With the rapid adoption of 5G and 6G in the works, there exists a wide range of possibilities for mobile networking in conjunction with principles such as Software Defined Networking (SDN), Network Function Virtualization (NFV), Time Sensitive Networking (TSN) and Network slicing, techniques which will enable the application of end-to-end virtual and logical networks and also enable more flexible and accurate management of network resources.

On the Radio Access Network (RAN) side, the 5G standard came with the implementation of Cloud-RAN architecture in a bid to centralize RAN functionality and processing in cloud-based environments in order to benefit from scalable and flexible functional slit optimized adaptation, resource pooling, joint radio resource allocation for enhance interference reduction and improved data rates. However, the implementation of all the above mentioned techniques including C-RAN present many issues related to RRH and BBU placement, fronthaul network interface which connects the distributed remote radio units to the centralized BBU pool, the power-distance budget for the FH as well as efficient resource allocation for processing real-time traffic in this new architecture.

To this end, this thesis investigates advancements to system wide issues related to networks of the future, bordering around latency reduction and optimal resource usage. We lay emphasis on novel techniques for Random Access resource usage and RA traffic scheduling (Chapter 3), load balancing, optimized cell traffic offloading and user/cell associations (Chapter 4) and finally TSN-based optimized scheduling for NR fronthaul networks (Chapter 5). All contributions were developed with focus

on latency reduction for URLLC applications as well as optimized resource management to accommodate the mixed-traffic analogous with 5G and future networks as a whole.

- The first contribution to our work in Chapter 3 addresses the problem of random access scheduling amidst high traffic contention. We design an innovative strategy which comprises of Random Access (RA) preamble reservation for 3 distinct traffic classes (High, Medium and Low), where the reserved resources for the High priority class remains static, while for the medium and low class, a conditional dynamic resource sharing mechanism is introduced. Access delay figures are then compared with standard LTE/LTE-A RA figures and simulation results show that our partitioned resource reservation mechanism significantly reduces control plane access delay figures in both priority and non-priority traffic, meeting the 5G ultra Reliable and Low Latency Communications (URLLCs) requirement of under 10ms. We also derived through analytical means, the optimum preamble partition for the reserved case which will also meet service requirements for low-priority traffic.

- The second contribution in Chapter 4 dealt with the problem of traffic load balancing and user association to cells in 5G and beyond heterogeneous cell type deployments in order to achieve E2E latency reduction. Our work in this Chapter comprises of a unique contribution to the standard LTE/LTE-A handover (HO) procedure by dynamically adjusting 3GPPs Almost Blank Subframe (ABS) ratio (ABS to non-ABS) of large macrocells within a multi-tiered network space, to the amount of HO requests received from nearby user equipment's (UEs) mostly attached to smaller cells. Our technique contributed towards encouraging attachment of devices to smaller powered nodes with less domineering reference signal received power (RSRP) as compared to higher powered nodes within the same coverage area. Properly balanced traffic load invariably leads to lower scheduling latency's and power savings, which are in turn enabling factors for a scalable and flexible virtual RAN. In addition to our load balancing technique, we proposed a cooperative gamified approach to RRH-BBU resource sharing. Finally, the performance of our joint optimization schemes were validated by simulations and the results demonstrated its effectiveness in load balancing and overall user throughput.

- Finally, a QoS based differential service scheduling technique for 5G an onwards Fronthaul networks using an IEEE 802.1 Time Sensitive Networking (TSN) mechanism to tackle the problem of robust traffic scheduling along the

additional fronthaul link introduce by the C-RAN architecture for future mobile networks. To address this issue we leverage on IEEE 802.1Q Burst Limiting Shaper algorithm and introduce our Dynamic Reserved Capacity (DRC) schedulability condition to it. The Key Idea to our proposal consists of dynamically scheduling some low priority traffic class flows within the reserved capacity for the Burst Limiting Shaped traffic (high priority URLLC flows) when the shaped traffic is underutilized (after long periods of inactivity) in every shaped traffic transmission window. The dynamic component of our scheme is therefore mapped to the periodic shaped traffic class load. We further derive the timing characteristics of both high and low priority flows within the switching component of the packet-based fronthaul architecture which we model, using Network Calculus framework. With our proposed scheduling strategy which we bench-marked against the WRR scheduling algorithm, our schemes performance was more robust compared to WRR with more noticeable improvements in output rates for the lower priority traffic flows.

From this thesis we were able to propose and implement novel traffic scheduling mechanisms in addition to a unique load balancing an resource usage strategies. Our work so far can be served as the foundation or contribution to a host of novel optimization techniques in 5G/6G heterogeneous mobile network systems.

6.2 5G and onwards to 6G

Currently adopted and future emerging techniques for 5G and 6G networks such as SDN, NFV and Artificial Intelligence (AI) are highlighted and can be relied upon to provide the adopted C-RAN system with much required flexibility and network service scalability. Through the foundation and synergy of the above mentioned techniques, a host of novel services can be developed in both industry and the academics to be able to meet the demands of 5G and beyond multi-tiered heterogeneous networks.

In line with this context, promising directions to further expand our work in this thesis are highlighted

Further RA procedure enhancements

In our work we considered the idealized interference-free RA scenario, which gives rise to convex optimization problems. The adoption of enhanced techniques for RA procedure such as parallel preamble transmissions and upgrade back-off schemes to

guarantee first attempt preamble success rates at the same time taking into consideration open- air interference should be investigated. This will go a long way in providing deterministic service guarantees for 5G URLLC applications in particular.

Load balancing and Interference management

In this thesis we considered the use of 3GPPs Almost Blank Frames (ABS) technique which is part of the eICIC standard to effectively balance traffic load by encouraging user association to smaller cells. However, with the rapid expansion and increasing heterogeneity of wireless mobile networks, low complexity algorithms which would also encourage low transmission latency would be required to further leverage on our technique

Limited Fronthaul Capacity

In our work, we considered a high-speed Ethernet based fronthaul link enabled with TSN based Ethernet switches and VLAN differential traffic tags. However, considering densely deployed C-RANs with delay bounded service requirements, cable-based FH installations might incur high CAPEX. Therefore, the investigation into mmWave transmission links are a promising research direction. The low attenuation associated with mmWaves frequencies should be factored in as well.

