TECHNISCHE UNIVERSITÄT DRESDEN

Analysis of Bandwidth and Latency Constraints on a Packetized Cloud Radio Access Network Fronthaul

Jay Kant Chaudhary

der Fakultät Elektrotechnik und Informationstechnik
der Technischen Universität Dresden

zur Erlangung des akademischen Grades eines

Doktoringenieurs
(Dr.-Ing.)

genehmigte Dissertation

Vorsitzender:	Prof. Dr.-Ing. habil. Christian Georg Mayr
Gutachter:	Prof. Dr.-Ing. Dr. h.c. Gerhard Fettweis
	Prof. Dr. Steinar Bjørnstad
Tag der Einreichung:	29. Oktober 2019
Tag der Verteidigung:	26. Februar 2020

Jay Kant Chaudhary
Analysis of Bandwidth and Latency Constraints on a Packetized Cloud Radio Access Network Fronthaul
Dissertation, 26. Februar 2020

Vodafone Chair Mobile Communications Systems
Institut für Nachrichtentechnik
Fakultät Elektrotechnik und Informationstechnik
Technische Universität Dresden
01062 Dresden

Abstract

Cloud radio access network (C-RAN) is a promising architecture for the next-generation RAN to meet the diverse and stringent requirements envisioned by fifth generation mobile communication systems (5G) and future generation mobile networks. C-RAN offers several advantages, such as reduced capital expenditure (CAPEX) and operational expenditure (OPEX), increased spectral efficiency (SE), higher capacity and improved cell-edge performance, and efficient hardware utilization through resource sharing and network function virtualization (NFV). However, these centralization gains come with the need for a *fronthaul*, which is the transport link connecting remote radio units (RRUs) to the base band unit (BBU) pool. In conventional C-RAN, legacy common public radio interface (CPRI) protocol is used on the fronthaul network to transport the raw, unprocessed baseband in-phase/quadrature-phase (I/Q) samples between the BBU and the RRUs, and it demands a huge fronthaul bandwidth, a strict low-latency, in the order of a few hundred microseconds, and a very high reliability.

Hence, in order to relax the excessive fronthaul bandwidth and stringent low-latency requirements, as well as to enhance the flexibility of the fronthaul, it is utmost important to redesign the fronthaul, while still profiting from the acclaimed centralization benefits. Therefore, a flexibly centralized C-RAN with different functional splits has been introduced. In addition, 5G mobile fronthaul (often also termed as an *evolved fronthaul*) is envisioned to be *packet-based*, utilizing the Ethernet as a transport technology.

In this thesis, to circumvent the fronthaul bandwidth constraint, a packetized fronthaul considering an appropriate functional split such that the fronthaul data rate is coupled with actual user data rate, unlike the classical C-RAN where fronthaul data rate is always *static* and *independent* of the traffic load, is justifiably chosen. We adapt queuing and spatial traffic models to derive the mathematical expressions for statistical multiplexing gains that can be obtained from the randomness in the user traffic. Through this, we show that the required fronthaul bandwidth can be reduced significantly, depending on the overall traffic demand, correlation distance and outage probability. Furthermore, an iterative optimization algorithm is developed, showing the impacts of number of pilots on a bandwidth-constrained fronthaul. This algorithm achieves additional reduction in the required fronthaul bandwidth.

Next, knowing the multiplexing gains and possible fronthaul bandwidth reduction, it is beneficial for the mobile network operators (MNOs) to deploy the optical transceiver (TRX) modules in C-RAN cost efficiently. For this, using the same framework, a cost model for fronthaul TRX cost optimization is presented. This is essential in C-RAN, because in a wavelength division multiplexing-passive optical network (WDM-PON) system, TRXs are generally deployed to serve at a peak load. But, because of variations in the traffic demands, owing to *tidal effect*, the fronthaul can be dimensioned requiring a lower capacity allowing a reasonable outage, thus giving rise to cost saving by deploying fewer TRXs, and energy saving by putting the unused TRXs in sleep mode.

The second focus of the thesis is the fronthaul latency analysis, which is a critical performance metric, especially for ultra-reliable and low latency communication (URLLC). An analytical framework to calculate the latency in the uplink (UL) of C-RAN massive multiple-input multiple-output (MIMO) system is presented. For this, a continuous-time queuing model for the Ethernet switch in the fronthaul network, which aggregates the UL traffic from several massive MIMO-aided RRUs, is considered. The closed-form solutions for the moment generating function (MGF) of sojourn time, waiting time and queue length distributions are derived using Pollaczek–Khinchine formula for our **M/HE/1** queuing model, and evaluated via numerical solutions. In addition, the packet loss rate – due to the inability of the packets to reach the destination in a certain time – is derived. Due to the slotted nature of the UL transmissions, the model is extended to a discrete-time queuing model. The impact of the packet arrival rate, average packet size, SE of users, and fronthaul capacity on the sojourn time, waiting time and queue length distributions are analyzed.

While offloading more signal processing functionalities to the RRU reduces the required fronthaul bandwidth considerably, this increases the complexity at the RRU. Hence, considering the 5G New Radio (NR) flexible numerology and XRAN functional split with a detailed radio frequency (RF) chain at the RRU, the total RRU complexity is computed first, and later, a tradeoff between the required fronthaul bandwidth and RRU complexity is analyzed.

We conclude that despite the numerous C-RAN benefits, the stringent fronthaul bandwidth and latency constraints must be carefully evaluated, and an optimal functional split is essential to meet diverse set of requirements imposed by new radio access technologies (RATs).

Kurzfassung

Ein cloud-basiertes Mobilfunkzugangsnetz (cloud radio access network, C-RAN) stellt eine vielversprechende Architektur für das RAN der nächsten Generation dar, um die vielfältigen und strengen Anforderungen der fünften (5G) und zukünftigen Generationen von Mobilfunknetzen zu erfüllen. C-RAN bietet mehrere Vorteile, wie z.B. reduzierte Investitions- (CAPEX) und Betriebskosten (OPEX), erhöhte spektrale Effizienz (SE), höhere Kapazität und verbesserte Leistung am Zellrand sowie effiziente Hardwareauslastung durch Ressourcenteilung und Virtualisierung von Netzwerkfunktionen (network function virtualization, NFV). Diese Zentralisierungsvorteile erfordern jedoch eine Transportverbindung (*Fronthaul*), die die Antenneneinheiten (remote radio units, RRUs) mit dem Pool an Basisbandeinheiten (basisband unit, BBU) verbindet. Im konventionellen C-RAN wird das bestehende CPRI-Protokoll (common public radio interface) für das Fronthaul-Netzwerk verwendet, um die rohen, unverarbeiteten Abtastwerte der In-Phase- und Quadraturkomponente (I/Q) des Basisbands zwischen der BBU und den RRUs zu transportieren. Dies erfordert eine enorme Fronthaul-Bandbreite, eine strenge niedrige Latenz in der Größenordnung von einigen hundert Mikrosekunden und eine sehr hohe Zuverlässigkeit.

Um die extrem große Fronthaul-Bandbreite und die strengen Anforderungen an die geringe Latenz zu lockern und die Flexibilität des Fronthauls zu erhöhen, ist es daher äußerst wichtig, das Fronthaul neu zu gestalten und dabei trotzdem von den erwarteten Vorteilen der Zentralisierung zu profitieren. Daher wurde ein flexibel zentralisiertes C-RAN mit unterschiedlichen Funktionsaufteilungen eingeführt. Außerdem ist das mobile 5G-Fronthaul (oft auch als *evolved Fronthaul* bezeichnet) als *paketbasiert* konzipiert und nutzt Ethernet als Transporttechnologie.

Um die Bandbreitenbeschränkung zu erfüllen, wird in dieser Arbeit ein paketbasiertes Fronthaul unter Berücksichtigung einer geeigneten funktionalen Aufteilung so gewählt, dass die Fronthaul-Datenrate mit der tatsächlichen Nutzdatenrate gekoppelt wird, im Gegensatz zum klassischen C-RAN, bei dem die Fronthaul-Datenrate immer *statisch* und *unabhängig* von der Verkehrsbelastung ist. Wir passen Warteschlangen- und räumliche Verkehrsmodelle an, um mathematische Ausdrücke für statistische Multiplexing-Gewinne herzuleiten, die aus der Zufälligkeit im Benutzerverkehr gewonnen werden können. Hierdurch zeigen wir, dass die erforderliche Fronthaul-Bandbreite abhängig von der Gesamtverkehrsnachfrage, der Korrelationsdistanz und der Ausfallwahrscheinlichkeit deutlich reduziert werden kann. Darüber hinaus wird ein iterativer Optimierungsalgorithmus entwickelt, der die Auswirkungen der Anzahl der Piloten auf das bandbreitenbeschränkte Fronthaul zeigt. Dieser Algorithmus erreicht eine zusätzliche Reduktion der benötigte Fronthaul-Bandbreite.

Mit dem Wissen über die Multiplexing-Gewinne und die mögliche Reduktion der Fronthaul-Bandbreite ist es für die Mobilfunkbetreiber (mobile network operators, MNOs) von Vorteil, die Module des optischen Sendeempfängers (transceiver, TRX) kostengün-

stig im C-RAN einzusetzen. Dazu wird unter Verwendung des gleichen Rahmenwerks ein Kostenmodell zur Fronthaul-TRX-Kostenoptimierung vorgestellt. Dies ist im C-RAN unerlässlich, da in einem WDM-PON-System (wavelength division multiplexing-passive optical network) die TRX im Allgemeinen bei Spitzenlast eingesetzt werden. Aufgrund der Schwankungen in den Verkehrsanforderungen (*Gezeiteneffekt*) kann das Fronthaul jedoch mit einer geringeren Kapazität dimensioniert werden, die einen vertretbaren Ausfall in Kauf nimmt, was zu Kosteneinsparungen durch den Einsatz von weniger TRXn und Energieeinsparungen durch den Einsatz der ungenutzten TRX im Schlafmodus führt.

Der zweite Schwerpunkt der Arbeit ist die Fronthaul-Latenzanalyse, die eine kritische Leistungskennzahl liefert, insbesondere für die hochzuverlässige und niedriglatente Kommunikation (ultra-reliable low latency communications, URLLC). Ein analytisches Modell zur Berechnung der Latenz im Uplink (UL) des C-RAN mit massivem MIMO (multiple input multiple output) wird vorgestellt. Dazu wird ein Warteschlangen-Modell mit kontinuierlicher Zeit für den Ethernet-Switch im Fronthaul-Netzwerk betrachtet, das den UL-Verkehr von mehreren RRUs mit massivem MIMO aggregiert. Die geschlossenen Lösungen für die momenterzeugende Funktion (moment generating function, MGF) von Verweildauer-, Wartezeit- und Warteschlangenlängenverteilungen werden mit Hilfe der Pollaczek-Khinchin-Formel für unser $\mathbf{M/HE/1}$-Warteschlangenmodell hergeleitet und mittels numerischer Verfahren ausgewertet. Darüber hinaus wird die Paketverlustrate derjenigen Pakete, die das Ziel nicht in einer bestimmten Zeit erreichen, hergeleitet. Aufgrund der Organisation der UL-Übertragungen in Zeitschlitzen wird das Modell zu einem Warteschlangenmodell mit diskreter Zeit erweitert. Der Einfluss der Paketankunftsrate, der durchschnittlichen Paketgröße, der SE der Benutzer und der Fronthaul-Kapazität auf die Verweildauer-, die Wartezeit- und die Warteschlangenlängenverteilung wird analysiert.

Während das Verlagern weiterer Signalverarbeitungsfunktionalitäten an die RRU die erforderliche Fronthaul-Bandbreite erheblich reduziert, erhöht sich dadurch im Gegenzug die Komplexität der RRU. Daher wird unter Berücksichtigung der flexiblen Numerologie von 5G New Radio (NR) und der XRAN-Funktionenaufteilung mit einer detaillierten RF-Kette (radio frequency) am RRU zunächst die gesamte RRU-Komplexität berechnet und später ein Kompromiss zwischen der erforderlichen Fronthaul-Bandbreite und der RRU-Komplexität untersucht.

Wir kommen zu dem Schluss, dass trotz der zahlreichen Vorteile von C-RAN die strengen Bandbreiten- und Latenzbedingungen an das Fronthaul sorgfältig geprüft werden müssen und eine optimale funktionale Aufteilung unerlässlich ist, um die vielfältigen Anforderungen der neuen Funkzugangstechnologien (radio access technologies, RATs) zu erfüllen.

Acknowledgement

This thesis is the result of continued effort of four years of a long journey at the Vodafone Chair Mobile Communication Systems at the TU Dresden. This journey seemed to me like a roller-coaster ride with many joys and a few hurdles in between. Despite me being the actor, I would like to take this opportunity to thank all the directors of this research work, who played quite an influential role, directly and indirectly.

First and foremost, I would like to express my deepest gratitude to my doctoral supervisor Prof. Gerhard Fettweis for having trust in my potential for this research, for providing his continuous support over the last years, and concrete guidelines and advices during my study and work at the Vodafone Chair Mobile Communication Systems at the TU Dresden. Particularly, I am overwhelmed by his capability of providing constructive ideas and immediate approaches towards problem solution, and not to forget, by his friendly and motivational behaviour.

Further, I would like to thank Prof. Steinar Bjørnstad, the second referee, for his time to review my thesis and providing valuable feedback and suggestion.

I am also grateful to my current group leader Dr. André Noll Barreto and former group leaders Dr. Dan Zhang and Dr. Meryem Simsek. Their willingness to help, detailed insights during numerous meetings, doksems and whiteboad brainstorming sessions provided fruitful outcomes, which paved a way in right direction for the completeness of this thesis.

A special thanks goes to my colleagues at the chair, most importantly Dr. Jobin Francis and Dr. Jens Bartelt for their productive collaboration as well fruitful discussions. Not to forget, in a non-exhaustive list, I would also like to thank other colleagues at the chair: Ahmad, Atul, Behnam, David, Henrik, Lucas, Max, Philipp, Tom, and Waqar. With many of them, I often had business travel, free time, and numerous technical and non-technical discussions. I owe you all guys tons of thanks.

During my PhD work, I had the pleasure of working with the EU projects: 5G-XHaul and 5G-PICTURE, and my research topic was mostly aligned with their objectives. I would like to thank all the involved project members, particularly Eckhard Grass, Jesús Gutiérrez and Jim Zou.

In addition, here at the chair many administrative tasks, especially applying for travel and reimbursements, contract extension, etc. would have been a big mess and troublesome without the continued support from Kathrin Fromke, Sylvia Steppat and Eva Bolza-Schünemann. Furthermore, I have received much help and support from our IT specialists: Raffael Kozerski and Rüdiger Hartmann. Hence, my heart cannot refrain me from offering them big thanks for their support.

My studies would not have been possible without the financial and moral support of my parents, since the time eight years ago when I moved to Finland for my master study, and later to Germany for a PhD. Especially, my father's desire that his son should have a doctoral degree always gave me a motivational and inspirational boost. I am truly indebted to them. Last but quite importantly, I am thankful to my wife Pinki, a.k.a., my better half, for her unconditional love, and unquestionable support and care. I still recall her motivation and encouragement, especially at the time when I was having many ups and downs. I am hopeful to have such a great journey ahead together with our little angel, Aayara. Dear wife, let me beside by you forever.

Dresden, October, 2019 Jay Kant Chaudhary

Contents

Chapter 1

Introduction

1.1 Motivation

Since the introduction of the first generation mobile communication systems (1G) in 1980's, cellular mobile communications have witnessed a phenomenal growth over the recent decades. While 1G was analog only, the second generation mobile communication systems (2G) extended its capability to digital with voice, short message service (SMS) and very limited data services. With the introduction of the third generation mobile communication systems (3G) and fourth generation mobile communication systems (4G), the circuit-switched network in 2G was shifted towards packet-switched network providing massive mobile broadband, ubiquitous connectivity and on-demand video streaming, to name a few. Ever increasing continued demand of mobile broadband services is surging, requiring yet another generation of mobile technology to address the emerging challenges. The current Cisco Visual Networking Index (VNI) forecasts in [Cis19] that the overall mobile data traffic will grow to 77.5 exabytes per month by 2022, showing nearly a threefold increase compared with that in 2019.

As the quest for quality of service (QoS) and quality of experience (QoE) is continuously evolving, standardization bodies and industry forums, such as the next generation mobile network (NGMN) [NGM15b], third generation partnership project (3GPP) [3GP17a] and fifth generation public private partnership (5GPPP) [5GP19] have started working on the next-generation mobile technology (5G), which is considered to be both evolutionary and revolutionary[1] from the state-of-the-art (SoTA) 4G technologies. 5G aims to provide not only massive capacity and massive connectivity needed for a new era of communication for enhanced mobile broadband (eMBB) but also new use cases and applications: massive machine-type communication (mMTC), and ultra-reliable and low latency communication (URLLC). While eMBB focuses mainly on providing a very high peak data rate, better spectral and energy efficiency, improved performance and increased seamless user experience, mMTC aims to support a very large number of connected devices that require relatively a lower bit rate but a better network energy efficiency, since longer battery life is vital for such devices (e.g., actuators, sensors). URLLC focuses on provid-

[1] Although 5G was initially hyped to be a revolutionary technology (refer to e.g., [Deu]) from the 4G technology, there are also some disagreements on this claim (refer to e.g., [Nok]).

ing highly reliable and low-latency communications for mission-critical applications, such as wireless industry automation, tactile internet, medical applications (e.g., telesurgery), augmented reality (AR), virtual reality (VR), smart grid and intelligent transport systems (ITSs). Thus, it is quite probable that 5G will offer a true potential to enable the connection of *everyone* to *everything*, transforming our digital lives and means of communication [Nok].

While the mobile internet traffic is continuously increasing, due to unprecedented penetration of smartphones, tablets, gadgets and machine-type devices, mobile network operators (MNOs) are compelled to increase capital expenditure (CAPEX) and operational expenditure (OPEX) in order to meet the users' requirements [WZHW15]. However, average revenue per unit (ARPU) generated is almost flat or even declining slowly, which has raised severe concerns among the MNOs, amidst the fierce competition environment [Chi13, HNHS19]. As the cost to build, operate and upgrade the radio access network (RAN) is becoming more and more expensive, MNOs must find efficient and economical solutions to enhance QoS and QoE, increase the spectral efficiency (SE), and maintain a healthy profit and sustained growth, while reducing the CAPEX and OPEX. Hence, future radio access technologies (RATs) have to meet these requirements: reduced cost (CAPEX and OPEX), lower energy consumption, higher spectral efficiency, flexibility and scalability for future expandability, easy system update and upgrade, and efficient platform for additional revenue generation services.

In order to meet the aforementioned requirements, C-RAN has been proposed by China Mobile Research Institute [Chi13] as a promising architecture to support use cases and application scenarios envisioned by 5G. Unlike the conventional RAN with standalone base station (BS) that performs complete protocol stack functions, in C-RAN all the baseband functionalities - from the PHY layer to higher layers - are offloaded from the BS and centralized into a common location known as BBU pool, while leaving aside only the radio frequency (RF) functionalities at the RRU. This simplifies the RRUs, as they are small form-factor and low-power devices. On the other hand, the BBU pool is dynamically shared among several RRUs, thus offering better spectral and energy efficiency, multiplexing gains, reduced CAPEX, and easy system operation and maintenance. In addition, C-RAN also offers advanced cooperation and coordinated signal processing capabilities.

Despite huge potentials of C-RAN, the transport link connecting the RRU to BBU, called *fronthaul* in C-RAN, suffers mainly from two strict requirements: huge fronthaul bandwidth and extremely low latency. Huge fronthaul bandwidth arises as the fronthaul transports digitized time-domain in-phase/quadrature-phase (I/Q) samples of each antenna carrier. Thus, the required bandwidth scales with the number of transmitting antenna elements at the RRU and the carrier bandwidth. This is a problem for 5G, because the required fronthaul bandwidth would be prohibitively too high, which is not an economical and viable option for operators to deploy, due to limited availability of fiber optics and expensive fiber costs. The latency constraint arises largely due to strict timing requirement imposed by hybrid automatic repeat request (HARQ). Despite these requirements, the good news is that the stringent bandwidth and low-latency requirements on the fronthaul placed by the CPRI protocol can be relaxed with an appropriate functional split, by moving a few or more signal processing functions to the RRU. However, this not only reduces the fronthaul bandwidth and relaxes latency requirement but also increases the complexity at the RRU. Hence, it is clear that an optimal[2] functional split is needed for future RATs.

The focus of this thesis lies in the fronthaul bandwidth- and latency constraints. Regarding bandwidth-constraint, this thesis studies means to reduce the required fronthaul bandwidth and obtains statistical multiplexing gains, develops an optimization algorithm based on the number of required pilots, and later studies transceiver cost optimization[3]. For latency-constraint, latency at an Ethernet switch is modelled first by numerical simulation and later evaluated by means of analytical solutions.

1.2 Contributions and Thesis Outline

This thesis is organized into seven chapters. It lists also the author's papers in each chapter. However, the paper details are presented in the respective chapters. The remainder of the thesis is structured as follows:

- Chapter 2 explains the basic underlying concepts of cloud radio access network (C-RAN), starting from the distributed radio access network (D-RAN) to C-RAN, building essential groundwork of this thesis. A detailed study on functional split, which is considered as an efficient mean for reducing the fronthaul bandwidth burden, is presented along with each functional split's data rate. Study of functional split is essential, due to the fronthaul limitations imposed mainly by the future RATs. In a packet-switched fronthaul, random packet delays due to queuing at switch can occur. Hence, the basics of a continuous- and a discrete-time queuing theory, which are essential for Chapter 4 and Chapter 5, respectively are presented.

- In Chapter 3, a system model is introduced with massive MIMO-based C-RAN considering Intra-PHY functional split. The notion behind using this split is that the required fronthaul data rate is much relaxed, as it is coupled with the actual user data rate. As the fronthaul data rate is coupled, spatial traffic maps and queuing theory are used to analyze statistical multiplexing gains, and mathematical expressions for these gains are presented. It is shown that assuming a reasonable outage, user-based fronthauling can reduce the required fronthaul bandwidth significantly. This chapter analyzes impacts of traffic density, correlation distance and outage probability, and shows that the relative fronthaul capacity in the fronthaul segments. Furthermore, the impact of the number of pilots on bandwidth-constrained fronthaul is shown. For this, an iterative pilot optimization algorithm is developed, which shows that an additional bandwidth reduction in the fronthaul segments can be achieved, thus providing a larger optimization gain (c.f. Section 3.5.2). At the

[2] Choice of an optimal split is largely dependent on use cases and application scenarios.

[3] Unlike the usual meaning of optimization, which refers to finding parameters that minimize or maximize a given target function under certain constraints, optimization here – without any strict sense – refers to improvement based on selection of rightly chosen parameters with respect to (w.r.t.) reference values.

end of this chapter, cost optimization of the optical transceiver modules is shown to demonstrate how the aforementioned system model with a given split can be used to lower the fronthaul deployment cost, particularly the transceiver cost (c.f. Section 3.6). The publications related to this chapter are:

  - J. K. Chaudhary, J. Bartelt and G. Fettweis, "Statistical multiplexing in fronthaul-constrained massive MIMO,"European Conference on Networks and Communications (EuCNC), Oulu, 2017, pp. 1-6.
  - J. K. Chaudhary, J. Zou and G. Fettweis, "Cost saving analysis in capacity-constrained C-RAN fronthaul,"IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, 2018, pp. 1-7.
  - J. K. Chaudhary, J. Bartelt and G. Fettweis, "Statistical multiplexing and pilot optimization in fronthaul-constrained massive MIMO,"EURASIP Journal of Wireless Communications and Networking, 2018, pp. 1-11.

- Chapter 4 extends the capacity-constraint discussion presented in Chapter 3 to latency-constraint. Latency in the fronthaul network is one of the critical performance metrics, especially for URLLC applications. This chapter models the access link traffic generated by massive MIMO-based RRUs, and maps the arrival process at the switch as Poisson process and the service process as a hyperexponential (HE) distribution, leading to an **M/HE/1** queuing model. As the traffic from several RRUs is aggregated at an Ethernet switch, user traffic is likely to experience some waiting time in the queue at the switch. Towards this end, this chapter first analyzes through simulation the sojourn time, waiting time and queue length distributions, which are later compared with their analytical results for the moment generating function (MGF) for general file size distribution. For analytical results, a tractable, closed-form expressions in terms of MGF for the steady-state queue length queue length, sojourn time and waiting time distributions at the output port of an Ethernet switch in the fronthaul network are derived with the help of Pollaczek–Khinchine formula. Moreover, this chapter presents the impact of file size, arrival rate, switch speed and spectral efficiency on the fronthaul latency, and provides insights for network dimensioning, particularly in terms of packet loss rate (PLR). The packet loss rate (PLR) arises due to inability of the transmitted packets to reach the destination within a certain time. The publications related to this chapter are:

  - J. K. Chaudhary, J. Francis, A. N. Baretto and G. Fettweis, “Latency in the uplink of massive MIMO C-RAN with packetized fronthaul: modeling and analysis”, IEEE Wireless Communications and Networking Conference (WCNC), Marrakech, April, 2019, pp. 1-7.
  - J. K. Chaudhary, J. Francis, A. N. Baretto and G. Fettweis, “Packet loss in latency-constrained Ethernet-based packetized C-RAN fronthaul”, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, September, 2019, pp. 1-6.

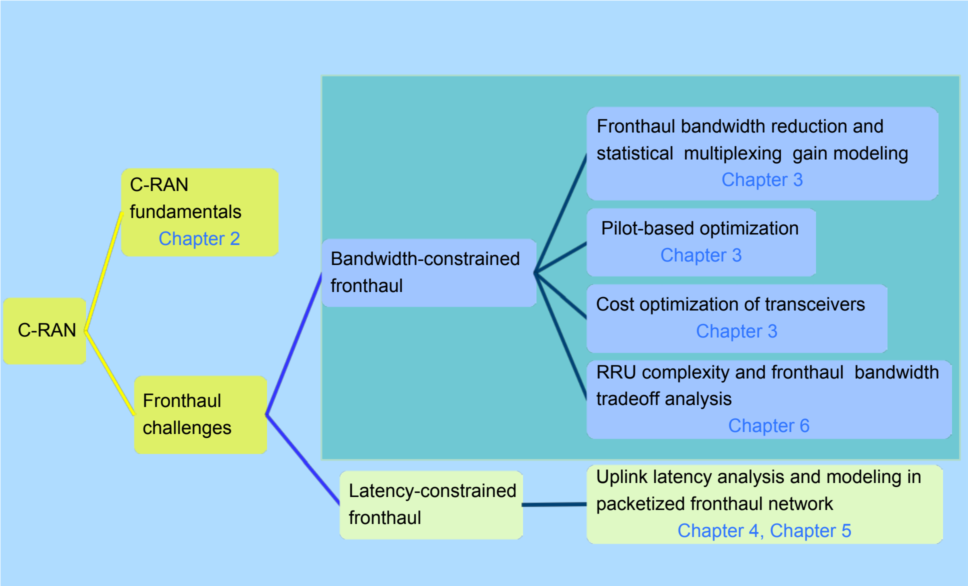

C-RAN
C-RAN fundamentals
Chapter 2
Fronthaul challenges
Bandwidth-constrained fronthaul
Fronthaul bandwidth reduction and statistical multiplexing gain modeling
Chapter 3
Pilot-based optimization
Chapter 3
Cost optimization of transceivers
Chapter 3
RRU complexity and fronthaul bandwidth tradeoff analysis
Chapter 6
Latency-constrained fronthaul
Uplink latency analysis and modeling in packetized fronthaul network
Chapter 4, Chapter 5

Fig. 1.1. Overview of thesis outline highlighting the main research problems and their associated chapters.

- Chapter 5 extends the model in Chapter 4 to a discrete-time queuing in order to account for the slotted nature of transmissions. The main reason for the discrete-time queuing is that the arrival process occurs only at the slot boundaries, and service time is also discrete, meaning that the service time requires just an integer number of the slot duration. This chapter presents a novel queuing model to characterize the distribution of queuing delays at an aggregation gateway in the uplink (UL). This yields tractable, closed-form expressions for the generating functions of steady-state queue length and sojourn time distributions. The analytical results are verified by numerical simulations. The proposed model is then used to study the probability of an outage, which occurs when the sojourn time exceeds the delay budget. It is illustrated that the outage probability decreases as the fronthaul capacity increases. Moreover, it shows that owing to statistical multiplexing, the fronthaul capacity per RRU required to meet a delay constraint significantly decreases when traffic is aggregated from a higher number of RRUs. The publication related to this chapter is:

  – J. Francis, J. K. Chaudhary, A. N. Baretto and G. Fettweis, "Uplink latency in massive MIMO-based C-RAN with Intra-PHY functional split", IEEE Communications Letters (CL), 2020, pp. 1-5.

- Chapter 3 shows that required fronthaul data can be significantly reduced by offloading a part of baseband functionalities to the RRU. However, this increases the complexity at the RRU. Chapter 6 analyzes the RRU complexity and fronthaul bandwidth tradeoff. In addition, functional split with a detailed RF chain is taken into account, together with 5G NR flexible numerology. The publication related to this chapter is:

  – J. K. Chaudhary, A. Kumar, J. Bartelt and G. Fettweis, "C-RAN employing xRAN functional split: complexity analysis for 5G NR remote radio unit,"European Conference on Networks and Communications (EuCNC), Valencia, 2019, pp. 1-6.

- Finally, Chapter 7 concludes the thesis summarizing the core findings of the works presented in this thesis. In addition, it provides possible future directions and some open research problems.

For better illustration, the overall thesis structure is illustrated in Fig. 1.1 correlating the main research problems and their associated chapters.

Chapter 2

Radio Access Network Architecture

As stated in Chapter 1, the traffic demand from the mobile users is surging. Some of the alternatives to cater increasing traffic demand are [WZHW15]: (1) employing advanced transmission techniques such as (massive) MIMO and beamforming; (2) using higher bandwidth channels with millimeter wave (mmWave); (3) exploiting spectrum holes through dynamic spectrum access technologies such as cognitive radio (CR), and (4) deploying a large number of small cells. The first approach has made a significant progress in the recent decades and is approaching a practical limit [WZHW15]. The second approach requires normally line-of-sight (LoS) communications. The third one cannot ensure consistent and reliable services [WZHW15]. The fourth one takes advantage of frequency reuse, which will introduce more interferences. However, the interference can be mitigated by advanced cell coordination and cooperation schemes. Introduction of advanced radio access techniques, particularly massive MIMO, mmWave, carrier aggregation, and deployment of small cells has placed a stringent requirement in the transport network to carry massive amounts of data with a minimum delay from hundreds of thousands of cells [RWN$^+$18]. Hence, the evolution of radio access networks needs to be complemented by the evolution of the transport network, thus demanding the redesign of the future transport technologies. To this direction, C-RAN architecture has been introduced [Chi13]. In C-RAN, the processing resources can be centralized (and even virtualized) at a pool and are shared among many RRUs. In addition, it features real-time cloud computing and power efficient infrastructure. In this chapter, we recap C-RAN architecture, present its potentials and the challenges of C-RAN deployment for 5G. Furthermore, we analyze how C-RAN challenges can be relaxed by means of functional splitting. This pushes the conventional C-RAN approach towards a packetized fronthaul network. In a packetized fronthaul, random packet delays due to queuing at switching/aggregation gateways can occur and it is necessary to characterize distribution of queuing delays. Hence, queuing theory is also presented. This chapter builds essential groundwork for the remaining chapters.

2.1 Cloud Radio Access Network Architectures

2.1.1 Radio Access Network Architecture Overview: From D-RAN to C-RAN

The conventional RAN architecture (also termed as D-RAN) shown in Fig. 2.1 (left) has a standalone BS, which performs all analog, digital and power functions at a dedicated location, and the RF signal generated by BS's RF unit is carried to and from antennas mounted on the rooftop through coaxial cables. As coaxial cables are lossy, signal is degraded before it reaches to antennas. Moreover, in order to accommodate more data traffic, many BSs need to be deployed. Although this increases wireless throughput per unit area, this might cause interference among the BSs, as BSs are closer to each other and they might be reusing the frequency. It is reported in [Chi13] that the majority of power consumption is coming from BSs, but the BS power efficiency is only 50% because inside the BS, only half of the power is used by RAN equipment and the remaining half is used by air conditioners or coolers and other facilitate equipments. Therefore, deployment of more BSs will cause more energy consumption, resulting in higher OPEX and a significant environmental impact. Furthermore, often the average utilization of the BSs is much lower than the peak utilization, which causes waste of the processing resources and power at idle times [Chi13]. Thus, the legacy networks are inefficient in handling spatio-temporal variations, known as *tidal effect*, of the underlying traffic demand [GRI+17]. Moreover, the system flexibility for easy updates or upgrades is very limited.

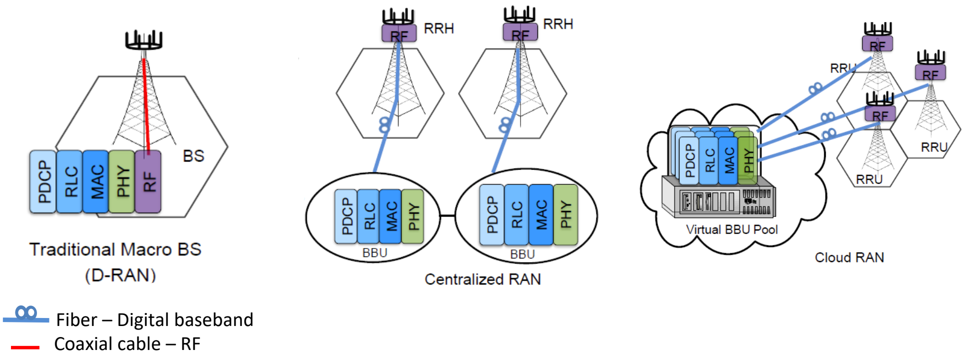


Fig. 2.1. RAN evolution from legacy D-RAN (left) to C-RAN (right) through centralized RAN (center).

D-RAN has very high CAPEX and OPEX, making it not an economical and viable solution for next-generation mobile networks. In order to overcome D-RAN disadvantages, a centralized RAN architecture evolved as shown in Fig. 2.1 (middle), where the RF part is separated from the BS, and moved to a low-cost, small and light-weight form-factor remote radio head (RRH) deployed at the antenna site, and the BS performs only the baseband signal processing functions at a central location, known as BBU. The RRH is connected to a BBU by means of a transport link, known as fronthaul[1], using radio over fiber (RoF) transmission technologies. The RoF can be digital or analog [FSM+15].

[1] In contrast to fronthaul, there is a transport link, known as *backhaul*, which connects the BBU to the core network (CN).

For the transport of fronthaul traffic, the main specifications defined by radio equipment manufactures based on digital radio over fiber (D-RoF) transmission technique are CPRI [CPR15], open base station architecture initiative (OBSAI) [OBS06] and open radio interface (ORI)[2] [ORI15], whereby the radio signal is sampled, quantized and encoded before being transmitted over the fronthaul. CPRI and OBSAI specifications differ[3] in the way how information is transmitted [dHLA16, SS14a]. The most widely adopted protocol by the vendors is the CPRI. Another D-RoF solution is an ongoing work in IEEE 1914.3 [NGFa] to define a radio over Ethernet (RoE) solution. The BBU transports or receives usually the digitized baseband samples, preferably[4] by means of dedicated optical fiber links. Generally, BBUs can transport or receive also the analog radio signal by means of analog radio over fiber (A-RoF) techniques. However, A-RoF techniques [HG14] are not as often deployed as their digital counterparts for the following reasons: firstly, A-RoF is not yet standardized and secondly, D-RoF is mostly preferred due to inherent advantages of digital solution, such as its immunity to noise and hardware impairments, and flexibility in the transport deployment. In the centralized RAN, each BBU connects only one RRH through a dedicated fiber. Hence, it lacks coordination among the BBUs and normally has no or only limited resource sharing. Inheriting the benefits of cloud computing, the centralized RAN architecture has evolved to a cloud-based RAN architecture, known as C-RAN, shown in Fig. 2.1 (right). In C-RAN, one or several of RRUs are connected to a pool of BBUs, often referred to as BBU pool, thus offering efficient utilization of BBU resources. Some tutorials and overview papers on C-RAN can be found in [CCY$^+$15, WZHW15, ATM18]. C-RAN offers several advantages, which are listed below:

C-RAN Advantages

- Reduction of total cost of ownership (TCO), and lower energy consumption owing to centralization of RAN functionalities at the BBU pool;
- Simpler and cheaper operations and maintenance, and easy centralized system updates and upgrades;
- Easier implementation of advance coordinated and cooperative signal processing functions, such as coordinated multi-point (CoMP), enhanced intercell interference coordination (eICIC), joint transmission (JT) and joint reception (JR), which are essential to improve the spectrum efficiency, link reliability, and the communication quality, particularly of the cell-edge users;

[2] ORI was introduced by the European Telecommunications Standards Institute (ETSI) to address CPRI comparability issue and to provide better compatibility among vendors.

[3] CPRI is a serial constant bitrate (CBR) interface, whereas OBSAI is a packet-based interface. OBSAI was established in 2002, before CPRI. The first version of CPRI specification was released at the end of 2003. It is worth noting that CPRI has less overhead compared to that in OBSAI, which makes it more advantageous to implement. Another significant advantage of CPRI is that the BER requirement in CPRI is 10^{-12}, which is less strict than the OBSAI BER requirement of 10^{-15}.

[4] In addition to optical fiber, mmWave, microwave or any other access link media can also be used. However, unlike optical fiber, mmWave and microwave are used for shorter distances and preferably in LoS communications.

- Efficient hardware utilization through resource sharing and network function virtualization (NFV) offering statistical multiplexing gains;
- Scalability to add/remove/upgrade services as required.

Challenges in Fronthaul

Implementing C-RAN architecture in the existing 4G mobile networks has several advantages, which are listed above, but implementing C-RAN in a 5G network is quite demanding. Particularly, deploying a reliable fronthaul network for future RATs in a cost- and energy-efficent way, while still satisfying the stringent fronthaul bandwidth and latency requirements is enormously challenging [RWN+18]. In order to enable efficient centralized and cooperative processing, the fronthaul links must offer huge bandwidth, very low-latency and jitter, and very tight synchronization. Unfortunately, the practical fronthaul is often capacity- or delay-constrained [PWLP15]. These two contraints – excessive fronthaul bandwidth and extremely low-latency – on the fronthaul are the major obstacles in the deployment of C-RAN architecture. The core of the thesis are these constraints, which are explained below.

- **Challenge 1: Fronthaul has a Capacity Bottleneck**

  The current C-RAN architecture is designed for 4G mobile networks and the fronthaul transports digitized I/Q- (in-phase/quadrature-phase) samples using the CPRI protocol. The required fronthaul data rate per sector (in bits/s), in general, is given by [DDM+13]

$$D_{\mathrm{CPRI}} = N_{\mathrm{Ant}} \cdot f_{\mathrm{s}} \cdot N_{\mathrm{Q,opt8}} \cdot 2 \cdot \zeta_{\mathrm{opt8}}, \tag{2.1}$$

  where N_{Ant} is the number of antennas at the RRU, f_{s} the sampling rate (in samples/s/carrier), $N_{\mathrm{Q,opt8}}$ the resolution of time-domain quantizer (in bits/sample), 2 is a multiplication factor to account for I/Q samples, and $\zeta_{\mathrm{opt8}} = \gamma_{\mathrm{CW}} \cdot \gamma_{\mathrm{LC}}$ is a CPRI specific overhead factor, where γ_{CW} represents the overhead introduced by CPRI control words[5] (typically one control word for 15 words of payload data), and γ_{LW} represents the line-coding overhead (e.g., 10/8 for 8B/10B coding or 66/64 for 64B/66B coding).

  As seen from (2.1), the required CPRI bandwidth scales linearly with the number of antennas and sampling frequency (and thus with the applied transmission bandwidth on the access link). The next-generation mobile systems are envisaged not only to support carrier aggregation and multi-band, but also to integrate new radio access techniques, such as massive MIMO and mmWave communications [ATM18, R+13]. Therefore, the next-generation mobile systems will induce huge fronthaul bandwidth demands, which makes fronthaul networks deployment non-affordable. As illustrated in Fig. 2.2, a 100 MHz 5G sub-6 GHz massive MIMO system employing 96 antennas

[5] A CPRI basic frame consists of 16 words, where the first word is used for a control word and the remaining 15 words are used for payload data.