Further delay sources

In this thesis we considered the delay associated with the RA mechanism, user-cell association and fronthaul link component delay. The total E2E delay including the added delay brought on by the cloud located BBU pool processing of queued baseband functions should be taken into consideration with reliable test-beds. The URLLC stringent delay requirement for 5G/6G networks would also require more advanced traffic scheduling algorithms to carter for the ever expanding network.

Appendix A

Acronyms

3GPP Third Generation Partnership Project

4G Fourth Generation

5G Fifth Generation

5GB Fifth Generation and Beyond

ABS Almost Blank Subframes

ACB Access Class Baring

ACK Acknowledgement

AI Artificial Intelligence

AMC Adaptive Modulation and Coding

ARoF Analog Radio over Fiber

BBU baseBand Unit

BLS Burst Limiting Shaper

BS Base Station

C-RAN Cloud Radio Access Network

CAPEX CApital Expenditure

CBS Credit Based Shaper

CIR Channel Impulse Response

CN Core network

CoMP Coordinated Multi-Point

CP Control Plane

CP Cyclix Prefix

CPRI Common Public Radio Interface

CRE Cell Range Expansion

CQI Channel Quality Indicator

CU Centralized Unit

D2D Device-to-Device

D-RAN Distributed Radio Access Network

DFT Discreet Fourier Transform

DL Downlink

DL-SCH Downlink shared channel

DRCS Dynamic Reserved Capacity Scheduler

DS Dynamic Scheduling

DU Distributed Unit

E2E End-to-End

eICIC Enhance Inter-Cell Interference Coordination

eMBB Enhanced Mobile BroadBand

eNB Evolved Node B

FBMC Filter Bank Multi Carrier

FDD Frequency Division Duplexing

FFT Fast Fourier Transform

FH Fronthaul

FIFO First-in-First-Out

FOFDM Filtered Orthogonal Frequency Division Multiplexing

FPGA Field Programmable Gate Array

GFDM Generalized Frequency Division Multiplexing

GI Guard Interval

gNB next Generation Node B

HARQ Hybrid Automatic Repeat Request

HO Hand Over

IA-PFT Interference Avoidance transmission by partitioned frequency and Time Domain

I/Q In-Phase and Quadrature

ICI Inter Cell Interference

IDFT Inverse Discrete Fourier Transform

IEEE Institute of Electrical and Electronics Engineers

IoT Internet of Things

IMT International Mobile Telecommunications

ITU-R International Telecommunication Union Radio communication sector

KPI Key Performance Indicator

LDPC Low Density parity Check

LTE Long Term evolution

LTE-A Long Term Evolution-Advanced

M2M Machine-to-Machine

MAC Medium Access Control

MANO Management and Orchestration

MEC Multi-access Edge Computing

MIMO Multiple-Input Multiple-Output

mMTC Massive Machine-Type Communication

NACK Negative ACKnowledgement

NC Network Calculus

NFV Network Function Virtualization

NGFI Next Generation Fronthaul Interface

OFDM Orthogonal Frequency-Division Multiplexing

OPEX OPerating EXpense

O-RAN Open Radio Access Network

PAPR Peak-to-Average Power Ratio

PDCP Packet Data Convergence Protocol

PDSCH Physical Downlink Shared Channel

PHY PHYsical layer

PF Proportional Fair

PS Priority Scheduler

PRACH Physical Random Access CHannel

PRB Physical Resource Block

QoE Quality of Experience

QoS Quality of Service

RA Random Access

RACH Random Access Channel

RAN Radio Access Network

RAR Random Access Response

RAT Radio Access Technology

RB Resource Block

RF Radio Frequency

RLC Radio Link Control

RoE Radio over Ethernet

RRC Radio Resource Control

RRH Remote Radio Head

RRU Remote Radio Unit

RR Round Robin

RSRP Reference Signal Received Power

SDN Software-Defined Networking

SIB2 System Information Block 2

SINR Signal to Interference plus Noise Ratio

TA Timing Advance

TN Transport Network

TSN Time Sensitive Networking

TTI Transmission Time Interval

UE User Equipment

UL Uplink

URLLC Ultra-Reliable and Low Latency Communication

V2X Vehicular-to-everything

VNF Virtual Network Function

WRR Weighted Round Robin

Appendix B

Publications

International Conferences

- Ogechi Akudo Nwogu, Gladys Diaz, Marwen Abdennebi, *A Combined Static/Dynamic Partitioned Resource Usage Approach for Random Access in 5G Cellular Networks*, **2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)**, September 2019 Split, Croatia.
- Ogechi Akudo Nwogu, Gladys Diaz, Marwen Abdennebi, *An Optimized Approach to Load Balancing and Resource Usage in 5G Multi-tiered Cellular Networks*, **2020 Global Information Infrastructure and Networking Symposium (GIIS)**, October 2020 Tunis, Tunisia.
- Ogechi Akudo Nwogu, Gladys Diaz, Marwen Abdennebi, *Differential Traffic QoS Scheduling for 5G/6G Fronthaul Networks*, accepted in **2021 International Telecommunication Networks and Applications Conference (ITNAC)**, November 2021 Sydney, Australia.

Bibliography

[1] Internet of things in the 5g era.

[2] Precoding.

[3] 3GPP. discussions on 2 steps rach procedure, 2017.

[4] Mohammed I Aal-nouman, Osamah Abdullah, and Noor Qusay A AlShaikhli. Inter-cell interference mitigation using adaptive reduced power subframes in heterogeneous networks. *International Journal of Electrical & Computer Engineering (2088-8708)*, 11(4), 2021.

[5] Javad Abdoli, Ming Jia, and Jianglei Ma. Filtered ofdm: A new waveform for future wireless systems. In *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 66–70. IEEE, 2015.

[6] Evolved Universal Terrestrial Radio Access. Study on ran improvements for machine-type communications,". Technical report, TR 37.868 V. 11.0. 0.

[7] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 18(3):1617–1655, 2016.

[8] Irfan Ahmed, Hedi Khammari, Adnan Shahid, Ahmed Musa, Kwang Soon Kim, Eli De Poorter, and Ingrid Moerman. A survey on hybrid beamforming techniques in 5g: Architecture and system model perspectives. *IEEE Communications Surveys & Tutorials*, 20(4):3060–3097, 2018.

[9] Nurzaman Ahmed, Debashis De, and Iftekhar Hussain. Internet of things (iot) for smart precision agriculture and farming in rural areas. *IEEE Internet of Things Journal*, 5(6):4890–4899, 2018.

[10] Adnan Aijaz, Mischa Dohler, A Hamid Aghvami, Vasilis Friderikos, and Magnus Frodigh. Realizing the tactile internet: Haptic communications over next generation 5g cellular networks. *IEEE Wireless Communications*, 24(2):82–89, 2016.

[11] Ian F Akyildiz, Pu Wang, and Shih-Chun Lin. Softair: A software defined networking architecture for 5g wireless systems. *Computer Networks*, 85:1–18, 2015.

[12] Sally R Aldaeabool and Maysam F Abbod. Reducing power consumption by dynamic bbus-rrhs allocation in c-ran. In *2017 25th Telecommunication Forum (TELFOR)*, pages 1–4. IEEE, 2017.

[13] Leandro Almeida, Paulo Ditarso Maciel Jr, and Fabio Luciano Verdi. Cloud network slicing: A systematic mapping study from scientific publications. 2020.

[14] Ghassan Alnwaimi and Hatem Boujemaa. Adaptive packet length and mcs using average or instantaneous snr. *IEEE Transactions on Vehicular Technology*, 67(11):10519–10527, 2018.

[15] Jeffrey G Andrews, Stefano Buzzi, Wan Choi, Stephen V Hanly, Angel Lozano, Anthony CK Soong, and Jianzhong Charlie Zhang. What will 5g be? *IEEE Journal on selected areas in communications*, 32(6):1065–1082, 2014.

[16] Emad Aqeeli, Abdallah Moubayed, and Abdallah Shami. Power-aware optimized rrh to bbu allocation in c-ran. *IEEE Transactions on Wireless Communications*, 17(2):1311–1322, 2017.

[17] Atefeh Hajijamali Arani, Abolfazl Mehbodniya, Mohammad Javad Omidi, Fumiyuki Adachi, Walid Saad, and Ismail Güvenç. Distributed learning for energy-efficient resource management in self-organizing heterogeneous networks. *IEEE Transactions on Vehicular Technology*, 66(10):9287–9303, 2017.

[18] IEEE Standards Association et al. Ieee standard for local and metropolitan area networks—timing and synchronization for time-sensitive applications in bridged local area networks. *IEEE Std*, 802.

[19] F Baker, J Polk, and M Dolly. A differentiated services code point (dscp) for capacity-admitted traffic. *Internet Engineering Task Force (IETF)*, 2010.

[20] Maurice Bellanger, Markku Renfors, Tero Ihalainen, and Carlos AF da Rocha. Ofdm and fbmc transmission techniques: a compatible high performance proposal for broadband power line communications. In *ISPLC2010*, pages 154–159. IEEE, 2010.

[21] Brahim Bensaou, Danny HK Tsang, and King Tung Chan. Credit-based fair queueing (cbfq): a simple service-scheduling algorithm for packet-switched networks. *IEEE/ACM Transactions on Networking*, 9(5):591–604, 2001.

[22] Sushmit Bhattacharjee, Robert Schmidt, Kostas Katsalis, Chia-Yu Chang, Thomas Bauschert, and Navid Nikaein. Time-sensitive networking for 5g fronthaul networks. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2020.

[23] Leonardo Bonati, Michele Polese, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. Open, programmable, and virtualized 5g networks: State-of-the-art and the road ahead. *Computer Networks*, 182:107516, 2020.

[24] David Boswarthick, Omar Elloumi, and Olivier Hersent. *M2M communications: a systems approach*. John Wiley & Sons, 2012.

[25] Anne Bouillard, Laurent Jouhet, and Eric Thierry. *Service curves in Network Calculus: dos and don'ts*. PhD thesis, INRIA, 2009.

[26] Karen Boulos, Melhem El Helou, and Samer Lahoud. Rrh clustering in cloud radio access networks. In *2015 International Conference on Applied Research in Computer Science and Engineering (ICAR)*, pages 1–6. IEEE, 2015.

[27] Karen Boulos, Kinda Khawam, Melhem El Helou, Marc Ibrahim, Steven Martin, and Hadi Sawaya. A hybrid approach for rrh clustering in cloud radio access networks based on game theory. In *Proceedings of the 16th ACM International Symposium on Mobility Management and Wireless Access*, pages 128–132, 2018.

[28] Karen Boulos, Kinda Khawam, Melhem El Helou, Marc Ibrahim, Hadi Sawaya, and Steven Martin. An efficient scheme for bbu-rrh association in c-ran architecture for joint power saving and re-association optimization. In *2018 IEEE 7th International Conference on Cloud Networking (CloudNet)*, pages 1–6. IEEE, 2018.

[29] Stefan Brueck, Lu Zhao, Jochen Giese, and M Awais Amin. Centralized scheduling for joint transmission coordinated multi-point in lte-advanced. In *2010 International ITG Workshop on Smart Antennas (WSA)*, pages 177–184. IEEE, 2010.

[30] Lauren Buckalew, Jeff Loucks, and James Macaulay. Internet of everything in the public sector: Generating value in an era of change.

[31] Yunlong Cai, Zhijin Qin, Fangyu Cui, Geoffrey Ye Li, and Julie A McCann. Modulation and multiple access for 5g networks. *IEEE Communications Surveys & Tutorials*, 20(1):629–646, 2017.

[32] Abdulkadir Celik, Redha M Radaydeh, Fawaz S Al-Qahtani, and Mohamed-Slim Alouini. Joint interference management and resource allocation for device-to-device (d2d) communications underlying downlink/uplink decoupled (dude) heterogeneous networks. In *2017 IEEE international conference on communications (ICC)*, pages 1–6. IEEE, 2017.

[33] Olfa Chabbouh, Sonia Rejeb, Zièd Choukair, and Nazim Agoulmine. A two-stage rrh clustering mechanism in 5g heterogeneous c-ran. In *5th International Workshop on ADVANCEs in ICT Infrastructures and Services (ADVANCE 2017)*, 2017.

[34] He Chen, Rana Abbas, Peng Cheng, Mahyar Shirvanimoghaddam, Wibowo Hardjawana, Wei Bao, Yonghui Li, and Branka Vucetic. Ultra-reliable low latency cellular networks: Use cases, challenges and approaches. *IEEE Communications Magazine*, 56(12):119–125, 2018.