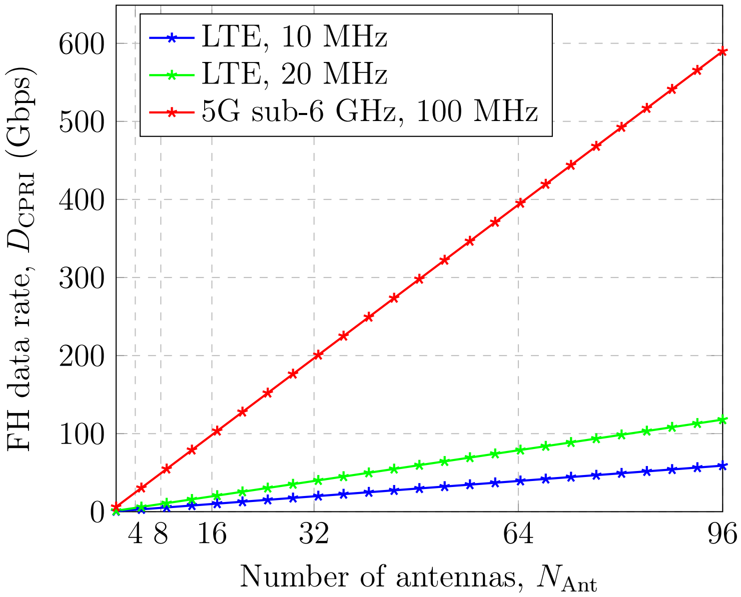


Fig. 2.2. Fronthaul data rate per sector in a 10 MHz, 20 MHz LTE and 100 MHz 5G sub-6 GHz with respective sampling frequency $f_{\mathrm{s}} = \{15.36, 30.72, 153.6\}$ MHz. ($N_{\mathrm{Q,T}} = 15$, $\gamma_{\mathrm{CW}} = 16/15$ and $\gamma_{\mathrm{LC}} = 10/8$)

requires roughly a 60 times larger bandwidth ($D_{\mathrm{CPRI}} \approx 590$ Gbps), when compared with a 20 MHz 8 × 8 LTE ($D_{\mathrm{CPRI}} \approx 9.8$ Gbps). The required fronthaul bandwidth is extremely high and possibly too expensive for practical fiber deployment. Hence, the fronthaul could become a bottleneck for the performance of the future mobile networks, if it is not dimensioned correctly [HR16, Chapter 4]. Furthermore, the bandwidth requirement of CPRI-based fronthaul is fixed for a given cell configuration[6] and does not depend on the amount of real traffic associated with the cell. As a result, for example, to support a peak UL user data rate of 150 Mbps in a 20 MHz single-sectored 8 × 8 LTE, UE Cat 8 [3GP19, Table 4.1-2], roughly 50 Gbps constant fronthaul bandwidth is required for CPRI-based fronthaul in a 5G network, irrespective of RRU's real traffic load.

Methods to Mitigate Fronthaul Capacity Bottleneck

In order to ease the challenging fronthaul data requirements, various solutions have been proposed, such as (i) decreasing the fronthaul required data rate [Chi13] e.g., by reducing the signal sampling rate, applying non-linear quantization, or using compression techniques (I/Q data compression, subcarrier compression) (ii) increasing the fronthaul capacity using single fiber bidirection (SFBD), wavelength division multiplexing passive optical network (WDM-PON) [CLHD+14, 5G-a], optical transport network (OTN), time shared optical networks (TSON) [ZTA+11, YQZ+12] or mmWave communication [PM17, PM17]. Vendors have shown that the fiber consumption in LTE deployments can be saved by half using CPRI compression techniques, such as 2:1 compression with lossless performance [LC13]. In addition, using a SFBD technology, which allows simultaneous UL and DL transmission on a single

[6] Even the most recent CPRI specification [CPR15] has a maximum supported data rate of about 24 Gbps per cell.

fiber, the fiber consumption can be further halved. Thus, combining CPRI compression and SFBD, fiber consumption can be saved threefold [Chi13]. However, fiber consumption reduction using these techniques are not enough for 5G fronthaul, mainly because the 5G fronthaul requires an enormous bandwidth and there is a growing complexity associated with the compression techniques. Alternatively, redefining the current functional split architecture between the BBU and RRU by splitting the signal processing functions in different ways has been considered as a promising architecture by several standardization bodies and forums, such as 3GPP [3GP17a], CPRI consortium [eCP19], next generation fronthaul interface (NGFI) [NGFb], NGMN Alliance [NGM15b] and Small Cell Forum [Sma15]. This approach moves the current CPRI architecture towards a packet-based network, such as Ethernet with new functional splits between BBU and RRU [DDM+13, WRB+14], as it will be explained in Section 3.1. Data needs to be encapsulated in the form of packets rather than a constant stream in CPRI. Hence, a packet switching protocol such as Ethernet can be used, which allows us to enjoy inherent benefits of Ethernet, such as its cost effectiveness, ubiquity and flexibility.

- **Challenge 2: Fronthaul is Latency-constrained**

The latency constraint in the fronthaul originates either from the timing requirement of the hybrid automatic repeat request (HARQ) or from use cases, such as Tactile Internet, autonomous driving or augmented and/or virtual reality. In the LTE MAC, the HARQ process is co-located[7] with a scheduler and it requires the acknowledgement signal to be sent within a pre-defined round-trip time (RTT) denoted as $T_{\text{max, delay}}$. Most of the RTT $T_{\text{max, delay}}$ is spent at the BBU and RRU for baseband signal and RF processing, respectively, and the remaining time T_{FH} is left for the fronthaul transport. In general, the latency budget left for the fronthaul with the HARQ process located at the BBU is a few hundreds of microseconds, typically $T_{\text{FH}} \leq 250$ μs [3GP17a, SS14b]. The main latency components in the fronthaul are analyzed here. The RTT delay between the BBU and RRU is considered, which consists of mainly three parts: delays in the access link, delays in the fronthaul network and delays in the RRU and baseband processing. These latency components can be broken down, for simplicity, into transmission delay, propagation delay, processing delay, packetization delay, fabric delay and queuing delay. Therefore, the round-trip fronthaul latency can be obtained as [3GP17a, SS14b]

$$T_{\text{FH}} = T_{\text{max, delay}} - 2(T_{\text{trans}} + T_{\text{Proc}}^{\text{RRU}} + T_{\text{Proc}}^{\text{BBU}} + T_{\text{prop}}^{\text{RAN}}), \tag{2.2}$$

where, $T_{\text{max, delay}}$ is 8 ms HARQ RTT in FDD LTE, T_{trans} = packet size/FH bitrate[8] the transmission delay, $T_{\text{Proc}}^{\text{RRU}}$ the processing delay at the RRU, $T_{\text{Proc}}^{\text{BBU}}$ the processing delay at the BBU and $T_{\text{prop}}^{\text{RAN}}$ the propagation delay in the RAN.

[7] The HARQ timing requirement is very critical if HARQ is located at the BBU, however, the timing requirement is much relaxed if the process is located at the RRU [3GP17a].

[8] The fronthaul bitrate value is different for each split and is explained in Section 2.1.2.

The signal is packetized in the fronthaul and it also requires certain time to propagate in the fronthaul. Hence, the round-trip fronthaul latency can be given by

$$T_{\mathrm{FH}} = 2\left(\sum_{i=1}^{N} T_{q,i} + T_{\mathrm{prop}}^{\mathrm{FH}} + N(T_{\mathrm{f}} + T_{\mathrm{pkt}})\right), \qquad (2.3)$$

where, T_{pkt} = packet size/switch speed the packetization delay, $T_{\mathrm{prop}}^{\mathrm{FH}}$ is the propagation delay in the fiber, T_{f} the fabric delay, $T_{q,i}$ the queuing delay for the i^{th} switch and N the number of switches. Packetization delay is the time required to packetize data samples into a packet. Fabric delay is the delay in the hardware of a switch and depends on the quality of the switch. Fabric delay value is very low, typically in the order of nanoseconds or a few microseconds. Note that fronthaul delay contains deterministic terms for a given scenario but the queuing delay will be variable. Focus of the thesis is to characterize the queuing delay distributions at the switch, which is presented in Chapter 4 and Chapter 5.

The low-latency requirements place a limit on the separation between the BBU and RRU. Thus, knowing the allowable fronthaul latency, the maximum (one way) distance between the BBU and RRU can be computed using $d_{\mathrm{FH,\,max}} = T_{\mathrm{FH}}/\Delta T_{\mathrm{P}}$, where, ΔT_{P} is the round-trip propagation delay per km, which is 10 μs/km for fiber-based fronthaul. Typically, the maximum separation between the BBU and RRU is 25 km.

2.1.2 Flexible RAN Architecture: Birth of Functional Split

Researchers must find new possibilities for lowering the stringent fronthaul requirements stated above, while still benefiting from the acclaimed C-RAN benefits stated in 2.1.1. One possibility is to revisit the traditional concept of C-RAN, by allocating more functions locally at the cell site, and, thus, processing the signal more before being transported to the BBU. However, the important question is how many functions should be kept locally at the cell cites and how many should be left for central processing? Well, to this end, several functional subdivisions, also called functional splits, are under consideration by 3GPP [3GP17a], CPRI consortium [eCP19], NGFI [NGFa], NGMN Alliance [NGM15b] and Small Cell Forum [Sma15]. Functional splitting refers to the division of signal processing functions between the BBU and the RRU, and each functional split corresponds to a split point (split option). 3GPP [3GP17a] has defined eight functional splits with sub-options for some of the splits. It is to be noted that although a higher-layer split (HLS) (Option 2) has already been agreed upon, still no consensus has been reached yet (at the time of thesis writing) for a lower-layer split (LLS) [AZH+18]. The functional split naming is varying among the organization bodies and forums, but we restrict ourselves to the 3GPP numerology. For the mapping of the numerologies, 3GPP and eCRPI are compared in Fig. 2.3, as these two have been widely presented. For numerologies mapping among all leading organization, refer to e.g., [Cha, Figure 2.2] or [LCC19, Figures 7 and 8]. LTE protocol stack is considered in this work and an overview of 3GPP splits with their required data-rate and latency requirements is presented. Note that in 5G RAN, a BBU

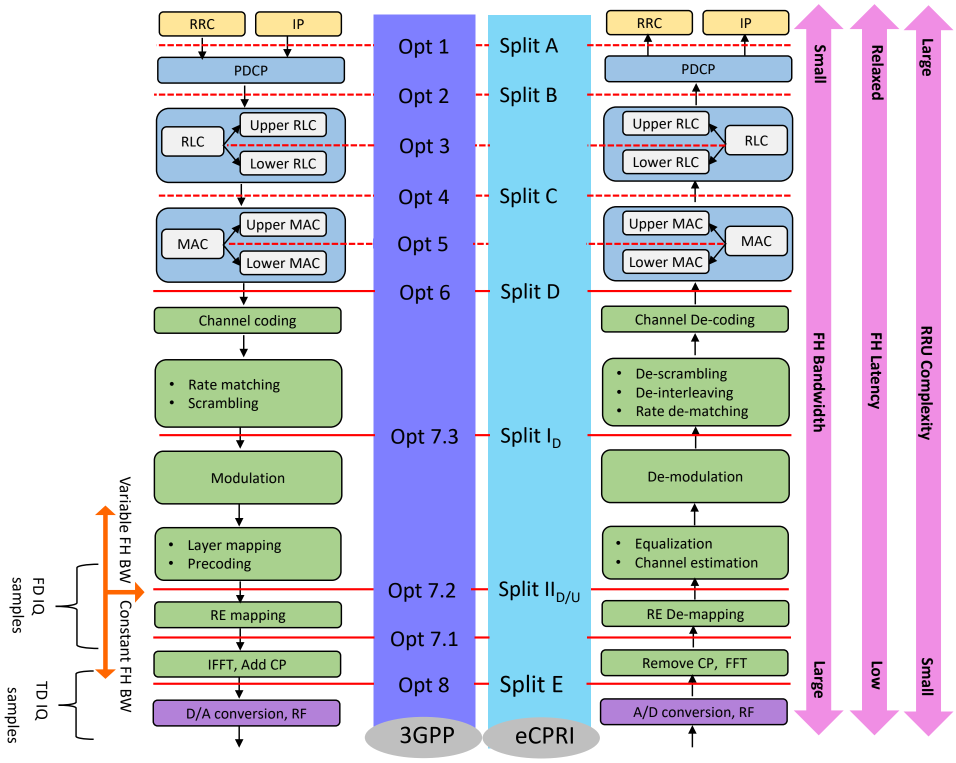


Fig. 2.3. LTE protocol stack with functional splits proposed by 3GPP[3GP17a], CPRI consortium [eCP19]. The RF, PHY layer (Layer 1), data link layer (Layer 2) and network layer (Layer 3) are represented by light purple, light green, light blue and light gold colored functional blocks, respectively.

is further divided into two segments: a distribution unit (DU) and a control unit (CU). The DU is located close to the user, and the CU is located in a datacenter and virtualized. The transport link connecting DU to CU is known as *midhaul*. Throughout this thesis, we interchangeably use BBU/CU for the baseband unit and RRH[9]/RRU/RU for distributed radio elements. Moreover, capacity-constrained fronthaul is loosely used in place of bandwidth-constrained fronthaul.

Fig. 2.3 illustrates the eight functional splits with LTE protocol stack as a reference. This figure shows how the received signal by the antennas are transmitted via the antennas ports for processing into the PHY layer, then to the network layer through the data link layer. Reverse operations occur in the downlink (DL). The red lines in the figure represent different functional splits names. They mean that the functions below the split line are executed locally at the RRU, while those above the split line are performed centrally at the BBU. Furthermore, as we move towards higher-layer splits (with smaller split numbering in 3GPP naming), a signal is processed more locally, before being transported to the BBU. The more the signal is processed locally at the RRU before the data is transported to or from the BBU, the lower the bitrate on the fronthaul.

In the UL, the radio signal received by the antenna is first filtered, amplified and then down-converted to a baseband signal, which is then sampled, quantized and encoded to get the time-domain signal. The digitized time-domain I/Q samples are sent for processing

[9] Strictly speaking, the RRH and BBU are terminologies used in LTE.

into the PHY layer. Next, the cyclic prefix (CP) is removed and the fast Fourier transform (FFT) operation is applied, resulting into the frequency-domain I/Q samples, i.e., subcarriers in frequency-domain. After the FFT, the guard subcarriers are removed. For example, in a 20 MHz LTE, where we have a total of 2048 subcarriers corresponding to 100 physical resource blocks (PRBs), 848 subcarriers are guard band subcarriers, which are removed, leaving only 1200 usable subcarriers. Hence, this brings a drop of approximately 40% fronthaul bitrate. Next, the subcarriers are demapped by the resource element (RE) demapper and only the allocated PRBs are transported in the fronthaul. This makes the fronthaul data rate vary with the actual user load. Thus, the RRU with included RE demapper/mapper makes the fronthaul data rate varying, which is illustrated by an arrow on the left hand side of Fig. 2.3. Note that the RRUs without demapper/mapper will have CBR traffic.

In the PHY layer, de-modulation, rate de-matching and de-scrambling are carried out, which further reduces the rate depending on the order of modulation. Next, we describe the individual fronthaul data rate for selected splits. The design of the fronthaul transport network is affected by the choice of a functional split, as each split has its own advantages and disadvantages, e.g., in terms of required bitrate and latency. In this work, only the default CPRI split (Option 8), PHY layer splits (Options 7.1,7.2 and 7.3) and MAC-PHY split (Option 6) are considered and their bitrate are calculated. For detailed fronthaul bitrate calculations, please refer to, e.g., [DDM+13], [Sma15, Appendix C], [R3-16b] and [R3-16a, Table 1]. Note that the peak fronthaul data rates are calculated for each sector. Hence, the net peak fronthaul data rate can be obtained by multiplying the per sector peak fronthaul data rate with the number of sectors. The description of the symbols in the rates calculation and their numerical values are listed in Table 2.1.

- **Option 8**:

  This split is similar to the classical CPRI split, where all the RF processing such as amplification, filtering, A/D or D/A conversion is performed at the RRU, while leaving aside all the baseband signal processing functions at the BBU. Time-domain I/Q samples are transported using a fronthaul interface. Being a conventional defacto split in C-RAN, this inherits all the benefits listed in Section 2.1.1. The major disadvantage of this split is that it requires a continuous bitrate transport, irrespective of whether user traffic is present or not. Thus, Option 8 seems impractical for 5G RATs, as it requires a prohibitively high bit rates and a very low-latency (as explained in Section 2.1.1). The fronthaul data rate for Option 8 can be calculated as

  $$D_{\text{opt8}} = N_{\text{Ant}} \cdot f_{\text{s}} \cdot N_{\text{Q,opt8}} \cdot 2 \cdot \zeta_{\text{opt8}}. \tag{2.4}$$

- **Option 7.1**:

  At this split, FFT/inverse fast Fourier transform (IFFT) and CP removal/addition are done at the RRU. Thus, frequency-domain I/Q samples of all PRBs are forwarded. The fronthaul data rate for Option 7.1 can be calculated as

Tab. 2.1. Parameters for fronthaul data rate calculation.

Parameters	**Symbol**	**Unit**	**LTE**	**sub-6 GHz**
Carrier frequency	f_{C}	GHz	2	2
Channel bandwidth	B	MHz	20	100
Sampling frequency	f_{C}	MHz	30.72	153.6
Number of resource blocks	N_{RB}	-	100	500
Number of active subcarriers per RB	$N_{\mathrm{SC}}^{\mathrm{RB}}$	-	12	12
Number of symbols per subframe	$N_{\mathrm{sym}}^{\mathrm{SF}}$	-	14	14
Subframe duration	T_{SF}^{-1}	s	1	1
Number of antennas	N_{Ant}	-	8	96
Number of antenna ports (ADC/DAC chains)	N_{Port}	-	4	8
Number of spatial layers	N_{Layer}	-	4	8
Quantizer bit resolution per I/Q dimension	$\{N_{\mathrm{Q,opt8}},\ N_{\mathrm{Q,opt7.1}},\ N_{\mathrm{Q,opt7.2}},\ N_{\mathrm{Q,opt7.3}},\ N_{\mathrm{Q,opt6}}\ \}$	bits	$\{15, 9, 9, 3, 1\}$	$\{15, 12, 12, 3, 1\}$
Maximum utilization	μ	-	1	1
Resource overhead	η	-	0.1	0.1
Fronthaul overhead	$\{\zeta_{\mathrm{opt8}},\ \zeta_{\mathrm{opt7.1}},\ \zeta_{\mathrm{opt7.2}},\ \zeta_{\mathrm{opt7.3}},\ \zeta_{\mathrm{opt6}}\}$	-	4/3	4/3
Modulation order	M_{mod}	-	64	256
Code rate	R_{c}	-	5/6	5/6

$$D_{\mathrm{opt7.1}} = N_{\mathrm{Ant}} \cdot N_{\mathrm{RB}} \cdot N_{\mathrm{SC}}^{\mathrm{RB}} \cdot N_{\mathrm{sym}}^{\mathrm{SF}} \cdot T_{\mathrm{SF}}^{-1} \cdot N_{\mathrm{Q,opt7.1}} \cdot 2 \cdot \zeta_{\mathrm{opt7.1}}. \tag{2.5}$$

Note that sampling frequency is now replaced by the product of the number of resource blocks N_{RB}, number of subcarriers per subframe $N_{\mathrm{SC}}^{\mathrm{RB}}$, number of symbols per subframe $N_{\mathrm{sym}}^{\mathrm{SF}}$ and the subframe duration T_{SF}^{-1}. In addition, it includes frequency-domain quantizer resolution $N_{\mathrm{Q,opt7.1}}$ instead of the time-domain quantizer resolution $N_{\mathrm{Q,opt8}}$. Note that different quantization resolution bits are used in different splits [DDM+13] depending on the dynamics of the signal. In Option 8, as signal signal has a higher dynamic range, a higher number of quantization bits (e.g., 15 bits per I/Q dimension) is used. After FFT and resource demapping, a fewer number of quantization bits are used.

- **Option 7.2**:

  The fronthaul data rate for Option 7.2 can be calculated as

$$D_{\mathrm{opt7.2}} = N_{\mathrm{Port}} \cdot N_{\mathrm{RB}} \cdot N_{\mathrm{SC}}^{\mathrm{RB}} \cdot N_{\mathrm{sym}}^{\mathrm{SF}} \cdot T_{\mathrm{SF}}^{-1} \cdot \mu \cdot N_{\mathrm{Q,opt7.2}} \cdot 2 \cdot \zeta_{\mathrm{opt7.2}}. \tag{2.6}$$

Resource element mapping/demapping is performed at the RRU. Only the allocated subcarriers are transported in the fronthaul. The fronthaul data rate is coupled with the actual cell load because only the resources occupied by the user data transmission need to be forwarded. Hence, this makes the resultant fronthaul data rate variable and this allows to obtain statistical multiplexing gains.

- **Option 7.3**:

  The fronthaul data rate for Option 7.3 can be calculated as

$$D_{\text{opt7.3}} = N_{\text{Layer}} \cdot N_{\text{RB}} \cdot N_{\text{SC}}^{\text{RB}} \cdot N_{\text{sym}}^{\text{SF}} \cdot T_{\text{SF}}^{-1} \cdot \mu \cdot (1-\eta) \cdot N_{\text{Q,opt7.3}} \cdot \log_2(M_{\text{mod}}) \cdot \zeta_{\text{opt7.3}}. \quad (2.7)$$

  This split resembles Option 7.3 in 3GPP and is defined only for the DL. This split exhibits asymmetry in the UL and DL because it differs how the information bits are transported in the UL and DL. In the DL, encoded user bits are forwarded, whereas in the UL, one log-liklihood ratio (LLR) value per information bit is forwarded. Each LLR is typically represented by 3 bits. Thus, depending on the used modulation scheme, it requires $N_{\text{Q,opt7.3}} \cdot \log_2(M_{\text{mod}})$ bits. Furthermore, the data rate depends now on the number of spatial streams (spatial layers) N_{L}, instead of on the number of antenna ports (number of ADC/DAC chains). The number of spatial streams and the modulation scheme depend on the user's channel quality. Suppose that the channel quality of a certain user is not good enough to have a permissible spatial separation between the independent streams, then only the supported streams need to be forwarded, instead of one stream for each antenna [HR16, Chapter 4]. Note that this dependency on the spatial layers, particularly for massive MIMO employing hundreds of antennas, is really significant, because the required fronthaul bitrate is reduced a lot. Otherwise the fronthaul would require a tremendously high data rate. Moreover, reference symbols no longer need to be forwarded, since the channel estimation and equalization are done at the RRU. Hence, the fronthaul bitrate is further reduced by a factor $1-\eta$, where η is the resource overhead. Thus, these dependencies make fronthaul traffic more coupled to the actual user traffic.

- **Option 6**:

  This split marks a separation point between the PHY and MAC layers. Thus, the RRU executes all the PHY as well as RF layer functions, whereas the BBU performs the data and the network layer functions. The data rate in this split is determined by the transport block sizes (TBSs) and bit-level user data is transported. Decoding/coding removes/add extra redundant bits from/to the actual information bits. No redundant bits are forwarded. Hence, the fronthaul bitrate is further reduced according to the code rate R_{c}. Decoder output are information bits. Hence, the quantizer resolution is one bit, i.e., $N_{\text{Q,opt6}} = 1$. The fronthaul data rate for Option 6 can be calculated as

$$D_{\text{opt6}} = N_{\text{Layer}} \cdot N_{\text{RB}} \cdot N_{\text{SC}}^{\text{RB}} \cdot N_{\text{sym}}^{\text{SF}} \cdot T_{\text{SF}}^{-1} \cdot \mu \cdot (1-\eta) \cdot N_{\text{Q,opt6}} \cdot \log_2(M_{\text{mod}}) \cdot R_{\text{c}} \cdot \zeta_{\text{opt6}}. \quad (2.8)$$

Fig. 2.4 shows illustrates fronthaul data rates for five splits (Options 8, 7.1,7.2, 7.3 and 6) and two RATs (LTE and 5G sub-6 GHz). The parameters are justifiably chosen to calculate the bitrates and are listed in Table 2.1. 5G sub-6 will have higher bandwidth with

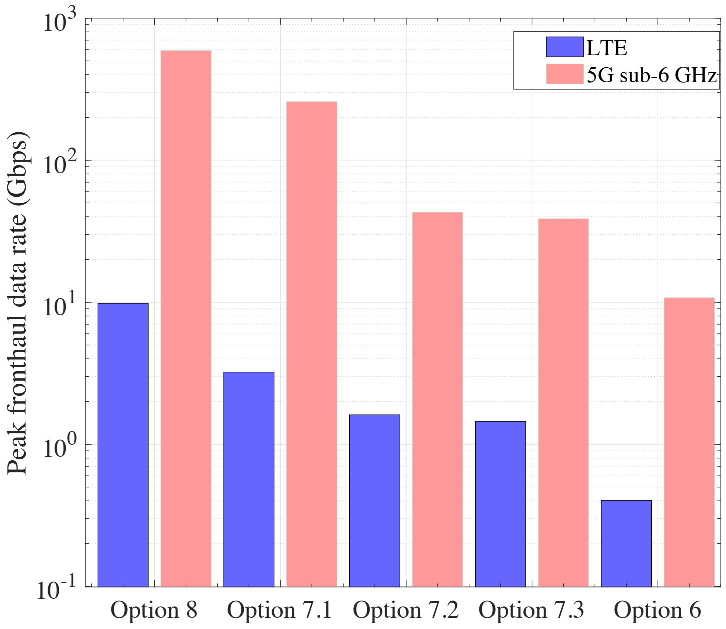


Fig. 2.4. Peak fronthaul data rates for Options 8, 7.1,7.2, 7.3 and 6 with SoTA LTE and 5G sub-6 GHz.

the aggregation of five carriers and higher-order modulation scheme. In [DDM^{+}13], 7-9 bits are used for frequency-domain bit resolution and 64 QAM is considered. Sub-6 will also have a 256 QAM or even a higher-order modulation. Hence, for sub-6 GHz, a higher number of quantization bits are used. Furthermore, the number of antennas, antenna ports and spatial layers are considered from AIRRAYS's massive MIMO RRU [5G-b], which have been tested with a field measurement in the 5G-XHaul project, where AIRRAYS GmbH (Xilinx Dresden GmbH) was an small and medium-sized enterprises (SME) partner. Note that the fronthaul overhead factor is different for different splits and depends e.g., on the transport medium and UL or DL direction. For example, if Ethernet is used as a fronthaul network, overhead is approximately 8% for Options 7.1 and 7.2, and 10.7% for Option 7.3, according to [NGM15a]. This overhead arises, e.g., due to synchronization, Ethernet frame and scheduling control. However, for the shake of comparison, the same overhead value is used for different splits, assuming fiber-based deployment. Moreover, maximum load utilization is considered, which yields peak fronthaul data rates.

Variable Fronthaul Data Rate: Enabler for Statistical Multiplexing Gain

Many benefits of Ethernet can be obtained if the fronthaul data rate is variable. So, the interesting questions are: when does the fronthaul data rate become variable and how does the fronthaul data rate depend on the user? Which split generates a variable fronthaul split? This chapter investigates that and considers a suitable split for statistical multiplexing gain.

Considering the UL, time-domain samples are converted to frequency-domain samples after FFT operation. This reduces fronthaul data rate. However the reduction is only due to discarding a fixed set of guard carriers and removal of cyclic prefixes. This means, fronthaul data rate is still static in Option 7.1. Next, if the resource mapping is also done at the RRU (refer to Option 7.2), only the used subcarriers from the user are forwarded. As the number of allocated subcarriers vary among the users, the required fronthaul data rate is now coupled with the actual user data rate. Thus, statistical multplexing gains can be obtained as the fronthaul bit rate is varying. This is explained in Section 3.3.3. Hence, the lowest functional split for a variable fronthaul data rate is Option 7.2. Furthermore, more RRUs can be accommodated for a given fronthaul bandwidth if the fronthaul data rate is reduced.

Functional split latency requirements

In Option 5, HARQ and other timing critical functions are located at lower MAC. Hence, Options 1 to 5 will have relaxed latency requirements on the fronthaul link, whereas as Options 6 to 8 will have strict fronthaul latency requirements [LCC19]. In [Sma15, Appendix B], one-way latency requirements on the fronthaul link are categorized in terms of ideal (0.25 ms), near-ideal (2 ms), sub-ideal (6 ms) and non-ideal (30 ms). In addition, 3GPP [3GP17a, Table A-1] has also proposed one-way fronthaul latency, which is maximum 0.25 ms for Options 6 to 8, hundreds of microseconds for Option 5, approximate 0.1 ms for Option 4, max 1.5 to 10 ms for Options 2 and 3, and max 10 ms for Option 1. From both 3GPP and SCF, it is clear that the most critical latency on the fronthaul is the latency requirement of 250 μs.

2.2 Packet-based Fronthaul Networks

2.2.1 Standardisation Activities

Research on C-RAN has been conducted by many research organization as well as standardization bodies. A few of them are recapped here.

The current fronthaul is based on the semi-proprietary CPRI protocol, which is by far the most adopted and the most popular RRU-BBU interface so far. However, CPRI is a serial CBR interface, and requires very high bandwidth and low-latency fronthaul links, which are foreseen to be challenging for 5G. To overcome these challenges, an evolved fronthaul is required to support 5G. In addition, as the demand of packet-switched technology is growing, recently a new industry standard enhanced CPRI, referred to as eCPRI, [eCP17, eCP19] has been released, which is designed for packet-based transport networks and supports real-time traffic. Moreover, a recent initiative IEEE 1914 working group for next generation fronthaul interface (NGFI) [NGFa] has shown that the stringent fronthaul requirements can be relaxed by using a packetized transport solution, such as Ethernet. IEEE 1914 work group has two active projects: (1) 1914.1 [IEEa], which is a standard packet-based fronthaul transport network, focusing on the architecture for the transport of mobile fronthaul traffic, and on defining the requirements (data rates, timing and synchronization, and QoS) for the fronthaul networks; and (2) 1914.3 [IEEb], which is a standard for RoE encapsulations and mappings, enabling encapsulation of digitized radio I/Q payload data as well as supporting the header format for both structure-aware and structure-agnostic encapsulation of existing digitized radio transport formats.

Ethernet is considered as a prime candidate technology for the evolved fronthaul due to its cost effectiveness through economies-of-scale, flexibility, and ubiquity. Furthermore, it supports NFV and software-defined networking (SDN) [SXT$^+$19], and takes advantage of statistical multiplexing gains. Thus, it is clear that future RATs require the redesign of the fronthaul, leading to an *evolved fronthaul*. Thus, the consensus is towards a packetized fronthaul with a variable bitrate (VBR) functional split.

2.2.2 Challenges in Ethernet Fronthaul Network

Although Ethernet is considered as a transport medium for mobile fronthaul for RoE, due to its inherent advantages stated earlier, Ethernet does suffer, especially from low delay, delay variation, packet loss and tight synchronization requirements [BCV18]. Ethernet systems were not originally designed for delay sensitive networks. Hence, fulfilling the delay requirements by packet-switched Ethernet systems is identified as one of the main challenging requirements [BCV18]. Nevertheless, Ethernet can be used for the functional splits where these requirements are relaxed [MMM19]. The recently created Time-Sensitive Networking (TSN) Task Group [IEE18], which is a part of IEEE 802.1 Working Group, is working to develop new extensions to support Ethernet traffic forwarding, while providing guaranteed packet transport with bounded low latency, low packet delay variation, and low packet loss. Ethernet is natively asynchronous, which makes it not suitable for the transport of CPRI traffic as such. Thus, additional mechanisms need to be applied to Ethernet to make it suitable for the mobile fronthaul. Furthermore, for multi-hop connection, it is mandatory to have synchronous Ethernet [dHLA16]. To this end, high Precision Time Protocol (PTP) over Ethernet, such as IEEE 1588 [IEE08] and Synchronous Ethernet (SyncE) [G.818] exist. PTP provides time synchronization, whereas SyncE provides frequency synchronization in packet-based Ethernet networks.

2.2.3 Ethernet Switch Structure

Fig. 2.5 shows a schematic illustration of an Ethernet switch and its output port structure. The switch consists of several input and output ports, a packet processor and buffer elements. The packets are received at the input ports of the switch and are later forwarded to the packet processor. The switch reads and processes the source and destination medium access control (MAC) addresses of the Ethernet frame. The packet processor looks at the destination address of the packets and routes them to the appropriate output ports. Each output port has a buffer element and the packets are queued at the buffer at the switch output port. Each port has several queues, where packets can be sorted according to the priority of the queues and/or their traffic classes. Traffic can consist of a single homogeneous traffic source or heterogeneous traffic sources with different QoS requirements, and different scheduling policies can be applied by the scheduler. The switch operation is controlled by the switch policy, which consists of a buffer management policy and a scheduling

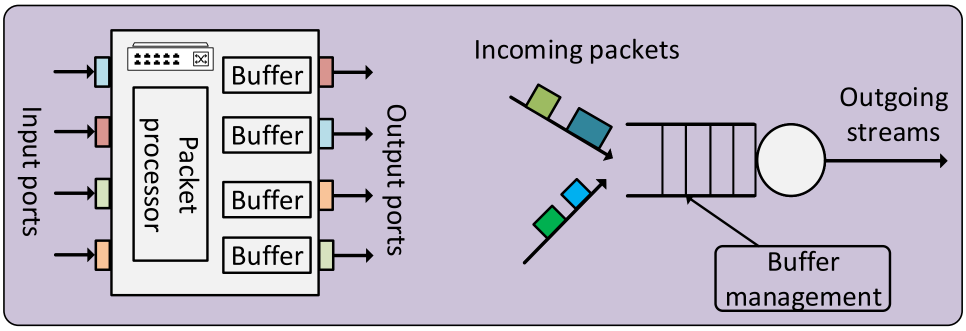


Fig. 2.5. Schematic of representation of an Ethernet switch structure: (left) Simplified structure of an Ethernet switch, (right) Output port structure.

policy [KR06]. The former controls the admission of the packets into the buffers, i.e., it decides whether to accept or drop a packet based mainly on the available buffer space at the switch. Note that even if a packet is accepted, it can be later preempted if preemption is allowed. Although in practice the buffer space at the switch can be limited, we, for simplicity, assume that the buffer space is sufficiently large so that packet dropping at the switch can be ignored. The scheduling policy selects packets to be transported from the input to outputs. A queuing discipline in the switch determines the order by which the waiting packets are served. In this work, we assume first in first out (FIFO), i.e., the packets are processed (outputted) in the order of their arrivals.

There are latency concerns owing to potential queuing delays at the Ethernet switch and we model the queuing delay at the switch in Chapter 4 and Chapter 5. In Chapter 4, a continuous-time queuing model is considered, which is extended to a discrete-time queuing model in Chapter 5.

2.3 Queuing System

The system model in the thesis considers a packetized fronthaul network, such as the Ethernet switch, which collects user traffic from several RRUs. Thus, the packets are buffered in the queue and might experience a waiting time. The queuing system is briefly described below.

Queuing models and Kendall's notation

A queue in the basic queuing model such as shown in Fig. 2.6 is characterized by the following properties:

- Arrival process:

  The arrival process is described by the distribution of the interarrival times between two consecutive arrivals. Let us consider that the packets arrive at times t_1, t_2 . . . , t_n. Then the interarrival times $\tau_n = t_n - t_{n-1}$, $n > 1$ are random variables (RVs), which can have many possible distributions. As in many practical situations, in this work it is assumed Poisson arrival, where the interarrival times are independent and

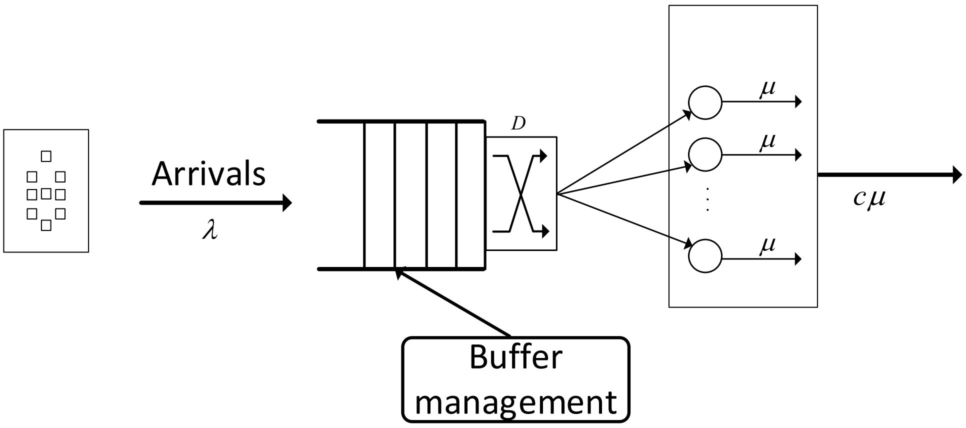


Fig. 2.6. A basic queuing system.

identically distributed (i.i.d.) and are exponentially distributed, that is, $P[t \leq \tau] = 1 - \exp(-\lambda t)$, where λ is the average arrival rate, denoted as $\lambda = \mathbb{E}[1/\tau]$.

- Service times:

  The service times s_n for the n^{th} arriving packets are generally assumed to be i.i.d. RVs and are independent of the interarrival times. For example, the service times can be deterministic, exponentially, hyperexponentially distributed or even follow a general distribution.

- Number of servers:

  The number of servers defines how many packets can be served simultaneously. The number of servers in a system in the context discussed here can be only one, which is the simplest queuing system that can serve only one customer at a time or can have c servers (a multiserver system), which can serve up to c packets simultaneously.

- System capacity:

  It defines the maximum number of packets in the system, including those already in the service. Unless stated, the system capacity is assumed to be infinite or unlimited. If a queue has reached the peak system capacity, no further arrivals will be permitted in the system.

- Population size:

  It refers to the total number of potential packets, which could be finite or infinite.

- Queue discipline

  The queue discipline determines the order in which the waiting packets are serviced. Many queuing disciplines such as FIFO, last in first out (LIFO), priority or processor sharing (PS) are possible. The most common one is FIFO, where the waiting packets are serviced in the order of their arrivals.

  To characterize a range of queuing modules, Kendall [Ken53] introduced a shorthand notation, referred to as Kendall's notation, of the form $A/S/c/K/N/D$, where the first three denote the interarrival time distribution, service time distribution and number of servers, respectively. The last three are generally the default parameters and specify the capacity of the queue (number of buffers), the population size and the queuing discipline, respectively. For example, the full notation for the most common M/M/1 queue is M/M/1/∞/∞/FIFO. The symbols used for A and S are:

- M Exponential time distribution (Markov or random times)
- G General (any distribution with mean and variance)
- D Deterministic
- E_k Erlang with parameter k
- H_k Hyperexponential with parameter k

In a queuing system, the network load utilization factor ρ, (sometimes also known as traffic intensity) is used to specify the percentage of the time the server is utilized and is defined as the ratio of the mean service time $\rho = \mathbb{E}[s]$ to the mean interarrival time $\mathbb{E}(\tau)$ as

$$\rho = \frac{\mathbb{E}[s]}{\mathbb{E}[\tau]} = \lambda \mathbb{E}[s]. \tag{2.9}$$

For the queuing system to be stable, it must be ensured that $\rho \leq 1$. In a multiserver system, the stability condition is $\rho/c \leq 1$[10]. The used queuing models in this work are M/G/m/m (c.f. Chapter 3), M/HE/1 (c.f. Chapter 4, and Chapter 5).

Discrete-time Queuing

Due to the ever increasing digitization trends in the telecommunication systems compared to their analog counterparts, discrete time analysis plays a crucial role due to its applications to slotted systems, such as LTE, slotted ALOHA, satellite and asynchronous transfer mode (ATM) network [Woo94]. The queuing theory is used to analyze the performance of a discrete-time (DT) system in Chapter 5. A DT system means that the system is observed for analysis only at pre-specified points that may be of equal or unequal intervals.