[35] Kulin Chen and Run Duan. C-ran the road towards green ran. *China Mobile Research Institute, white paper*, 2, 2011.

[36] Xiaoming Chen, Derrick Wing Kwan Ng, Wei Yu, Erik G Larsson, Naofal Al-Dhahir, and Robert Schober. Massive access for 5g and beyond. *arXiv preprint arXiv:2002.03491*, 2020.

[37] Yu-Jia Chen, Li-Yu Cheng, and Li-Chun Wang. Prioritized resource reservation for reducing random access delay in 5g urllc. In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–5. IEEE, 2017.

[38] Jen-Po Cheng, Chia-han Lee, and Tzu-Ming Lin. Prioritized random access with dynamic access barring for ran overload in 3gpp lte-a networks. In *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, pages 368–372. IEEE, 2011.

[39] I Chih-Lin, Jinri Huang, Yannan Yuan, Shijia Ma, and Ran Duan. Ngfi, the xhaul. In *2015 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2015.

[40] I Chih-Lin, Yannan Yuan, Jinri Huang, Shijia Ma, Chunfeng Cui, and Ran Duan. Rethink fronthaul for soft ran. *IEEE Communications Magazine*, 53(9):82–88, 2015.

[41] Divya Chitimalla, Koteswararao Kondepu, Luca Valcarenghi, Massimo Tornatore, and Biswanath Mukherjee. 5g fronthaul–latency and jitter studies of cpri over ethernet. *Journal of Optical Communications and Networking*, 9(2):172–182, 2017.

[42] Seunghyun Choi, Wonbo Lee, Dongmyoung Kim, Kyung-Joon Park, Sunghyun Choi, and Ki-Young Han. Automatic configuration of random access channel parameters in lte systems. In *2011 IFIP Wireless Days (WD)*, pages 1–6. IEEE, 2011.

[43] Michal Cierny, Risto Wichman, and Zhi Ding. Impact of base station time synchronization mismatch on almost blank subframes. *IEEE communications letters*, 17(11):2092–2095, 2013.

[44] Carlos A Coello Coello, Gary B Lamont, David A Van Veldhuizen, et al. *Evolutionary algorithms for solving multi-objective problems*, volume 5. Springer, 2007.

[45] Li Da Xu, Wu He, and Shancang Li. Internet of things in industries: A survey. *IEEE Transactions on industrial informatics*, 10(4):2233–2243, 2014.

[46] Aleksandar Damnjanovic, Juan Montojo, Joonyoung Cho, Hyoungju Ji, Jin Yang, and Pingping Zong. Ue's role in lte advanced heterogeneous networks. *IEEE Communications Magazine*, 50(2):164–176, 2012.

[47] Aleksandar Damnjanovic, Juan Montojo, Yongbin Wei, Tingfang Ji, Tao Luo, Madhavan Vajapeyam, Taesang Yoo, Osok Song, and Durga Malladi. A survey on 3gpp heterogeneous networks. *IEEE Wireless communications*, 18(3):10–21, 2011.

[48] Panagiotis Demestichas, Andreas Georgakopoulos, Dimitrios Karvounas, Kostas Tsagkaris, Vera Stavroulaki, Jianmin Lu, Chunshan Xiong, and Jing Yao. 5g on the horizon: Key challenges for the radio-access network. *IEEE vehicular technology magazine*, 8(3):47–53, 2013.

[49] Zhiguo Ding, Xianfu Lei, George K Karagiannidis, Robert Schober, Jinhong Yuan, and Vijay K Bhargava. A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends. *IEEE Journal on Selected Areas in Communications*, 35(10):2181–2195, 2017.

[50] Amruta Sarvajeet Dixit, Sumit Kumar, Shabana Urooj, and Areej Malibari. A highly compact antipodal vivaldi antenna array for 5g millimeter wave applications. *Sensors*, 21(7):2360, 2021.

[51] Behrouz Farhang-Boroujeny. Ofdm versus filter bank multicarrier. *IEEE signal processing magazine*, 28(3):92–112, 2011.

[52] Anaïs Finzi, Emmanuel Lochin, Ahlem Mifdaoui, and Fabrice Frances. Improving rfc5865 core network scheduling with a burst limiting shaper. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pages 1–6. IEEE, 2017.

[53] Anaïs Finzi, Ahlem Mifdaoui, Fabrice Frances, and Emmanuel Lochin. Network calculus-based timing analysis of afdx networks with strict priority and tsn/bls shapers. In *2018 IEEE 13th International Symposium on Industrial Embedded Systems (SIES)*, pages 1–10. IEEE, 2018.

[54] GMDT Forecast. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. *Update*, 2017:2022, 2019.

[55] Cesar A Garcia-Perez and Pedro Merino. Enabling low latency services on lte networks. In *2016 IEEE 1st International Workshops on Foundations and Applications of Self* Systems (FAS* W)*, pages 248–255. IEEE, 2016.

[56] Liljana Gavrilovska, Valentin Rakovic, and Vladimir Atanasovski. Visions towards 5g: Technical requirements and potential enablers. *Wireless Personal Communications*, 87(3):731–757, 2016.

[57] Liljana Gavrilovska, Valentin Rakovic, and Daniel Denkovski. From cloud ran to open ran. *Wireless Personal Communications*, pages 1–17, 2020.

[58] Liljana Gavrilovska, Valentin Rakovic, Aleksandar Ichkov, Davor Todorovski, and Simona Marinova. Flexible c-ran: Radio technology for 5g. In *2017 13th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, pages 255–264. IEEE, 2017.

[59] Abiodun Omowunmi Gbenga-Ilori and Olufunmilayo Idayat Sanusi. Sequentially distributed coalition formation game for throughput maximization in c-rans. *International Journal of Electronics and Telecommunications*, 64(4):505–512, 2018.

[60] Franz-Josef Gotz. Traffic shaper for control data traffic (cdt). In *IEEE 802 AVB Meeting*, 2012.

[61] Ismail Guvenc. Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination. *IEEE Communications Letters*, 15(10):1084–1087, 2011.

[62] Johanna Heinonen, Tapio Partti, Marko Kallio, Kari Lappalainen, Hannu Flinck, and Jarmo Hillo. Dynamic tunnel switching for sdn-based cellular core networks. In *Proceedings of the 4th workshop on All things cellular: operations, applications, & challenges*, pages 27–32, 2014.

[63] Ekram Hossain and Monowar Hasan. 5g cellular: key enabling technologies and research challenges. *IEEE Instrumentation & Measurement Magazine*, 18(3):11–21, 2015.

[64] J Huang and Y Yuan. White paper of next generation fronthaul interface. *labs. chinamobile. com/cran, ver*, 1, 2015.

[65] Junwei Huang, Pengguang Zhou, Kai Luo, Zhiming Yang, and Gongcheng He. Two-stage resource allocation scheme for three-tier ultra-dense network. *China Communications*, 14(10):118–129, 2017.

[66] Common Public Radio Interface. Common public radio interface: ecpri interface specification. Technical report, Technical Report eCPRI specification, 2018.

[67] SM Riazul Islam, Daehan Kwak, MD Humaun Kabir, Mahmud Hossain, and Kyung-Sup Kwak. The internet of things for health care: a comprehensive survey. *IEEE access*, 3:678–708, 2015.

[68] Hyoungju Ji, Sunho Park, Jeongho Yeo, Younsun Kim, Juho Lee, and Byonghyo Shim. Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects. *IEEE Wireless Communications*, 25(3):124–130, 2018.

[69] Yinghao Jin and Ling Qiu. Joint user association and interference coordination in heterogeneous cellular networks. *IEEE Communications Letters*, 17(12):2296–2299, 2013.

[70] Ivan Jovović, Ivan Forenbacher, and Marko Periša. Massive machine-type communications: An overview and perspectives towards 5g. In *Proc. 3rd Int. Virtual Res. Conf. Tech. Disciplines*, volume 3, 2015.

[71] Megumi Kaneko, Toshihiko Nakano, Kazunori Hayashi, Takuya Kamenosono, and Hideaki Sakai. Distributed resource allocation with local csi overhearing and scheduling prediction for ofdma heterogeneous networks. *IEEE Transactions on Vehicular Technology*, 66(2):1186–1199, 2016.

[72] Mohammad T Kawser, HMAB Farid, Abduhu R Hasin, Adil MJ Sadik, and Ibrahim K Razu. Performance comparison between round robin and proportional fair scheduling methods for lte. *International Journal of Information and Electronics Engineering*, 2(5):678–681, 2012.

[73] Sam Kayyali. Resource management and quality of service provisioning in 5g cellular networks. *arXiv preprint arXiv:2008.09601*, 2020.

[74] Petteri Kela, Jani Puttonen, Niko Kolehmainen, Tapani Ristaniemi, Tero Henttonen, and Martti Moisio. Dynamic packet scheduling performance in utra long term evolution downlink. In *2008 3rd International Symposium on Wireless Pervasive Computing*, pages 308–313. IEEE, 2008.

[75] M Khan, RS Alhumaima, and HS Al-Raweshidy. Quality of service aware dynamic bbu-rrh mapping in cloud radio access network. In *2015 International Conference on Emerging Technologies (ICET)*, pages 1–5. IEEE, 2015.

[76] M Khan, Zainab H Fakhri, and Hamed S Al-Raweshidy. Semistatic cell differentiation and integration with dynamic bbu-rrh mapping in cloud radio access network. *IEEE Transactions on Network and Service Management*, 15(1):289–303, 2017.

[77] Muhammad Khan, Raad S Alhumaima, and Hamed S Al-Raweshidy. Qos-aware dynamic rrh allocation in a self-optimized cloud radio access network with rrh proximity constraint. *IEEE Transactions on Network and Service Management*, 14(3):730–744, 2017.

[78] Sajjad Ahmad Khan, Adnan Kavak, Kerem Küçük, et al. A novel fractional frequency reuse scheme for interference management in lte-a hetnets. *IEEE Access*, 7:109662–109672, 2019.

[79] Bernhard Korte and Jens Vygen. The knapsack problem. In *Combinatorial Optimization*, pages 397–406. Springer, 2000.

[80] Gwanmo Ku and John MacLaren Walsh. Resource allocation and link adaptation in lte and lte advanced: A tutorial. *IEEE communications surveys & tutorials*, 17(3):1605–1633, 2014.

[81] Pardeep Kumar, Sanjeev Kumar, and Chetna Dabas. Comparative analysis of downlink scheduling algorithms for a cell affected by interference in lte network. *Annals of Data Science*, 3(2):135–153, 2016.

[82] Quang Duy La, Yong Huat Chew, and Boon-Hee Soong. Potential games. In *Potential Game Theory*, pages 23–69. Springer, 2016.

[83] Eeva Lähetkangas, Kari Pajukoski, Esa Tiirola, Jyri Hämäläinen, and Zhong Zheng. On the performance of lte-advanced mimo: How to set and reach beyond 4g targets. In *European Wireless 2012; 18th European Wireless Conference 2012*, pages 1–6. VDE, 2012.

[84] Line MP Larsen, Aleksandra Checko, and Henrik L Christiansen. A survey of the functional splits proposed for 5g mobile crosshaul networks. *IEEE Communications Surveys & Tutorials*, 21(1):146–172, 2018.

[85] George Lawton. Machine-to-machine technology gears up for growth. *computer*, 37(9):12–15, 2004.

[86] Trung-Kien Le, Umer Salim, and Florian Kaltenberger. An overview of physical layer design for ultra-reliable low-latency communications in 3gpp releases 15, 16, and 17. *IEEE Access*, 2020.

[87] Jun Li, Lei Chen, and Jiajia Chen. Enabling technologies for low-latency service migration in 5g transport networks. *Journal of Optical Communications and Networking*, 13(2):A200–A210, 2021.

[88] Qian Clara Li, Huaning Niu, Apostolos Tolis Papathanassiou, and Geng Wu. 5g network capacity: Key elements and technologies. *IEEE Vehicular Technology Magazine*, 9(1):71–78, 2014.

[89] Yue Li, Yue Ma, Zhenyu Yin, Ai Gu, and Si Sun. A time-sensitive streams management method based on ieee 802.1 qat srp for industrial internet. In *2019 1st International Conference on Industrial Artificial Intelligence (IAI)*, pages 1–5. IEEE, 2019.

[90] Shao-Yu Lien, Kwang-Cheng Chen, and Yonghua Lin. Toward ubiquitous massive accesses in 3gpp machine-to-machine communications. *IEEE Communications Magazine*, 49(4):66–74, 2011.