In a DT queuing system, the time axis is divided into a sequence of time intervals, known as slots, as shown in Fig. 2.7 and their end points are called slot boundaries. For simplicity, without loss of generality, we consider that the time intervals are equidistant. Let the slots be numbered sequentially such that the m^{th} slot lies in the interval $[m-1, m)$, where $m \geq 1$. Let, T denotes the slot duration. For such a system, it is assumed that all the activities such as arrivals, departures and service to a packet occur only at the slot boundaries and that they might occur at the same time. In such systems, arriving packets could occur either at beginning of the slot or at the end of the slot and hence, accordingly are termed as early arrival or late arrival method. For mathematical convenience, unless

[10] Note that in many queuing systems even a stricter stability condition, i.e. $\rho/c < 1$ should be fulfilled because the mean queue length explodes when $\rho/c = 1$. For the system with no randomness at all, i.e. for D/D/1 system, even $\rho/c = 1$ guarantees the stability. [AR15]

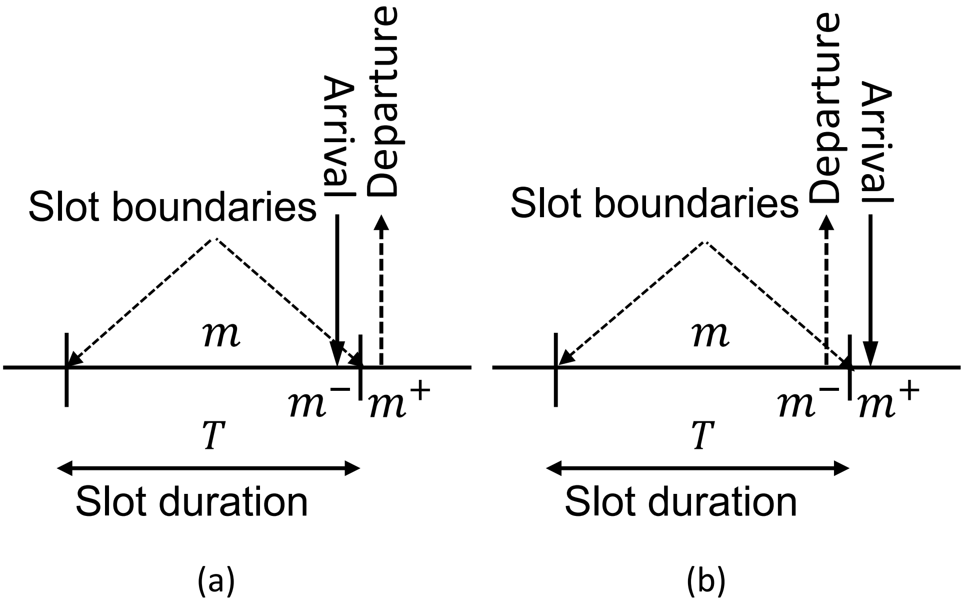


Fig. 2.7. Arrival methods: (a) Late arrival method (b) Early arrival method.

otherwise stated, late arrival method is followed in Chapter 5, i.e., arrivals occur at an instance $m^- = m - \Delta t$ just before the slot boundaries m, and departures occur at the instance $m^+ = m + \Delta t$ immediately after the slot boundaries m [AFSP13], where Δt is accordingly an infinitesimal small time period just before or soon after the slot boundary m. The mean interarrival time between the slots is geometrically distributed with rate $p_{\text{arr}}, 0 < p_{\text{arr}} < 1$. In other words, the packet arrivals occur in a slot according to a geometrical arrival process with success probability p_{arr}.

We assume that the number of arrivals in successive slots are i.i.d. RVs. The service times (in terms of number of slots, i) required to process packets are i.i.d. RVs, (say S) with probability distribution $p_S(i) = \mathbb{P}(S = i)$ for $i \geq 1$. Therefore, its MGF is $S(z) = \sum_{i=1}^{\infty} p_S(i) z^i$. Notice that at least one slot, i.e., $S = 1$ is needed for service time, if there is an arrival. This discussion will be used for the latency analysis in Chapter 5.

2.4 Chapter Summary

C-RAN concepts and its need, advantages and disadvantages are described. Moreover, challenges introduced by legacy CPRI protocol on the fronthaul are presented, which are motivations for the thesis. Functional splits, which are considered as a promising approach to relax the fronthaul challenges, are described focussing on their data rate calculations. Variable fronthaul data is required to obtain statistical multiplexing gain. For this, it is explained how and under which conditions the fronthaul data rate becomes variable. At the end, continuous- and discrete-time queuing systems are explained, which will be used for latency analysis in Chapters 4 and 5, respectively. To sum up, foundations of the thesis work and research problems are presented in this chapter.

Chapter 3

Bandwidth-Constrained C-RAN Fronthaul

The C-RAN architecture, which was initially envisioned for 4G, is problematic for 5G RATs, because the conventional C-RAN is a fully centralized architecture and is based on the legacy CPRI protocol, which requires a serial CBR fronthaul traffic and a very low-latency. However, with a suitable functional split, the fronthaul does not need to forward the data continuously, but only when there is actual user traffic in the cell. This alleviates the burden on the fronthaul bandwidth and makes the fronthaul data rate variable, because only the allocated subcarriers to the users will be forwarded (refer to Section 2.1.2). Thus, this dependency of the fronthaul data rate on the actual user traffic can be used to obtain statistical multiplexing gains. In this chapter, we explain how to obtain statistical multiplexing gains by exploiting the randomness of the user traffic. Moreover, user data rate also depends on the number of pilots, as a certain number of pilots are used for channel estimation. Hence, a simple iterative optimization algorithm is presented in Section 3.4 to adapt the number of pilots to the fronthaul capacity. We term this optimization approach as a *pilot-based optimization* and analyze the impact of the number of pilots on fronthaul bandwidth. This is based on the assumption that the number of pilots that can be assigned to the active users is limited. Thus, if the fronthaul cannot incorporate more user traffic, it does not need to support more users in terms of number of pilots.

Next, it is essential for the MNOs to dimension the fronthaul based on the actual traffic demands. This means that the number of fronthaul transceivers (TRXs) need not necessarily be dimensioned for the peak load, as full the fronthaul bandwidth utilization occurs only occasionally. Hence, assuming a reasonable outage, fronthaul can be dimensioned cost efficiently. For this, a simple cost model is presented in Section 3.6 to optimize the fronthaul TRX cost.

Contributions

The contributions in this chapter are briefly summarized below:

1. An analysis on achieving statistical multiplexing gain is provided exploiting the

randomness of user traffic using spatial traffic maps and queuing theory;

2. A simple, iterative optimization algorithm is developed to analyze the impacts of pilots on statistical multiplexing gain;

3. A simple optimization method is provided to save WDM-PON transceiver cost.

3.1 Packet-switched C-RAN Fronthaul

A recent initiative IEEE 1914 working group for NGFI [NGFa] has shown that the stringent fronthaul data rate requirement, as discussed in Section 2.1.1, can be relaxed using a packet-based transport solution such as Ethernet. Besides that, recently, an eCPRI specification [eCP19] has been released, which also enables packet-based transport and supports real-time traffic.

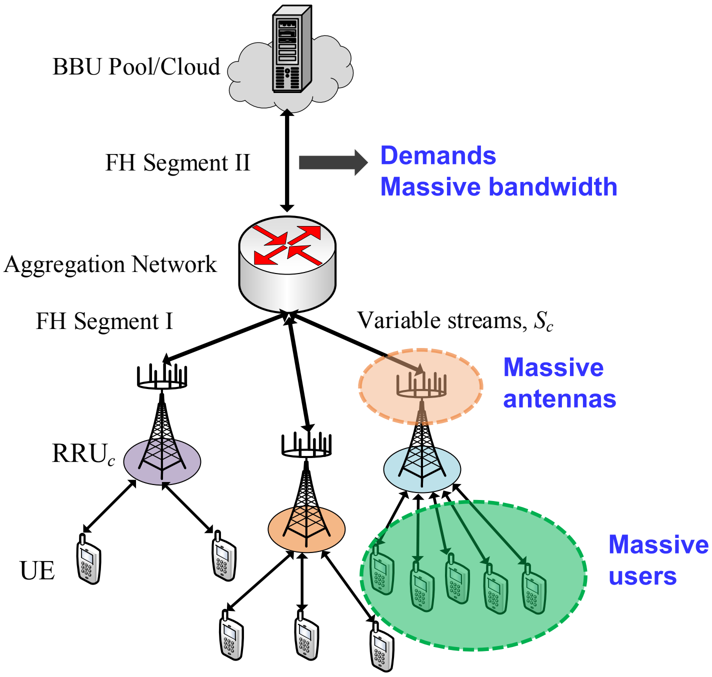


Fig. 3.1. Packetized C-RAN Fronthaul with an Ethernet switch connecting two fronthaul segments: FH Segment I and FH Segment II.

In this work, we consider a packetized fronthaul network with a basic schematic as illustrated in Fig. 3.1. The fronthaul network consists of an Ethernet switch and two fronthaul segments: *FH Segment I* and *FH Segment II*. FH Segment I is the direct link between the RRU and the aggregation network, also termed as *last mile*, and FH Segment II is the link between the aggregation network and the BBU pool. The advantage of having FH Segment I is that it allows users to have data delivery with shorter cable lengths, which otherwise would have been a single, long dedicated fiber between each RRU and BBU. On the other hand, FH Segment II requires higher bandwidth and more protection against the link failure. The traffic from multiple RRUs via Segment I is multiplexed at an aggregation network (e.g., an Ethernet switch), and then the aggregated traffic is transported to the BBU via Segment II in the UL and vice versa in the DL [LXZN15, BRW$^+$15],[HR16, Chapter 4]. Ethernet allows multiple RRUs to share a common fronthaul resource through virtualization technologies, takes advantage of statistical multiplexing, and provides considerable bandwidth saving, which is explained in Section 3.2.

3.2 Statistical Multiplexing Motivation

Unlike packetized transport networks, in a fully centralized C-RAN with CPRI-like split, fronthaul data rate is always *static* and *independent* of the traffic load, i.e., full fronthaul data rate needs to be forwarded even when there is no user connected to the RRU. However, with the appropriate RRU-BBU functional split such as intra-PHY split (refer to 2.1.2), where resource mapping and precoding operations are executed at the RRU instead of centrally at the BBU, fronthaul data rate now depends on the user traffic. This allows fronthaul data rate to be more closely coupled with the actual user traffic. Hence, statistical multiplexing gains can be obtained by using traffic randomness.

Precoding[1] at the RRU enables to transmit one stream per user instead of one stream per transceiver. As the streams from users are varying, we can observe two methods to lower the required fronthaul data rate: first, accepting a certain outage (which can be applied to both FH Segments I and II), and second, accounting for the effect of statistical multiplexing in the aggregation part (which is only possible in Segment II). In this work, we describe a methodology how to evaluate these gains and quantify the benefits for different scenarios.

Performing precoding at the RRU gives rise to two advantages: First, the number of spatial streams will vary according to the users currently served. Hence, by allowing a certain outage probability within the limits of acceptable QoS, i.e., by dimensioning the fronthaul capacity only for a reasonable percentile of the traffic distribution, the required fronthaul capacity can be reduced considerably. Second, the variable streams of different RRUs can be combined in the aggregation segment, resulting in statistical multiplexing, which further lowers the required fronthaul capacity. In Section 3.3.3, we explain this concept. The possible factors to exploit statistical multiplexing gain are: variable fronthaul streams, aggregation of the transport streams from different cells and a reasonable outage probability. We combine these factors, and use queuing and spatial traffic models to derive the mathematical expressions for statistical multiplexing gains that can be obtained from the randomness in the user traffic.

[1] For simplicity, we assume that the RRU generates the beamforming weights locally at the RRU after obtaining the perfect channel state information (CSI) from the UL pilots, which means that there is, generally, no beamforming signaling overhead on the fronthaul.

3.3 System model

3.3.1 MIMO Rate

Fig. 3.2 shows the frame structure of a massive MIMO system. The time-frequency plane is divided into coherence blocks of length $\tau_{\mathrm{c}} = \tau_{\mathrm{coh}} B_{\mathrm{coh}}$ symbols over which the channels can be approximated as time-invariant and frequency-flat, τ_{coh} and B_{coh} being the coherence time and coherence bandwidth of the channel, respectively [BLD16]. Out of τ_{c} symbols, $1 \leq \tau_{\mathrm{p}} < \tau_{\mathrm{c}}$ symbols in each frame are reserved for UL pilot signaling, and the remaining $(\tau_{\mathrm{c}} - \tau_{\mathrm{p}})$ symbols are allocated for payload data and are split between UL and DL transmission. Let $\zeta^{(\mathrm{ul})}$ and $\zeta^{(\mathrm{dl})}$ are the fractions of the UL and DL data transmissions, respectively with the constraint $\zeta^{(\mathrm{ul})} + \zeta^{(\mathrm{dl})} = 1$. Then $\zeta^{(\mathrm{ul})}(\tau_{\mathrm{c}} - \tau_{\mathrm{p}})$ and $\zeta^{(\mathrm{dl})}(\tau_{\mathrm{c}} - \tau_{\mathrm{p}})$ denote fraction of symbols used for the UL and DL data transmission, respectively. We assume that the massive MIMO systems operate in the time division duplexing (TDD) mode and the channel is reciprocal, which means that the RRU estimates the DL channel using the UL pilots[2] and then the RRUs exploit the channel knowledge for DL transmission [Mar10, R$^+$13, LETM14]. The number of orthogonal pilots that can be assigned to users is always limited [BLD16]. We focus on the DL and assume for simplicity that $\zeta^{(\mathrm{ul})} = 0$, i.e., there is only DL data transmission and in the UL only pilots are transmitted.

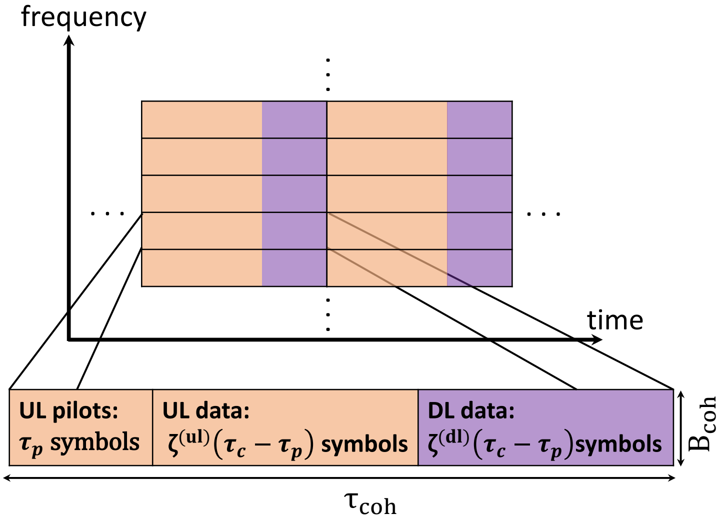

frequency
time
UL pilots:
τ_p symbols
UL data:
$\zeta^{(\mathrm{ul})}(\tau_c - \tau_p)$ symbols
DL data:
$\zeta^{(\mathrm{dl})}(\tau_c - \tau_p)$symbols
B_{coh}
τ_{coh}

Fig. 3.2. Time-frequency plane divided into coherence blocks of length $\tau_{\mathrm{c}} = \tau_{\mathrm{coh}} B_{\mathrm{coh}}$ symbols in which the channels are time-invariant and frequency-flat. Out of the total τ_{c} symbols, τ_{p} symbols are used for UL pilot signaling and the remaining $(\tau_{\mathrm{c}} - \tau_{\mathrm{p}})$ symbols are used for UL and DL data transmission.

Let us now consider the DL transmission of a massive MIMO system, where the cellular network consists of a set $\mathcal{C} = \{1, 2, ..., C\}$ of C cells. Each cell has its own RRU equipped with an array of M_c antennas that can communicate with K_c single-antenna users at a time, out of a set of K_{max} users. Usually for massive MIMO system it is assumed that the number of antennas at each RRU is much larger than the number of served users,

[2] It is assumed in massive MIMO that the channel hardening suppresses small scale fading, which makes the power allocation easier and eliminates the need for the DL pilots [BL15].

i.e., $M_c >> K_c >> 1$. We assume that the transmissions between all the RRUs and UEs are perfectly synchronized, the UEs share the same time-frequency resource, and transmissions take place over Rayleigh block fading channels.

The area served by the RRUs is denoted $\mathcal{A}$, with a single location indicated by its coordinates (x, y). The path-loss factor, defined here as a ratio of received power to transmitted power, between RRU c and location (x, y) is denoted $\alpha_c(x, y)$ and modeled according to the urban microcellular path-loss model defined in [IR09]. Users are associated with the RRU providing the lowest path loss, and, hence, the serving area $\mathcal{A}_c$ of an RRU c is given as

$$(x, y) \in \mathcal{A}_c \text{ if } c = \arg\max_{\mathcal{C}} \alpha_c(x, y). \tag{3.1}$$

Let us consider that the total transmit power of an RRU c is pM_c, where p is the average power per antenna, which is considered to be same[3] for all the antennas. Each cell c receives interference from the active antennas in all other cells. Let M_d be the number of active antennas in any other cell $d \neq c$. Hence, the total transmitting power of the corresponding cell is pM_d. Then, the signal-to-interference-plus-noise ratio (SINR) at a location (x, y) can be obtained as:

$$\gamma(x, y) = \frac{pM_c\alpha_c(x, y)}{\sigma^2 + \sum_{d\in\mathcal{C}\setminus c} pM_d\alpha_d(x, y)}, \tag{3.2}$$

where σ^2 denotes the noise power.

We use Poisson point process to describe the traffic demand in the network, compare e.g. [BB09]. For this, we define for each location a user arrival rate per area $\Lambda(x, y)$ (in 1/s/km^2), and a corresponding traffic density (in Mbps/km^2) is

$$\Omega(x, y) = \Lambda(x, y) \cdot \overline{F}, \tag{3.3}$$

where $\overline{F}$ (in bits) is the mean file size requested per user. The mean traffic density of the overall area $\mathcal{A}$ we denote by $\bar{\Omega}$. For the serving area of an RRU this results in user arrivals with arrival rate (in 1/s)

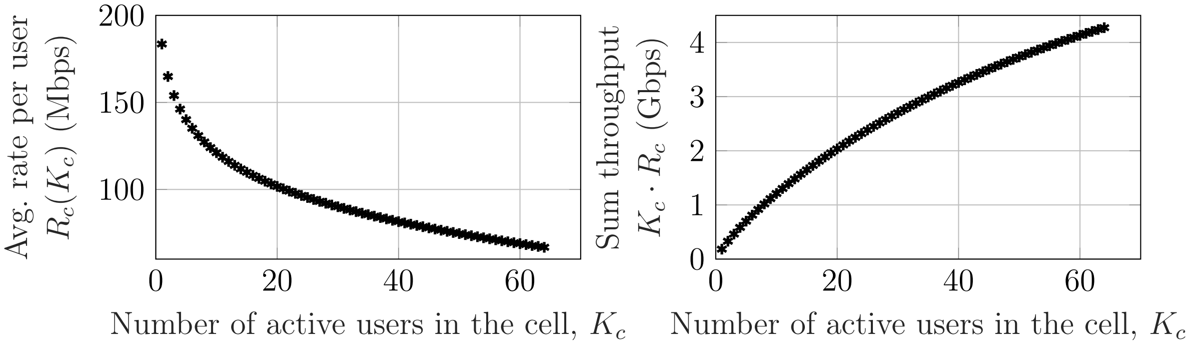
$$\lambda_c = \int_{\mathcal{A}_c} \Lambda(x, y)\, dx\, dy. \tag{3.4}$$

Next, we define the average SINR in the serving area of an RRU c as the expected value of the SINRs weighted according to the traffic distribution $\Omega(x, y)$, i.e.,

$$\bar{\gamma}_c = \mathbb{E}[\gamma(x, y)] = \frac{\int_{\mathcal{A}_c} \gamma_c(x, y)\Omega(x, y)\, dx\, dy}{\int_{\mathcal{A}_c} \Omega(x, y)\, dx\, dy}. \tag{3.5}$$

Furthermore, for simplicity, we assume each RRU has obtained perfect CSI from its users, which is a reasonable assumption for low mobility scenario. In addition, we consider that RRUs employ zero-forcing (ZF) precoding in order to cancel out the intracell interference

[3] The same average power per antenna is based on the assumption of independent fading and the fact that power gets averaged over many subcarriers [HCBJ18].



(a) **(b)**

Fig. 3.3. Illustration of massive MIMO impact: (a) Average data rate per user and (b) total sum throughput [CBF17, CBF18].

and adapt power allocation such that each of the K_c user achieves the same average data rate R_c (in bps), given by [BJDO14, HCBJ18]

$$R_c(K_c) = W\left(1 - \frac{\beta K_{\max}}{\tau_c}\right)\log_2\left(1 + \frac{\bar{\gamma}_c}{K_c}(M_c - K_c)\right), \tag{3.6}$$

where W is the channel bandwidth, β the pilot reuse factor, $\tau_{\mathrm{c}} = \tau_{\mathrm{coh}}B_{\mathrm{coh}}$ the length of channel coherence interval, and $K_{\max}$ the maximum number of users, which is, for now, assumed to be same for all cells. $M_c - K_c$ is the effective array gain, and the factor $1/K_c$ accounts for the fact that the total transmit power is split between all users. The pre-log factor $\left(1 - \frac{\beta K_{\max}}{\tau_c}\right)$ accounts for the necessary overhead for channel estimation and it will play an important role, which is explained in Section. 3.4. Note that the above rate expression is a lower bound on the average Ergodic rate of a user obtained using Jensen's inequality[4]. (refer to [HCBJ18] for the proof).

Exemplary plots of the average data rate per user, R_c and total sum throughput, $K_c \cdot R_c(K_c)$ are shown in Fig. 3.3a and Fig. 3.3b, respectively. It is obvious from Fig. 3.3a that higher per user average rate is achievable with fewer active users. On the other hand, Fig. 3.3b shows that the sum throughput increases when increasing the number of active users, which highlights the general benefit of massive MIMO in terms of capacity. The simulation parameters for all the figures are listed in Table 3.1.

3.3.2 Queueing model

In order to get the user distribution in the cell, i.e., in order to find the probability of a certain number of users served by each cell, we utilize queueing theory results from [CS07, CS94] and model the load at each MIMO RRU as a state-dependent $M/G/m/m$ queue, where M indicates the arrival process is Markovian or memoryless, G indicates distribution of service time is general and the last two m indicate that the number of servers and number of places in the system are equal (m servers and no waiting). In this

[4] Although Jensen's inequality is generally a loose bound, a pretty tight bound in this case is obtained because of the channel hardening that averages out the small-scale fading variations [HCBJ18].

work, we assume that the maximum number of users that a BS can serve is $m = K_{\max}$. Under $M/G/m/m$ state-dependent queue, the steady-state probabilities of the number of users served by an RRU c, $\pi_c(n) \equiv P_r[K_c = n]$ are given by [CS07]

$$\pi_c(n) = \left[\frac{\left[\lambda_c \frac{\overline{F}}{R_c(1)}\right]^n}{n! f(n) f(n-1) \dots f(2) f(1)} \right] \pi_c(0), \qquad n \in \{1, 2, \dots K_{\max}\} \tag{3.7}$$

where $\pi_c(0)$ denotes probability that there is no user in cell c and is given by

$$\pi_c(0) = \left[1 + \sum_{i=1}^{K_{\max}} \left(\frac{\left[\lambda_c \frac{\overline{F}}{R_c(1)}\right]^i}{i! f(i) f(i-1) \dots f(2) f(1)} \right) \right]^{-1},$$

where $f(n) = R_c(n)/R_c(1)$ is the normalized rate per user, R_c the average data rate per

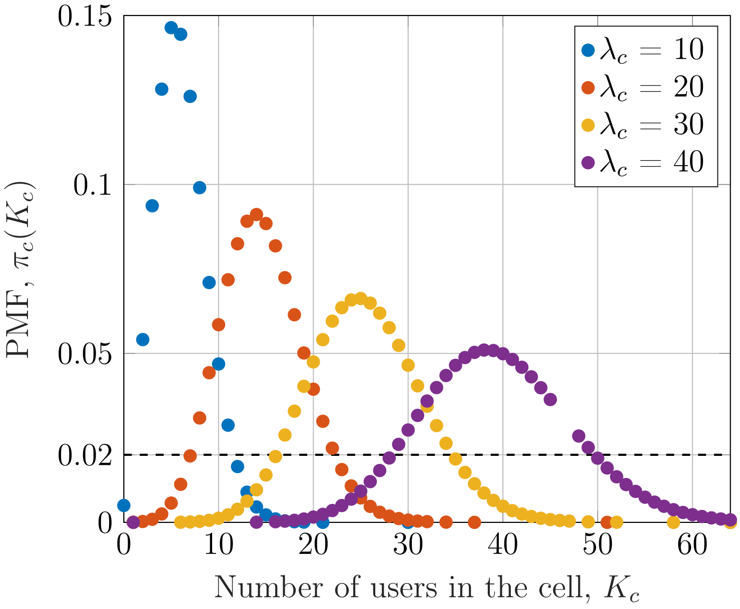


Fig. 3.4. Massive MIMO user distribution: PMF of the number of concurrently active users in massive MIMO cell with $K_{\max} = 64$ and different arrival rates [CBF17, CBF18].

user given by (3.6), $\overline{F}$ the average file size requested by each user, $\overline{F}/R_c(n)$ the average service time and λ_c is the arrival rate from (3.4). The corresponding CDF we denote as $\Pi_c(n) = \sum_{i \leq n} \pi_c(i)$.

An example of user distributions defined by (3.7) is given in Fig. 3.4 for different values of λ_c. Fig. 3.4 implies that as the arrival rate increases, there are more flows per second from the users and the number of users attempting to get the resources is also increasing. For example, at 2% probability, the number of active users for arrival rate of $\lambda_c = 10$ flows per second is 12, which increases to 50 users for arrival rate of $\lambda_c = 40$ flows per second.

3.3.3 Fronthaul Capacity, Outage and Multiplexing

In literature, there are two main different definitions of the fronthaul data rate [TTQL17] requirement: The first one states that fronthaul data rate is defined as the maximum sum data rate transmitted on each fronthaul. In this case, the authors always implicitly assume each fronthaul can serve unlimited number of users. However, this assumption can not hold in real systems. On the other hand, the second definition states that fronthaul data rate is defined as maximum number of users that can be served on each fronthaul. In this work, we adopt the second definition.

Conventionally, each fronthaul Segment I needs to be dimensioned to serve its maximum number of users, i.e., for $S_{c,\max} = K_{\max}$ streams. Similarly, Segment II could be dimensioned for $S_{\mathcal{C}} = C \cdot K_{\max}$ streams. Such a dimensioning is common in conventional CPRI-based fronthaul networks, which require a static and constant fronthaul data rate per RRU. However, this high and constant data rate in the fronthaul makes the fronthaul bandwidth-constrained, which eventually will be bottleneck for massive MIMO systems, as explained in Section 2.1.1. From network operators perspective, it would be beneficial for them to constrain that capacity to lower deployment cost. As the traffic is varying due to varying number of user-streams, we can assume a certain outage probability[5] P_{O} on each link according to some QoS requirements. Hence, the fronthaul in Segment I can be dimensioned with the outage capacity

$$S_{c,\mathrm{O}} = n \quad \text{such that } 1 - \Pi_c(n) < P_{\mathrm{O}}, \tag{3.8}$$

where $\Pi_c(n) = \sum_{i \leq n} \pi_c(i)$ is the CDF of each individual cell, as defined in Section 3.3.2. Furthermore, the streams from multiple RRUs are aggregated (summed up) at the switch before being transported to the FH Segment II. As the streams from different RRUs are independent, their summation leads to a convolution of the corresponding probability distributions. Hence, the distribution of user-streams on FH Segment II becomes

$$\pi_{\mathcal{C}} = \pi_1 * \pi_2 * \cdots * \pi_c, \tag{3.9}$$

with combined CDF $\Pi_{\mathcal{C}}$. Hence, the outage capacity in the FH Segment II becomes

$$S_{\mathcal{C},\mathrm{O}} = n \quad \text{such that } 1 - \Pi_{\mathcal{C}}(n) < P_{\mathrm{O}}. \tag{3.10}$$

The convolution gives the distribution of the cell $\mathcal{C}$. Note that we will have a truncated distribution if there is an outage in the FH Segment I. We obtain a statistical mutiplexing gain G, as $S_{\mathcal{C},\mathrm{O}} \leq \sum_{\mathcal{C}} S_{c,\mathrm{O}}$. In order to assess the benefit of the statistical multiplexing, we define the relative required fronthaul capacity in Segment I and II as:

$$S_1 = \frac{\sum_{\mathcal{C}} S_{c,\mathrm{O}}}{C \cdot S_{c,\max}} \qquad \text{for FH Segment I,} \tag{3.11}$$

$$S_2 = \frac{S_{\mathcal{C},\mathrm{O}}}{C \cdot S_{c,\max}} \qquad \text{for FH Segment II.} \tag{3.12}$$

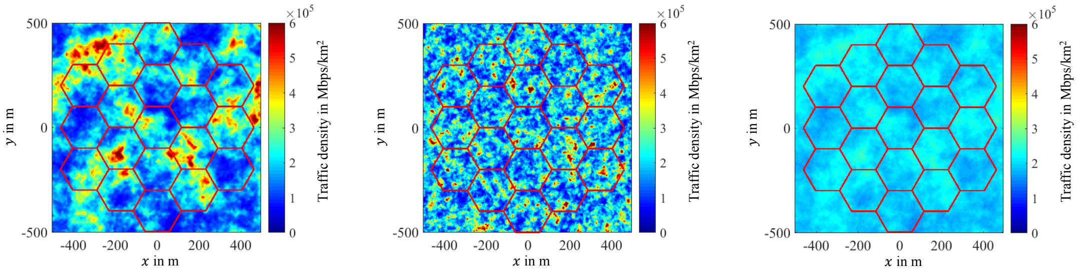
[5] The outage probability in this case refers to a probability when the offered fronthaul traffic, which is random, exceeds the average fronthaul traffic.

and, statistical mutiplexing gain G as

$$G = 1 - \frac{S_{\mathcal{C},\mathrm{O}}}{C \cdot S_{c,\mathrm{max}}}. \tag{3.13}$$

(3.11) and (3.12) calculate the fraction of the total fronthaul traffic to be transported in the Segment I and II, respectively, and (3.11) gives the relative bandwidth savings due to statistical multiplexing.

3.3.4 Traffic Model



(a) $\bar{\Omega}$ = 200 Gbps/km^2, σ_Ω = 100 Gbps/km^2, $d_{\mathrm{corr}} = 50$ m.

(b) $\bar{\Omega}$ = 200 Gbps/km^2, σ_Ω = 100 Gbps/km^2, $d_{\mathrm{corr}} = 10$ m.

(c) $\bar{\Omega}$ = 200 Gbps/km^2, σ_Ω = 20 Gbps/km^2, $d_{\mathrm{corr}} = 50$ m.

Fig. 3.5. Exemplary traffic distributions: Fig. (b) exhibits a lower correlation distance, and Fig. (c) a lower standard deviation compared to Fig. (a).

In general, the statistical multiplexing gain will depend on the variance of the total number of streams. This variance is affected both by the (temporal) variation of users from (3.7), and by the different (spatial) variation of users among different cells based on the traffic distribution $\Omega(x, y)$. In order to model $\Omega(x, y)$, we utilize a traffic model developed in [LZZ$^+$14, KSF15]. This traffic model allows to create random spatial traffic maps via log-normal distributed random fields defined by three statistical parameters: mean traffic density $\bar{\Omega}$, traffic density standard deviation σ_Ω, and a correlation distance d_{corr}. Three different examples of such traffic maps are given in Fig. 3.5. The parameter $\bar{\Omega}$ controls the overall traffic demand, σ_Ω controls the ratio between traffic demand in hot spots and low-traffic areas, and d_{corr} controls the size of the hotspots. With the traffic maps generated based on this model, statistical multiplexing gains are averaged over random scenarios without having to rely on just a single scenario, leading to more consistent results and more general conclusions for real scenarios. For more details on this traffic model see [KSF15].

Now, in order to illustrate the underlying concepts, a layout of 19 homogeneous hexagonal cells each having inter site distance of $d_{\mathrm{ISD}} = 200$ m are plotted as shown in Fig. 3.5. It is common practice to assume regular cells, in particular hexagonal cells to establish the general properties, although the practical deployments have irregular cells. The innermost cell is surrounded by a tier of six cells, which in turn are surrounded by additional tier of 12 cells. Fig. 3.6a and Fig. 3.6b respectively illustrates the PMF and CCDF of the number of users in each cell. The total traffic from $C = 19$ such cells, assuming

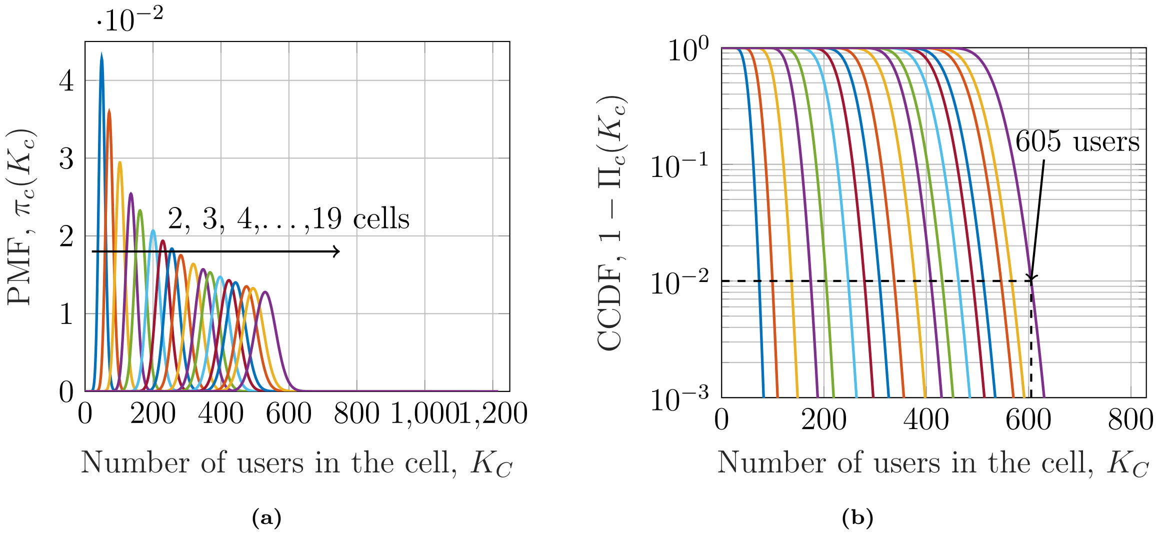


Fig. 3.6. Statistical multiplexing: Exemplary distribution of served users/required user streams for $C = 19$ aggregated RRUs. $\bar{\Omega} = 45$ Gbps/km^2, $\sigma_\Omega = 0.25\,\bar{\Omega}$, $d_{\text{corr}} = 10$ m. (a) PDF and (b) CCDF. [CBF17, CBF18]

each cell in its peak load can serve $S_{c,\max} = K_{\max} = 64$ users, demands to have total $S_\mathcal{C} = C \cdot K_{\max} = 64 \times 19 = 1216$ user streams to be forwarded. However, assuming a reasonable $P_\text{O} = 1\%$ outage on the FH Segment II, we need to transport only 605 users as shown in Fig. 3.6b, which means less fronthaul capacity demand. This results up to 50% fronthaul capacity saving.

3.4 Pilot-based Optimization

In (3.6), the pre-log scaling factor $\left(1 - \frac{\beta K_{\max}}{\tau_c}\right)$ is the channel estimation overhead. $K_{\max}$ is the maximum number of users that an RRU can support due to the number of transmitted pilots. This number $K_{\max}$ is, in general, a system design parameter and would be set according to an *expected* general peak demand. However, as we have shown in the previous section in the case of fronthaul, it can be more efficient to design a system based on the *actual* demand. If a lower number of pilots is used, it would increase the rate of all active users according to (3.6). What is even more important is that we already dimension the fronthaul to support only a limited number of users based on an acceptable outage probability. Hence. it does not make sense to support more users in terms of pilots if they cannot be served by the fronthaul anyway. We can, hence, derive a simple, iterative optimization algorithm to adapt the number of pilots to the fronthaul capacity, which is in turn based on the outage probability. For this, we assume from now on that the number of pilots (number of supported users) can be different for each RRU c and is denoted $K_{\max,c}$. The algorithm to find the optimal number of pilots $K^*_{\max,c}$ is depicted in Algorithm 1.

The algorithm can be explained as follows. We set the number of pilots to a starting value $K_{\max}$ for all RRUs c. We then calculate the number of Segment I outage fronthaul streams $S_{c,\text{O}}$ according to (3.8). This is the number of fronthaul streams/users we support

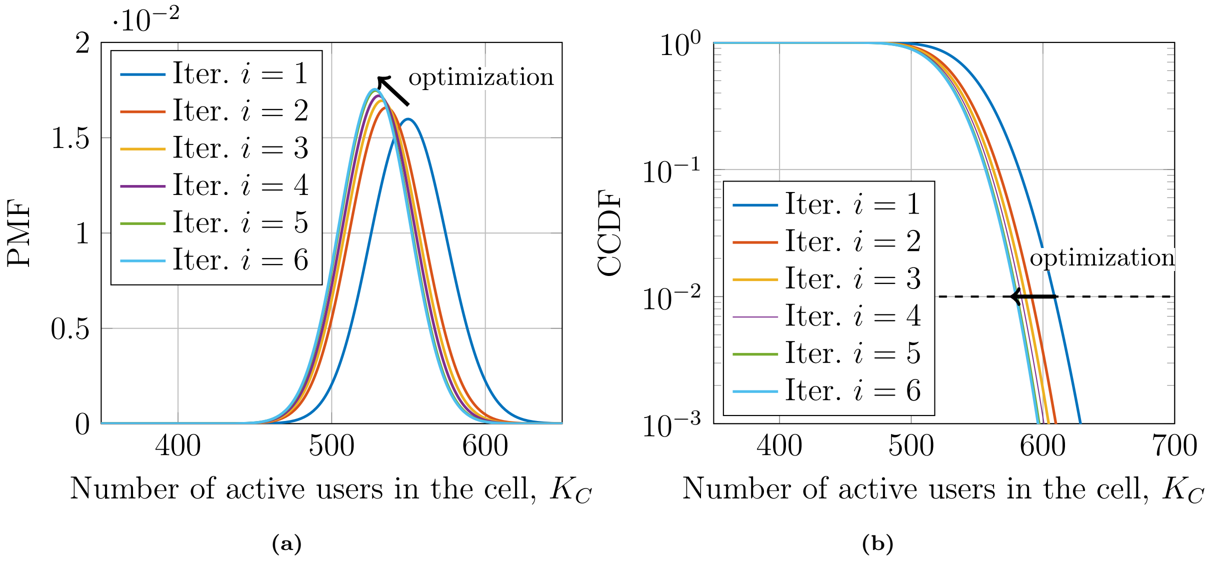


Fig. 3.7. Pilot-based optimization: Example of the pilot-based optimization for $K_{\max} = 64$ users. (a) PDF and (b) CCDF. [CBF18]

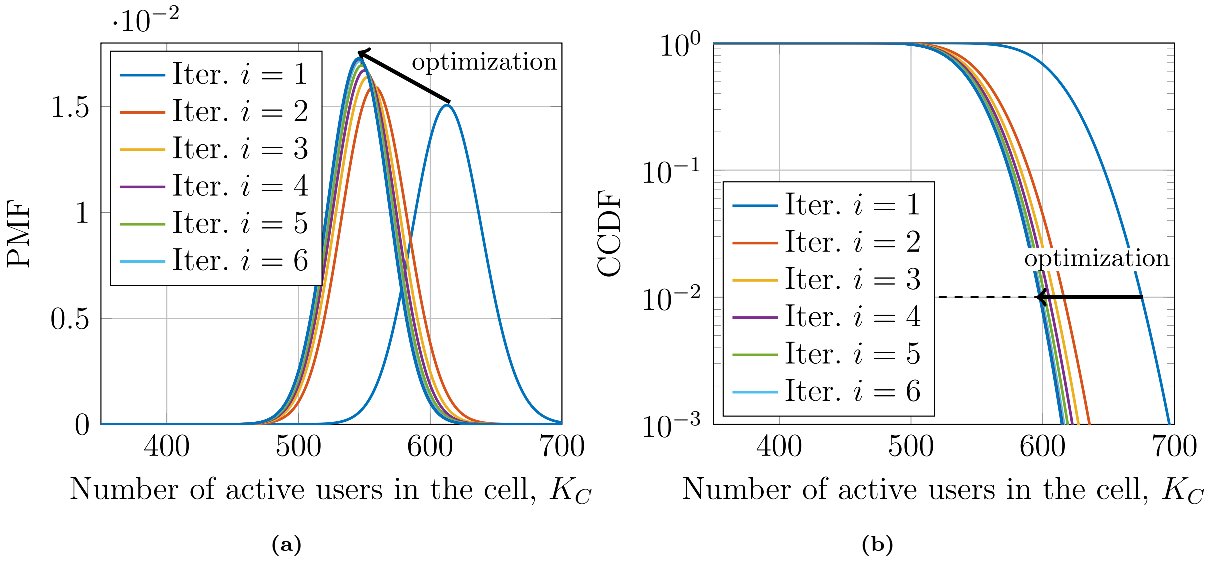


Fig. 3.8. Pilot-based optimization: Example of the pilot-based optimization for $K_{\max} = 128$ users. (a) PDF and (b) CCDF.[CBF18]

for each RRU in Segment I. Now we can use that value as the new number of pilots in each RRU. This will increase the rate of all users according to (3.6), which in turn leads to less number of users in the queue according to (3.7) (illustrated in Figs. 3.7a and 3.8a) which in turn again may reduce the number of outage streams in (3.8) (illustrated in Figs. 3.7b and 3.8b). Note that optimization gain in Fig. 3.8b is more compared to that in Fig. 3.7b as the starting point $K_{\max}$ is set higher. The algorithm terminates when the increase in rate is so low that it does not lead to a reduction of n in (3.6) by at least one. The convergence is hence guaranteed by the limitation of n to integer values.

After the final iteration, the optimal number of pilots will be equal to the fronthaul capacity in Segment I. This also means that the outage no longer occurs in Segment I

Algorithm 1 Pilot optimization.

```
$i = 0$
$K^{(0)}_{\max,c} = K_{\max} \ \forall \ c \in \mathcal{C}$
repeat
    for all $c \in \mathcal{C}$ do
        calculate (3.6), (3.7), (3.8),(3.10)
        $K^{(i)}_{\max,c} = S_{c,\mathrm{O}}$
    end for
    $i = i + 1$
until $K^{(i)}_{\max} = K^{(i-1)}_{\max} \ \forall \ c \in \mathcal{C}$
$K^{*}_{\max,c} = K^{(i)}_{\max,c}$, $S^{*}_{c,\mathrm{O}} = S^{(i)}_{c,\mathrm{O}}$, $S^{*}_{\mathcal{C},\mathrm{O}} = S^{(i)}_{\mathcal{C},\mathrm{O}} \ \ \forall \ c \in \mathcal{C}$
```

but already during user admission in the wireless link, as no more pilots are available. To illustrate the additional benefit achieved by the optimization, we define the optimization gain in terms of fronthaul capacity as:

$$g_1 = 1 - \frac{\sum_{\mathcal{C}} S^{*}_{c,\mathrm{O}}}{\sum_{\mathcal{C}} S_{c,\mathrm{O}}} \qquad \text{for FH Segment I,} \tag{3.14}$$

$$g_2 = 1 - \frac{S^{*}_{\mathcal{C},\mathrm{O}}}{S_{\mathcal{C},\mathrm{O}}} \qquad \text{for FH Segment II.} \tag{3.15}$$

3.5 Numerical Results

Scenario

To evaluate the fronthaul capacity reduction, we utilize an exemplary setup in Fig. 3.5 consisting of 19 uniformly placed hexagonal cells with inter site distance $d_{\mathrm{ISD}} = 200$ m and RRUs placed at a height of $h_{\mathrm{RRU}} = 12$ m. At first, we generate the random traffic maps according to Section 3.3.4 and place the cells on those traffic maps. Then, the relative required fronthaul data rates according to Eqs. (3.11), (3.12) are evaluated, and the results are averaged over 25 instances of random traffic maps in order to get more accurate results and to make more general conclusions for real scenarios.

3.5.1 Statistical Multiplexing Gains

Figs. 3.9 - 3.11 illustrate the reduction in the required relative fronthaul capacity that is achieved by accepting an outage and utilizing statistical multiplexing in Segment II, first without pilot optimization. As it can be seen, the relative fronthaul capacity mainly scales with the mean traffic density $\bar{\Omega}$. In addition, the required capacity in FH Segment II is always lower, as here the additional effect of statistical multiplexing comes into effect. The difference between the fronthaul Segment I and II is more pronounced towards higher traffic densities. Here, clearly the statistical multiplexing effect is more dominant compared to the reduction possible by accepting outage. Furthermore, it can be seen that higher values of traffic variance (refer to Fig. 3.9) and correlation distance (refer to Fig. 3.10)

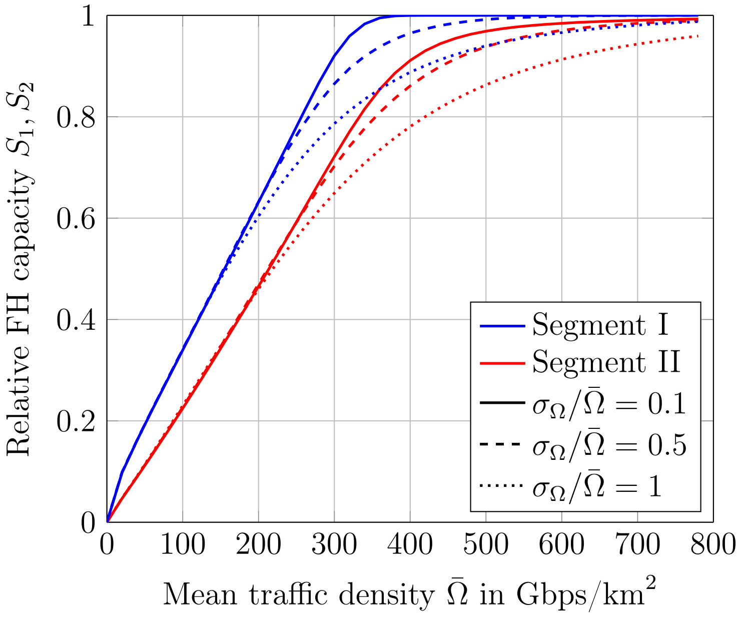


Fig. 3.9. Impact of standard deviation: Relative fronthaul capacity for different relative standard deviations of the traffic density ($\bar{\Omega}$), $d_{\mathrm{corr}} = 50$ m, $P_{\mathrm{O}} = 0.01$ [CBF17, CBF18].

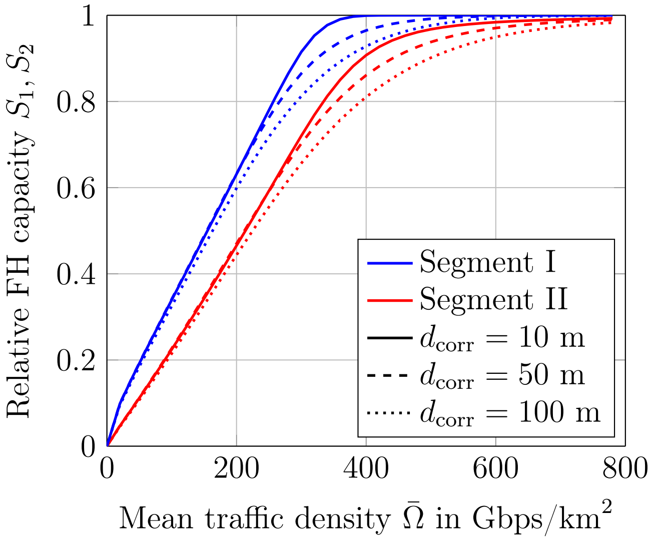


Fig. 3.10. Impact of traffic correlation: Relative fronthaul capacity for different traffic correlation distances (d_{corr}), $\sigma_\Omega/\Omega = 0.5$, $P_{\mathrm{O}} = 0.01$ [CBF17, CBF18].

lead to lower relative fronthaul capacities, as both parameters lead to a higher variability in total cell traffic among the different RRUs, hence resulting in a higher multiplexing gain G. Finally, it can be seen in Fig. 3.11 that a higher outage probability leads to an lower required fronthaul capacity, as expected. Here, especially Segment I profits, that is, there is more reduction (more gaps) between the outage percentiles. In Segment II, due to the statistical multiplexing effect, the probability distribution converges towards the mean traffic, and hence the difference between the outage percentiles is less pronounced.

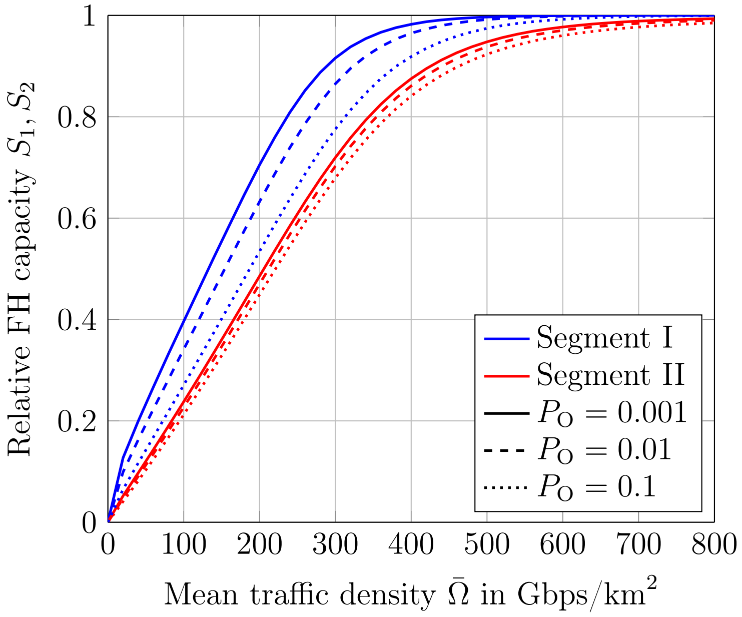


Fig. 3.11. Impact of outage probability: Relative fronthaul capacity for different outage probabilities (P_0), $\sigma_\Omega/\Omega = 0.5$, $d_{\text{corr}} = 50$ m [CBF17, CBF18].

Tab. 3.1. Simulation parameters for statistical multiplexing gain analysis, pilot-based optimization and transceiver cost optimization.

Parameters	**Symbol**	**Value**
Number of cells	C	19
intersite distance	d_{ISD}	200 m
RRH height	h_{RRU}	12 m
Bandwidth	W	20 MHz
Coherence bandwidth	B_{coh}	200 kHz
Coherence time	T_{coh}	5 ms
Number of transmitting antennas	M_c	256
Maximum users	$K_{\max}$	64
Total transmit power	pM_c	23 dBm
Noise figure	$\mathcal{F}$	5 dB
Noise power	σ^2	-101 dBm
Average file size	$\overline{F}$	10 MB
Pilot reuse factor	β	1, 4

3.5.2 Pilot Optimization

Next, Fig. 3.12 shows the relative fronthaul capacity before and after pilot optimization according to Sec. 3.4. Further, the additional reduction provided by the optimization according to Eqs. (3.14), (3.15) is illustrated. As it can be seen, the optimization achieves an additional reduction in required fronthaul capacity of approximately 10-15% in both segments. Moreover, we see that the gain is of course higher when the starting point $K_{\max}^{(0)}$ is chosen larger (refer to Fig 3.12), as in this case there is more room for improvement.

Lastly, in order to show the impact of pilot reuse factor β, we consider pilot reuse factor $\beta = 4$ in Fig. 3.13, unlike all other simulated figures, where we assumed $\beta = 1$ for

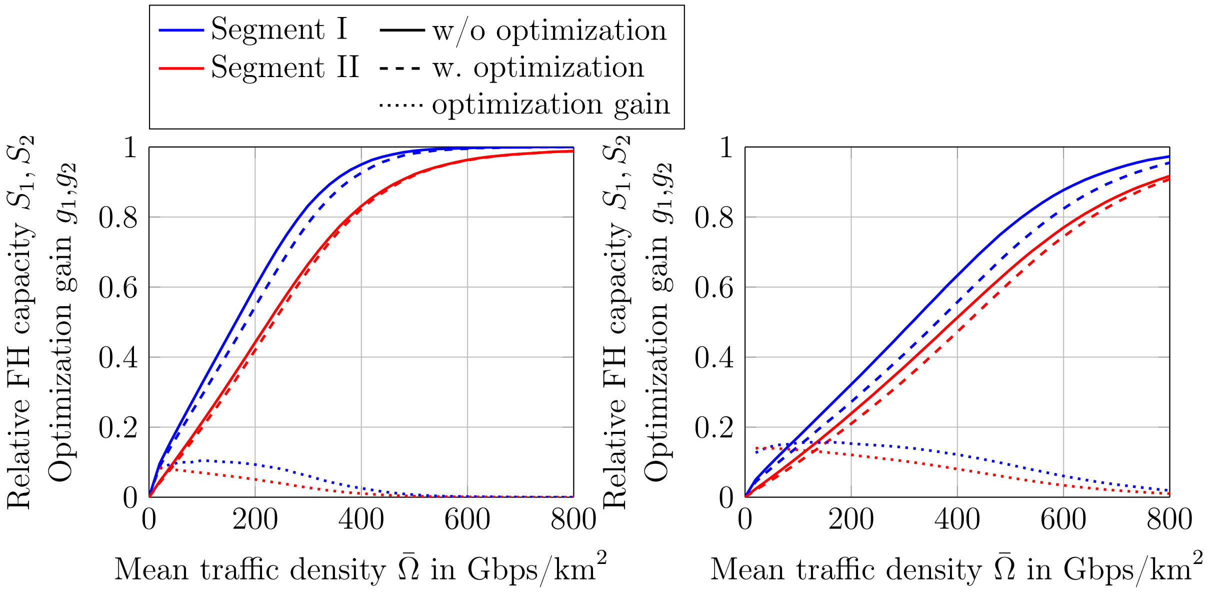


Fig. 3.12. Impact of pilot optimization: Relative fronthaul capacity after optimization for $K_{\max}^{(0)} = 64$ (left) and $K_{\max}^{(0)} = 128$ (right) [CBF18].

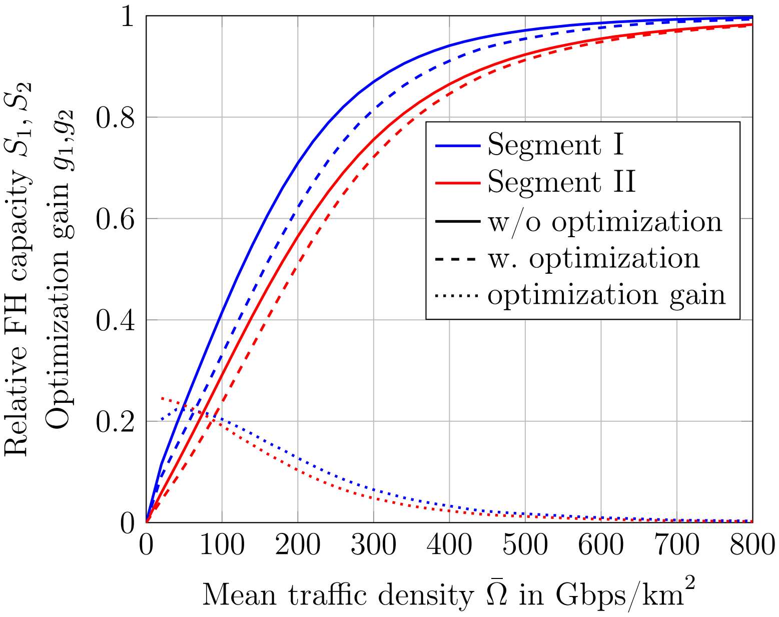


Fig. 3.13. Impact of pilot optimization: Relative fronthaul capacity after optimization for $K_{\max}^{(0)} = 64$ and $\beta = 4$ [CBF18].

simplicity. Higher value of β allows to have sufficient pilot resources to be shared among the users and hence, it helps in mitigating the pilot contamination [BLD16]. It is to be noted that proper choice of β depends on various factors, such as number of transmitting antennas, SINR values, number of allocated users and spectral efficiency. As there are more pilot resources for the users, the optimization gain has now significantly improved to roughly 25% in Fig. 3.13 compared with Fig. 3.12, where the gain is only about 10%.

After analyzing statistical multiplexing gains, we present in the next session how to dimension the fronthaul transceiver cost efficiently, which is beneficial for the MNOs. For this, WDM-PONs system is considered.

3.6 Transceiver Cost Saving Analysis

More than 2/3 of CAPEX of a mobile operator is spent in the RAN, which mainly involves site acquisition, site support, and equipment purchase [Chi13]. In terms of cost, one of the important factors affecting economical deployment of C-RAN is the availability of low-cost, low-latency and high-bandwidth optical modules for fronthaul. In general, fronthaul is dimensioned to support peak load. However, in many cases, the probability of full fronthaul capacity usage at peak load is very low. Hence, it is important to dimension the fronthaul efficiently and economically in order to reduce the TRX cost.

For this, we use the spatial traffic model and queuing theory explained in Section 3.3.4 and 3.3.2 to calculate the required number of TRX to be deployed at a particular scenario at a given outage probability, and then calculate the TRX cost saving. To our knowledge, we are the first to exploit traffic randomness using spatial traffic model and queuing theory to obtain the TRX cost saving considering practical TRX cost data. We, at first, analytically obtain the essential TRX cost saving equations, and then compute the savings using simulation. We show through the numerical results a significant cost saving of about 50% at a moderately low traffic density of 200 Gbps/km^2 compared with the case when full fronthaul capacity utilization is considered. It is shown in Section 3.6.3 that the cost saving varies with the traffic densities, outage probabilities and correlation distances.

3.6.1 WDM-PON System

A fronthaul network normally uses WDM-PON systems, as shown in Fig. 3.14, due to their high flexibility, infrastructure sharing, high capacity and low latency advantages [Chi13, ZWE17, ATM18]. A WDM-PON system is mainly composed of a central node, called an optical line terminal (OLT) at the service provider's central office, and a number of user nodes, called optical network units (ONUs) located at the customer premises.

WDM-PON systems may be designed to satisfy the highest possible required data rate, but, because of variations in the traffic demands, this situation occurs with a very low probability. Therefore, allowing an acceptable outage, the fronthaul could be dimensioned to serve fewer simultaneous links, requiring a lower capacity and, consequently, fewer WDM TRXs. This presents two advantages: firstly the deployment expenditures can be lower, and secondly, during low traffic periods, the unused ONUs or OLTs can be put into sleep mode, thus leading to energy saving. In this work, we consider only the cost saving, which is explained in Section 3.6.2.

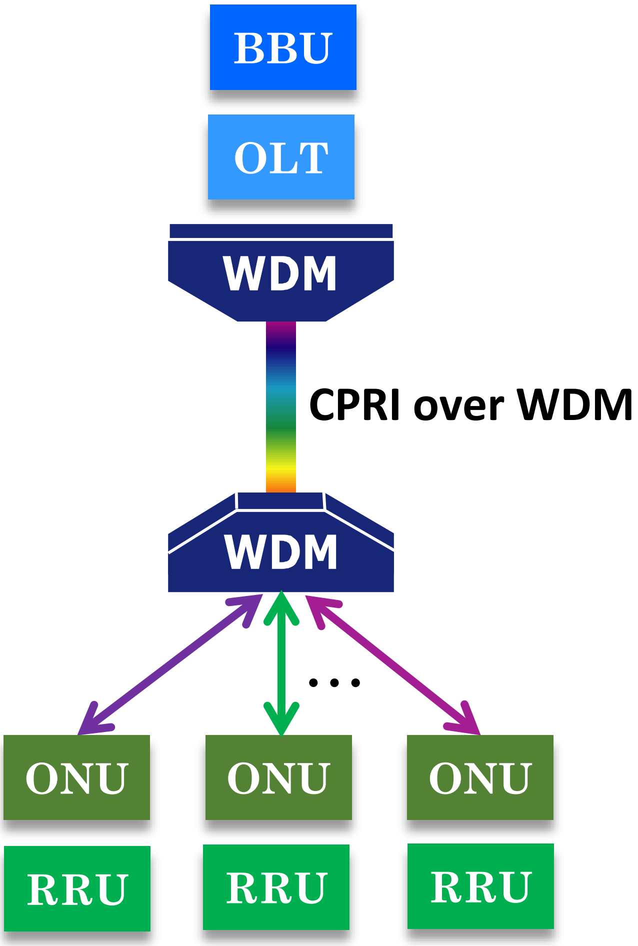


Fig. 3.14. Implementation of fronthaul network using WDM-PON system.

3.6.2 Transceiver Cost Saving Model

Let R_1, R_2, $\cdots$, R_T denote the capacities (in Gbps) of T different types of TRXs and Ψ_1, Ψ_2, $\cdots$, Ψ_T denote the corresponding associated cost factor of these TRXs. The cost factor here refers to the relative cost of a TRX compared with the standard 10 Gbps TRX, whose relative cost factor is 1. Further, let us consider we need to deploy a TRX with maximum Γ Gbps for a particular scenario, where we might require w_1, w_2,$\cdots$, w_T number of R_1, R_2, $\cdots$, R_T Gbps TRX, respectively. Then, the total required TRX capacity Γ (in Gbps) can be obtained as

$$\Gamma = \sum_{i=1}^{T} R_i \cdot w_i, \tag{3.16}$$

where w_1, w_2,$\cdots$, w_T are integers ≥ 0. In addition, we define a cost function as

$$\zeta = \sum_{i=1}^{T} w_i \cdot \Psi_i. \tag{3.17}$$

Our objective is to minimize the cost function ζ. An analysis in [COM, GE16] shows the cost for different WDM TRX capacities in rural, urban and suburban scenarios. However, their analysis is limited up to the 40 Gbps TRX. We adopt their TRX cost data in urban scenarios and calculate the cost of TRX capacities higher than 40 Gbps, with the constraint that the required TRX capacity has lower cost compared to other possible TRX combinations.

Let us consider we have $T = 4$ types TRX with capacities $R_1 = 10$, $R_2 = 20$, $R_3 = 28$, $R_4 = 40$ Gbps and their associated cost factors are $\Psi_1 = 1$, $\Psi_2 = 1.8$, $\Psi_3 = 2.5$ and $\Psi_4 = 3.2$, respectively [COM]. In order to give an example, suppose we need to deploy $\Gamma = 60$ Gbps TRX, then it is more cost efficient to deploy a 40 Gbps TRX and another 20 Gbps TRX (total relative cost factor $3.2 + 1.8 = 5$), instead of six 10 Gbps TRX (total relative cost factor $1 \times 6 = 6$) or three 20 Gbps (total relative cost factor $1.8 \times 3 = 5.4$). Using this optimization approach, we have shown TRX capacity-cost model e.g. up to 120 Gbps TRx in Fig. 3.15.

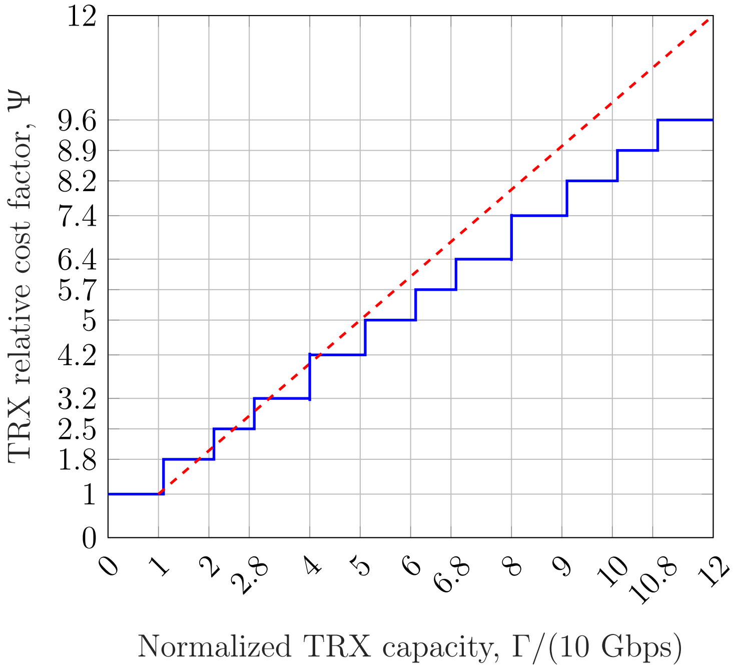


Fig. 3.15. Normalized (w.r.t 10 Gbps) TRX capacity verses relative cost factor, Ψ. The red dashed line is drawn to depict that TRX cost is not increasing in proportion with the increase in the TRX capacity.

Under the assumption that one stream per user can be transmitted as explained in Section 3.2, we need less fronthaul bandwidth. Hence, the required fronthaul bandwidth $D_{\text{req,FH}}$ corresponding to $S_{C,O}$ users in (3.10) can be obtained as [B$^+$17]

$$D_{\text{req,FH}} = S_{C,O} \cdot f_{\text{s}}^{*} \cdot \mu \cdot N_{\text{Q,F}} \cdot 2 \cdot \gamma, \tag{3.18}$$

where $f_{\text{s}}^{*} = N_{\text{RB}} \cdot N_{\text{SC}}^{\text{RB}} \cdot N_{\text{sym}}^{\text{SF}} \cdot T_{\text{SF}}^{-1}$, is the sampling rate, which is the product of the number of resource blocks N_{RB}, number of subcarriers per subframe $N_{\text{SC}}^{\text{RB}}$, number of symbols per subframe $N_{\text{sym}}^{\text{SF}}$ and the subframe duration T_{SF}^{-1}. $N_{\text{Q,F}}$ is the bit resolution of frequency domain quantizer, μ the utilization of subcarriers, i.e., the load and γ the CPRI overhead. The required bandwidth, $D_{\text{req,FH}}$ needs to be fulfilled by suitable cost-efficient TRX equipment, whose capacity is given by (3.16) with minimum cost function, ζ in (3.17).

A lower data rate requirement at the fronthaul, as mentioned in Section 3.6.1, corresponds to the deployment of TRX with less capacity, and hence, allows cost saving. The cost saving in fronthaul can be calculated in terms of relative cost saving ϕ_{rel}, absolute cost saving ϕ_{abs}, and percentage cost saving ϕ_{per}. The relative cost saving ϕ_{rel} is calculated by multiplying the required number of TRX by its relative cost factor Ψ. Next, in order to calculate the absolute cost ϕ_{abs} we first find the relative cost ϕ_{rel} and then multiply the result by the number of cost unit n_{CU} and actual cost of each cost unit P_{CU}. The equations used in calculating these savings are follows:

$$\phi_{\mathrm{rel}} = \left(\left\lceil \frac{C \cdot k_{\max} \cdot f_{\mathrm{s}}^{*} \cdot \mu \cdot N_{\mathrm{Q,F}} \cdot 2 \cdot \gamma}{\Gamma} \right\rceil - \left\lceil \frac{D_{\mathrm{req,FH}}}{\Gamma} \right\rceil \right) \cdot \Psi, \tag{3.19}$$

$$\phi_{\mathrm{abs}} = \phi_{\mathrm{rel}} \cdot n_{\mathrm{CU}} \cdot P_{\mathrm{CU}}, \tag{3.20}$$

$$\phi_{\mathrm{per}} = \left(1 - \frac{\left\lceil \frac{D_{\mathrm{req,FH}}}{\Gamma} \right\rceil}{\left\lceil \frac{C \cdot k_{\max} \cdot f_{\mathrm{s}}^{*} \cdot \mu \cdot N_{\mathrm{Q,F}} \cdot 2 \cdot \gamma}{\Gamma} \right\rceil} \right) \times 100\%, \tag{3.21}$$

where Γ is the TRX capacity in Gbps, Ψ the TRX relative cost factor, n_{CU} the number of cost units, P_{CU} the price per cost unit, $S_{\mathcal{C},O}$ the outage capacity obtained from (3.10), and the other parameters are defined already in the previous sections. The $\lceil\ \rceil$ symbol is used to denote the ceiling function, since the number of required TRXs needs to be rounded up in order to incorporate all users.

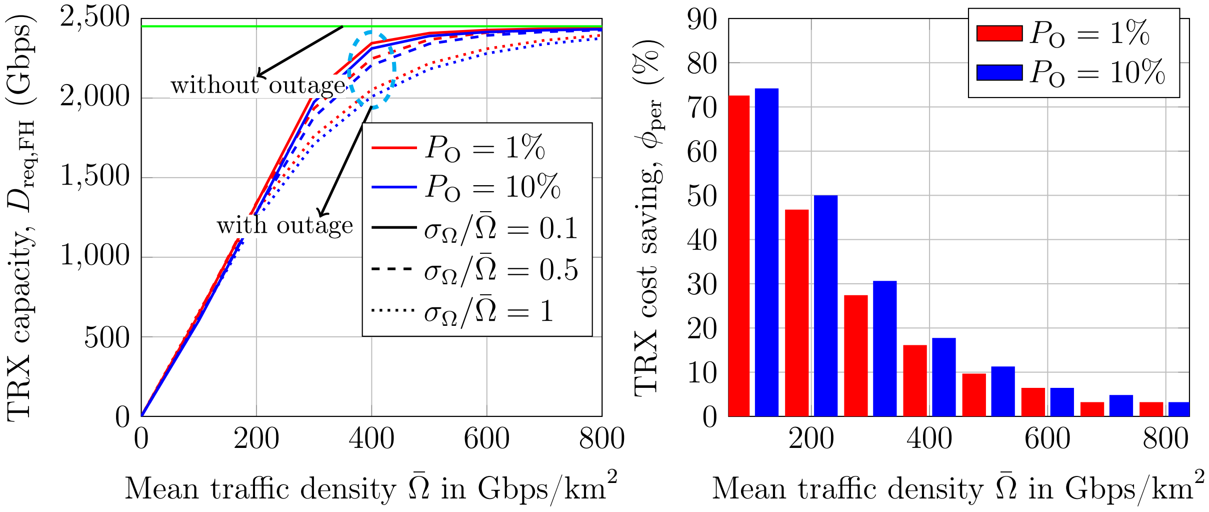
In [COM], it is calculated that $n_{\mathrm{CU}} \approx 20$ and $P_{\mathrm{CU}} \approx 100$ € for WDM-PON system in urban scenario. However, we consider these values only as a reference, since the price of the TRX equipment seems to gradually decrease due to advancement in the technologies. Furthermore, in the result, the cost saving only in terms of percentage ϕ_{per} is shown, as it gives more insights than the relative ϕ_{rel} or absolute ϕ_{abs} cost saving values.

3.6.3 Results

We consider a 5G sub-6 GHz system with 100 MHz bandwidth and sampling frequency $fs = 153.6$ MHz. In order to calculate the required fronthaul data rate, first the traffic map as illustrated in Fig. 3.5 is generated according to Section 3.3.4. Then, 19 hexagonal cells with intersite distance $d_{\mathrm{ISD}} = 200$ m are placed in a square grid of size 1000 m × 1000 m. The RRUs are put at a height of $h_{\mathrm{RRU}} = 12$ m. The parameters used for simulation are listed in Table 3.1.

We get temporal variations from the users using (3.7) and find CDF of the number of users in each individual cell. Next, using (3.10), we calculate the number of users, $S_{\mathcal{C},\mathrm{O}}$ at a given outage probability P_{O}. The required fronthaul capacity $D_{\mathrm{req,FH}}$ corresponding to $S_{\mathcal{C},\mathrm{O}}$ is calculated using (3.18) for different traffic densities. The results are averaged over 25 instances of random traffic maps in order to get more accurate results, and to make more general conclusions for real scenarios.

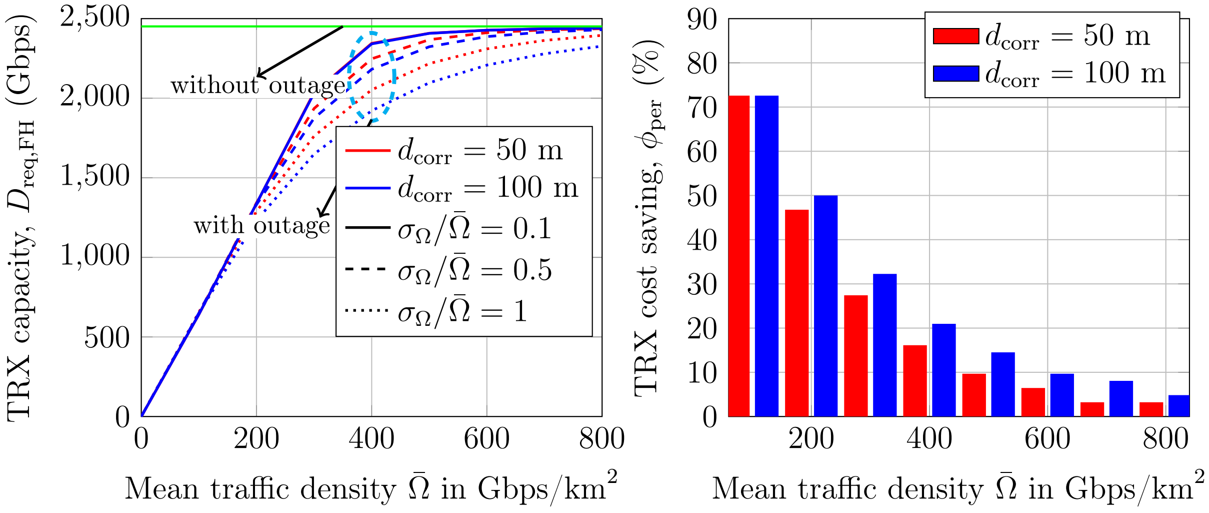
Fig. 3.16a shows the required aggregated capacity in the fronthaul Segment II for all the 19 cells at $P_{\mathrm{O}} = 1\%$ and $P_{\mathrm{O}} = 10\%$ outage probability at peak load for three



(a) fronthaul capacity in Segment II for different relative standard deviations of traffic density, $d_{\mathrm{corr}} = 50$ m.

(b) TRX cost saving at $\sigma_\Omega/\bar{\Omega} = 1$, $d_{\mathrm{corr}} = 50$ m.

Fig. 3.16. Required TRX capacity and TRX cost saving at different outage probabilities.



(a) Fronthaul capacity in Segment II for different relative standard deviations of traffic density, $P_{\mathrm{O}} = 1\%$.

(b) TRX cost saving at $\sigma_\Omega/\bar{\Omega} = 1$, $P_{\mathrm{O}} = 1\%$.

Fig. 3.17. Required TRX capacity and TRX cost saving at different correlation distances.

normalized standard deviations $\sigma_\Omega/\bar{\Omega} = 0.1$, $\sigma_\Omega/\bar{\Omega} = 0.5$ and $\sigma_\Omega/\bar{\Omega} = 1$. As the standard deviation σ_Ω increases, it causes more variability in the total cell traffic and, hence, less capacity is required. The horizontal line (green line) is the total required capacity for all 19 cells at peak load without any outage. Furthermore, it is evident that the gap between the fronthaul capacity requirement without outage (green line) and with outage (blue and red lines) narrows down as the mean traffic density $\bar{\Omega}$ increases. This means that fronthaul needs to be dimensioned close to full fronthaul capacity for higher traffic densities. The difference in the capacities between the horizontal line and any of the curved lines indicates the capacity saving that one can achieve by dimensioning fronthaul appropriately at a given outage probability.

Fig. 3.16b illustrates the percentage TRX cost saving at $P_{\mathrm{O}} = 1\%$ and $P_{\mathrm{O}} = 10\%$ outage probability for $\sigma_{\Omega}/\bar{\Omega} = 1$ at peak load. As seen from this figure, TRX cost saving is much higher at lower traffic, and decreases gradually to a much lower value at medium and high traffics. Furthermore, we see that at a higher outage probability, fewer user streams need to be transmitted, which requires less fronthaul capacity, and, hence, fronthaul cost saving in this case is higher when compared with that with a lower outage probability. It is to be noted that in practical deployment, fronthaul is normally dimensioned for peak traffic. However, such a dimensioning can lead to waste of allocated resources, since full TRX capacity utilization occurs only occasionally. Nevertheless, by allowing for a certain outage probability given by the QoS requirements, we can still achieve a significant cost saving.

Fig. 3.17a shows required capacity for correlation distances, $d_{\mathrm{corr}} = 50$ m and $d_{\mathrm{corr}} = 100$ m for three normalized standard deviations $\sigma_{\Omega}/\bar{\Omega} = 0.1$, $\sigma_{\Omega}/\bar{\Omega} = 0.5$ and $\sigma_{\Omega}/\bar{\Omega} = 1$. As seen from this figure, higher correlation leads to lesser fronthaul capacity demand and vice versa for all the normalized standard deviations. This is because higher correlation distance causes more traffic variations. The gap between curved lines to that of a horizontal line shrinks with the increase of traffic density similar to that in 3.16a. Fig. 3.17b shows percentage cost saving corresponding to Fig. 3.17a at $\sigma_{\Omega}/\bar{\Omega} = 1$ for $P_{\mathrm{O}} = 1\%$. It is clear from this figure that TRX cost reduction is higher at $d_{\mathrm{corr}} = 100$ m compared to that at $d_{\mathrm{corr}} = 50$ m.

3.7 Chapter Summary

The stringent fronthaul-bandwidth constraint in C-RAN can be mitigated by implementing per-user fronthauling. Using the queueing and spatial traffic models, mathematical expressions are derived to analyze the statistical multiplexing gains. We illustrated the impacts of traffic density, correlation distance and outage probability, and showed that the relative fronthaul capacity in the FH Segment II is always lower than that in the FH Segment I. Furthermore, a simple iterative pilot-based optimization algorithm is developed to show the impact of the number of pilots. It is showed that additional reduction in fronthaul bandwidth can be achieved, which leads larger optimization gain up to 25% in the investigated scenarios. Owing to the fronthaul bandwidth reduction, it is desirable to deploy the transceiver optical modules cost efficiently, as the probability of full fronthaul capacity utilization is very low. For this a simple tranceiver cost model in the WDM-PON system is developed and it is shown that fronthaul cost saving up to 50% (refer to Fig. 3.6b) can be achieved for the investigated scenario at a moderately low traffic density of 200 Gbps/km^2, compared with the case when full fronthaul bandwidth utilization is considered. given. Note that although the cost saving analysis is presented for the switched fronthaul technique like WDM-PON, a similar analysis can be conducted for the alternative fronthaul techniques to the switched fronthaul, like e.g., dense wavelength division multiplexing (DWDM) fronthaul or dedicated fiber fronthaul.

Chapter 4

Latency-Constrained C-RAN Fronthaul with Continuous-Time Queuing Model

After studying the bandwidth-constraint aspects on the C-RAN fronthaul and analyzing the statistical multiplexing gains in Chapter 3, our next focus is on the fronthaul latency, which is a critical performance metric, especially for URLLC applications. Radio over Ethernet [IEEb] is being considered for packetized fronthaul network solutions, such as the Ethernet due to its cost-effectiveness and widespread deployment in core networks and data centers. However, packetized transport over fronthaul introduces latency concerns, owing to potential queuing delays at the Ethernet switch. Therefore, this chapter concentrates on the fronthaul latency constraint, and computes and analyzes the fronthaul latency in the UL of a C-RAN system in massive MIMO-based RRUs considering the 3GPP functional split 7 (refer to Section 2.1.2). The tractable, closed-form expressions in terms of the generating functions of the queue length steady-state probabilities, sojourn time and waiting time distributions at the output port of an Ethernet switch in the fronthaul network are derived, and are verified via numerical evaluation. In addition, transport network dimensioning insights in terms of fronthaul latency and PLR are provided.

Introduction

As stated in Chapter 2, prohibitively high fronthaul bandwidth and low latency requirements by CPRI standard in a C-RAN system make it not suitable for massive MIMO-based RRUs. Functional splitting [DDM$^+$13, 3GP17a] is used relax the stringent fronthaul requirements. In this work, 3GPP intra-PHY split is considered, as this split is expected to be suitable for massive MIMO applications [3GP17a]. At this split, precoding in the DL and equalization in the UL are offloaded to the RRU[1]. As a result, the fronthaul bandwidth requirement is lowered, because the required fronthaul bandwidth now scales with the number of spatial streams, unlike in CPRI protocol, where it scales with the number

[1] In fact, there is tradeoff (e.g., in terms of channel capacity, forwarding beamforming weights, CSI estimation, complexity etc.) whether the UL channel estimation and DL channel precoding are performed at RRU compared with the case when they are performed at the BBU. However, their tradeoff investigation (refer to e.g. [SKKS16, PCB13, PWLP15]) is beyond the scope of this thesis work.

of antennas. In addition, the latency requirement is also relaxed, as it is determined by the HARQ process and not by the CPRI protocol.

The latency constraint in the fronthaul originates either from the timing requirement of the HARQ or from use cases, such as Tactile Internet, autonomous driving, AR and/or VR. In the LTE MAC, the HARQ process is co-located[2] with a scheduler and it requires an acknowledgement (ACK) or negative-acknowledgement (NACK) signal must be sent within a pre-defined RTT. Most of the RTT is spent at the BBU and RRU for baseband signal and RF processing. Thus, the latency on the fronthaul must be less than the round trip time minus the time for signal processing. In general, the latency budget left for the fronthaul with the HARQ process located at the BBU is a few hundreds of microseconds, typically 250 μs [3GP17a, SS14b].

Ethernet-based packetized transmission is considered for fronthaul by the next generation fronthaul interface (NGFI) [NGFb] and eCPRI [eCP17] standardization bodies. Despite the aforementioned advantages of Ethernet, providing latency guarantees on the Ethernet-based fronthaul is difficult due to the randomness in latency caused by queuing at the Ethernet switches. Hence, the focus of this work is to develop an analytical model to characterize the delay in the fronthaul for a massive MIMO C-RAN system with functional split 7.