[91] Hyung-Taek Lim, Daniel Herrscher, Martin Johannes Waltl, and Firas Chaari. Performance analysis of the ieee 802.1 ethernet audio/video bridging standard. *SimuTools*, 3:27–36, 2012.

[92] Dantong Liu, Lifeng Wang, Yue Chen, Maged Elkashlan, Kai-Kit Wong, Robert Schober, and Lajos Hanzo. User association in 5g networks: A survey and an outlook. *IEEE Communications Surveys & Tutorials*, 18(2):1018–1044, 2016.

[93] Deming Liu and Yann-Hang Lee. An efficient scheduling discipline for packet switching networks using earliest deadline first round robin. *Telecommunication Systems*, 28(3):453–474, 2005.

[94] Jingchu Liu, Tao Zhao, Sheng Zhou, Yu Cheng, and Zhisheng Niu. Concert: a cloud-based architecture for next-generation cellular systems. *IEEE Wireless Communications*, 21(6):14–22, 2014.

[95] Rui Liu, Qimei Chen, Guanding Yu, and Geoffrey Ye Li. Joint user association and resource allocation for multi-band millimeter-wave heterogeneous networks. *IEEE Transactions on Communications*, 67(12):8502–8516, 2019.

[96] Anthony Lo, Yee Wei Law, Martin Jacobsson, and Michal Kucharzak. Enhanced lte-advanced random-access mechanism for massive machine-to-machine (m2m) communications. In *27th World Wireless Research Forum (WWRF) Meeting*, volume 2011. Dusseldorf, Germany Dusseldorf, Germany, 2011.

[97] M Luby, A Shokrollahi, M Watson, et al. Qualcomm incorporated. 2011.

[98] Hua Luo. System capacity enhancement for 5g network and beyond. 2017.

[99] Mohammed Yazid Lyazidi, Nadjib Aitsaadi, and Rami Langar. Dynamic resource allocation for cloud-ran in lte with real-time bbu/rrh assignment. In *2016 IEEE international conference on communications (ICC)*, pages 1–6. IEEE, 2016.

[100] Mohammed Yazid Lyazidi, Nadjib Aitsaadi, and Rami Langar. A dynamic resource allocation framework in lte downlink for cloud-radio access network. *Computer Networks*, 140:101–111, 2018.

[101] Mohammed Yazid Lyazidi, Lorenza Giupponi, Josep Mangues-Bafalluy, Nadjib Aitsaadi, and Rami Langar. A novel optimization framework for c-ran bbu selection based on resiliency and price. In *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pages 1–6. IEEE, 2017.

[102] Mukesh Kumar Maheshwari, Mamta Agiwal, Navrati Saxena, and Abhishek Roy. Directional discontinuous reception (ddrx) for mmwave enabled 5g communications. *IEEE Transactions on Mobile Computing*, 18(10):2330–2343, 2018.

[103] Sulastri Manap, Kaharudin Dimyati, Mhd Nour Hindia, Mohamad Sofian Abu Talip, and Rahim Tafazolli. Survey of radio resource management in 5g heterogeneous networks. *IEEE Access*, 8:131202–131223, 2020.

[104] Silvano Martello. Knapsack problems: algorithms and computer implementations. *Wiley-Interscience series in discrete mathematics and optimiza tion*, 1990.

[105] Barbara Martini, Federica Paganelli, Paola Cappanera, Stefano Turchi, and Piero Castoldi. Latency-aware composition of virtual functions in 5g. In *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*, pages 1–6. IEEE, 2015.

[106] Maximilian Matthe, Luciano Leonel Mendes, Nicola Michailow, Dan Zhang, and Gerhard Fettweis. Widely linear estimation for space-time-coded gfdm in low-latency applications. *IEEE Transactions on Communications*, 63(11):4501–4509, 2015.

[107] Albena Mihovska, Ramjee Prasad, et al. Spectrum sharing and dynamic spectrum management techniques in 5g and beyond networks: A survey. *Journal of Mobile Multimedia*, pages 65–78, 2021.

[108] Rashid Mijumbi, Joan Serrat, Juan-Luis Gorricho, Niels Bouten, Filip De Turck, and Raouf Boutaba. Network function virtualization: State-of-the-art and research challenges. *IEEE Communications surveys & tutorials*, 18(1):236–262, 2015.

[109] Debashisha Mishra, PC Amogh, Arun Ramamurthy, A Antony Franklin, and Bheemarjuna Reddy Tamma. Load-aware dynamic rrh assignment in cloud radio access networks. In *2016 IEEE Wireless Communications and Networking Conference*, pages 1–6. IEEE, 2016.

[110] Bertrand Muquet, Zhengdao Wang, Georgios B Giannakis, Marc De Courville, and Pierre Duhamel. Cyclic prefixing or zero padding for wireless multicarrier transmissions? *IEEE Transactions on communications*, 50(12):2136–2148, 2002.

[111] Diala Naboulsi, Assia Mermouri, Razvan Stanica, Herve Rivano, and Marco Fiore. On user mobility in dynamic cloud radio access networks. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1583–1591. IEEE, 2018.

[112] Wooseok Nam, Dongwoon Bai, Jungwon Lee, and Inyup Kang. Advanced interference management for 5g cellular networks. *IEEE Communications Magazine*, 52(5):52–60, 2014.

[113] Ahmed Nasrallah, Akhilesh S Thyagaturu, Ziyad Alharbi, Cuixiang Wang, Xing Shao, Martin Reisslein, and Hesham ElBakoury. Ultra-low latency (ull) networks: The ieee tsn and ietf detnet standards and related 5g ull research. *IEEE Communications Surveys & Tutorials*, 21(1):88–145, 2018.

[114] Solmaz Niknam, Abhishek Roy, Harpreet S Dhillon, Sukhdeep Singh, Rahul Banerji, Jeffery H Reed, Navrati Saxena, and Seungil Yoon. Intelligent o-ran for beyond 5g and 6g wireless networks. *arXiv preprint arXiv:2005.08374*, 2020.

[115] Homayoun Nikookar. *Wavelet radio: adaptive and reconfigurable wireless systems based on wavelets*. Cambridge University Press, 2013.