Related Literature

The authors in [PHL17] study delay constraints imposed by CPRI-like traffic in a ring-star topology and based on their results, propose a packetization strategy for fronthaul traffic to reduce average aggregated queuing delay. Work in [PHL17] is extended to [PHL18, OLH19], where authors provide a network planning and dimensioning based on Kingman's exponential law of congestion for G/G/1 queuing model assuming eCPRI-like traffic. They consider a functional split with equalization at the BBU pool, leading to a CBR fronthaul traffic, and present the tradeoff between the fronthaul delay and frame loss rate. However, the works in [PHL17, PHL18, OLH19] require to have aggregation of a large number of fronthaul traffic flows, which cannot be ensured in all scenarios. Unlike our work, their works consider functional split with CBR, present only approximate results, and do not model the air-interface between users and RRUs. Feasibility of Options 2, 6 and 7 have been analyzed in [MRMD17, MCM$^+$17] for URLLC support through experimental testbed using software defined radio (SDR) and open air interface (OAI). The impact of packetization and scheduling policies on latency at the aggregation gateway is experimentally studied in [CNS16]. However, no analytical results are presented in [MRMD17, MCM$^+$17, CNS16]. Work in [SBMD17] analyzes tradeoff between transport network dimensioning and jitter. However, their system models are limited to a single cell and lack a closed-form solution. [KAP$^+$17] develops a general C-RAN queuing model for Poisson arrivals and investigates statistical multiplexing of BBU's computational resources analyzing tradeoff between latency and energy. We note that none of the above

[2] The HARQ timing requirement is very critical if HARQ is located at the BBU, however, the timing requirement is much relaxed if the process is located at the RRU [3GP17a].

contributions considered massive MIMO, which will cause a significant impact on the required bandwidth and latency of the fronthaul segment.

Contributions

Our contribution for this chapter can be briefly summarized as:

1. We consider a practical massive MIMO scenario and calculate the latency in a packetized fronthaul network for intra-PHY split. We model the access link traffic generated by massive MIMO-based RRUs, and map the arrival process at the switch as Poisson arrivals and service process as a hyperexponential (HE) distribution, leading to an **M/HE/1** queuing model;

2. With the help of Pollaczek–Khinchine formula for M/G/1 queue, tractable, closed-form expressions for the MGF of the queue length steady-state probabilities, sojourn time and waiting time distributions at the output port of an Ethernet switch for M/HE/1 queue are derived;

3. The analytical results are verified by means of numerical simulations, not only for exponential file size distribution but also for general file size distribution, such as gamma distributed file size;

4. We show through numerical results that the file size and spectral efficiency of the users are critical in determining the fronthaul latency. In addition, it is shown that speed of the switch can be reduced without causing significant increase in the fronthaul latency, which further reveals the benefits of possible statistical multiplexing;

5. We present the impact of file size, arrival rate, switch speed and spectral efficiency on waiting time, and provide insights for network dimensioning, particularly in terms of PLR in a latency-constrained fronthaul.

4.1 System Model

The traffic from the users is likely to experience some waiting time in the queue at the switch. Hence, we first need to model such traffic from the users. For this purpose, we consider the C-RAN with massive MIMO-based RRUs as a system model.

A schematic diagram of the system model is shown in Fig.4.1. It consists of a massive MIMO access network, an Ethernet-based fronthaul transport network, and a BBU. Further, the fronthaul network consists of two segments: FH Segment I and FH Segment II, and an Ethernet switch. They are described in detail in Section 4.1.2. For such a system, we need to model user arrival traffic. However, before modeling the user traffic, we calculate the SE of each user and, consequently, the number of UL channel resources required to send files for each user.

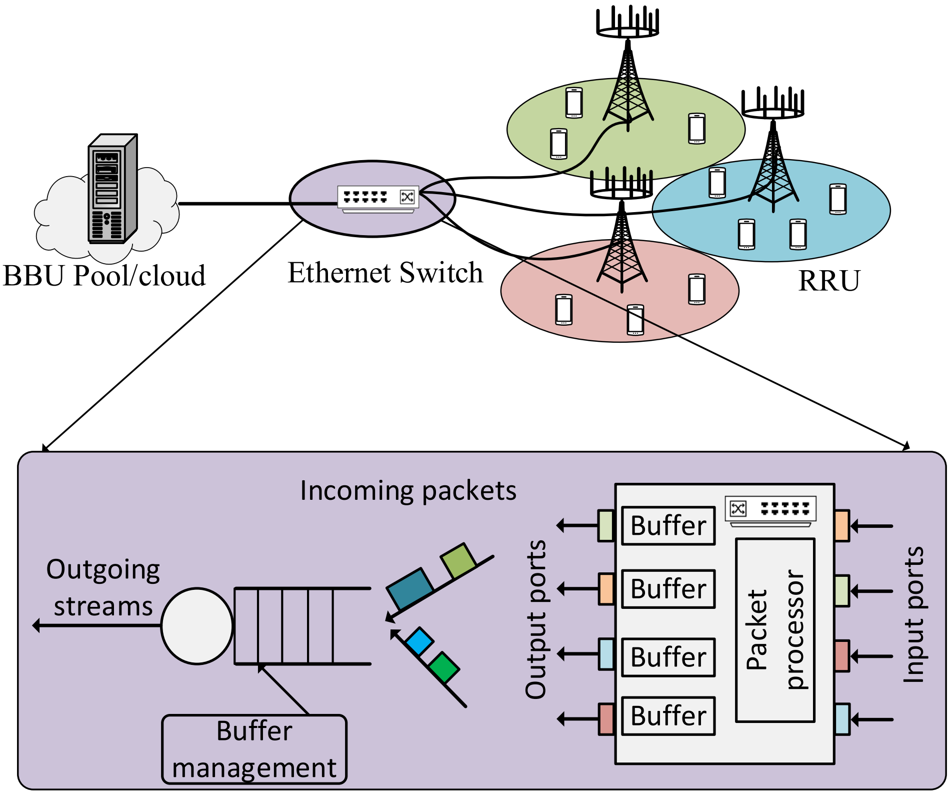


Fig. 4.1. Packetized C-RAN Fronthaul: (a) C-RAN with an Ethernet switch (b) Simplified structure of an Ethernet switch and its output port structure.

4.1.1 Massive MIMO Access Network

We consider the UL of a cellular network that consists of L cells with massive MIMO RRUs. The RRUs are located at the center of the cell and equipped with M antennas that serve K single-antenna users, which are spatially multiplexed onto the same time-frequency resource. We assume that the network operates in time division duplex (TDD) mode such that the RRU obtains the CSI from UL pilots. The RRU exploits them for DL data transmission assuming that the channel is reciprocal. Further, we assume that the channel between the users and RRUs is time-invariant and frequency-flat in a coherence interval of $\tau_{\mathrm{c}} = B_{\mathrm{coh}}\tau_{\mathrm{coh}}$ transmission symbols, where B_{coh} is the coherence bandwidth and τ_{coh} is the coherence time (refer to Fig. 3.2 for UL time-frequency resources).

We describe below the UL training and data transmission. The data transmission in the DL is not discussed as our focus is on the latency analysis in the UL.

Uplink Pilot Training

Let us consider, in a coherence interval τ_{c} transmission symbols, τ_{p} symbols are utilized for UL pilot signaling. Hence, the remaining $\tau_{\mathrm{c}} - \tau_{\mathrm{p}}$ symbols will be used for payload data. To be precise, let $\zeta^{(\mathrm{ul})}(\tau_{\mathrm{c}} - \tau_{\mathrm{p}})$ and $\zeta^{(\mathrm{dl})}(\tau_{\mathrm{c}} - \tau_{\mathrm{p}})$ are proportion of symbols be used for UL data and DL data transmission, respectively, where $\zeta^{(\mathrm{ul})} + \zeta^{(\mathrm{dl})} = 1$ and $1 \leq \tau_{\mathrm{p}} < \tau_{\mathrm{c}}$. Here, $\zeta^{(\mathrm{ul})}$ and $\zeta^{(\mathrm{dl})}$ denote fraction of time-frequency resources allocated for UL and DL data transmission, respectively. We consider the worst case scenario, i.e., we assume full pilot reuse and random pilot assignment to the users. Therefore, pilots are reused in every cell and are assigned randomly to the users in a cell. Full pilot reuse results in pilot contamination. Let $\mathcal{B}_{it}$ denote the set of users that use the same pilot sequence as user t in cell i. We assume that there is no pilot power control and all the users transmit at the maximum power P_{ue}.

Uplink Data Transmission

Let $\mathbf{h}_{it}^{l} \in \mathbb{C}^{M}$ be the channel response between RRU l and user t in cell i, and is modelled as block fading. Hence, $\mathbf{h}_{it}^{l}$ is constant within a block and takes independent realizations across blocks. More precisely, circular symmetric Gaussian realizations $\mathbf{h}_{it}^{l} \sim \mathcal{CN}(\mathbf{0}, \beta_{it}^{l}\mathbf{I}_M)$ is assumed, where β_{it}^{l} is the large-scale fading [BL15]. As in the UL pilot training, there is no power control for UL data transmission and users transmit at a full power P_{ue}. Then, the received signal $\mathbf{y}_l$ at the RRU l is obtained by the superposition of the transmitted signals x_{it} from all the users in the network $\mathcal{S} = \{(it) : i \in \{1, \cdots, L\}, t \in \{1, \cdots, K\}\}$ and is given by

$$\mathbf{y}_l = \sum_{it \in \mathcal{S}} \sqrt{P_{\text{ue}}}\mathbf{h}_{it}^{l} x_{it} + \mathbf{n}_l, \tag{4.1}$$

where $\mathbf{n}_l \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_M)$ is the additive white Gaussian noise (AWGN) and σ^2 denotes its noise variance.

We assume matched filtering at the RRU. That is, $\left(\mathbf{h}_{lk}^{l}\right)^{H} \mathbf{y}_l$ is used to recover signal x_{lk}. Then, the signal-to-interference-plus-noise-ratio (SINR) γ_{lk} can be obtained as [CBL18]:

$$\gamma_{lk} = \frac{\left(\frac{P_{\text{ue}} M (\beta_{lk}^{l})^2}{\sum_{it \in \mathcal{B}_{lk}} \beta_{it}^{l} + \frac{\sigma^2}{\tau_p P_{\text{ue}}}}\right)}{\sigma^2 + P_{\text{ue}} \sum_{it \in \mathcal{S}} \beta_{it}^{l} + P_{\text{ue}} \left(\frac{M \sum_{it \in \mathcal{B}_{lk} \setminus (l,k)} \left(\beta_{it}^{l}\right)^2}{\sum_{it \in \mathcal{B}_{lk}} \beta_{it}^{l} + \frac{\sigma^2}{\tau_p P_{\text{ue}}}}\right)}. \tag{4.2}$$

The numerator in (4.2) is the received signal power. The first, second, and third terms in the denominator can be identified as noise power, multiuser interference, and interference due to pilot contamination, respectively. Note that the pilot contamination term persists even if the number of antennas M grows to infinity. This shows that pilot contamination becomes the limiting factor when the number of antennas is large. The UL SE, R_{lk} (in bits/s/Hz) of transmission is then given by

$$R_{lk} = \zeta^{(\text{ul})} \left(\tau_{\text{c}} - \tau_{\text{p}}\right) \log_2 \left(1 + \gamma_{lk}\right). \tag{4.3}$$

We note that the expression in (4.2) has not accounted for the dynamic interference resulting from dynamic user traffic. This simplification has been made to have the latency analysis tractable, as it avoids the coupling between the arrival and service processes of different users. Note also that the above spectral efficiency expression is a lower bound on spectral efficiency with dynamic traffic. The spectral efficiency of a user determines the number of I/Q symbols needed to send its file. Each received I/Q symbol after equalization is quantized to N_q bits at the RRU. The quantized bits corresponding to a user file are encapsulated in an Ethernet packet and are sent over the fronthaul network.

4.1.2 User Traffic Model and Ethernet-based Fronthaul Network

Our focus is now to describe the dynamics of the UL data traffic from different users. Let λ_{lk} denotes the arrival rate for user k in cell l. The file arrival process from a user is a Poisson point process. The file size $F_{lk} \in [0, \infty)$ for user k in cell l is a random variable and we assume, for simplicity, that it is exponentially distributed with mean $\mathbb{E}[F_{lk}] = \overline{F}$. Hence, its probability distribution function (PDF) is $f_{F_{lk}}(F_{lk}, \overline{F}) = (1/\overline{F})\ \exp(-F_{lk}/\overline{F})$. The extension to the case of general file size distribution is discussed in Section 4.2.3.

As stated earlier, the fronthaul network in Fig. 4.1 (a) consists of two segments: FH Segment I and FH Segment II, and an Ethernet switch. FH Segment I connects the RRUs to the input ports of the switch and FH Segment II connects the output port of the switch to the BBU pool. A schematic diagram of a switch is shown in Fig. 4.1 (b). The switch consists of input-output ports, packet processor and buffer elements. The switch is configured as a multiplexer, and the traffic from the users in the access network is multiplexed at the switch and forwarded to the BBU for further processing. The switch reads and processes the source and destination MAC addresses of the Ethernet frame. The packet processor looks at the destination address of the packets and routes them to the appropriate output ports. Thus, the packet is queued at the switch before it is transmitted. We assume that the switch speed is matched to the capacity C_{FH} of the FH segment II. Hence, the packet is pushed out of the queue as fast as possible. Further, we assume that buffer space at the switch is sufficiently large so that packet dropping at the switch is ignored.

4.2 Queuing Theoretic Modeling and Steady-state Analysis

In this section, we model the queue at the switch, which requires us to have information about the arrival and service processes. In addition, we have to ensure that the stability condition of the switch is fulfilled. Later, we derive the closed-form expressions for MGF of queue length steady-state probabilities, sojourn time and waiting time distributions.

4.2.1 Queue Model

Arrival Process
Recall that the I/Q streams of the users are recovered after equalization at the RRU. Since the users' I/Q streams are generated from their Poisson file arrival process, the I/Q streams at the RRU for each user are also Poisson processes. Slotted nature of UL transmission in the access network is ignored in this work, since the symbol duration is small. For example, in LTE, it is $66.7\,\mu$s (without cyclic prefix), which is an order of magnitude lower than the time scale of interest (ms). In practice, slotted nature of the UL transmission cannot be ignored. Hence, a discrete-time queuing model will be considered and explained in Chapter 5. The aggregate arrival process from an RRU to the switch is also Poisson, as it is the sum of K independent Poisson processes. That is, the arrival

process from RRU l is Poisson with arrival rate $\sum_{k=1}^{K} \lambda_{lk}$. Then, the overall arrival process at the queue is the sum of independent Poisson arrival processes from different RRUs. It is a Poisson process with sum arrival rate Λ as

$$\Lambda = \sum_{l=1}^{L} \sum_{k=1}^{K} \lambda_{lk} \tag{4.4}$$

Service Process

The process time of a file depends on its size. Since file sizes are independent across different arrivals and users, service processes are i.i.d.. The marginal service time distribution is computed as follows. For user k in cell l with SE R_{lk}, the number of time-frequency REs needed to send a file F_{lk} in the UL is F_{lk}/R_{lk}. This is also the number of received I/Q samples from user k after equalization at RRU l. At the RRU, each I/Q symbol is quantized to $2N_q$ bits before being sent over the FH[3]. Then, the number of quantized fronthaul bit streams corresponding to file F_{lk} are $N_{\text{bitsteams},lk} = 2N_q F_{lk}/R_{lk}$. These bits are from the packet. The time required by the switch to forward this packet is the number of bits divided by the switch speed, as the switch is operating at a constant speed. Hence, the service time required to process the packet corresponding to file F_{lk} can be obtained by

$$S_{lk} = \frac{N_{\text{bitsteams},lk}}{C_{\text{FH}}} = \frac{2N_q F_{lk}}{R_{lk} C_{\text{FH}}},$$

where C_{FH} is the constant operating speed of the switch. As F_{lk} is exponentially distributed with mean $\overline{F}$, the service time is also exponentially distributed but with mean $\mu_{lk} = \mathbb{E}[S_{lk}] = 2N_q \overline{F}/(R_{lk} C_{\text{FH}})$. Therefore, the PDF of S_{lk} is $f_{S_{lk}}(S_{lk}, \mu_{lk}) = (1/\mu_{lk}) \exp(-S_{lk}/\mu_{lk})$. Note that the mean of the service time distribution is different for different users and depends on their SEs.

Any random packet arriving at the switch could be from any one of the LK users in the network. Since the arrival process at the switch is the superposition of LK independent Poisson processes, as discussed above, a packet arriving at the switch is from user k in cell l with probability (w.p.) $p_{lk} = \lambda_{lk}/\Lambda$ [SD07]. Hence, the service time RV S is given by

$$S = \begin{cases} S_{11}, & \text{w. p.} \quad p_{11}, \\ \vdots & \\ S_{LK}, & \text{w. p.} \quad p_{LK}. \end{cases} \tag{4.5}$$

The RV S has a mixture distribution with probability density function (PDF) $f_S(S = x)$ given by

$$f_S(x) = \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} f_{S_{lk}}(S_{lk}), \tag{4.6}$$

where $f_{S_{lk}}(S_{lk})$ is the PDF of S_{lk}. Because S_{lk} is exponentially distributed, the distribution

[3] A factor of 2 because both I- and Q-symbols are quantized to N_q bits.

of S is known as hyperexponential distribution [SD07]. The mean service time is given by

$$\mathbb{E}[S] = (2N_q\overline{F}/C_{\mathrm{FH}}) \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk}/R_{lk}. \tag{4.7}$$

Queue Model

From the above discussion, it follows that the queue at the switch has Poisson arrivals, hyperexponential service time distribution. Further, we assume the first come first serve (FCFS) principle and an infinite buffer. Therefore, as per Kendall's notation, the queue is represented as M/HE/1.

4.2.2 Steady-state Analysis

We now use the results available in the literature for M/G/1 queue to obtain closed-form expressions for MGF of the steady-state queue length and sojourn time distributions of an M/HE/1 queue [Vir]. These results follow from the embedded Markov chain at instances when a packet leaves the queue.

Stability of Queue

The stability of the queue requires that the load ρ, which is defined as the product of arrival rate and average service time, is less than 1. That is, $\Lambda\mathbb{E}[S] < 1$. Substituting for $\mathbb{E}[S]$ of the hyperexponential distribution, the criterion for queue stability is

$$\rho = \frac{2N_q\overline{F}}{C_{\mathrm{FH}}} \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{\lambda_{lk}}{R_{lk}} < 1. \tag{4.8}$$

This equation brings out how the different network parameters affect the stability of the queue.

Steady-state Queue Length Probabilities

Let π_i denotes the steady state probability of the queue length being equal to i, for $i = 0, 1, \ldots \infty$. As shown in [Vir], these steady-state probabilities satisfy the following recursion:

$$\pi_i = \frac{1}{k_0} \left(a_{i-1}\pi_0 + \sum_{j=1}^{i-1} a_{i-j}\pi_j \right). \tag{4.9}$$

The recursion begins with $\pi_0 = 1 - \rho$. For $i = 0, 1, \ldots, \infty$, k_i denotes the probability of i arrivals in the service time of a packet. It is given by

$$k_i = \int_0^{\infty} \frac{(\lambda x)^i}{i!} \exp(-\lambda x) f_S(x) dx. \tag{4.10}$$

When S has hyperexponential distribution, evaluating k_i yields (refer to A.1 for the derivation)

$$k_i = \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(\frac{\Lambda}{\Lambda + {\mu_{lk}}^{-1}} \right)^i \left(\frac{{\mu_{lk}}^{-1}}{\Lambda + {\mu_{lk}}^{-1}} \right), \tag{4.11}$$

where $\mu_{lk} = 2N_q\overline{F}/(R_{lk}C_{\text{FH}})$.

Sojourn Time Distribution

Let T and W denote the sojourn time and waiting time, respectively. S is the service time, defined previously. Then sojourn time, $T = W + S$, is the time spent in the switch. Assuming that the waiting time and the service time are independent, the PDF of the sojourn time is obtained by convolving the PDF of the waiting time with the PDF of the service time as $f_T(x) = f_W(x) * f_S(x)$. In order to compute sojourn time distribution, we employ the Pollaczek–Khinchine formula [Vir] for M/G/1 queue and derive the relation to M/HE/1 queue model. The Pollaczek–Khinchine formula expresses the MGF of the sojourn time RV T in terms of the MGF of the service time RV S.

For random variable (RV) X, the MGF is $\Psi_X(s) = \mathbb{E}[\exp(-sX)]$. MGF of a RV X is in fact the Laplace transform of its PDF. Hence, taking the Laplace transform of $f_T(x)$, we get $\Psi_T(s) = \Psi_W(s) \cdot \Psi_S(s)$, where $\Psi_T(s)$, $\Psi_W(s)$ and $\Psi_S(s)$ denote MGF of sojourn time, waiting time and service time, respectively. Employing the Pollaczek–Khinchine formula to M/HE/1 queue model, the sojourn time MGF can be obtained as [Vir]

$$\Psi_T(s) = \frac{s\,(1-\rho)\,\Psi_S(s)}{s - \Lambda + \Lambda\Psi_S(s)}. \tag{4.12}$$

Now our aim is to find $\Psi_S(s)$, which can be obtained by taking the Laplace transform of $f_S(x)$ as

$$\begin{aligned}\Psi_S(s) = \mathcal{L}\left\{f_S(x)\right\} &= \int\limits_{-\infty}^{+\infty} \exp(-sx)\left\{f_S(x)\right\}dx \\ &= \int\limits_{0}^{+\infty} \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\mu_{lk}^{-1}\,\exp\left(-(s+\mu_{lk}^{-1})x\right)dx \\ \Rightarrow \Psi_S(s) &= \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{{\mu_{lk}}^{-1}}{s+{\mu_{lk}}^{-1}}\right).\end{aligned} \tag{4.13}$$

Substituting $\Psi_S(s)$, Λ and ρ in (4.12), we get the final expression of the MGF of the sojourn time as

$$\Psi_T(s) = \frac{s\left(1 - \frac{2N_q\overline{F}}{C_{\text{FH}}}\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\lambda_{lk}}{R_{lk}}\right)}{s - \sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_{lk} + \sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_{lk}\left(\frac{{\mu_{lk}}^{-1}}{s+{\mu_{lk}}^{-1}}\right)}\left(\sum_{l=1}^{L}\sum_{k=1}^{K}p_{lk}\left(\frac{{\mu_{lk}}^{-1}}{s+{\mu_{lk}}^{-1}}\right)\right). \tag{4.14}$$

The PDF of the sojourn time can be obtained by taking the inverse Laplace transform of $\Psi_T(s)$. However, as we could not evaluate it in closed-form and numerical techniques are used to evaluate the inverse Laplace transform.

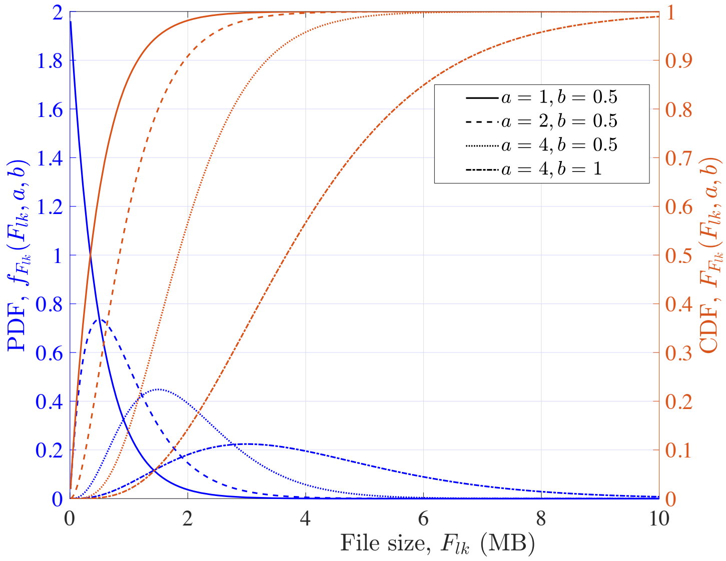


Fig. 4.2. PDF and CDF of gamma distributed file size, F_{lk}.

4.2.3 Extension to General File Size Distribution

The above analysis is not specific to the exponential file size distribution. Hence, it can be extended to any general file size distribution. Accordingly, we will have different expressions for $f_S(x)$, k_i and $\Psi_S(s)$ from (4.6), (4.11) and (4.13), respectively.

In order to show the generality, we consider a more general case when the file size is gamma distributed as shown in Fig. 4.2. Therefore, $F_{lk} \sim \Gamma(a,b)$ for $F_{lk} > 0$ and $a, b > 0$, where a and b are the shape and scale parameters, respectively, and $\Gamma(a)$ is a gamma function. Then, its mean is $\mathbb{E}[F_{lk}] = \overline{F} = ab$ and the PDF is $f_{F_{lk}}(F_{lk}, a, b) = {F_{lk}}^{a-1} e^{-F_{lk}/b} / (b^a \Gamma(a))$. It is to be noted that the exponential distribution, Erlang distribution and Chi-squared distribution are special cases of the gamma distribution. Following the discussion in Section 5.2.1, the service time S_{lk} of user lk, $S_{lk} = 2N_q F_{lk}/(R_{lk} C_{\mathrm{FH}})$ is also gamma distributed, i.e., $S_{lk} \sim \Gamma(a, c_{lk})$, where $c_{lk} = 2N_q b/(R_{lk} C_{\mathrm{FH}})$. Hence, the mean of S_{lk} is $\mu_{lk} = \mathbb{E}[S_{lk}] = 2N_q \overline{F}/(R_{lk} C_{\mathrm{FH}})$. The PDF of S_{lk} is given by $f_{S_{lk}}(S_{lk}, a, c_{lk}) = {S_{lk}}^{a-1} e^{-S_{lk}/c_{lk}} / (c_{lk}^a \Gamma(a))$. Using this and (4.6), the PDF of the service time RV S can be obtained. Since $\mathbb{E}[S_{lk}]$ is same as before, $\mathbb{E}[S]$ and the queue stability condition in (4.8) are also the same. Substituting the $f_S(x)$ in (4.10), k_i can be evaluated as(refer to A.2 for the derivation)

$$k_i = \frac{\Gamma(a+1)}{i!\Gamma(a)} \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(\frac{\Lambda}{\Lambda + {c_{lk}}^{-1}} \right)^i \left(\frac{{c_{lk}}^{-1}}{\Lambda + {c_{lk}}^{-1}} \right)^a . \tag{4.15}$$

Further, the MGF of the RV S is given by (refer to A.3 for derivation)

$$\Psi_S(s) = \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(\frac{{c_{lk}}^{-1}}{s + {c_{lk}}^{-1}} \right)^a . \tag{4.16}$$

Using these new expressions for k_i and $\Psi_S(s)$, the steady-state queue length probabilities and sojourn time distribution can be evaluated as in Section 4.2.2. It is to be noted that the gamma distribution becomes an exponential distribution if $a = 1$.

Waiting Time Distribution

Similar to sojourn time, the waiting time can be computed employing Pollaczek–Khinchine formula [Vir] as

$$\Psi_W(s) = \frac{s\left(1 - \frac{2N_q\overline{F}}{C_{\text{FH}}}\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\lambda_{lk}}{R_{lk}}\right)}{s - \sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_{lk} + \sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_{lk}\left(\frac{{\mu_{lk}}^{-1}}{s+{\mu_{lk}}^{-1}}\right)}. \tag{4.17}$$

We take the inverse Laplace transform of (4.17) to get the distribution of the waiting time, which we later evaluate against the simulation results.

4.2.4 Packet Loss Rate

In latency critical applications, the transmitted packets must reach the destination within a certain time defined by the network or use case. Packets exceeding the allowed time result in packet drops. The packet loss rate accounts for packet loss due to the reason that packets are either erroneous, lost or arriving too late. Here, we define the packet loss rate (PLR) as

$$\text{PLR} = \mathbb{P}(W > T_{\text{FH}}), \tag{4.18}$$

where, T_{FH} is the FH latency threshold obtained using (2.3).

4.3 Numerical Results

4.3.1 Access Link Throughput

We consider a C-RAN system with massive MIMO RRUs employing $M = 300$ antennas in each cell. The cellular layout is 7-cell hexagonal with wrap-around implementation. We drop $K = 10$ users in each cell such that no user lies at a distance of $d_{\text{min}} \leq 35$ m from the center of the cell. The pilot and data transmission powers are set to $P_{\text{ue}} = 23$ dBm. The remaining simulation parameters are listed in Table 4.1. Using the 3GPP LTE model [3GP10], we compute the large-scale fading coefficient β_{it}^{l} in dB as

$$\beta_{it}^{l} = -148.1 - 37.6\log_{10}(d_{it}^{l}) + X_{\sigma,it}^{l} \text{ dB}, \tag{4.19}$$

where d_{it}^{l} is the distance in km between the user t in cell i and the RRU l, and $X_{\sigma,it}^{l}$ describes lognormal shadowing with zero mean and $\sigma = 7$ dB standard deviation.

Fig. 4.3 shows that the cumulative distribution function (CDF) plot of SE (in b/s/Hz) for three values of M with $K = 10$ and $\tau_{\text{p}} = 10$. The plot shows that the SE increases with increase in M. User locations are averaged over 10,000 realizations. Some users, especially at the cell-edges, might experience extremely low data rates, which can occur due to bad channel conditions or due to severe multiuser interference and pilot contamination. In general, users having lower SE will require more resources. However, since the practical

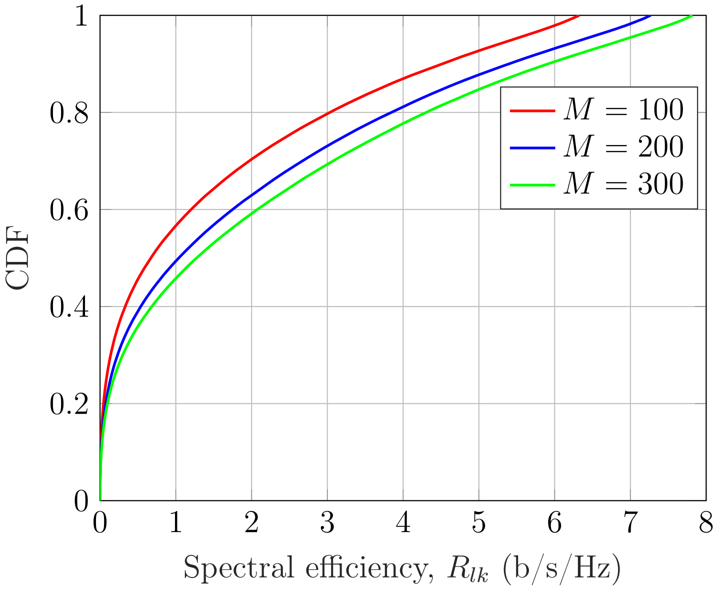


Fig. 4.3. CDF plot of spectral efficiency, $K = 10$, $\tau_p = 10$ [CFBF19a].

communication systems will have a limited number of resources, we assume a minimum bitrate of 5 Mbps for each user in order to guarantee that the load is less than one, thereby ensuring the stability of the queue. This choice is justified as more than 75% of the users had SE higher than this value for all user drops.

4.3.2 Sojourn Time and Queue Length

In order to compare the simulation results with the analytical solutions, we take the inverse Laplace transform of (4.12) using a built-in MATLAB function. We compare the results with varying file sizes and different arrival rates. Fig. 4.4 shows simulation and analytical results of sojourn time distribution for $\lambda = 1$ and $\lambda = 5$. As we see, both the simulation and analytical results match quite well. However, we have some mismatch around zero. This occurs possibly due to MATLAB's lack of precision in handling the inverse Laplace transform at the vicinity of zero. The impact of this mismatch is noticeable at very low latency only. But, for the interest in large delays, the impact of the mismatch is negligible. Further, notice that the higher value of arrival rates, stretches the curve reducing the PDF peaks. Hence, latency increases with the higher values of arrival rates.

Fig. 4.5 plots the queue length distribution. As in the previous case, the analytical result follows the simulation result. More than half of the time, the queue length is zero and in the remaining time it lies between 1 and 5. The queue length probabilities decay quicker as C_{FH} increases because the packets will be processed quickly. Moreover, the queue length probabilities increase with the higher values of the arrival rates and larger file sizes for a given switch speed. Higher values of arrival rates will increase the queue lengths at the switch, and larger file sizes demand more resources, thus increasing the required time.

Next, in order to illustrate that the presented model works for any general file size

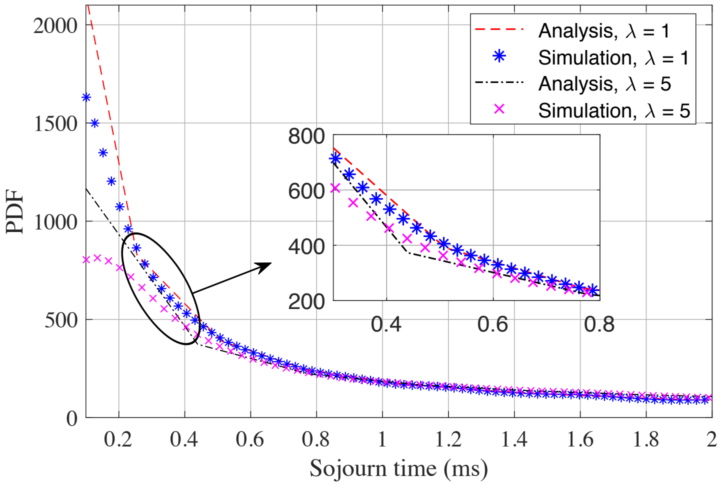


Fig. 4.4. Sojourn time distribution, $\overline{F} = 0.5$ MB, $C_{\text{FH}} = 100$ Gbps [CFBF19a].

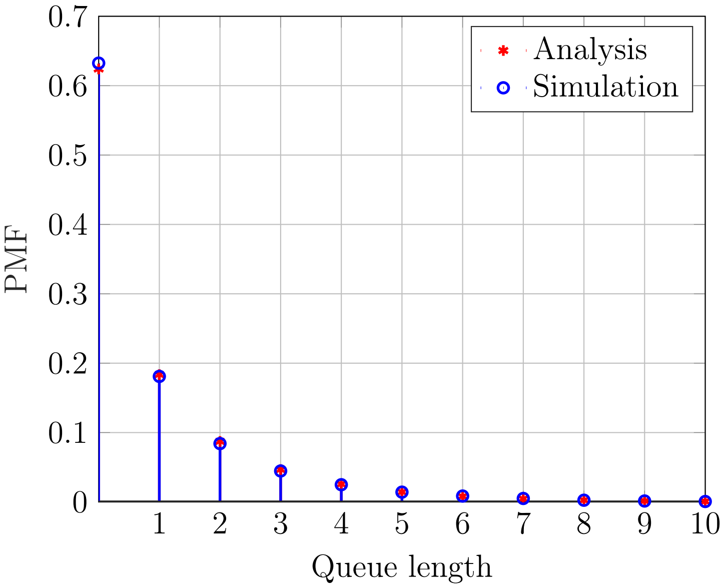


Fig. 4.5. Queue length distribution, $\lambda = 5, \overline{F} = 0.5$ MB, $C_{\text{FH}} = 100$ Gbps [CFBF19a].

distribution, we consider that the file size is gamma distributed with the values of a and b fulfilling the stability condition in (4.8). Fig. 4.6 shows the results for two values of scale parameter, $a = 2$ and $a = 3$ for fixed b. Depending upon different values of the shape parameter a[4], the shape of the distribution will have different forms for given b. This is especially apparent in comparison to Fig. 4.4, where a is set to 1. Contrary to a, which changes shape of the distribution, the scale parameter b for a given a has the effect of stretching or shrinking the distribution shape. Fig. 4.7 shows the results for different values of b while keeping a fixed. Notice that the peak value of the distribution curve in Fig. 4.7 decreases when the value of b increases. As illustrated in Figs. 4.6 and 4.7, both results also match each other when the file size has general distribution.

[4] The computation time of the sojourn time increases with a because the second term of the MGF of S in (4.16) is raised to exponent a.

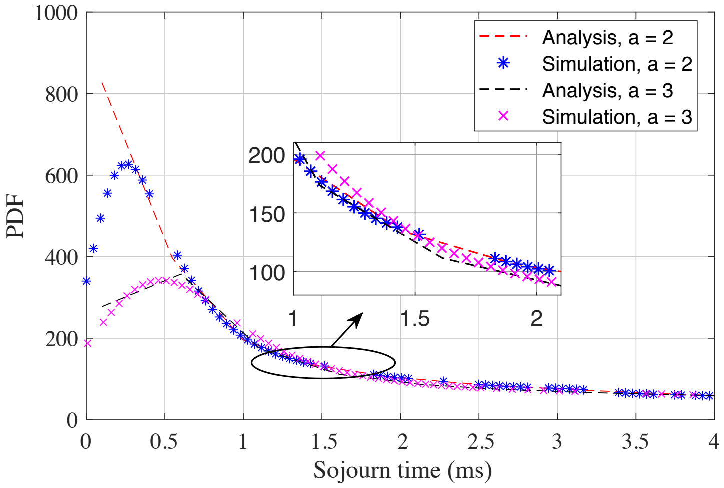


Fig. 4.6. Sojourn time distribution for gamma distributed file size, $\lambda = 1, b = 0.5$ MB, $C_{\text{FH}} = 100$ Gbps [CFBF19a].

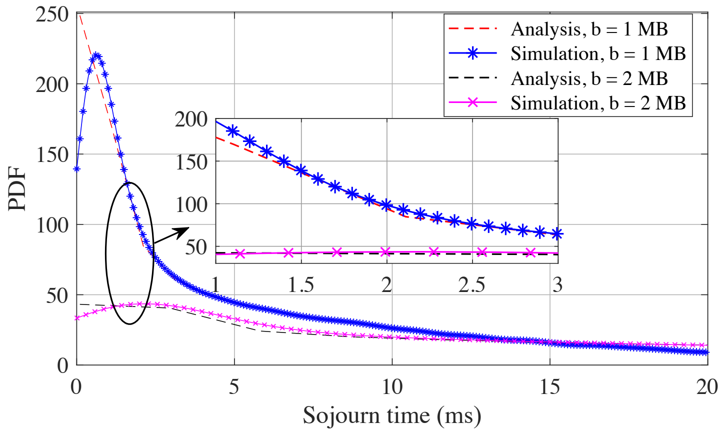


Fig. 4.7. Sojourn time distribution for gamma distributed file size, $\lambda = 1$, $a = 2$, $C_{\text{FH}} = 100$ Gbps [CFBF19a].

4.3.3 Packet Size Impact

Now, we are interested to know the lowest achievable latency for different packet sizes at different percentiles. Depending upon the use cases and application, the fronthaul will have its own latency threshold T_{FH}. This value can be as low as some hundreds of μs, typically it is assumed to be 250 μs. In order to guarantee such a low FH latency requirement, we assume the file size is small such that it contains only a single packet. According to [GPR16], we assume a packet size of 500 B for URLLC and 1500 B for eMBB. Fig. 4.8 illustrates the 99th, 90th, 50th percentiles of the simulated sojourn time for 500 B and 1500 B packet sizes. The following observations can be made from Fig. 4.8. First, the sojourn time increases significantly with the increase in packet size as it requires more resources to process it. Second, with the faster switch speed, sojourn time decreases. For the slower switch speed, sojourn time grows abruptly, and given a $T_{\text{FH}} = 250$ μs latency budget cannot be guaranteed. Hence, in order to meet the URLLC performance metric, one needs to have smaller packet sizes and the switch needs to operate at reasonably higher speeds.

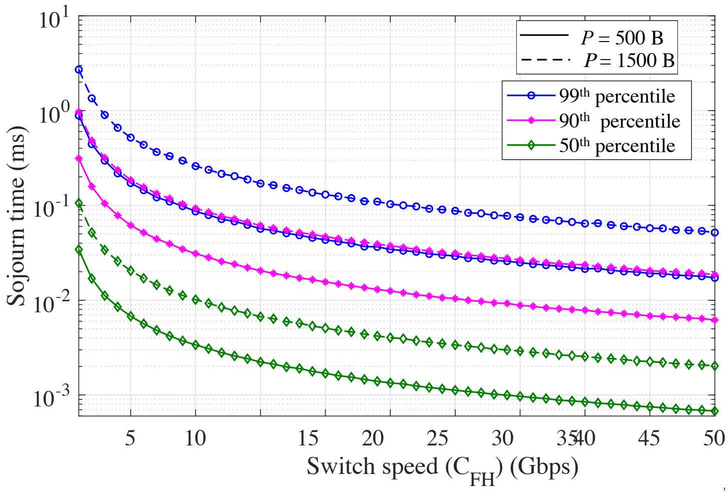


Fig. 4.8. %tile of sojourn time for 500 B and 1500 B packet sizes, $\lambda = 1$ [CFBF19a].

Third, the switch speed can be decreased without increasing the FH latency significantly, which means we can benefit from a statistical multiplexing gain as well.

4.3.4 Waiting Time

The literatures on latency analysis of C-RAN [LBC18, LCC19] have either ignored the waiting time at the switch or assumed, for simplicity, some deterministic value for delays at the switch. However, waiting time will play a significant role, particularly for heavily loaded system such as massive MIMO RRUs, where the switch has several arrivals from different users in the network with varying requirements. Hence, it is important to model and analyze the effects of waiting time in real scenarios. In this work, we attempt to do exactly that. We extended the model presented in [CFBF19a] for waiting time calculation and thereby compute the waiting time distribution at the switch for random packet arrivals from the users in the network considering massive MIMO RRUs.

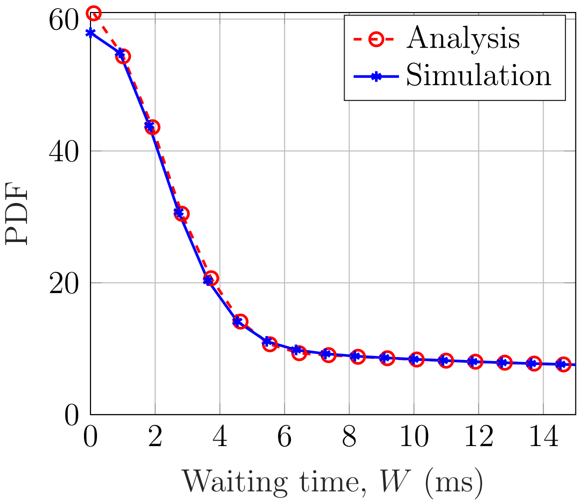


Fig. 4.9. Waiting time distribution, $\overline{F} = 0.5$ MB, $\lambda = 1$, $C_{\text{FH}} = 10$ Gbps [CFBF19b].

Fig. 4.9 plots the waiting distribution for simulated and analytical results. We consider an average file size, $\overline{F} = 0.5$, mean arrival rate, $\lambda = 1$ and switch speed of 10 Gbps. We observe that both the simulated and analytical results match well. A slight deviation of the analytical solution occurs in the vicinity of zero, as in sojourn time distribution, which is probably due to MATLAB's precision for inverse Laplace transform at the vicinity of zero. Waiting time is impacted mainly by the file size and switch speed. A bigger file size takes more resources and hence, more time to process for a given switch speed. On the other hand, even a bigger file size could be processed much quicker if the switch is operating at faster speeds. Note that frequency of arrival of a big file size will be less compared to smaller file sizes.

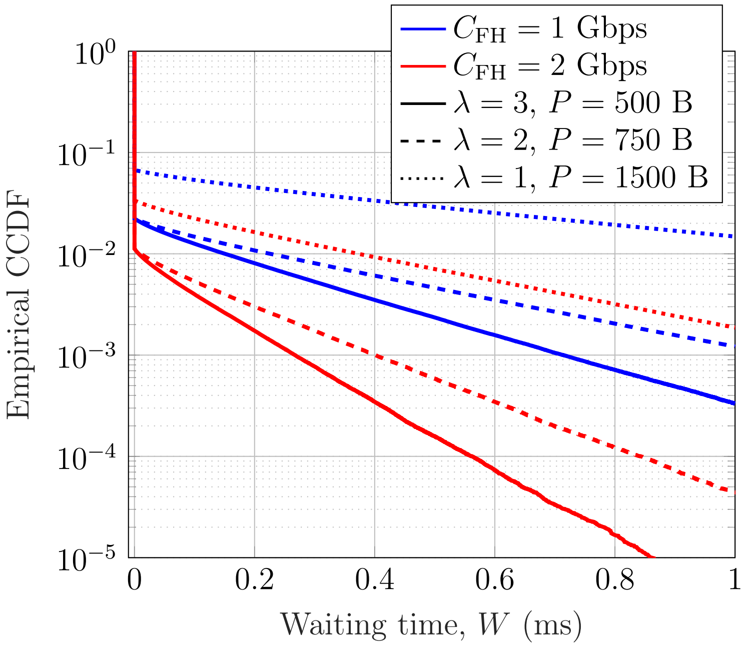


Fig. 4.10. CCDF plot of waiting time, W [CFBF19b].

Fig. 4.10 plots the simulation results for the empirical complementary cumulative distribution function (CCDF) of the waiting time. Practically, the waiting time at the switch will be much smaller. Hence, we compare three smaller but fixed packet sizes (P) of 500 B, 750 B and 1500 B with corresponding mean arrival rates $\lambda = 3$, $\lambda = 2$ and $\lambda = 1$, while keeping the load at the switch constant.

4.3.5 Packet Loss Rate

Fig. 4.11 plots simulated PLR for different switch speeds for given packet sizes. As an example, the probability of a waiting time of 0.25 ms when the switch operates at 2 Gbps is 2%, 0.2% and 0.1% for $P = 1500$ B, $P = 750$ B and $P = 500$ B packet sizes, respectively. This shows that PLR increases if larger packet sizes are used. Their corresponding PLRs are much lower if the switch operates at faster speeds. For a fixed packet size, we can also infer that PLR increases with the higher values of mean arrival rates. This occurs because higher values of arrival rates increases the waiting time at the switch for a given switch

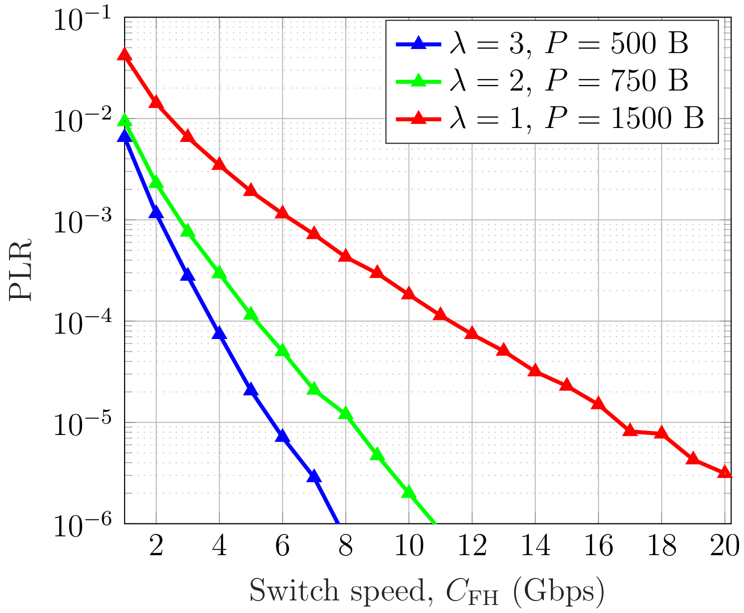


Fig. 4.11. Packet loss rate for varying switch speeds for different packet sizes and arrival rates [CFBF19b].

Tab. 4.1. Simulation parameters for UL latency analysis in C-RAN fronthaul.

Parameters	**Symbol**	**Value**
Intersite distance (ISD)	d_{ISD}	500 m
Number of pilots	τ_{p}	10
Channel bandwidth	B	20 MHz
Coherence interval	τ_{c}	200 symbols
Noise figure	$\mathcal{F}$	5 dB
Noise power	σ^2	-101 dBm
Average file size	$\overline{F}$	0.5 MB
Quantizer resolution	N_{q}	8 bit

speed. Generally, the Ethernet switch are over provisioned to operate at faster switch speeds compared to the incoming traffic from the RRUs such that PLR is extremely low or no PLR. This is because packets are processed much quicker and hence, their waiting times in the queue are much smaller.

4.4 Chapter Summary

In this work, an analytical framework to calculate the latency in the UL of C-RAN massive MIMO system with intra-PHY functional split is presented. We considered both the access and FH networks in the analysis. We showed that the output port of an Ethernet switch can be modelled as an **M/HE/1** queue when the file arrival process is Poisson and the file sizes are exponentially distributed. This allowed us to derive the tractable, closed-form expressions for MGF of the sojourn time, waiting time and queue length distributions. The simulation results corroborated the correctness of our analytical results. We showed that the analysis presented in this paper applies to general file size distribution and we illustrated this by presenting the results for the gamma distribution. Our analysis also revealed the impact of different parameters such as average file size, arrival rate for the users, spectral efficiency of the users, and switch speed on the fronthaul latency. It is shown that the average file size, arrival rate, and spectral efficiency played a critical role. Furthermore, it is observed that the switch speed can be reduced without incurring a notable increase in the fronthaul latency, which enables to exploit the benefits of statistical multiplexing. In addition, the inability of the transmitted packets to reach the destination within a certain time causes packet loss, which we presented in our results in terms of PLR.

Chapter 5

Latency-Constrained C-RAN Fronthaul with Discrete-Time Queuing Model

In the packetized C-RAN fronthaul, random packet delays due to queuing at switching/aggregation gateways occur and it is necessary to characterize distribution of queuing delays. The framework developed in Chapter 4 analyzed the UL latency in massive MIMO-based C-RAN system and derived the closed-form expressions for the MGF of queuing delays and queue length distributions. However, the slotted nature of the UL transmissions from users to RRUs is not modeled. However, in reality, this assumption is not practical as it assumes the UL transmissions can occur at any arbitrary time. This leads to a continuous-time queuing model, unlike in this chapter, which is a discrete-time queuing model. Hence, in this chapter, the earlier work is extended in order to account for slotted transmission and a novel discrete-time queuing model for an RRU gateway in the fronthaul network is presented. Further, the closed-form expressions for the generating functions of steady-state queue length and sojourn time probability mass functions (PMFs) are derived, and the analytical results are verified via numerical simulations. Moreover, the proposed model is then used to study the probability of an outage, which occurs when the sojourn time exceeds a delay budget.

5.1 System Model

The schematic diagram of the system model is shown in Fig. 5.1, which consists of L cells, K single-antenna users in each cell, a two-hop Ethernet-based FH with an RRU gateway, and a BBU pool. Each cell is equipped with M antennas and is located at the cell center. It is assumed that $M >> K$, thus, the cells consist of massive MIMO-based RRUs.

Uplink Transmission Model

It is assumed that the network operates in the TDD mode. The UL/DL transmissions to the K users in a cell are spatially multiplexed onto the same time-frequency resource, referred to as RE as in long-term evolution (LTE). Let $\mathbf{h}_{ik,l} \in \mathbb{C}^M$, for $k = 1, \dots, K$ and $i = 1, \dots, L$, denotes the complex baseband channel gain vector from user k in cell i to RRU l. We assume spatially uncorrelated Rayleigh fading [BL15], i.e., $\mathbf{h}_{ik,l} \sim \mathcal{CN}(\mathbf{0}, \beta_{ik,l}\mathbf{I}_M)$,

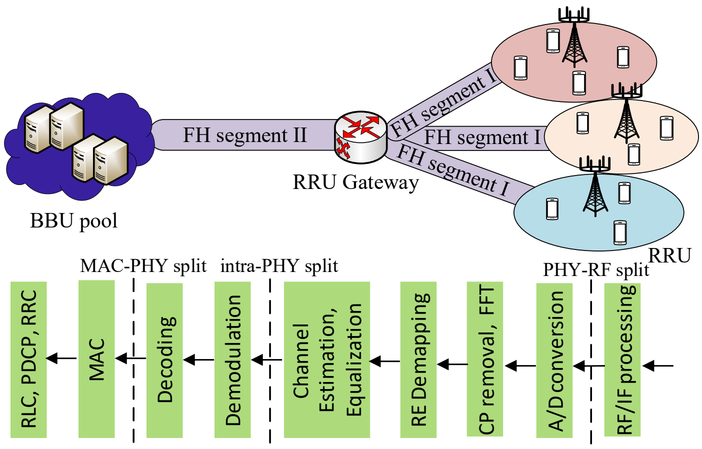


Fig. 5.1. C-RAN with Ethernet-based FH and intra-PHY split.

where $\beta_{ik,l}$ denotes the large-scale fading coefficient.

The channel gain vectors are estimated via UL training, which is repeated every coherence interval of τ_c REs. Out of τ_c REs, τ_p Orthogonal pilot sequences are used for training with pilot reuse. Hence, this results in pilot contamination. Let $\mathcal{B}_{ik}$ denotes the set of users that reuse the pilot sequence used by user k in cell i. Each RRU generates the minimum mean square estimates of the UL channel gain vectors from users in its cell. That is, RRU l estimates $\mathbf{h}_{lk,l}$, for $k = 1, \ldots, K$.

The estimated UL channel gains are used for matched filter equalization during UL data transmission. Note that only users which have data to transmit are active in a transmission slot and cause interference to transmissions from other active users. Let p_{lk} and α_{lk} denote the UL transmit power and activity factor of user k in cell l, respectively. The activity factor is the probability of a user being active in a transmission slot. Then, the UL SINR to user k in cell l when active is given by

$$\gamma_{lk} = \frac{M\eta_{lk}p_{lk}\beta_{lk,l}^2}{\sigma^2 + p_{lk}\beta_{lk,l} + \sum_{it\in\mathcal{K}_{lk}} \alpha_{it}p_{it}\beta_{it,l} + M\eta_{lk}\sum_{it\in\mathcal{K}_{lk}} \alpha_{it}p_{it}\beta_{it,l}^2}, \tag{5.1}$$

where $\eta_{lk} = \left(\sum_{it\in\mathcal{B}_{lk}} \beta_{it,l} + \sigma^2/\left(P_{\mathrm{ue}}\tau_p\right)\right)^{-1}$, P_{ue} is the maximum UL transmit power, $\mathcal{K}_{lk} = \{(it) \neq (lk) : i \in \{1, \cdots, L\}, t \in \{1, \cdots, K\}\}$ the set of all users in the network except user k in cell l, and σ^2 the AWGN variance. The SINR expression is derived in a manner similar to that in [BL15]. The UL spectral efficiency R_{lk} is then given by $R_{lk} = \zeta^{(\mathrm{ul})}\left(1 - \tau_{\mathrm{c}/\tau_{\mathrm{p}}}\right)\log_2\left(1+\gamma_{lk}\right)$ bits/RE, where $\zeta^{(\mathrm{ul})}$ is the fraction of time-frequency resources allocated for UL data transmission. Note that the worst-case SE, which does not account for the dynamic UL interference due to random user activity is used in (4.2) in Chapter 4. But, the spectral efficiency expression (5.1), unlike (4.2), accounts for the dynamic nature of interference owing to the random user activity. A min-max UL power optimization is carried out to ensure a minimum spectral efficiency $R_{\min}$ for all the users [VBL16].

User Traffic Model

The UL user traffic is generated as Poisson process. Let user k in cell l generates packet for UL transmission as a Poisson point process with arrival rate λ_{lk} packets/s. The interarrival time between two consecutive packet arrivals is exponentially distributed with rate λ_{lk}. Let $\mathcal{T}$ be the equidistant time interval, known as slot duration and j the number of packets arriving in a slot, $(0, \mathcal{T}]$. A slot is basic unit of time in discrete-time system, which could be, e.g., based on 5G NR flexible numerology that defines the slot duration as $\mathcal{T} = 2^{-\mu}$, compared with the LTE transmission time interval (TTI) of 1 ms or the orthogonal frequency division multiplexing (OFDM) symbol duration $T_{\text{OFDM}} = 2^{-\mu}$ compared with the LTE OFDM symbol duration of 66.67 μs, where $\mu \in \{0, 1, 2, 3, 4, 5\}$ is an integer. Thus, $\mathcal{T} = 1/0.5/0.25/0.125/0.0625$ ms in terms of TTI, or $\mathcal{T} = 66.7/33.3/16.7/8.3/4.2$ μs in terms of OFDM symbol duration [3GP18a].

These $j \geq 0$ arrivals in a slot, $(0, \mathcal{T}]$ follow Poisson distribution with arrival rate λ_{lk} packets/s, that is, for any $\mathcal{T} > 0$,

$$\mathbb{P}(\text{Number of arrivals} = j) = \frac{\exp(-\lambda_{lk}\mathcal{T})(\lambda_{lk}\mathcal{T})^j}{j!} \quad \sim \mathcal{P}(\lambda\mathcal{T}) \tag{5.2}$$

Let q_{lk} denotes the probability of no arrival, i.e., $q_{lk} = \mathbb{P}(\text{j}= 0)$. Then, $q_{lk} = \exp(-\lambda_{lk}\mathcal{T})$. Hence, probability of at most one arrival is $\alpha_{lk} = 1 - q_{lk} = 1 - \exp(-\lambda_{lk}\mathcal{T})$, which we call the activity factor due to Poisson arrivals.

Let F_{lk} denotes the packet size for user k in cell l. It is modeled as an exponential RV with mean $\overline{F}$. Note that there could be multiple packet arrivals in a slot for any user. In such cases, packets are transmitted together to the RRU in the next slot. While the packets can arrive at any time due to Poisson arrivals, the UL transmissions start only at the slot boundaries. This models the slotted nature of resource grants and transmission in practical systems, such as LTE and 5G NR. The user's SE determines the number of REs, needed to transmit the packets that arrived in a slot.

At the RRU, the received signals at different antennas are equalized to recover the spatially multiplexed symbols on an RE. Each of these symbols is quantized to $2N_q$ bits. The bits from all the users in the cell are then encapsulated in an Ethernet frame for transmission over FH.

The Ethernet aggregates packets from multiple RRUs at the RRU gateway, which does the switching of traffic between the RRUs and BBU pool. The Ethernet frames from RRUs are stored in a FIFO queue until FH segment II is available. Since the capacity C_{FH} of FH segment II is finite, queuing delays are present. This can result in outages if these delays exceed the budget D for the fronthaul. The fronthaul delay budget D is computed by deducting from the one-way HARQ trip time, which is 3 ms in LTE, the fixed delays involved in RRU and BBU pool processing, packetization, and propagation [CNS16].

5.2 Queue Modeling and Steady-state Analysis

We first develop a model for queuing at the RRU gateway and then derive novel closed-form expressions for steady-state queue length and sojourn time distributions.

5.2.1 Queue Model

In order to study the queuing dynamics at the RRU gateway, the arrival and service processes of the queue need to be characterized[1]. This is done below.

Arrival Process

We first note that the Ethernet frames arrive at the RRU gateway only at slot boundaries. This is because the UL transmissions from users last for the duration of a slot and only at the end of the slot the digitized I/Q samples are encapsulated in Ethernet frames. Further, no frame arrival occurs at a slot boundary if no user transmits in the previous slot. Thus, the arrivals at RRU follow a Bernoulli process with probability of no arrival given by $\prod_{l=1}^{L}\prod_{k=1}^{K}\exp(-\lambda_{lk}\mathcal{T}) = \exp\left(-\Lambda\mathcal{T}\right)$, where $\Lambda = \sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_{lk}$. It is the joint probability of no packet arrival in a slot duration for all the users. Hence, the probability of arrival p_{arr} is $p_{\text{arr}} = 1 - \exp\left(-\Lambda\mathcal{T}\right)$. Thus, the interarrival time expressed in number of slots is geometrically distributed with success probability p_{arr}.

Service Process

The service time is independent across Ethernet frame arrivals, as the number of packet arrivals in a slot and their packet sizes are assumed to be independent across slots and users. Thus, only the marginal distribution of service time conditioned on an arrival to the queue is needed. Towards this end, the distribution of the number of bits B added to the queue upon an arrival is required. Here, B is the sum of the frame sizes in bits from L RRUs, i.e., $B = \sum_{l=1}^{L} B_l$, where B_l denotes the frame size in bits from RRU l.

In order to find B_l, we start by considering a user k in cell l. Let N_{lk} denote the number of packet arrivals in a slot for user k in cell l. It is a Poisson RV with rate $\lambda_{lk}\mathcal{T}$ since the packet arrival process is Poisson. Further, let $F_{lk}^{(n)}$ denote the packet size in bits for the n^{th} arrival. Thus, the total number of bits to be transmitted in the UL is $\sum_{n=1}^{N_{lk}} F_{lk}^{(n)}$. The number of time-frequency resources needed to transmit these bits is $\sum_{n=1}^{N_{lk}} F_{lk}^{(n)}/R_{lk}$. This is also the number of symbols after equalization at RRU l for user k. Thus, the total number of received symbols at RRU l is obtained by summing over K users in cell l and is $\sum_{k=1}^{K}\sum_{n=1}^{N_{lk}} F_{lk}^{(n)}/R_{lk}$. Since each received symbols is quantized into $2N_q$ bits , the frame size B_l of RRU l is $2N_q\sum_{k=1}^{K}\sum_{n=1}^{N_{lk}} F_{lk}^{(n)}/R_{lk}$. Therefore, the number of bits B added to

[1] The model in this work is based on the submitted manuscript for [FCBF20] and it now differs from – after addressing the reviewers' comments– the final and accepted work in [FCBF20]. In this thesis work, the frames from different RRUs are aggregated to form a bigger Ethernet frame. However, Ethernet frames from different RRUs can arrive simultaneously at the RRU gateway. Thus, in [FCBF20], the frames from different RRUs are not aggregated and are treated as multiple Ethernet frames. This models the behaviour of an Ethernet switch better [FCBF20].

the queue in an arrival event is

$$B = 2N_q \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{1}{R_{lk}} \sum_{n=1}^{N_{lk}} F_{lk}^{(n)}. \tag{5.3}$$

Since $F_{lk}^{(n)}$ is an exponentially distributed RV with mean $\overline{F}$, $G_{lk}^{(n)} = 2N_q F_{lk}^{(n)}/R_{lk}$ is exponentially distributed with mean $\mu_{lk} = 2N_q\overline{F}/R_{lk}$. Therefore, $\sum_{n=1}^{N_{lk}} G_{lk}^{(n)}$ is Erlang distributed with shape parameter N_{lk} and scale parameter μ_{lk}. Thus, the RV B is the sum of LK Erlang RVs with different shape and scale parameters. The following result gives its distribution.

Result 1. *The cumulative distribution function (CDF) $F_B(x)$ of B conditioned on the event that there is an arrival, i.e., $N = \sum_{l=1}^{L}\sum_{k=1}^{K} N_{lk} > 0$, is given by*

$$F_B(x) = \sum_{m=1}^{\infty} \frac{\mathcal{T}^m \exp(-\Lambda\mathcal{T})}{1 - \exp(-\Lambda\mathcal{T})} \sum_{\substack{n_{11},\dots,n_{LK}\geq 0 \\ \sum_{l=1}^{L}\sum_{k=1}^{K} n_{lk}=m}} \left[\prod_{l=1}^{L}\prod_{k=1}^{K} \frac{\lambda_{lk}^{n_{lk}}}{n_{lk}!}\right] \times \left(1 - \boldsymbol{\vartheta}^T \exp\left(x\boldsymbol{M}\right)\mathbf{1}\right), \tag{5.4}$$

where $\boldsymbol{\vartheta} = [1, 0, \dots, 0]^T$ and $\mathbf{1} = [1, \dots, 1]^T$ are $m \times 1$ vectors, and $\exp(x\boldsymbol{M})$ is the matrix exponential of $x\boldsymbol{M}$ for $x \geq 0$. Here, $\boldsymbol{M}$ is an $m \times m$ block-diagonal matrix with entries $\boldsymbol{M}_{11}, \dots, \boldsymbol{M}_{LK}$, where $\boldsymbol{M}_{lk}$ is an $n_{lk} \times n_{lk}$ matrix with $-1/\mu_{lk}$ in the main diagonal, $1/\mu_{lk}$ in the superdiagonal[2] and the remaining coefficients of M are zero.

Proof. The proof is relegated to Appendix A.4. □

The service time (in terms of number of slots) S needed for the RRU gateway to forward the B bits to segment II is $\lceil B/(C_{\text{FH}}\mathcal{T})\rceil$, where $\lceil\cdot\rceil$ denotes the ceil operation[3]. Using (5.4), the PMF $p_S(i)$ of S, for $i = 1, \dots, \infty$, is

$$p_S(i) = \mathbb{P}(S = i) = F_B\left(iC_{\text{FH}}\mathcal{T}\right) - F_B\left((i-1)C_{\text{FH}}\mathcal{T}\right). \tag{5.5}$$

5.2.2 Steady-state Analysis

For a discrete-time queue with Bernoulli arrivals and service time PMF given in (5.5), we now present results for stability, queue length, and sojourn time distributions. These results are adapted from [Bos02], which provides the results for general service time distribution.

[2] A superdiagonal of a square matrix is a set of elements directly above and to the right of the main diagonal.

[3] For analytical tractability, we assume that the RRU gateway and UL transmissions have slot as the same basic unit of time.

Stability

The load ρ is defined as the product of arrival rate and the average service time. The arrival rate for the Bernoulli arrival process is p_{arr}. The average service time for the service time distribution in (5.5) is $\overline{S} = \sum_{i=1}^{\infty} i p_S(i)$. Thus, the load is given by

$$\rho = p_{\mathrm{arr}} \sum_{i=1}^{\infty} i p_S(i). \tag{5.6}$$

The stability of the queue is ensured when the offered load is less than one, i.e. $\rho < 1$.

Queue Length and Sojourn Time Distributions

We use the Pollaczek-Khinchine formula [Bos02] to express the generating functions of queue length and sojourn time in terms of the generating functions of number of arrivals in a slot duration and service time. The generating function of number of arrivals in a slot duration is $A(z) = 1 - p_{\mathrm{arr}} + z p_{\mathrm{arr}}$ and that of service time is $S(z) = \sum_{i=1}^{\infty} p_S(i) z^i$. Then, the generating function $Q(z)$ of queue length is given by

$$Q(z) = \frac{(1-\rho)(1-z)S(1-p_{\mathrm{arr}}+zp_{\mathrm{arr}})}{S(1-p_{\mathrm{arr}}+zp_{\mathrm{arr}})-z}, \tag{5.7}$$

where $S(1-p_{\mathrm{arr}}+zp_{\mathrm{arr}})$ is the generating function of the number of arrivals during a service time.
Further, the generating function $T(z)$ of the sojourn time is given by

$$T(z) = \frac{(1-\rho)(1-z)S(z)}{1-p_{\mathrm{arr}}-z+p_{\mathrm{arr}}S(z)}. \tag{5.8}$$

Finally, we take the inverse Z-transforms of $Q(z)$ and $T(z)$ in order to find the queue length and sojourn time PMFs, respectively.

The mean sojourn time $\overline{T}$ is computed using (5.8) and Little's law [Bos02]. It is given by

$$\overline{T} = \overline{S} + \frac{\Lambda \mathbb{E}[S^2] - \rho}{2(1-\rho)}, \tag{5.9}$$

where $\mathbb{E}[S^2]$ is the second moment of service time distribution.

Efficient Computation of Inverse Z-transform

In order to efficiently compute the inverse Z-transform, the long-division method [BC59] is used. This involves expressing the Z-transform as a ratio of two polynomials. Note that $[p_S(0), \ldots, p_S(s_{\max})]$ are the coefficients of the polynomial $S(z)$, where $s_{\max}$ is the maximum service time beyond which the PMF values are negligible. Then, the coefficients for the numerator polynomial of $T(z)$ are $(1-\rho)[p_S(0), p_S(1)-p_S(0), \ldots, p_S(s_{\max})-p_S(s_{\max}-1), p_S(s_{\max})]$ and for the denominator polynomial of $T(z)$ are $[1-p_{\mathrm{arr}}+p_S(0), p_S(1)-1, \ldots, p_S(s_{\max})]$. The long-division method can then be used to find the quotient polynomial, whose coefficients yield the sojourn time PMF values.

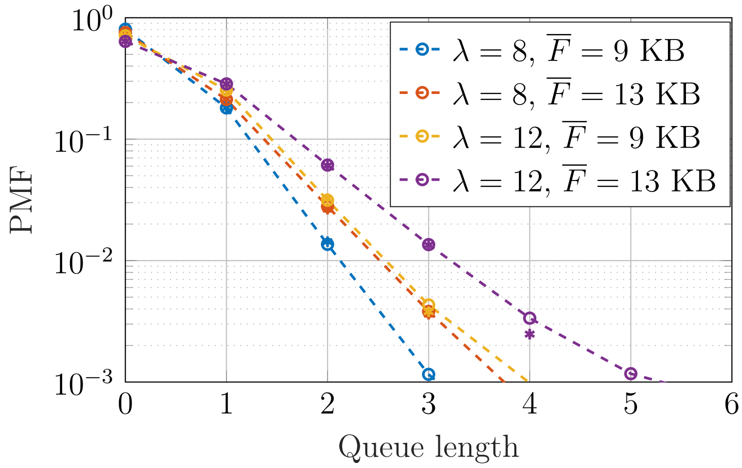


Fig. 5.2. Queue length PMF for different arrival rate λ and average packet size $\overline{F}$ when $C_{\text{FH}} = 1$ Gbps. Circle and star markers denote analytical and simulation results, respectively. [FCBF20]

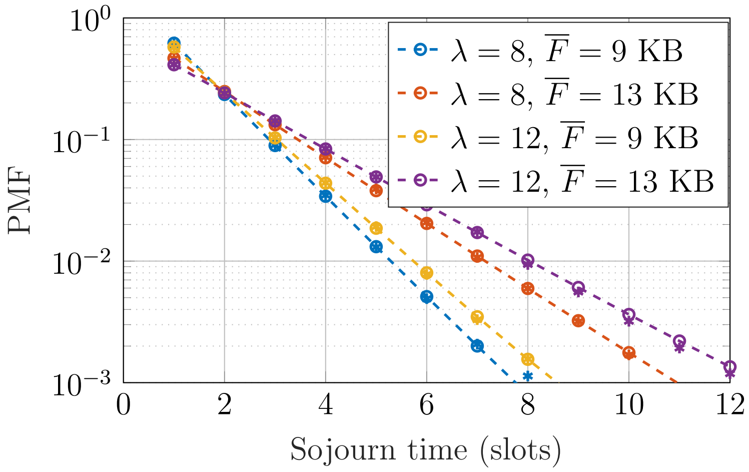


Fig. 5.3. Sojourn time PMF for different arrival rate λ and average packet size $\overline{F}$ when $C_{\text{FH}} = 1$ Gbps. Circle and star markers denote analytical and simulation results, respectively. [FCBF20]

This procedure can be repeated to evaluate the queue length PMFs. However, evaluating the coefficients of $S(1 - p_{\text{arr}} + zp_{\text{arr}}) = \sum_{i=0}^{s_{\max}} p_S(i)(1 - p_{\text{arr}} + zp_{\text{arr}})^i$ is slightly more involved. This can be done efficiently by using repeated convolution to compute the coefficients of $(1 - p_{\text{arr}} + zp_{\text{arr}})^i$ and then appropriately summing the coefficients for different i.

5.3 Numerical Results

We consider a 7-cell hexagonal cellular layout with wrap-around. The cell radius is 500 m and $K = 10$ users are randomly dropped in each cell. The RRUs are equipped with $M = 200$ antennas. The maximum transmit power $P_{\text{ue}} = 23$ dBm and noise variance $\sigma^2 = -174$ dBm. Pilot sequences are uniquely assigned to users. Hence, $\tau_p = 70$. The coherence interval is set to $\tau_c = 200$ REs. The large-scale fading coefficient in dB is

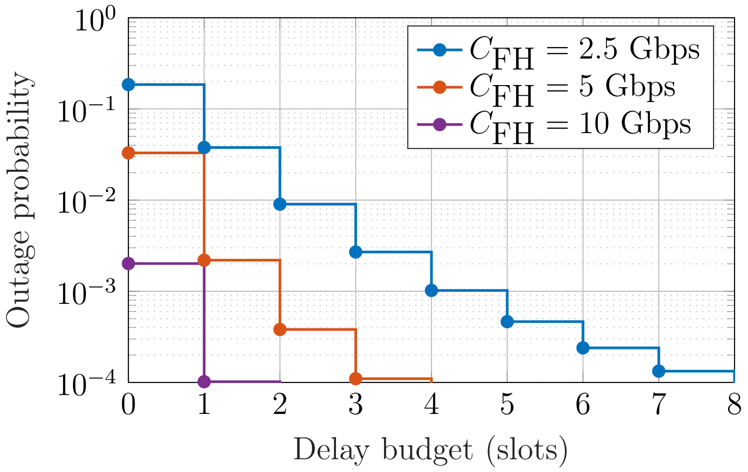


Fig. 5.4. Outage probability for different values of FH capacity C_{FH} when $\lambda = 8$ packets/s [FCBF20].

[CFBF19a] [CBF18] $\beta_{ik,l} = -128.1 + 37.6\log_{10}(d_{ik,l}/d_0) + \Psi_{\mathrm{shad}}$, where $d_{ik,l}$ is the distance between user k in cell i and RRU l, and $d_0 = 100$ m is the reference distance. Here, Ψ_{shad} is a Gaussian RV with zero mean and standard deviation 8 dB, which models lognormal shadowing. Minimum user SE is set to $R_{\mathrm{min}} = 1$ bit/symbol. The packet arrival rate λ is the same for all users. We choose the slot duration to be $\mathcal{T} = 0.25$ ms, which is one of the possible TTIs in 5G. The average packet size $\overline{F}$ is set to be in the range of maximum transport block size 12.9 KB in 5G NR [3GP18a]. For these simulation parameters, we have observed that the first 3 terms of the series in (5.5) are sufficient to ensure numerical accuracy.

Fig. 5.2 plots the queue length PMF for different values of arrival rate and average packet size for a random realization of the large-scale fading coefficients. We see an excellent match between the analysis and simulation curves, which validates our analytical results. The PMF value for $\lambda = 8$ is higher than that for $\lambda = 12$ at queue length equal to zero. This is expected as the queue is empty more often at lower arrival rates. For larger queue lengths, the PMF values are lower for $\lambda = 8$ when compared with $\lambda = 12$. This is because large queue lengths occur less often with the former value, given its lower arrival rate. Similar trends are observed when $\overline{F}$ increases. The PMF value at queue length equal to zero is higher for $\overline{F} = 9$ KB when compared with that for $\overline{F} = 13$ KB. The reverse is true at higher queue lengths.

The sojourn time PMFs for different arrival rates and average packet sizes are plotted in Fig. 5.3. Note that the PMF value is zero at sojourn time equal to zero because a minimum of one slot is needed to service an arrival. We again observe an excellent match between analysis and simulation results. The trends exhibited by the curves for $\lambda = 8$, 12 and $\overline{F} = 9$ KB, 13 KB are similar to those of the queue length PMFs in Fig. 5.2.

Fig. 5.4 plots the outage probability as a function of delay budget for different values of C_{FH}. These curves are the complementary CDFs of sojourn time averaged over large-scale fading. These results can be used to appropriately dimension FH; for example, $C_{\mathrm{FH}} = 10$ Gbps ensures that the outage probability is less than 10^{-4} when $\lambda = 8$ packets/s. We see that the outage probability is lower for higher C_{FH}. This is expected as the service

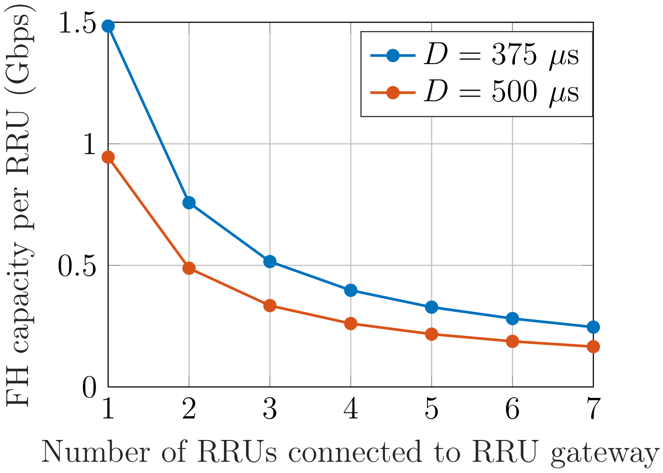


Fig. 5.5. FH capacity per cell as a function of the number of cells connected to the RRU gateway for $\lambda = 8$ packets/s.

time decreases as C_{FH} increases.

We now study the statistical multiplexing gains possible by aggregating traffic from multiple RRUs at the RRU gateway. This is done as follows [Maz]. First, the FH capacity needed to ensure that the average sojourn time $\overline{T}$ is below a threshold D is computed. It is then divided by the number of RRUs connected to the RRU gateway to determine the required FH capacity per RRU. This is repeated for different number of RRUs connected to the RRU gateway. These results are plotted in Fig. 5.5 for two different D. We see that the required fronthaul capacity per RRU decreases as the number of RRUs connected to the RRU gateway increases. This saving in the FH capacity is the statistical multiplexing gain. It is 83% for $D = 500$ μs when the number of connected RRUs is increased from one to seven. We also see that the FH capacity per RRU is lower when D is higher, i.e., for a more relaxed delay requirement.

5.4 Chapter Summary

We proposed a novel queuing model for the RRU gateway with UL traffic from users to RRUs and then to a BBU pool through an Ethernet-based FH. We derived closed-form expressions for the steady-state queue length and sojourn time distributions. The analysis took into account the user activity factors, users' SEs to massive MIMO-aided RRUs, the slotted nature of UL transmissions, and FH capacity of a two-hop packetized FH network. The analytical results were validated through numerical simulations. We saw that the probability of higher queuing delays decreased as either the arrival rate or average file size was decreased, or when the FH capacity was increased. We also studied the FH capacity savings possible through statistical multiplexing. In the investigated scenario, the statistical multiplexing gains are as high as 83%.

Chapter 6

RRU Computational Complexity Analysis in C-RAN

6.1 Introduction and Motivation

By offloading more signal processing functionalities to the RRU, the required fronthaul bandwidth is significantly decreased (refer to, Section 2.1.1), which is desirable to lower the network deployment cost. However, this reduces not only the acclaimed C-RAN centralization/virtualization benefits, but also demands more processing power at the RRU. Considering practical implementation aspects, particularly size, cost, weight and power consumption, it is often desirable to keep the RRU as simple, yet efficient, as possible. In this chapter, we calculate the computational complexity of the RRU with regards to the 5G NR flexible numerology when employing the recently standardized xRAN functional split 7.2. Unlike CPRI and eCPRI, which do not deliver a full interface standardization that would allow a true interoperability among different vendors, the recently formed xRAN Fronthaul Working Group supports an *open*, *interoperable* and *efficient* fronthaul interface. xRAN Forum, now known as O-RAN Alliance [ORA], has identified a single split point, known as Option 7.2 [Gro18], and has delivered an extensive interface specification that will enable *true interoperability* between RRUs and BBUs of different vendors. We compare suitability in terms of power efficiency and flexibility of the RRU being implemented using either field programmable gate array (FPGA) or general purpose processor (GPP) (e.g., x86) considering their computational requirements. Based on the complexity analysis, we calculate the required number of FPGAs or GPPs to support the complexity of the RRU. We show that FPGA is more feasible[1] option compared to x86 in terms of power consumption, particularly for rooftop-mounted RRU.

In C-RAN, it is often assumed that a GPP [NMM+14], e.g., x86 can be utilized, benefiting from economies of scale and pooling gains. In principle, all of the PHY layer functions can be performed on GPP hardware. However, for future RATs employing, e.g., carrier aggregation and mmWave communication with larger bandwidths, PHY processing will be challenging on GPP. Moreover, PHY layer processing functions are usually fixed,

[1] This analysis is very implementation specific, and depends on the processor architecture, manufactures and model. Nevertheless, it shows comparative insights with some trends.

and require less flexibility or programmability.

Contribution

The main contributions in this work can be summarized as follows:

1. We find out the computational requirements of the C-RAN focusing on the functional split 7.2 with detailed PHY functions performed at the RRU in the DL;
2. We consider 5G NR flexible numerology and compute its complexity in order to account for different use cases and application scenarios;
3. For the offered net computational requirement, we calculate the required number of FPGAs or GPPs and compare the deployment of the RRUs using FPGA or GPP in terms of flexibility and power efficiency.