[116] Bruno Astuto A Nunes, Marc Mendonca, Xuan-Nam Nguyen, Katia Obraczka, and Thierry Turletti. A survey of software-defined networking: Past, present, and future of programmable networks. *IEEE Communications surveys & tutorials*, 16(3):1617–1634, 2014.

[117] Ogechi Akudo Nwogu, Gladys Diaz, and Marwen Abdennebi. An optimized approach to load balancing and resource usage in 5g multi-tiered cellular networks. In *2020 Global Information Infrastructure and Networking Symposium (GIIS)*, pages 1–5. IEEE, 2020.

[118] Jinyoung Oh and Youngnam Han. Cell selection for range expansion with almost blank subframe in heterogeneous networks. In *2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications-(PIMRC)*, pages 653–657. IEEE, 2012.

[119] Jannis Ohms, Martin Böhm, and Diederich Wermser. Concept of a tsn to real-time wireless gateway in the context of 5g urllc. In *2020 8th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 1–6. IEEE, 2020.

[120] Kenta Okino, Taku Nakayama, Chiharu Yamazaki, Hirotaka Sato, and Yoshimasa Kusano. Pico cell range expansion with interference mitigation toward lte-advanced heterogeneous networks. In *2011 IEEE International Conference on Communications Workshops (ICC)*, pages 1–5. IEEE, 2011.

[121] Jérémy Pagé and Jean-Michel Dricot. Software-defined networking for low-latency 5g core network. In *2016 International Conference on Military Communications and Information Systems (ICMCIS)*, pages 1–7. IEEE, 2016.

[122] Imtiaz Parvez, Ali Rahmati, Ismail Guvenc, Arif I Sarwat, and Huaiyu Dai. A survey on low latency towards 5g: Ran, core network and caching solutions. *IEEE Communications Surveys & Tutorials*, 20(4):3098–3130, 2018.

[123] Gabriel Otero Pérez, David Larrabeiti López, and José Alberto Hernández. 5g new radio fronthaul network design for ecpri-ieee 802.1 cm and extreme latency percentiles. *IEEE Access*, 7:82218–82230, 2019.

[124] Javier Pérez Santacruz, Simon Rommel, Ulf Johannsen, Antonio Jurado-Navas, and Idelfonso Tafur Monroy. Candidate waveforms for arof in beyond 5g. *Applied Sciences*, 10(11):3891, 2020.

[125] Giuseppe Piro, Luigi Alfredo Grieco, Gennaro Boggia, Francesco Capozzi, and Pietro Camarda. Simulating lte cellular systems: An open-source framework. *IEEE transactions on vehicular technology*, 60(2):498–513, 2010.

[126] Guillermo Pocovi, Hamidreza Shariatmadari, Gilberto Berardinelli, Klaus Pedersen, Jens Steiner, and Zexian Li. Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements. *IEEE Network*, 32(2):8–15, 2018.

[127] Manli Qian, Wibowo Hardjawana, Jinglin Shi, and Branka Vucetic. Baseband processing units virtualization for cloud radio access networks. *IEEE Wireless Communications Letters*, 4(2):189–192, 2015.

[128] Olav Queseth, Ömer Bulakci, Panagiotis Spapis, Pascal Bisson, Patrick Marsch, Paul Arnold, Peter Rost, Qi Wang, Rolf Blom, Stefano Salsano, et al. 5g ppp architecture working group: View on 5g architecture (version 2.0, december 2017). 2017.

[129] Huda Adibah Mohd Ramli, Riyaj Basukala, Kumbesan Sandrasegaran, and Rachod Patachaianand. Performance of well known packet scheduling algorithms in the downlink 3gpp lte system. In *2009 IEEE 9th Malaysia international conference on communications (MICC)*, pages 815–820. IEEE, 2009.

[130] Peter Rost, Carlos J Bernardos, Antonio De Domenico, Marco Di Girolamo, Massinissa Lalam, Andreas Maeder, Dario Sabella, and Dirk Wübben. Cloud technologies for flexible 5g radio access networks. *IEEE Communications Magazine*, 52(5):68–76, 2014.

[131] Ibrahim Saidu, Shamala Subramaniam, Azmi Jaafar, and Zuriati Ahmad Zukarnain. A load-aware weighted round-robin algorithm for ieee 802.16

networks. *EURASIP Journal on Wireless Communications and Networking*, 2014(1):1–12, 2014.

[132] Tabinda Salam, Waheed Ur Rehman, and Xiaofeng Tao. Data aggregation in massive machine type communication: Challenges and solutions. *IEEE Access*, 7:41921–41946, 2019.

[133] Mahadev Satyanarayanan, Paramvir Bahl, Ramón Caceres, and Nigel Davies. The case for vm-based cloudlets in mobile computing. *IEEE pervasive Computing*, 8(4):14–23, 2009.

[134] Frank Schaich, Thorsten Wild, and Yejian Chen. Waveform contenders for 5g-suitability for short packet and low latency transmissions. In *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2014.

[135] M Series. Future technology trends of terrestrial imt systems. *Int. Telecommun. Union, Geneva, Switzerland, Rep. ITU*, pages 2320–0, 2015.

[136] M Series. Imt vision–framework and overall objectives of the future development of imt for 2020 and beyond. *Recommendation ITU*, 2083, 2015.

[137] Tshiamo Sigwele, Atm Shafiul Alam, Prashant Pillai, and Y Fun Hu. Evaluating energy-efficient cloud radio access networks for 5g. In *2015 IEEE International Conference on Data Science and Data Intensive Systems*, pages 362–367. IEEE, 2015.

[138] David Soldani and Antonio Manzalini. Horizon 2020 and beyond: On the 5g operating system for a true digital society. *IEEE Vehicular Technology Magazine*, 10(1):32–42, 2015.

[139] Beatriz Soret, Preben Mogensen, Klaus I Pedersen, and Mari Carmen Aguayo-Torres. Fundamental tradeoffs among reliability, latency and throughput in cellular networks. In *2014 IEEE Globecom Workshops (GC Wkshps)*, pages 1391–1396. IEEE, 2014.

[140] IHS Statista. Internet of things (iot) connected devices installed base worldwide from 2015 to 2025 (in billions), 2018.

[141] Withawat Tangtrongpairoj, Takeshi Higashino, and Minoru Okada. Handover reduction using optical matrix switch for centralized radio access network. *IEICE Technical Report; IEICE Tech. Rep.*, 116(204):65–68, 2016.