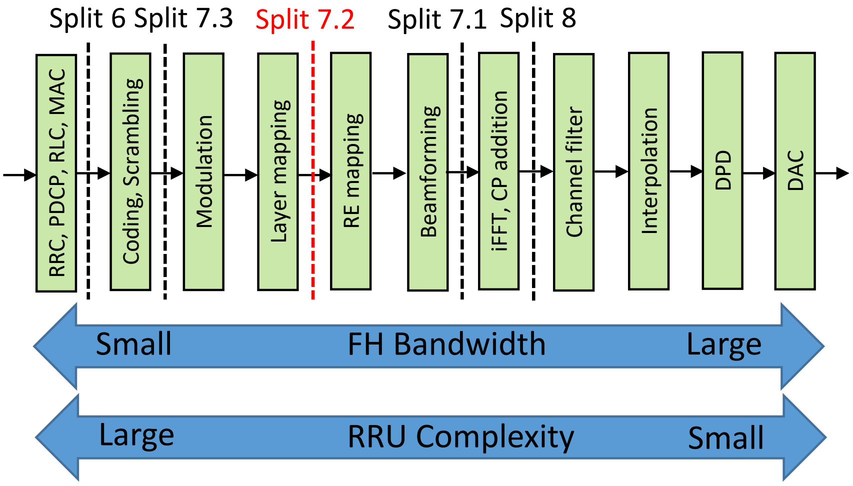


Fig. 6.1. xRAN functional split 7.2 showing full processing chain. The fronthaul bandwidth is reduced and the RRU complexity is increased if the functional split is moved from the right to left towards lower split namings.

6.2 Flexible 5G New Radio

3GPP with Release 15 has finalized the 5G NR specification [3GP18b] and it supports operation in a wide range of frequency bands, ranging from sub-1 GHz to mmWave bands. It has defined two operating frequency ranges (FRs): FR1: 450 MHz - 6 GHz (commonly referred to as sub-6) and FR2: 24.25 GHz - 52.6 GHz (also referred to as mmWave). In FR1 and FR2, the maximum bandwidth is 100 MHz and 400 MHz, respectively, which are much greater than the maximum LTE bandwidth of 20 MHz. In order to support a wide range of use cases and application scenarios, 5G NR supports flexible subcarrier spacing given by [3GP18b]

$$\Delta f = 2^{\mu} \times 15 \text{ kHZ},$$

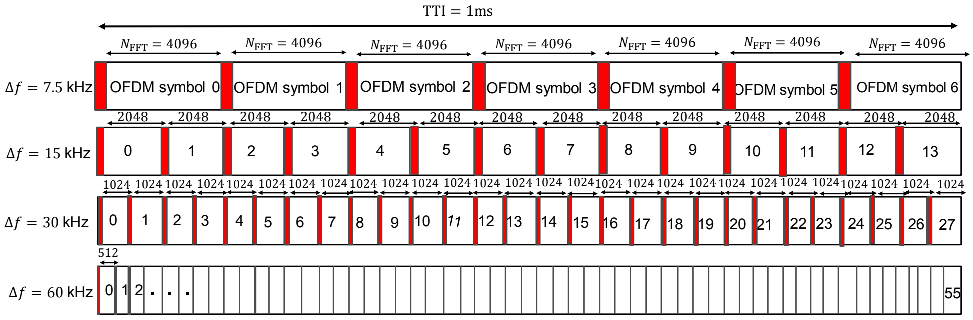


Fig. 6.2. 5G NR flexible numerology.

where $\mu \in \{-1, 0, 1, 2, 3, 4, 5\}$ is an integer. Accordingly, the slot duration is scaled by a factor of $\mathcal{T} = 2^{-\mu}$, when compared with the 1 ms LTE TTI, and each slot now contains 14 OFDM symbols, unlike a LTE slot, which contains 7 OFDM symbols. This means that the slot duration, $\mathcal{T}$, and hence, the cyclic prefix (CP) length, $T_{\text{CP}} = 2^{-\mu} \times 4.7\mu$s and the OFDM symbol duration, $T_{\text{OFDM}} = 1/\Delta f$ are reduced as the subcarrier spacing increases, as illustrated in Fig. 6.2.

The bandwidth can be divided into different blocks using different numerologies. However, as only the subcarriers within a numerology are orthogonal to each other, the subcarriers from one numerology interfere with subcarriers from neighbouring numerology causing inter-numerology interference (INI) [YA18]. The effect of INI can be mitigated, e.g. by inserting guard tones between numerologies [3GP17c] or by applying time-domain filtering per numerology (sub-band) or time-domain windowing [3GP17b]. It is claimed in [3GP17b] that windowing is an efficient and simple tool to control INI, and windowing has extremely low complexity. Hence, for simplicity, we do not account for windowing in our complexity analysis.

Computational Complexity

The computational complexity of signal processing operations in literature is often stated in terms of operations per seconds (OPS). In this work, we quantify the computational complexity in terms of the total number of real multiplications and real additions per symbol. We specify this in terms of floating point operations per second (FLOPS). The signal processing operations in each functional block have different complexity. Hence, depending upon the the employed functional split, the RRU net complexity differs. Unlike the simple 3GPP functional block diagram, Fig. 6.1 shows the C-RAN architecture with full processing chain including the detailed RF chain, which is important for practical implementation aspects. The necessary additions arise from two main practical aspects [5G-18]: carrier aggregation and massive MIMO. Firstly, the RRUs need to support several carriers at the same time, which requires digital channel filtering, up-conversion and carrier mixing to create a composite signal of many carriers. Secondly, in order to provide sufficient transmit power, RRUs have to use non-linear power amplifiers (PAs). In order to avoid distortion by the non-linearity of the PAs, digital predistortion (DPD) needs to be applied. These processing steps significantly increase the required computational complexity and in this work, we compute their complexity considering Split 7.2. In general, it is quite involved to calculate the exact computational complexity, as computational complexity of a certain function depends very much on the specific implementation (e.g., depends on processor architecture, model and manufacturing company). Nevertheless, based on the basic operations to be performed, the order of magnitude can be estimated for comparison.

6.3 C-RAN Complexity Analysis

Computational Complexity for Beamforming

In order to compute the beamforming complexity, we consider a beamformer, which converts N_L data streams (layers/beams) into N_{Ant} antenna streams. For a given signal carrier bandwidth with $N_{\text{sub,act}}$ utilized subcarriers, and $N_{\text{sym}}^{\text{SF}}$ OFDM symbols per T_{SF} second, the output is produced at a rate of (samples/sec)

$$R_{\text{BF},L} = N_{\text{Ant}} \cdot N_{\text{sub,act}} \cdot N_{\text{sym}}^{\text{SF}} \cdot T_{\text{SF}}^{-1}. \tag{6.1}$$

One antenna sample corresponding to N_L layers requires N_{CMA} complex multiplications and additions. Assuming that multiplications and additions can be performed within one clock cycle, and that one complex multiplication requires $N_{\text{OP,CMA}} = 3$ non-complex operations, the total rate of operations can be obtained as

$$R_{\text{BF,tot}} = R_{\text{BF},L} \cdot N_{\text{CMA}} \cdot N_{\text{OP,CMA}}. \tag{6.2}$$

5G NR Flexible Numerology and its Complexity Analysis

Computational complexity of flexible numerology based on the TTI frame structure is considered here. For this, 100 MHz bandwidth with and without flexible numerology configurations is considered, as shown in Fig. 6.3, where Fig. 6.3a represents a uniform division of the 100 MHz bandwidth into five different sub-bands of 20 MHz each, and the FFT computation of the each sub-band is performed independently based on the TTI. Fig. 6.3b represents non-uniform distributions of the 100 MHz bandwidth, and Fig. 6.3c represents a full 100 MHz bandwidth with 15 kHz subcarrier spacing and FFT size of 8192[2].

After fixing the FFT length of each band based on the configurations shown in Fig. 6.3, our goal is to compute the short FFT length of each band independently. After short FFT calculation, we perform first the interpolation and then mixing of the different frequency bands in order to match one of them with the same sampling rate, as shown in Fig. 6.4. Therefore, the total complexity associated with flexible numerology can be obtained as

$$C_{\text{Total}} = C_{\text{FFT}} + C_{\text{Interpolation}} + C_{\text{Mixer}}, \tag{6.3}$$

[2] Although in the standard, 100 MHz bandwidth with 15 kHz subcarrier spacing is not specified, this spacing is considered as a baseline, since the theoretical analysis is still valid.

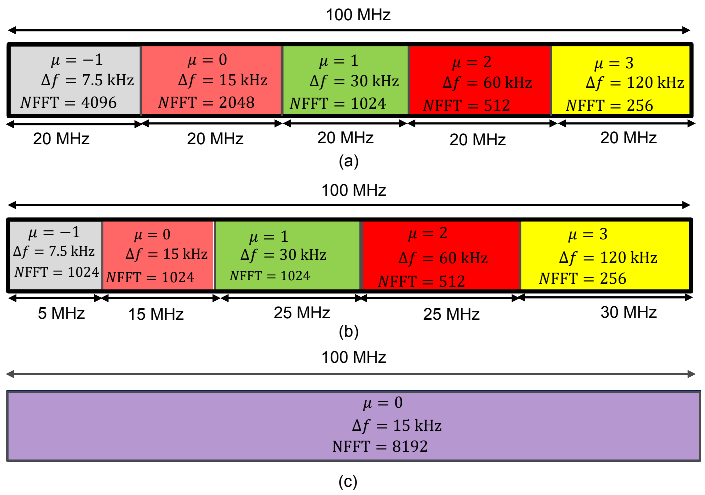

100 MHz
$\mu = -1$ Δf = 7.5 kHz NFFT = 4096
$\mu = 0$ Δf = 15 kHz NFFT = 2048
$\mu = 1$ Δf = 30 kHz NFFT = 1024
$\mu = 2$ Δf = 60 kHz NFFT = 512
$\mu = 3$ Δf = 120 kHz NFFT = 256
20 MHz
20 MHz
20 MHz
20 MHz
20 MHz
(a)
100 MHz
$\mu = -1$ Δf = 7.5 kHz NFFT = 1024
$\mu = 0$ Δf = 15 kHz NFFT = 1024
$\mu = 1$ Δf = 30 kHz NFFT = 1024
$\mu = 2$ Δf = 60 kHz NFFT = 512
$\mu = 3$ Δf = 120 kHz NFFT = 256
5 MHz
15 MHz
25 MHz
25 MHz
30 MHz
(b)
100 MHz
$\mu = 0$ Δf = 15 kHz NFFT = 8192
(c)

Fig. 6.3. FFT length calculation comparison of the 100 MHz bandwidth: (a) Flexible numerology based on uniform division of the bandwidth (b) Flexible numerology based on non-uniform division of the bandwidth (c) Long FFT calculation based on the uniform subcarrier spacing of 15 kHz of 8192 FFT length.

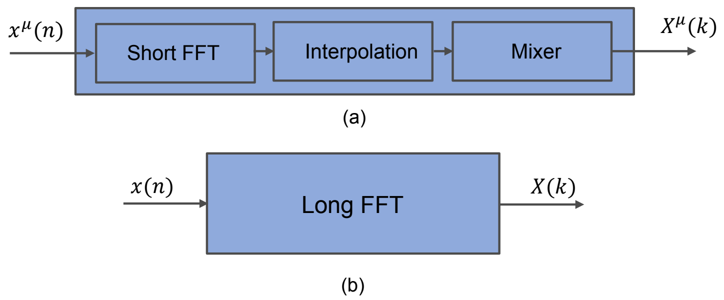


Fig. 6.4. Total computational complexity associated with flexible numerology of the short FFT length (a), where μ represents the different flexible numerology, $x(n)$ and $X(k)$ denote the time-domain and frequency-domain signal, respectively (b) Long FFT.

where C_{FFT} is the complexity associated with FFT calculation, $C_{\mathrm{Interpolation}}$ the complexity of the interpolation filter, and C_{Mixer} accounts for the complexity associated with mixing mixes different frequency bands.

A detailed summary of the different methods for computing the FFT complexity is given in Table 6.1. Now, considering Fig. 6.4, we describe the associated complexity with

Tab. 6.1. Different methods for computing the FFT complexity [CLW67]-[KM19].

Complexity	**Adder**	**Multiplier**
Radix-2	$N_{\mathrm{FFT}}\log_2(N_{\mathrm{FFT}})$	$\frac{N_{\mathrm{FFT}}}{2}\log_2(N_{\mathrm{FFT}})$
Split-radix algorithm	$3(N_{\mathrm{FFT}}(\log_2(N_{\mathrm{FFT}})-1)+4)$	$N_{\mathrm{FFT}}(\log_2(N_{\mathrm{FFT}})-3)+4$

the following steps [YKLV$^+$17]:

1. The complexity of N_{FFT} FFT can be computed in terms of number of addition and multiplication operations from Table 6.1;

2. Insertion of the CP for each sub-band of the flexible numerology;
3. Complexity associated with the low-pass interpolation filter, whose filter length is approximately $N_{\text{FIR}}N_{\text{FFT}}/N$, is $N_{\text{FIR}}N_{\text{FFT}}(N_{\text{FFT}} + N_{\text{CP}})/N$, where N_{FIR} is the order of the filter required to match the output-sampling rate of 30.72 MHz, N_{FFT} is the required FFT length for each numerology, and N is the total FFT length for uniform subcarrier spacing;
4. Finally, mixing at the output with the same sampling rate requires $N^{\mu}(N + N_{\text{CP}}N/N_{\text{FFT}})$ multiplications per symbol, where N^{μ} is the total number of numerology sub-bands.

Channel Filter

In order to model the channel filter, we consider an FIR filter implementation of the channel. The total complexity of the FIR filter is $C_{\text{FIR}} = C_{\text{FIR}}^{\text{Add}} + C_{\text{FIR}}^{\text{Mult}}$ flop, where

$$C_{\text{FIR}}^{\text{Add}} = N_{\text{taps}} \times (N_{\text{FFT}} + N_{\text{CP}} - 1), \tag{6.4}$$

$$C_{\text{FIR}}^{\text{Mult}} = N_{\text{taps}} \times (N_{\text{FFT}} + N_{\text{CP}}), \tag{6.5}$$

where N_{taps} is the number FIR filter taps, and $N_{\text{FFT}} + N_{\text{CP}}$ the total number of input samples.

Digital Pre-Distortion (DPD)

It is well known that PAs cause non-linear distortions in OFDM system and OFDM systems suffer from high peak-to-average power ratio (PAPR) and out-of-band (OOB) emission. In literature, several techniques for linearization of the PA behaviour [YSN+18] are available. Among them, DPD is one of the most cost-effective linearization techniques. DPD adds an extra non-linear function before the PA to process the input signal. The behavior of the resulting cascade is linear [YSN+18]. For K^{th} order, DPD filter complexity is $4KQN_{\text{Ant}}$ flop [YSN+18], where Q the memory depth, and N_{Ant} the number of transmit antennas in the DL. The DPD complexity can be obtained as $C_{\text{DPD}} = C_{\text{DPD}}^{\text{Add}} + C_{\text{DPD}}^{\text{Mult}}$ flop, where

$$C_{\text{DPD}}^{\text{Add}} = (N_{\text{FFT}} + N_{\text{CP}} - 1) \times 4 \times K \times N_{\text{Ant}}, \tag{6.6}$$

$$C_{\text{DPD}}^{\text{Mult}} = (N_{\text{FFT}} + N_{\text{CP}}) \times 4 \times K \times N_{\text{Ant}}. \tag{6.7}$$

6.4 Results

Depending on the processor characteristics, every function of Fig. 6.1 can show different computational complexity for different processors. The required computational power has to be computed and compared in relation with the processor available from different platforms. There are different families of FPGAs with different sizes and power requirements. As a reference, we use a Xilinx Ultrascale+ [Xil] for FPGA processing, and Xeon

6140 [Xeo] for GPP.

Xilinx FPGA

A Xilinx ultrascale FPGA features $N_{\text{DSP}} = 4272$ DSP48s, which can perform one operation each and can run approximately up to $C_{\text{FPGA}} = 500$ MHz. However, usually an FPGA cannot utilise all resources at maximum clock speed. Therefore, assuming a maximum utilization of $\mu_{\text{FPGA}} = 0.7$, the total computational power is estimated as

$$R'_{\text{FPGA}} = N_{\text{DSP}} \cdot C_{\text{FPGA}} \cdot \mu_{\text{FPGA}} \tag{6.8}$$

Each DSP can perform both the multiplication and addition operations. Hence, each operation requires 2 flops. Thus, the total computational power of FPGA (in Gflops) can be obtained as

$$\begin{aligned} R_{\text{FPGA}} &= 2 \cdot R'_{\text{FPGA}} \\ &= 2990 \text{ Gflops.} \end{aligned} \tag{6.9}$$

Power consumption of FPGAs is not straightforward, as it usually depends many factors, such as on the overall utilisation. Typical FPGA power consumption from [Xil] is $P_{\text{FPGA}} = 30$ W.

Xeon 6140

Xeon 6140 is a GPP, which features up to $N_{\text{cores}} = 18$ cores. Each core can perform $N_{\text{flop}} = 32$ flops per cycle and it runs at a maximum clock speed of $C_{\text{Xeon}} = 2.6$ GHz. Assuming a processor utilization of $\mu_{\text{Xeon}} = 0.7$, total computational power can be calculated as

$$\begin{aligned} R_{\text{Xeon}} &= N_{\text{cores}} \cdot C_{\text{Xeon}} \cdot N_{\text{flop}} \cdot \mu_{\text{Xeon}} \\ &= 1048 \text{ Gflops.} \end{aligned} \tag{6.10}$$

The power consumption of the Xeon 6140 is approximately $P_{\text{Xeon}} = 140$ W.

Flexible Numerology Complexity Analysis

Here, our goal is to compute and compare complexity between the different scalable numerologies. Fig. 6.5 shows the complexity associated with 100 MHz bandwidth, where 100 MHz bandwidth is equally split into five different sub-bands each of 20 MHz (refer to Fig. 6.3a). We group individual sub-bands complexity corresponding to sub-bands colors. We can see that the sum complexity of all individual sub-bands, denoted by $\sum C_{\mu}^{\text{uniform}}$, is comparable to that of a single 100 MHz band, denoted by $C_{100\text{ MHz}}^{\Delta f=15\text{ kHz}}$.

Similarly, Fig. 6.6 depicts the complexity of each sub-band for non-uniform carrier spacing shown in Fig. 6.3b. Unlike for uniform bandwidth division, for non-uniform bandwidth division, the sum of all individual complexities, denoted by $\sum C_{\mu}^{\text{nonuniform}}$ is less than that of the long FFT case in Fig. 6.3c. This is because larger bandwidth associated with a higher subcarrier spacing gives smaller size FFT length and hence, FFT complexity is reduced.

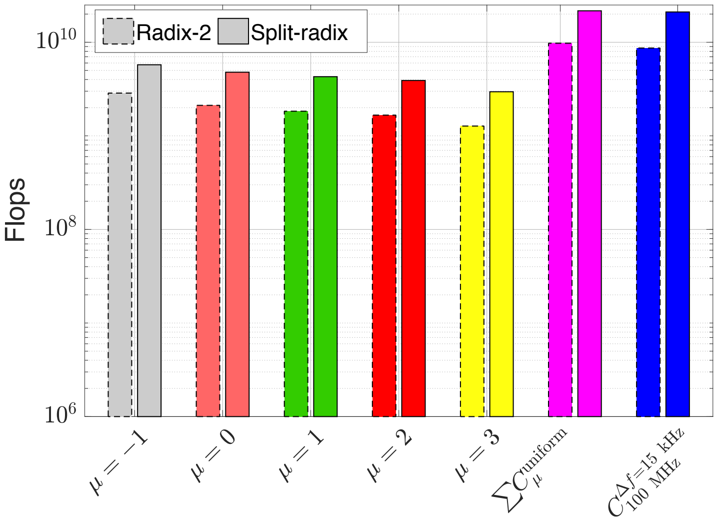


Fig. 6.5. Complexity comparison between uniform subcarrier spacing given in Fig. 6.3a with long FFT calculation in Fig. 6.3c [CKBF19].

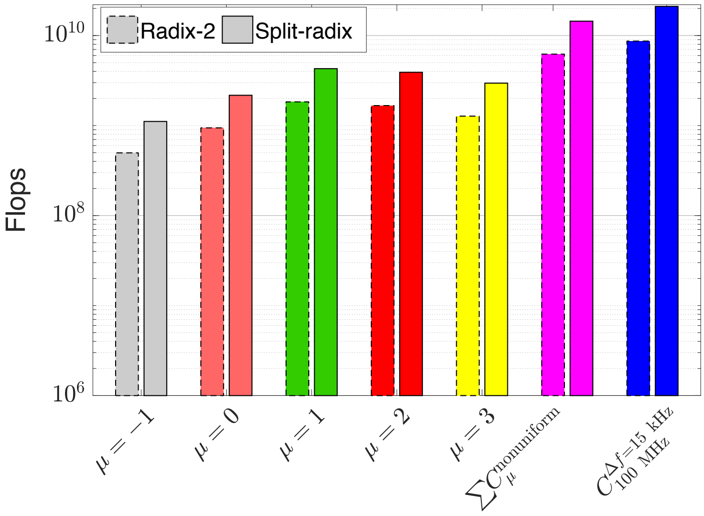


Fig. 6.6. Complexity comparison between non-uniform subcarrier spacing given in Fig. 6.3a with long FFT calculation in Fig. 6.3c [CKBF19].

Computational Complexity for Beamforming

As explained in Sec. 6.3, the beamforming complexity depends on the available subcarriers, which in turn depends on the carrier aggregation bandwidth. Fig. 6.7 shows beamforming complexity for different carrier aggregation techniques for four antenna numbers, $N_{\text{Ant}} = 32, 64, 100, 128$. We assume here $N_L = 16$, $T_{\text{SF}} = 1$ ms and $N_{\text{sub,act}} = 1200 \times n$, where $n \in \{1, 2, 3, 4, 5\}$ for different number of carrier aggregations. It is obvious that beamforming complexity increases with the number of antennas at the RRU and carrier signal bandwidth.

Fig. 6.8 shows the individual complexity of each block based on Split 7.2. Following the LTE specification, a channel FIR filter with $N_{\text{taps}} = 81$ paths is assumed. Similarly, complexity calculation of the DPD filter is also based on the FIR filter, but a $K = 4^{\text{th}}$

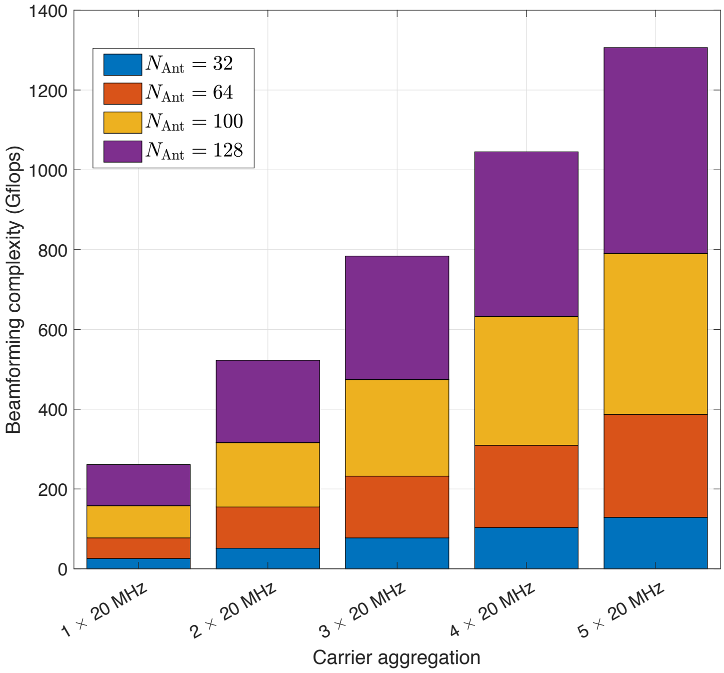


Fig. 6.7. Beam forming complexity for different number of carrier aggregation. $N_L = 16$ streams.

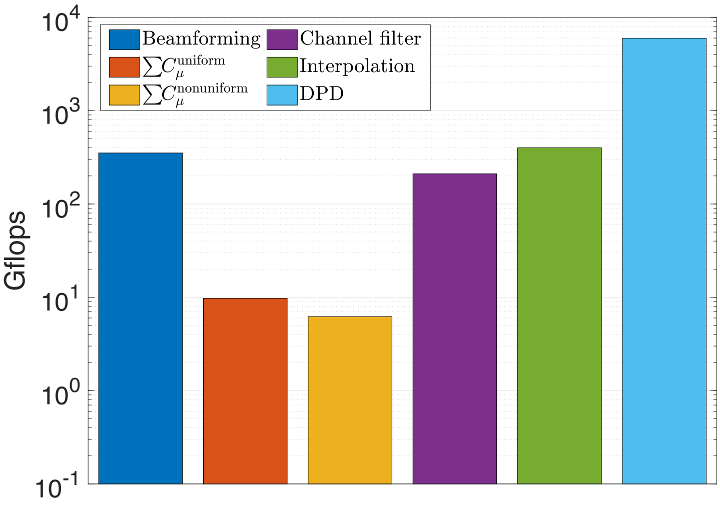


Fig. 6.8. Individual complexity for each functional block in the DL for the RRU employing Split 7.2 [CKBF19].

order filter with $N_{\text{Ant}} = 64$ antenna ports is considered. From Fig. 6.8, is clearly visible that DPD consumes most of the computation. This arises mainly due to more processing involved in compensating the non-linear distortion of the PAs for making them to operate in a linear region.

Next, we are interested to calculate the number of required devices, which can be obtained by dividing the total RRU complexity with the individual FPGA or Xeon complexity as

$$N_{\text{FPGA/Xeon}} = C_{\text{total}}/R_{\text{FPGA/Xeon}}, \tag{6.11}$$

where C_{total} is the the total complexity of the signal processing chain and $R_{\text{FPGA/Xeon}}$

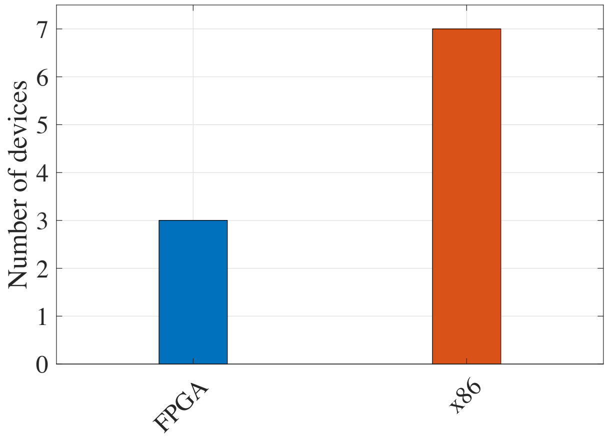


Fig. 6.9. Number of devices required for Split 7.2 [CKBF19].

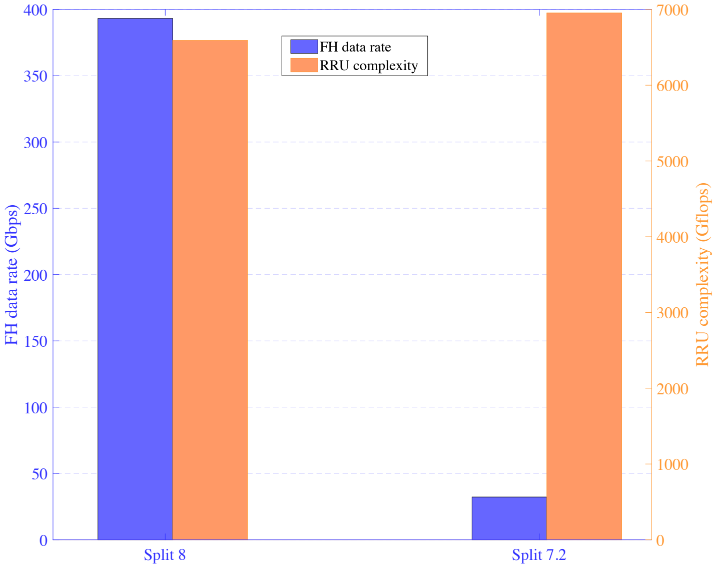


Fig. 6.10. Required FH data rate verses RRU complexity for Split 8 and Split 7.2 [CKBF19].

refers to the rate of operation of FPGA from (6.9) or GPP from (6.10). Fig. 6.9 shows the required number of FPGA or x86 devices based on the net complexity obtained from Fig. 6.8. We can observe that three FPGA devices and seven x86 devices are required for 100 MHz bandwidth. The corresponding power for FPGA would be $3 * P_{\mathrm{FPGA}} =$ 90 W and that of Xeon would be $7 * P_{\mathrm{Xeon}} =$ 980 W. From this calculation, it can be concluded that FPGAs are the only feasible option, especially considering the form-factor and power consumption. On the other hand, we would need $7 \times 18 = 126$ Xeon x86 cores corresponding to a powerful data center of 7 Xeon devices, which is not feasible for a rooftop mounted RRU. Although we calculated the required number of devices for 100 MHz bandwidth, similar analysis can be carried out for any other bandwidth and for other functional splits. However, we want to stress again that these numbers are mostly valuable for the sake of comparison and should not be seen as the definitive absolute values of practical implementations.

Fig. 6.10 depicts the RRU complexity and required fronthaul bandwidth (refer to, Section 2.1.2 for fronthaul bandwidth calculation) for Split 8 and Split 7.2. It is clear that RRU complexity increases for Split 7.2 compared to Split 8. However, at the same time, the required fronthaul bandwidth is significantly reduced, which is one of the main motivations for using XRAN functional Split 7.2.

6.5 Chapter Summary

In this work, complexity of the RRU using xRAN fictional Split 7.2 with detailed RF chain is analyzed. We presented complexity results for flexible numerology based on uniform and non-uniform bandwidth divisions and compared them with a single 100 MHz system bandwidth based on long FFT. In addition, based on the total computed RRU complexity, the required number of FPGAs or GPPs is calculated. We showed that FPGA is more feasible option compared with Xeon x86 processors. Moreover, we found that the most computational complexity entity is DPD, as it requires more processing power in order to compensate non-linear distortion of the PA. Although in practice it is quite involved to calculate the exact number of the required FPGA or Xeon devices, detailed insights to approximate their numbers are presented. Note that results are implementation specific. However, they provide comparative insights and the trends are likely to follow the presented results. The analysis can be extended to any other splits and for different platforms with varying processor architectures and models.

Chapter 7

Conclusion and Outlook

7.1 Core Findings and Summary

C-RAN offers several promising centralization/virtualization benefits. However, these advantages are threatened by the stringent fronthaul bandwidth and low-latency requirements, which are foreseen to be even more challenging for future RATs employing massive MIMO, mmWave or carrier aggregation techniques. This thesis solely focused on these two constraints and provided valuable contributions. The main contributions along with the key results and conclusions are briefly summarized below.

For bandwidth-constrained fronthaul, statistical multiplexing gains were investigated. For this, an appropriate functional split that is closely associated with actual user data rate, unlike a constant bitrate CPRI-like split, was justifiably chosen, and spatial traffic model and queuing theory were employed. The impacts of the statistical parameters – traffic density, correlation distance and outage probability – on the statistical multiplexing gain and the required fronthaul bandwidth were shown. In addition, an iterative pilot optimization algorithm was developed to show the impact of number of pilots on statistical multiplexing gain and it was shown that an additional reduction in fronthaul segments can be achieved, which leads a larger optimization gain up to 25%. At the end, a cost optimization of fronthaul transceiver module was developed, whereby it was shown that fronthaul transceiver cost saving up to 50% be obtained at a moderately low traffic density in the investigated scenarios.

The next focus of the thesis was fronthaul latency. An analytical framework is presented to analyze the UL latency in the packetized C-RAN fronthaul. A continuous-time queuing model for the Ethernet switch in the fronthaul network, which aggregates the UL traffic from several massive MIMO-aided RRUs, is presented. The closed-form expressions for the sojourn time, waiting time and queue length distributions were derived using Pollaczek–Khinchine formula for our M/HE/1 queuing model. In addition, insights on fronthaul network dimensioning was provided in terms of packet loss rate and fronthaul latency. Due to the slotted nature of UL transmissions, the earlier work with a continuous-time queuing model was extended to a novel discrete-time queuing model, whereby closed-form expressions for the generating functions of steady-state queue length and sojourn time distributions were derived. The impact of the packet arrival rate, average packet size, SE of users, and fronthaul capacity on the sojourn time, waiting time and queue length distributions are analyzed.

Finally, a tradeoff analysis between the required fronthaul bandwidth and RRU complexity was performed, considering the 5G NR flexible numerology and XRAN functional split. The required number of FPGAs or GPPs to support the overall RRU complexity was calculated. It was found that FPGA is more feasible option compared with x86 in terms of power consumption, particularly for rooftop-mounted RRU and the most computationally complex entity is the DPD filter.

While keeping the acclaimed C-RAN benefits but on the other hand, mitigating the foreseen challenging fronthaul requirements, it is expected that future RATs will need an optimal functional split. However, choice of an optimal split is largely dependent on the use cases and application scenarios.

To sum up, the main contribution of the thesis was to analyze the bandwidth and latency constrained C-RAN fronthaul. To this direction, we analyzed the possible statistical multiplexing gains in the fronthaul, presented analytical framework to compute the queuing delays at the Ethernet switch. In addition, a tradeoff between the required fronthaul bandwidth and the RRU complexity was presented.

7.2 Recommendations for Future Work

Although we have analysed fronthaul bandwidth and latency issues in the packet-based fronthaul, some associated extensions can be made to consider future works, focusing on the following research directions:

- The current implementation dealt with a homogeneous traffic source. However, this can be extended to include heterogeneous traffic sources for multiple applications with different QoS requirements and prioritizations, and in addition, different scheduling policies can be also added. This extension will provide more in-depth insights. In addition to the multiplexing gain in the fronthaul segment, a similar concept can be used to study multiplexing gain in the BBU pool. Due to tidal effect, since the average utilization of the RRU is much less than the peak utilization, a few RRUs can be turned off, leading to energy savings. Hence, the fronthaul transceiver cost saving analysis in the thesis can be complemented with the energy saving. Both the cost and energy savings will provide more fronthaul dimensioning insights, which will be a good contribution for the mobile network operators.

- The closed-form solution results for UL latency modeling, which were validated numerically and analytically in this thesis work can be further extended to complement their validation by means of experimental testbed, e.g., using software-defined networking and open air interface [MRMD17, MCM$^+$17, CNS16, CSN$^+$16]. Another possible extension would be study on the choice of a higher queuing delay percentile, considering the worst case delay compared with the average values, as the average values do not provide sufficient information. Furthermore, to account for the end-to-end delay, the uplink fronthaul latency modelling can be extended to account for downlink delay in fronthaul as well.

- The analytical framework presented the closed-form expressions in terms of the moment generating functions. This still requires to use the inverse Laplace transform or inverse Z-transform to find out the distributions of queue length and sojourn time. Hence, a possible extension would to find the analytical closed-form expressions directly in terms of their distributions.

- Although Ethernet is preferred due to it cost effectiveness and widespread use, Ethernet requires very tight synchronization and low-jitter. Future works could consider the extension of this work for jitter and synchronization issues.

- Each functional split has its own delay requirements. On the hand, the 5G traffic classes eMBB, mMTC and URLLC have a varying requirements on the reliability that should also be fulfilled by the fronthaul. Hence, a good extension would be to investigate the fronthaul delay and reliability tradeoff for different functional splits and traffic classes.

- The required fronthaul bandwidth and RRU complexity analysis can be further extended with other functional splits. LTE eNB complexity can be further compared with that of 5G gNB. Moreover, additional insights can be obtained by comparing the RRU complexity with FPGA, GPP and application-specific integrated circuit (ASIC).

Appendix A

Appendix

A.1 Derivation of K_i for exponentially distributed file size

$$k_i = \int_0^\infty \frac{(\lambda x)^i}{i!} \exp(-\lambda x) f_S(x) dx \tag{A.1}$$

$$= \int_0^\infty \frac{(\lambda x)^i}{i!} \exp(-\lambda x) \left(\sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} f_{S_{lk}}(x) \right) dx$$

$$= \int_0^\infty \frac{(\lambda x)^i}{i!} \exp(-\lambda x) \left(\sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(1/\mu_{lk} \, \exp(-x/\mu_{lk})\right) \right) dx$$

$$= \frac{\lambda^i}{i!} \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{p_{lk}}{\mu_{lk}} \int_0^\infty (x)^i \exp(-(\lambda + {\mu_{lk}}^{-1})x) dx \tag{A.2}$$

$$\text{Let } (\lambda + {\mu_{lk}}^{-1})x = y \Rightarrow dx = \frac{1}{\lambda + {\mu_{lk}}^{-1}} dy$$

$$= \frac{\lambda^i}{i!} \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{p_{lk}}{\mu_{lk}} \int_0^\infty \left(\frac{y}{\lambda + {\mu_{lk}}^{-1}} \right)^i \exp(-y) \frac{1}{\lambda + {\mu_{lk}}^{-1}} dy$$

$$= \frac{\lambda^i}{i!} \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{p_{lk}}{\mu_{lk}} \frac{(\lambda + {\mu_{lk}}^{-1})^{-1}}{(\lambda + {\mu_{lk}}^{-1})^i} \underbrace{\int_0^\infty y^i \exp(-y) dy}_{i!}$$

$$\Rightarrow k_i = \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(\frac{\Lambda}{\Lambda + {\mu_{lk}}^{-1}} \right)^i \left(\frac{{\mu_{lk}}^{-1}}{\Lambda + {\mu_{lk}}^{-1}} \right). \tag{A.3}$$

A.2 Derivation of K_i for gamma distributed file size

$$k_i = \int_0^\infty \frac{(\lambda x)^i}{i!} \exp(-\lambda x) f_S(x) dx \tag{A.4}$$

$$= \int_0^\infty \frac{(\lambda x)^i}{i!} \exp(-\lambda x) \left(\sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} f_{S_{lk}}(x) \right) dx$$

$$= \int_0^\infty \frac{(\lambda x)^i}{i!} \exp(-\lambda x) \left(\sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(\frac{x^{a-1} \exp(-x/c_{lk})}{{c_{lk}}^a \Gamma(a)} \right) \right) dx$$

$$= \frac{\lambda^i}{i!} \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{p_{lk}}{{c_{lk}}^a \Gamma(a)} \int_0^\infty (x)^{i+a-1} \exp(-(\lambda + {c_{lk}}^{-1})x) dx \tag{A.5}$$

$$\text{Let } (\lambda + {c_{lk}}^{-1})x = y \Rightarrow dx = \frac{1}{\lambda + {c_{lk}}^{-1}} dy$$

$$= \frac{\lambda^i}{i!} \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{p_{lk}}{{c_{lk}}^a \Gamma(a)} \int_0^\infty \left(\frac{y}{\lambda + {\mu_{lk}}^{-1}} \right)^{i+a-1} \exp(-y) \frac{1}{\lambda + {c_{lk}}^{-1}} dy$$

$$= \frac{\lambda^i}{i!} \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{p_{lk}}{{c_{lk}}^a \Gamma(a)} \frac{1}{(\lambda + {c_{lk}}^{-1})^{i+a}} \underbrace{\int_0^\infty y^{i+a-1} \exp(-y) dy}_{\Gamma(i+a)}$$

$$\Rightarrow k_i = \frac{\Gamma(i+a)}{i!\Gamma(a)} \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(\frac{\Lambda}{\Lambda + {c_{lk}}^{-1}} \right)^i \left(\frac{{c_{lk}}^{-1}}{\Lambda + {c_{lk}}^{-1}} \right)^a . \tag{A.6}$$

When a is a positive integer, $\Gamma(a) = (a-1)!$

$$\text{Hence, } k_i = \frac{(i+a-1)!}{i!(a-1)!} \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk} \left(\frac{\Lambda}{\Lambda + {c_{lk}}^{-1}} \right)^i \left(\frac{{c_{lk}}^{-1}}{\Lambda + {c_{lk}}^{-1}} \right)^a . \tag{A.7}$$

A.3 Derivation of $\Psi_S(s)$ for gamma distributed file size

$$\begin{aligned}
\Psi_S(s) &= \mathcal{L}\left\{f_S(x)\right\} && \text{(A.8)}\\
&= \int_{-\infty}^{+\infty} f_S(x)\exp(-sx)dx\\
&= \int_{0}^{\infty}\left(\sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk} f_{S_{lk}}(x)\right)\exp(-sx)dx\\
&= \int_{0}^{\infty}\left(\sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{x^{a-1}\exp(-x/c_{lk})}{{c_{lk}}^{a}\Gamma(a)}\right)\right)\exp(-sx)dx\\
&= \sum_{l=1}^{L}\sum_{k=1}^{K}\frac{p_{lk}}{{c_{lk}}^{a}\Gamma(a)}\int_0^\infty (x)^{a-1}\exp(-(s+{c_{lk}}^{-1})x)dx && \text{(A.9)}\\
\text{Let } (s+{c_{lk}}^{-1})x = y \Rightarrow dx &= \frac{1}{s+{c_{lk}}^{-1}}dy\\
&= \sum_{l=1}^{L}\sum_{k=1}^{K}\frac{p_{lk}}{{c_{lk}}^{a}\Gamma(a)}\int_0^\infty\left(\frac{y}{s+{c_{lk}}^{-1}}\right)^{a-1}\exp(-y)\frac{1}{s+{c_{lk}}^{-1}}dy\\
&= \sum_{l=1}^{L}\sum_{k=1}^{K}\frac{p_{lk}}{{c_{lk}}^{a}\Gamma(a)}\frac{1}{(s+{c_{lk}}^{-1})^{a}}\underbrace{\int_0^\infty y^{a-1}\exp(-y)dy}_{\Gamma(a)}\\
\Rightarrow \Psi_S(s) &= \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{{c_{lk}}^{-1}}{s+{c_{lk}}^{-1}}\right)^{a}. && \text{(A.10)}
\end{aligned}$$

A.4 Derivation of $F_B(x)$

The CDF $F_B(x)$ of B conditioned on an arrival event $\{N > 0\}$ is the probability $F_B(x) = \mathbb{P}(B < x|N > 0) = \mathbb{P}(B < x, N > 0)/\mathbb{P}(N > 0)$. The probability of an arrival $\mathbb{P}(N > 0) = 1 - \exp(-\Lambda\mathcal{T})$. To evaluate $\mathbb{P}(B < x, N > 0)$, we use the law of total probability to get $\mathbb{P}(B < x, N > 0) = \sum_{m=1}^{\infty}\mathbb{P}(B < x, N = m)$. Enumerating over all the possibilities of m arrivals from LK users, we get

$$\begin{aligned}
\mathbb{P}(B<x,N>0) = &\sum_{m=1}^{\infty}\sum_{\substack{n_{11},\dots,n_{LK}\geq 0\\ \sum_{l=1}^{L}\sum_{k=1}^{K} n_{lk}=m}}\mathbb{P}\left(\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{n=1}^{n_{lk}} G_{lk}^{(n)} < x\right)\\
&\times\mathbb{P}\big(N_{11}=n_{11},\dots,N_{LK}=n_{LK}|N=m\big)\mathbb{P}\big(N=m\big). \qquad \text{(A.11)}
\end{aligned}$$

The first probability term in the summation is the CDF of a sum of Erlang RVs with scale parameters $\mu_{11},\dots,\mu_{LK}$ and shape parameters $n_{11},\dots,n_{LK}$. As shown in [LJ15], it is given by $1 - \boldsymbol{\vartheta}^T\exp(x\boldsymbol{M})\mathbf{1}$. The second term is the probability of partitioning m Poisson arrivals among LK users and is given by $\binom{m}{n_{11}\cdots n_{LK}}\prod_{l=1}^{L}\prod_{k=1}^{K}(\lambda_{lk}/\Lambda)^{n_{lk}}$. Lastly,

$\mathbb{P}\big(N = m\big) = (\Lambda \mathcal{T})^m \exp(-\Lambda \mathcal{T})/m!$. Putting all these together, we get the expression for $F_B(x)$ in (5.4).

List of Abbreviations

1G	first generation mobile communication systems
2G	second generation mobile communication systems
3G	third generation mobile communication systems
3GPP	third generation partnership project
4G	fourth generation mobile communication systems
5G	fifth generation mobile communication systems
5GPPP	fifth generation public private partnership
ACK	acknowledgement
ADC	analog to digital converter
AR	augmented reality
A-RoF	analog radio over fiber
ARPU	average revenue per unit
ASIC	application-specific integrated circuit
ATM	asynchronous transfer mode
AWGN	additive white Gaussian noise
BBU	base band unit
BC	bandwidth-constrained
BER	bit error rate
BS	base station
CA	carrier aggregation
CAPEX	capital expenditure
CBR	constant bitrate

CCDF	complementary cumulative distribution function
CDF	cumulative distribution function
CN	core network
CoMP	coordinated multi-point
CPRI	common public radio interface
CP	cyclic prefix
CP-OFDM	cyclic-prefix orthogonal frequency division multiplexing
CR	cognitive radio
C-RAN	cloud radio access network
CSI	channel state information
CT	continuous-time
CTQ	continuous-time queuing
CU	control unit
DA	digital to analog
DAC	digital to analog converter
DL	downlink
DPD	digital predistortion
D-RAN	distributed radio access network
D-RoF	digital radio over fiber
DSP	digital signal processor
DT	discrete-time
DTQ	discrete-time queuing
DU	distribution unit
DWDM	dense wavelength division multiplexing
E2E	end-to-end
eCPRI	enhanced CPRI
eICIC	enhanced intercell interference coordination

eMBB	enhanced mobile broadband
eNB	evolved NodeB
ETSI	European Telecommunications Standards Institute
FDD	frequency division duplexing
FFT	fast Fourier transform
FH	fronthaul
FIFO	first in first out
FLOPS	floating point operations per second
FPGA	field programmable gate array
FR	frequency range
gNB	next generation NodeB
GPP	general purpose processor
HARQ	hybrid automatic repeat request
HLS	higher-layer split
INI	inter-numerology interference
I/Q	in-phase/quadrature-phase
IEEE	institute of electrical and electronics engineers
IFFT	inverse fast Fourier transform
i.i.d.	independent and identically distributed
IoT	Internet of Things
ISD	intersite distance
ITS	intelligent transport system
JR	joint reception
JT	joint transmission
LoS	line-of-sight
LIFO	last in first out
LLR	log-liklihood ratio

LLS	lower-layer split
LTE	long-term evolution
MAC	medium access control
MGF	moment generating function
MIMO	multiple-input multiple-output
MNO	mobile network operator
mMTC	massive machine-type communication
mmWave	millimeter wave
NACK	negative-acknowledgement
NFV	network function virtualization
NGFI	next generation fronthaul interface
NGMN	next generation mobile network
NR	New Radio
OAI	open air interface
OBSAI	open base station architecture initiative
OFDM	orthogonal frequency division multiplexing
OLT	optical line terminal
ONU	optical network unit
OOB	out-of-band
OPEX	operational expenditure
OPS	operations per seconds
ORI	open radio interface
OTN	optical transport network
PA	power amplifier
PAPR	peak-to-average power ratio
PDF	probability density function
PDCP	packet data convergence control

PHY	physical layer
PLR	packet loss rate
PMF	probability mass function
PON	passive optical network
PRB	physical resource block
PS	processor sharing
PTP	Precision Time Protocol
QAM	quadrature amplitude modulation
QoE	quality of experience
QoS	quality of service
RAN	radio access network
RAT	radio access technology
RE	resource element
RF	radio frequency
RLC	radio link control
RoE	radio over Ethernet
RoF	radio over fiber
RRC	radio resource control
RRH	remote radio head
RRU	remote radio unit
RTT	round-trip time
RU	remote unit
RV	random variable
TBS	transport block size
TDD	time division duplexing
TSN	Time-Sensitive Networking
SDN	software-defined networking

SE	spectral efficiency
SFBD	single fiber bidirection
SINR	signal-to-interference-plus-noise ratio
SME	small and medium-sized enterprises
SMS	short message service
SoTA	state-of-the-art
SyncE	Synchronous Ethernet
TCO	total cost of ownership
TDD	time division duplexing
TRX	transceiver
TSON	time shared optical networks
TTI	transmission time interval
UE	user equipment
UL	uplink
URLLC	ultra-reliable and low latency communication
VBR	variable bitrate
VR	virtual reality
WDM	wavelength division multiplexing
WDM-PON	wavelength division multiplexing passive optical network
ZF	zero-forcing
a.k.a.	also known as
c.f.	compare
i.e.	that is
e.g.	for example
w.p.	with probability
w.r.t.	with respect to

List of Symbols

Units of measure

bps	bits per second
B	Byte
dB	Decibel
dBm	Decibel (related to 10^{-3} W)
Hz	Hertz (1Hz = s^{-1})
m	Meter
s	Second
µs	Microsecond (10^{-6} s)
W	Watt

Number Fields

$\mathbb{N}$	Natural numbers
$\mathbb{R}$	Real numbers

Operators and Functions

$\cdot^*$	Complex conjugate operator
$*$	Convolution operator
$\in$	Element of operator
!	factorial, n! $= \prod_{i=1}^{n} i$
$(\cdot)^H$	Hermitian operator (conjugate transpose)
$(\cdot)^T$	Transpose operator
arg max$[f(x)]$	value of x that maximizes the function $f(x)$
$\mathbb{P}[\cdot]$	Probability operator
$\mathbb{E}[\cdot]$	Expectation operator
f_X	PDF of a RV X
F_X	CDF of a RV X
$\ln(x)$	Natural log of x
$\log_x(y)$	log, base x, of y
$\mathcal{N}(\mu, \sigma^2)$	Gaussian (normal) distribution with mean μ and variance σ^2

$\Psi_X(s)$	Moment generating function for random variable X
$\lceil a \rceil$	Represents the smallest integer greater than or equal to a

Variables and Symbols

$\mathcal{A}_c$	Area served by cell c
B_{coh}	Coherence bandwidth
$\mathcal{B}_{ik}$	Set of users that use the same pilot sequence as user t in cell i
$\mathcal{C}$	A set of C RRUs
C_{Total}	Total complexity associated with flexible numerology
C_{FFT}	Complexity associated with FFT computation
C_{FPGA}	Maximum clock speed of FPGA
$C_{\mathrm{Interpolation}}$	Complexity of interpolation filter
C_{Mixer}	Complexity associated with mixer
$C_{100\ \mathrm{MHz}}^{\Delta f=15\ \mathrm{kHz}}$	Complexity of a single 100 MHz bandwidth
$\sum C_{\mu}^{\mathrm{nonuniform}}$	Sum complexity of all individual sub-bands with nonuniform bandwidth division
$\sum C_{\mu}^{\mathrm{uniform}}$	Sum complexity of all individual sub-bands with uniform bandwidth division
C_{Xeon}	Maximum clock speed of Xeon
d_{corr}	Correlation distance
$D_{\mathrm{req,FH}}$	Required fronthaul bandwidth corresponding to $S_{\mathcal{C},\mathrm{O}}$
D_{CPRI}	CPRI data rate for Option 8 (Split A)
G	Statistical multiplexing gain
f_{C}	Carrier frequency
f_n	Normalized rate per user
f_{s}	Sampling frequency
$\overline{F}$	Mean file size
$\mathcal{F}$	Noise figure
$\mathbf{h}_{lk,l}$	Complex baseband channel gain vector from user k in cell i to RRU l
g_1	Optimization gain in FH Segment I
g_2	Optimization gain in FH Segment II
K_{c}	Number of UE Antennas
$K_{\mathrm{max},c}$	Maximum number of users in a cell c
M_{c}	Number of Antennas in cell c
M_{d}	Number of Antennas in neighbouring cell d
n_{CU}	Number of cost unit
N_{Ant}	Number of antennas
N_{CMA}	Number of complex multiplications and additions
N_{cores}	Number of GPP cores

N_{CP}	Cyclic prefix length
N_{FIR}	Order of FIR filter
N_{FPGA}	Number of FPGA devices
$N_{\mathrm{OP,CMA}}$	Number of non-complex operations
N_{Layer}	Number of spatial layers
N_{Port}	Number of ports
$N_{\mathrm{Q,opt8}}$	Quantizer bit resolution per I/Q dimension for Option 8
$N_{\mathrm{Q,opt7.1}}$	Quantizer bit resolution per I/Q dimension for Option 7.1
$N_{\mathrm{Q,opt7.2}}$	Quantizer bit resolution per I/Q dimension for Option 7.2
$N_{\mathrm{Q,opt7.3}}$	Quantizer bit resolution per I/Q dimension for Option 7.3
$N_{\mathrm{Q,opt6}}$	Quantizer bit resolution per I/Q dimension for Option 6
N_{RB}	Number of resource blocks
$N_{\mathrm{sub,act}}$	Number of active subcarriers
$N_{\mathrm{SC}}^{\mathrm{RB}}$	Number of subcarriers per RB
$N_{\mathrm{sym}}^{\mathrm{SF}}$	Number of OFDM symbols per subframe
N_{taps}	Number of FIR filter taps
N_{Xeon}	Number of Xeon devices
p	Average power per antenna
P_{CU}	Actual cost of each cost unit
P_{FPGA}	Power consumption of FPGA
p_{lk}	Uplink transmit power of user k in cell l
p_{ue}	Maximum UL transmit power of user
P_{O}	Outage probability
P_{Xeon}	Power consumption of Xeon
R_c	Average rate per user in cell c
R_{FPGA}	Computational complexity of FPGA
R_i	Capacity of i^{th} transceiver
R_{Xeon}	Computational complexity of Xeon
$R_{\mathrm{BF,tot}}$	Total rate of operation of a beamformer
S_1	Relative capacity in FH Segment I
S_2	Relative capacity in FH Segment II
$S_{c,\mathrm{O}}$	Outage capacity in FH Segment I
$S_{\mathcal{C},\mathrm{O}}$	Outage capacity in FH Segment II
T_{CP}	Cyclic prefix duration
T_{OFDM}	OFDM symbol duration
T_{SF}	Subframe duration
$\mathcal{T}$	Slot duration
W	Channel bandwidth

α_{ik}	Activity factor of user k in cell l
$\alpha_c(x,y)$	Pathloss factor at a location (x,y)
β	Pilot reuse factor
$\beta_{ik,l}$	Large-scale fading coefficient
η	resource overhead
$\mathcal{K}_{lk}$	Set of all users in the network except user k in cell l
μ	Utilization of subcarriers, i.e., load
μ_{FPGA}	Maximum load utilization of FPGA
μ_{Xeon}	Maximum load utilization of Xeon
$\gamma(x,y)$	Signal-to-noise-plus-interference ratio at a location (x,y)
γ	CPRI overhead
γ_{CW}	CPRI control word overhead
γ_{LW}	Line coding overhead
τ_{coh}	Coherence time
τ_{c}	Coherence interval length
τ_{p}	Number of pilots
ζ_{opt8}	Overhead for Option 8
$\zeta_{\text{opt7.1}}$	Overhead for Option 7.1
$\zeta_{\text{opt7.2}}$	Overhead for Option 7.2
$\zeta_{\text{opt7.3}}$	Overhead for Option 7.3
ζ_{opt6}	Overhead for Option 6
$\zeta^{(\text{ul})}$	Fraction of UL data transmission
$\zeta^{(\text{dl})}$	Fraction of DL data transmission
$\bar{\gamma}_c$	Average SINR in the serving area of cell c
λ_c	User arrival rate
$\Omega(x,y)$	Traffic density at a location (x,y)
$\bar{\Omega}$	Mean Traffic density at a location (x,y)
σ_Ω	Standard deviation of traffic density
$\pi_c(n)$	Steady state probabilities of the number of users served by an RRU c
$\pi_c(0)$	Steady state probability that there is no user in a cell
$\Pi_c(n)$	CDF of number of users in each cell
$\pi_{\mathcal{C}}$	PDF of user streams in FH Segment II
$\Pi_{\mathcal{C}}$	CDF of user streams in FH Segment II
Ψ	Cost factor of i^{th} transceiver
Γ	Maximum transceiver bandwidth
w_i	Number of i^{th} transceiver
ζ	Cost Optimization function
ϕ_{abs}	Relative transceiver cost saving
ϕ_{abs}	Absolute transceiver cost saving
ϕ_{per}	Percentage transceiver cost saving
Δf	Flexible subcarrier spacing

List of Figures

List of Tables

Bibliography

[3GP10] 3GPP. 3GPP TR 36.814 v9.0.0 (2010-03): Further advancements for E-UTRA physical layer aspects (release 9). 2010.

[3GP17a] 3GPP. 3GPP TR 38.801 v14.0.0 (2017-03): Study on new radio access technology: Radio access architecture and interfaces (Release 14). 2017.

[3GP17b] 3GPP. 3GPP TR R1-163224: Feasibility of mixing numerology in an OFDM system (Release 15). 2017.

[3GP17c] 3GPP. 3GPP TR R1-164623: Mixed numerology in an OFDM system (Release 15). 2017.

[3GP18a] 3GPP. Technical report, 3rd Generation Partnership Project (3GPP), 2018.

[3GP18b] 3GPP. 3GPP TR 38.912 v15.0.0 (2018-06): Study on new radio (NR) access technology (Release 15). 2018.

[3GP19] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio access capabilities (Release 15). Technical Specification (TS) 36.306, 3rd Generation Partnership Project (3GPP), June 2019. Version 15.5.0.

[5G-a] 5G-XHaul Project, Deliverable D2.1. Requirements specification and KPIs document. [Online]. Available: `http://http://www.5g-xhaul-project.eu/download/5G-XHaul_D_21.pdf`. Accessed Feb. 10, 2017.

[5G-b] 5G-XHaul Project, Deliverable D4.12. Advanced antenna systems for radio access with integrated L1 processing.

[5G-18] 5G-PICTURE Project, Deliverable D3.2. Intermediate report on data plane programmability and infrastructure components, Nov 2018.

[5GP19] 5GPPP. 5G PPP architecture working group view on 5G architecture (version 3.0), white paper, June 2019. [Online]. Available: `https://5g-ppp.eu/wp-content/uploads/2019/07/5G-PPP-5G-Architecture-White-Paper_v3.0_PublicConsultation.pdf`. Accessed July 10, 2019.

[AFSP13] I. Atencia, I. Fortes, S. Sánchez, and A. Pechinkin. A discrete-time queueing system with different types of displacement. In *ECMS*, 2013.

[AR15] I. Adan and J. Resing. Lecture notes on Queueing Systems, department of mathematics and computing science, eindhoven university of technology. Online: https://www.win.tue.nl/ iadan/queueing.pdf. Accessed July 19, 2019, March 2015.

[ATM18] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro. Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions. *IEEE Communications Surveys Tutorials*, 20(1):708–769, Firstquarter 2018.

[AZH$^+$18] P. Assimakopoulos, J. Zou, K. Habel, J. Elbers, V. Jungnickel, and N. J. Gomes. A converged evolved ethernet fronthaul for the 5G era. *IEEE Journal on Selected Areas in Communications*, 36(11):2528–2537, Nov 2018.

[B$^+$17] J. Bartelt et al. 5G transport network requirements for the next generation fronthaul interface. *EURASIP Journal on Wireless Communications and Networking*, 2017(1):89, May 2017.

[BB09] F. Baccelli and B. Blaszczyszyn. *Stochastic Geometry and Wireless Networks, Volume I - Theory*, volume 1 of *Foundations and Trends in Networking Vol. 3: No 3-4, pp 249-449*. NoW Publishers, 2009. Stochastic Geometry and Wireless Networks, Volume II - Applications; see http://hal.inria.fr/inria-00403040.

[BC59] S. Barnard and J. M. Child. *Higher Algebra.* Macmillan & Co. Ltd., 1959.

[BCV18] S. Bjørnstad, D. Chen, and R. Veisllari. Handling delay in 5G Ethernet mobile fronthaul networks. In *European Conference on Networks and Communications (EuCNC)*, pages 1–9, June 2018.

[BJDO14] E. Björnson, E. A. Jorswieck, M. Debbah, and B. Ottersten. Multiobjective signal processing optimization: The way to balance conflicting metrics in 5G systems. *IEEE Signal Processing Magazine*, 31(6):14–23, Nov 2014.

[BL15] E. Björnson and E. G. Larsson. Three practical aspects of massive MIMO: Intermittent user activity, pilot synchronism, and asymmetric deployment. In *Proc. Globecom Workshops*, pages 1–6, Dec. 2015.

[BLD16] E. Björnson, E. G. Larsson, and M. Debbah. Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated? *IEEE Transactions on Wireless Communications*, 15(2):1293–1308, Feb 2016.

[Bos02] S. K. Bose. *An Introduction to Queueing Systems.* Kluwer/Plenum, 2002.

[BRW$^+$15] J. Bartelt, P. Rost, D. Wübben, J. Lessmann, B. Melis, and G. Fettweis. Fronthaul and backhaul requirements of flexibly centralized radio access networks. *IEEE Wireless Communications*, 22(5):105–111, October 2015.

[CBF17] J. K. Chaudhary, J. Bartelt, and G. Fettweis. Statistical multiplexing in fronthaul-constrained massive MIMO. In *European Conference on Networks and Communications (EuCNC)*, June 2017.

[CBF18] J. K. Chaudhary, J. Bartelt, and G. Fettweis. Statistical multiplexing and pilot optimization in fronthaul-constrained massive MIMO. *EURASIP Journal on Wireless Communications and Networking*, 2018(1):204, Aug 2018.

[CBL18] T. Van Chien, E. Björnson, and E. G. Larsson. Joint pilot design and uplink power allocation in multi-cell massive MIMO systems. *IEEE Transactions on Wireless Communications*, 17(3):2000–2015, March 2018.

[CCY$^+$15] A. Checko, H.L. Christiansen, Ying Yan, L. Scolari, G. Kardaras, M.S. Berger, and L. Dittmann. Cloud RAN for mobile networks—a technology overview. *Communications Surveys Tutorials, IEEE*, 17(1):405–426, Firstquarter 2015.

[CFBF19a] J. K. Chaudhary, J. Francis, A. N. Barreto, and G. Fettweis. Latency in the uplink of massive MIMO C-RAN with packetized fronthaul: Modeling and analysis. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–7, April 2019.

[CFBF19b] J. K. Chaudhary, J. Francis, A. N. Barreto, and G. Fettweis. Packet loss in latency-constrained Ethernet-based packetized C-RAN fronthaul. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2019.

[Cha] C. Chang. *Cloudification and Slicing in 5G Radio Access Network*. PhD thesis, Sorbonne University, Doctoral School of Informatics, Telecommunications and Electronics of Paris EURECOM.

[Chi13] China Mobile Research Institute. C-RAN - The road towards green RAN (version 3.0). *White Paper*, Dec. 2013.

[Cis19] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022, white paper, 2019. [Online]. Available: `https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.pdf`. Accessed July 10, 2019.

[CKBF19] J. K. Chaudhary, A. Kumar, J. Bartelt, and G. Fettweis. C-RAN employing xRAN functional split: Complexity analysis for 5G NR remote radio unit. In *European Conference on Networks and Communications (EuCNC)*, pages 580–585, June 2019.

[CLHD$^+$14] I. Chih-Lin, J. Huang, R. Duan, C. Cui, J. X. Jiang, and L. Li. Recent progress on C-RAN centralization and cloudification. *IEEE Access*, 2:1030–1039, 2014.

[CLW67] J. W. Cooley, P. Lewis, and P.D. Welch. *The Fast Fourier Transform algorithm and its applications.* IBM Watson Research Center, 1967.

[CNS16] C. Chang, N. Nikaein, and T. Spyropoulos. Impact of packetization and scheduling on C-RAN fronthaul performance. In *IEEE Global Communications Conference (GLOBECOM)*, pages 1–7, Dec 2016.

[COM] COMBO Project, Deliverable D3.3. Analysis of transport network architectures for structural convergence.

[CPR15] CPRI. Common Public Radio Interface (CPRI); Interface Specification (V7.0). Technical report, October 2015.

[CS94] J. Y. Cheah and J. M. Smith. Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows. *Queueing Systems*, 15(1):365–386, 1994.

[CS07] F.R.B. Cruz and J. M. Smith. Approximate analysis of M/G/c/c state-dependent queueing networks. *Computers and Operations Research*, 34(8):2332 – 2344, 2007.

[CSN+16] C. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos, and C. Bonnet. Impact of packetization and functional split on C-RAN fronthaul performance. In *IEEE International Conference on Communications (ICC)*, pages 1–7, May 2016.

[DDM+13] U. Dötsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier. Quantitative analysis of split base station processing and determination of advantageous architectures for LTE. *Bell Labs Technical Journal*, 18(1):105–128, June 2013.

[Deu] Deutsche Telekom. 5G-revolution – not evolution. Online: https://www.telekom.com/en/company/details/5g-revolution-not-evolution-481778. Accessed July 19, 2019.

[dHLA16] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra. An overview of the CPRI specification and its application to C-RAN-based LTE scenarios. *IEEE Communications Magazine*, 54(2):152–159, February 2016.

[eCP17] eCPRI. Common Public Radio Interface (CPRI); eCPRI Interface Specification (V1.0). Technical report, Aug. 2017. [Online]. Available: `http://www.cpri.info/downloads/eCPRI_v_1_0_2017_08_22.pdf`. Accessed Sept. 25, 2017.

[eCP19] eCPRI. Common Public Radio Interface (CPRI); eCPRI Interface Specification (V2.0). Technical report, May 2019. [Online]. Available: `http://www.cpri.info/downloads/eCPRI_v_2.0_2019_05_10c.pdf`. Accessed July 19, 2019.

[FCBF20] J. Francis, J. K. Chaudhary, A. N. Barreto, and G. Fettweis. Uplink latency in massive MIMO-based C-RAN with intra-PHY functional split. pages 1–5, 2020.

[FSM+15] M. Fiorani, B. Skubic, J. Mårtensson, L. Valcarenghi, P. Castoldi, L. Wosinska, and P. Monti. On the design of 5G transport networks. *Photonic Network Communications*, 30(3):403–415, Dec 2015.

[G.818] ITU-T Recommendation G.8262. Timing characteristics of synchronous Ethernet equipment slave clock, Nov. 2018.

[GE16] K. Grobe and J. P. Elbers. Analysis of WDM-PON for next-generation back- and fronthaul. In *Broadband Coverage in Germany; 10. ITG-Symposium*, pages 1–5, April 2016.

[GPR16] Traffic model for legacy GPRS MTC. GP 160060, 3GPP GERAN meeting 69. February 2016.

[GRI+17] L. Gavrilovska, V. Rakovic, A. Ichkov, D. Todorovski, and S. 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, Oct 2017.

[Gro18] xRAN Fronthaul Working Group. XRAN-FH.CUS.0-v02.01: Control, user and synchronization plane specification. 2018.

[HCBJ18] M. M. A. Hossain, C. Cavdar, E. Björnson, and R. Jäntti. Energy saving game for massive MIMO: Coping with daily load variation. *IEEE Transactions on Vehicular Technology*, 67(3):2301–2313, March 2018.

[HG14] A. Haddad and M. Gagnaire. Radio-over-fiber (RoF) for mobile backhauling: A technical and economic comparison between analog and digitized RoF. pages 132–137, May 2014.

[HNHS19] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten. A comprehensive survey of RAN architectures toward 5g mobile communication system. *IEEE Access*, 7:70371–70421, 2019.

[HR16] K. M. S. Huq and J. Rodriguez. *Backhauling/Fronthauling for Future Wireless Systems*. John Wiley & Sons, 2016.

[IEEa] IEEE 1914.1 Task Force. Standard for packet-based fronthaul transport networks. [Online]. Available: `http://sites.ieee.org/sagroups-1914/p1914-1/`.

[IEEb] IEEE 1914.3 Task Force. Standard for radio over Ethernet encapsulations and mappings. [Online]. Available: `http://sites.ieee.org/sagroups-1914/p1914-3/`.

[IEE08] IEEE 1588-2008 - IEEE standard for a precision clock synchronization protocol for networked measurement and control systems, July 2008.

[IEE18] IEEE Std 802.1CM. Time sensitive networking for fronthaul, May 2018.

[IR09] ITU-R. ITU recommendation ITU-R M.2135-1: Guidelines for evaluation of radio interface technologies for IMT-Advanced, Dec. 2009.

[KAP$^+$17] A. E. Kalør, M. I. Agurto, N. K. Pratas, J. J. Nielsen, and P. Popovski. Statistical multiplexing of computations in C-RAN with tradeoffs in latency and energy. *CoRR*, abs/1703.04995, 2017.

[Ken53] D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *Ann. Math. Statist.*, 24(3):338–354, Sept. 1953.

[KM19] A. Kumar and M. Magarini. Symbol error probability analysis of DFrFT-based OFDM systems with CFO and STO in frequency selective rayleigh fading channels. *IEEE Transactions on Vehicular Technology*, 68(1):64–81, 2019.

[KR06] A. Kesselman and A. Rosén. Scheduling policies for CIOQ switches. volume 60, pages 60 – 83, 2006.

[KSF15] H. Klessig, M. Soszka, and G. Fettweis. Multi-cell flow-level performance of traffic-adaptive beamforming under realistic spatial traffic conditions. In *2015 International Symposium on Wireless Communication Systems (ISWCS)*, pages 726–730, Aug 2015.

[LBC18] M. P. Larsen, M. S. Berger, and H. L. Christiansen. Fronthaul for Cloud-RAN enabling network slicing in 5G mobile networks. *Wireless Communications and Mobile Computing*, 2018(3), 2018.

[LC13] J. Lorca and L. Cucala. Lossless compression technique for the fronthaul of lte/lte-advanced cloud-ran architectures. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–9, June 2013.

[LCC19] L. M. P. Larsen, A. Checko, and H. L. Christiansen. A survey of the functional splits proposed for 5g mobile crosshaul networks. *IEEE Communications Surveys Tutorials*, 21(1):146–172, Firstquarter 2019.

[LETM14] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, February 2014.

[LJ15] B. Legros and O. Jouini. A linear algebraic approach for the computation of sums of Erlang random variables. *Applied Mathematical Modelling*, 39(16):4971–4977, 2015.

[LXZN15] J. Liu, S. Xu, S. Zhou, and Z. Niu. Redesigning fronthaul for next-generation networks: beyond baseband samples and point-to-point links. *IEEE Wireless Communications*, 22(5):90–97, October 2015.

[LZZ$^+$14] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang. Spatial modeling of the traffic density in cellular networks. *IEEE Wireless Communications*, 21(1):80–88, February 2014.

[Mar10] T. L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 9(11):3590–3600, November 2010.

[Maz] R. R. Mazumdar. Notes on statistical multiplexing. `https://ece.uwaterloo.ca/~mazum/ECE610/statmux.pdf`.

[MCM$^+$17] G. Mountaser, M. Condoluci, T. Mahmoodi, M. Dohler, and I. Mings. Cloud-RAN in support of URLLC. In *2017 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6, Dec 2017.

[MMM19] G. Mountaser, M. Mahlouji, and T. Mahmoodi. Latency bounds of packet-based fronthaul for Cloud-RAN with functionality split. In *IEEE International Conference on Communications (ICC)*, pages 1–6, May 2019.

[MRMD17] G. Mountaser, M. L. Rosas, T. Mahmoodi, and M. Dohler. On the feasibility of MAC and PHY split in Cloud RAN. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, March 2017.

[NGFa] IEEE Next Generation Fronthaul Interface (1914) working group. http://sites.ieee.org/sagroups-1914/.

[NGFb] NGFI. Next Generation Fronthaul Interface. http://sites.ieee.org/sagroups-1914/.

[NGM15a] NGMN. Further study on critical C-RAN technologies, version 1.0, March 2015. White Paper.

[NGM15b] NGMN Alliance. 5G white paper, Feb. 2015. [Online]. Available: `https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf`. Accessed July 10, 2019.

[NMM$^+$14] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet. Openairinterface: A flexible platform for 5G research. *ACM SIGCOMM Computer Communication Review*, 44(5):33–38, 2014.

[Nok] Nokia Blog. 5G enabled by massive capacity, connectivity. Online: https://www.nokia.com/blog/5g-enabled-massive-capacity-connectivity/. Accessed July 19, 2019, 04/20/2016.

[OBS06] OBSAI. BTS System Reference Document version 2.0, 2006.

[OLH19] G. Otero Pérez, D. Larrabeiti López, and J. A. Hernández. 5G new radio fronthaul network design for eCPRI-IEEE 802.1CM and extreme latency percentiles. *IEEE Access*, 7:82218–82230, 2019.

[ORA] Operator defined next generation RAN architecture and interfaces. `https://www.o-ran.org/`. Accessed: 2019-10-25.

[ORI15] ORI. ORI Interface Specification; Part 1: Low Layers (Release 4), Oct. 2015.

[PCB13] S. Park, C. B. Chae, and S. Bahk. Before/after precoding massive MIMO systems for cloud radio access networks. *Journal of Communications and Networks*, 15(4):398–406, Aug 2013.

[PHL17] G. O. Pérez, J. A. Hernández, and D. L. López. Delay analysis of fronthaul traffic in 5G transport networks. In *2017 IEEE 17th International Conference on Ubiquitous Wireless Broadband (ICUWB)*, pages 1–5, Sept 2017.

[PHL18] G. O. Pérez, J. A. Hernández, and D. Larrabeiti. Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G. *IEEE/OSA Journal of Optical Communications and Networking*, 10(6):573–581, Jun. 2018.

[PM17] Ping-Heng Kuo and A. Mourad. Millimeter wave for 5G mobile fronthaul and backhaul. In *European Conference on Networks and Communications (EuCNC)*, pages 1–5, 2017.

[PWLP15] M. Peng, C. Wang, V. Lau, and H. V. Poor. Fronthaul-constrained cloud radio access networks: insights and challenges. *IEEE Wireless Communications*, 22(2):152–160, April 2015.

[R$^+$13] F. Rusek et al. Scaling up MIMO: Opportunities and challenges with very large arrays. *IEEE Signal Processing Magazine*, 30(1):40–60, Jan 2013.

[R3-16a] R3-161813. Transport requirement for cu&du functional splits options. Technical report, Aug. 2016. CMCC.

[R3-16b] R3-162102. CU-DU split: Refinement for annex A (transport network and RAN internal functional split). Technical report, Oct. 2016. NTT DOCOMO, INC.

[RWN$^+$18] C. Ranaweera, E. Wong, A. Nirmalathas, C. Jayasundara, and C. Lim. 5G C-RAN with optical fronthaul: An analysis from a deployment perspective. *Journal of Lightwave Technology*, 36(11):2059–2068, June 2018.

[SBMD17] P. Sehier, A. Bouillard, F. Mathieu, and T. Deiss. Transport network design for fronthaul. In *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pages 1–5, Sept 2017.

[SD07] L. N. Singh and G. R. Dattatreya. Estimation of the hyperexponential density with applications in sensor networks. *International Journal of Distributed Sensor Networks*, 3(3):311–330, 2007.

[SKKS16] O. Simeone, J. Kang, J. Kang, and S. Shamai. Cloud radio access networks: Uplink channel estimation and downlink precoding. *CoRR*, abs/1608.07358, 2016.

[Sma15] Small Cell Forum. Small cell virtualization functional splits and use cases. Technical report, June 2015. Document 159.05.1.01.

[SS14a] S. M. Shin and H. J. Son. Cpri(1): Emergence of C-RAN/Fronthaul and CPRI overview. March 2014.

[SS14b] H. Son and S. M. Shin. Fronthaul size: Calculation of maximum distance between RRH and BBU. April 2014.

[SXT$^+$19] S. Su, X. Xu, Z. Tian, M. Zhao, and W. Wang. 5G fronthaul design based on software-defined and virtualized radio access network. In *28th Wireless and Optical Communications Conference (WOCC)*, pages 1–5, May 2019.

[TTQL17] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang. System cost minimization in cloud RAN with limited fronthaul capacity. *IEEE Transactions on Wireless Communications*, 16(5):3371–3384, May 2017.

[VBL16] T. Van Chien, E. Björnson, and E. G. Larsson. Joint power allocation and user association optimization for massive MIMO systems. *IEEE Transactions on Wireless Communications*, 15(9):6384–6399, Sep. 2016.

[Vir] J. Virtamo. 38.3143 queueing theory/the M/G/1 queue.

[Woo94] M.E. Woodward. *Communication and computer networks: modelling with discrete-time queues*. Systems Series. IEEE Computer Society Press, 1994.

[WRB$^+$14] D. Wübben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis. Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN. *IEEE Signal Processing Magazine*, 31(6):35–44, Nov 2014.

[WZHW15] J. Wu, Z. Zhang, Y. Hong, and Y. Wen. Cloud radio access network (c-ran): a primer. *IEEE Network*, 29(1):35–41, Jan 2015.

[Xeo] Intel, “Intel Xeon Gold 6140 Processor”, Q3 2017. https://ark.intel.com/products/120485.

[Xil] Zynq UltraScale+ RFSoC: Product Tables and Product Selection Guide. https://www.xilinx.com/support/documentation/selection-guides/zynq-usp-rfsoc-product-selection-guide.pdf.

[YA18] A. Yazar and H. Arslan. Flexible multi-numerology systems for 5G new radio. *Journal of Mobile Multimedia*, 14(4):367–394, 2018.

[YKLV+17] J. Yli-Kaakinen, T. Levanen, S. Valkonen, K. Pajukoski, J. Pirskanen, M. Renfors, and M. Valkama. Efficient fast-convolution-based waveform processing for 5G physical layer. *IEEE Journal on Selected Areas in Communications*, 35(6):1309–1326, 2017.

[YQZ+12] Y. Yan, Y. Qin, G. Zervas, B. Rofoee, and D. Simeonidou. High performance and flexible fpga-based time shared optical network (tson) metro node. In *2012 38th European Conference and Exhibition on Optical Communications*, pages 1–3, Sept 2012.

[YSN+18] M. Yao, M. Sohul, R. Nealy, V. Marojevic, and J. Reed. A digital predistortion scheme exploiting degrees-of-freedom for massive mimo systems. In *IEEE International Conference on Communications (ICC)*, pages 1–5. IEEE, 2018.

[ZTA+11] G. S. Zervas, J. Triay, N. Amaya, Y. Qin, C. Cervello-Pastor, and D. Simeonidou. Time shared optical network (TSON): A novel metro architecture for flexible multi-granular services. In *2011 37th European Conference and Exhibition on Optical Communication*, pages 1–3, Sept 2011.

[ZWE17] J. Zou, C. Wagner, and M. Eiselt. Optical fronthauling for 5G mobile: A perspective of passive metro WDM technology. In *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, pages 1–3, March 2017.

Publications of the Author

Journal and Magazine Publications

[1] **J. K. Chaudhary**, J. Bartelt, and G. Fettweis. Statistical multiplexing and pilot optimization in fronthaul-constrained massive MIMO. *EURASIP Journal on Wireless Communications and Networking*, page 204, 2018.

[2] J. Francis, **J. K. Chaudhary**, A. N. Barreto, and G. Fettweis. Analyzing Uplink Latency of Massive MIMO-based C-RAN with Intra-PHY Functional Split. In *IEEE Communications Letters (CL'19)*, pages 1—-5, 2020.

[3] D. Camps-Mur, J. Gutiérrez, E. Grass, A. Tzanakaki, P. Flegkas, K. Choumas,D. Giatsios, A. F. Beldachi, T. Diallo, J. Zou, P. Legg, J. Bartelt, **J. K. Chaudhary**, A. Betzler, J. J. Aleixendri, R. González, and D. Simeonidou. 5G-XHaul: A Novel Wireless-Optical SDN Transport Network to Support Joint 5G Backhaul and Fronthaul Services. *IEEE Communications Magazine*, pages 1—-6, 2019.

[4] J. Gutiérrez, N. Maletic, D. Camps-Mur, E. García, I. Berberana, M. Anastasopoulos, A. Tzanakaki,V. Kalokidou, P. Flegkas, D. Syrivelis, T. Korakis, P. Legg, D. Markovic, G. Limperopoulos, J. Bartelt, **J. K. Chaudhary**, M. Grieger, N. Vucic, J. Zou, and E. Grass. 5G-XHaul: A Converged Optical and Wireless Solution for 5G Transport Networks. *Transactions on Emerging Telecommunications Technologies*, pages 1—-8, 2016.

Conference Publications

[1] **J. K. Chaudhary**, J. Bartelt, and G. Fettweis. Statistical multiplexing in fronthaul-constrained massive MIMO. In *European Conference on Networks and Communications (EuCNC'17)*, pages 1—-6, Oulu, Finland, 2017.

[2] **J. K. Chaudhary**, J. Zou, and G. Fettweis. Cost Saving Analysis in Capacity-Constrained C-RAN Fronthaul. In *2018 IEEE Globecom Workshops (GC Wkshps)*, pages 1—-7, Abudhabi, UAE, 2018.

[3] **J. K. Chaudhary**, J. Francis, A. N. Barreto, and G. Fettweis. Latency in the Uplink of massive MIMO CRAN with Packetized Fronthaul: Modeling and Analysis. In

IEEE Wireless Communications and Networking Conference(WCNC'19), pages 1—-7, Marrakech, Morocco, 2019.

[4] **J. K. Chaudhary**, A. Kumar J. Bartelt, and G. Fettweis. C-RAN Employing xRAN Functional Split: Complexity Analysis for 5G NR Remote Radio Unit. In *European Conference on Networks and Communications (EuCNC'19)*, pages 1—-6, Valencia, Spain, 2019.

[5] **J. K. Chaudhary**, J. Francis, A. N. Barreto, and G. Fettweis. Packet Loss in Latency-constrained Ethernet-based Packetized C-RAN Fronthaul. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'19)*, pages 1—-6, Istanbul, Turkey, 2019.