[142] Luis Tello-Oquendo, José-Ramón Vidal, Vicent Pla, and Luis Guijarro. Dynamic access class barring parameter tuning in lte-a networks with massive m2m traffic. In *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, pages 1–8. IEEE, 2018.

[143] Sivakumar Thangamuthu, Nicola Concer, Pieter JL Cuijpers, and Johan J Lukkien. Analysis of ethernet-switch traffic shapers for in-vehicle networking applications. In *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 55–60. IEEE, 2015.

[144] Daniel Thiele and Rolf Ernst. Formal worst-case timing analysis of ethernet tsn's burst-limiting shaper. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 187–192. IEEE, 2016.

[145] Joe F Thompson, Zahir UA Warsi, and C Wayne Mastin. *Numerical grid generation*. Number BOOK. North Holland, 1985.

[146] Jayashree Thota and Adnan Aijaz. On performance evaluation of random access enhancements for 5g urllc. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–7. IEEE, 2019.

[147] Matteo Vincenzi, Angelos Antonopoulos, Elli Kartsakli, John Vardakas, Luis Alonso, and Christos Verikoukis. Cooperation incentives for multi-operator c-ran energy efficient sharing. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[148] Feng Wang, Wen Chen, Hongying Tang, and Qingqing Wu. Joint optimization of user association, subchannel allocation, and power allocation in multi-cell multi-association ofdma heterogeneous networks. *IEEE Transactions on Communications*, 65(6):2672–2684, 2017.

[149] Gang Wang, Gang Feng, Shuang Qin, and Ruihan Wen. Efficient traffic engineering for 5g core and backhaul networks. *Journal of Communications and Networks*, 19(1):80–92, 2017.

[150] Kaiwei Wang, Ming Zhao, and Wuyang Zhou. Traffic-aware graph-based dynamic frequency reuse for heterogeneous cloud-ran. In *2014 IEEE Global Communications Conference*, pages 2308–2313. IEEE, 2014.

[151] Klaus Wehrle, Mesut Günes, and James Gross. *Modeling and tools for network simulation*. Springer Science & Business Media, 2010.

[152] Dimas Tribudi Wiriaatmadja and Kae Won Choi. Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks. *IEEE Transactions on Wireless Communications*, 14(1):33–46, 2014.

[153] Bei Xie, Zekun Zhang, Rose Qingyang Hu, Geng Wu, and Apostolos Papathanassiou. Joint spectral efficiency and energy efficiency in ffr-based wireless heterogeneous networks. *IEEE Transactions on Vehicular technology*, 67(9):8154–8168, 2017.

[154] Yongjun Xu, Yuan Hu, Qianbin Chen, Tiecheng Song, and Rong Lai. Robust resource allocation for multi-tier cognitive heterogeneous networks. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[155] Chungang Yang, Jia Xiao, Jiandong Li, Xiaoqiang Shao, Alagan Anpalagan, Qiang Ni, and Mohsen Guizani. Disco: Interference-aware distributed cooperation with incentive mechanism for 5g heterogeneous ultra-dense networks. *IEEE Communications Magazine*, 56(7):198–204, 2018.

[156] Mohamad Yassin. *Inter-Cell Interference Coordination in Wireless Networks.(Coordination des interférences intercellulaires dans les réseaux sans-fil).* PhD thesis, University of Rennes 1, France, 2015.

[157] Fang Ye, Jing Dai, and Yibing Li. Hybrid-clustering game algorithm for resource allocation in macro-femto hetnet. *TIIS*, 12(4):1638–1654, 2018.

[158] Qiaoyang Ye, Mazin Al-Shalash, Constantine Caramanis, and Jeffrey G Andrews. On/off macrocells and load balancing in heterogeneous cellular networks. In *2013 IEEE global communications conference (GLOBECOM)*, pages 3814–3819. IEEE, 2013.

[159] Turker Yilmaz, Naveed A Abbasi, and Ozgur B Akan. Millimeter-wave 5g-enabled internet of things. *5G-Enabled Internet of Things*, page 163, 2019.

[160] Ali A Zaidi, Robert Baldemair, Hugo Tullberg, Hakan Bjorkegren, Lars Sundstrom, Jonas Medbo, Caner Kilinc, and Icaro Da Silva. Waveform and numerology to support 5g services and requirements. *IEEE Communications Magazine*, 54(11):90–98, 2016.

[161] Zainab Zaidi, Vasilis Friderikos, Zarrar Yousaf, Simon Fletcher, Mischa Dohler, and Hamid Aghvami. Will sdn be part of 5g? *IEEE Communications Surveys & Tutorials*, 20(4):3220–3258, 2018.

[162] Anna Zakrzewska, Sarah Ruepp, and Michael S Berger. Towards converged 5g mobile networks-challenges and current trends. In *Proceedings of the 2014 ITU kaleidoscope academic conference: Living in a converged world-Impossible without standards?*, pages 39–45. IEEE, 2014.

[163] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. Internet of things for smart cities. *IEEE Internet of Things journal*, 1(1):22–32, 2014.

[164] Nawel Zangar, Sami Gharbi, and Marwen Abdennebi. Service differentiation strategy based on macb factor for m2m communications in lte-a networks. In *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 693–698. IEEE, 2016.

[165] Gongzheng Zhang, Tony QS Quek, Aiping Huang, Marios Kountouris, and Hangguan Shan. Backhaul-aware base station association in two-tier heterogeneous cellular networks. In *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 390–394. IEEE, 2015.

[166] Jiayi Zhang, Emil Björnson, Michail Matthaiou, Derrick Wing Kwan Ng, Hong Yang, and David J Love. Prospective multiple antenna technologies for beyond 5g. *IEEE Journal on Selected Areas in Communications*, 38(8):1637–1660, 2020.

[167] Xi Zhang, Ming Jia, Lei Chen, Jianglei Ma, and Jing Qiu. Filtered-ofdm-enabler for flexible waveform in the 5th generation cellular networks. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.

[168] Zhao Zhao, Malte Schellmann, Qi Wang, Xitao Gong, Ronald Boehnke, and Wen Xu. Pulse shaped ofdm for asynchronous uplink access. In *2015 49th Asilomar Conference on Signals, Systems and Computers*, pages 3–7. IEEE, 2015.