IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 6, 2021, accepted October 25, 2021, date of publication October 29, 2021, date of current version November 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3124374

Flexible Functional Split and Fronthaul Delay: A Queuing-Based Model

LUIS DIEZ[1], ALBERTO MARTÍNEZ ALBA[2], (Graduate Student Member, IEEE), WOLFGANG KELLERER[2], (Senior Member, IEEE), AND RAMÓN AGÜERO[1], (Senior Member, IEEE)

[1]Communications Engineering Department, University of Cantabria, 29005 Santander, Spain
[2]Chair of Communication Network, Technical University of Munich, 80333 Munich, Germany

Corresponding author: Luis Diez (ldiez@tlmat.unican.es)

This work was supported in part by the Spanish Government, Ministerio de Economía y Competitividad, Fondo Europeo de Desarrollo Regional (MINECO-FEDER), through the Project FIERCE: Future Internet Enabled Resilient smart CitiEs under Grant RTI2018-093475-AI00; and in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (FlexNets) under Grant 647158.

ABSTRACT We study the delay over virtual RAN (vRAN) topologies, entailing base stations that are divided into centralized and distributed units, as well as the packet-switched fronthaul network that connects them. We consider the use of flexible functional split, where the functions that are executed at each of these two entities can be dynamically shifted. We propose a queuing-based model, which is able to precisely mimic the behavior of such nodes, and we validate it by means of extensive simulations. We also exploit Jackson Networks theory to establish the end-to-end delay over the fronthaul network, allowing us to assess the impact of having different networking policies and conditions (for instance, background traffic or heterogeneous technologies). Thanks to the simulator we can also broaden the analysis, by studying the delay variability. In addition, we conduct an in-depth analysis of the performance exhibited by a realistic network setup, whose particular characteristics might hinder the services performance, due to the longer dwell times at each split configuration. The results evince the validity of the proposed model, even under realistic conditions. We show that it might not be enough to guarantee an average stable operation of the centralized/distributed units, but the traffic load should remain below the slowest service rate, to avoid reaching unacceptable delays. An increase of $>100\times$ is observed in the delay, using the realistic network setup, when these conditions do not hold.

INDEX TERMS 5G, beyond 5G, functional split, virtual RAN (vRAN), Markov chain, quasi-birth-death (QBD) process, Jackson theory.

I. INTRODUCTION

It is well known that one of the most stringent requirements for forthcoming cellular systems (5G and Beyond 5G) comes from the so-called Ultra-Reliable Low Latency Communication (URLLC). End-to-end delay needs to be kept at very low values, to enable an appropriate provisioning of some of the envisaged services (augmented reality, autonomous driving, etc). On the other hand, and considering an architectural look, Radio Access Networks (RAN) have also witnessed a strong shift, and thanks to the SDN and NFV paradigms, operators are deploying many of the functions that were traditionally co-located in the base station (BS) in centralized controllers. While this leverages several advantages (CAPEX/OPEX reduction, cooperation between access elements, etc.), it also imposes new challenges, in particular considering the delay [1].

In this regard, the Cloud-RAN (C-RAN) solution [3], [4] starts from a single centralized entity that controls a number of Remote Radio Heads (RRH). The former executes many of the functions of traditional BSs, and the RRHs implement the radio-frequency (RF) tasks. Although this approach might yield important capacity improvements in the RAN, by exploiting tight cooperation techniques (i.e. CoMP), it also imposes stringent requirements in the network connecting the different entities, both in terms of throughput and delay. In order to overcome the limitations coming from fully centralized solutions, functional-split architectures were proposed to permit the use of multiple centralization levels [5], which could be adapted to the particular network

The associate editor coordinating the review of this manuscript and approving it for publication was Meng-Lin Ku.

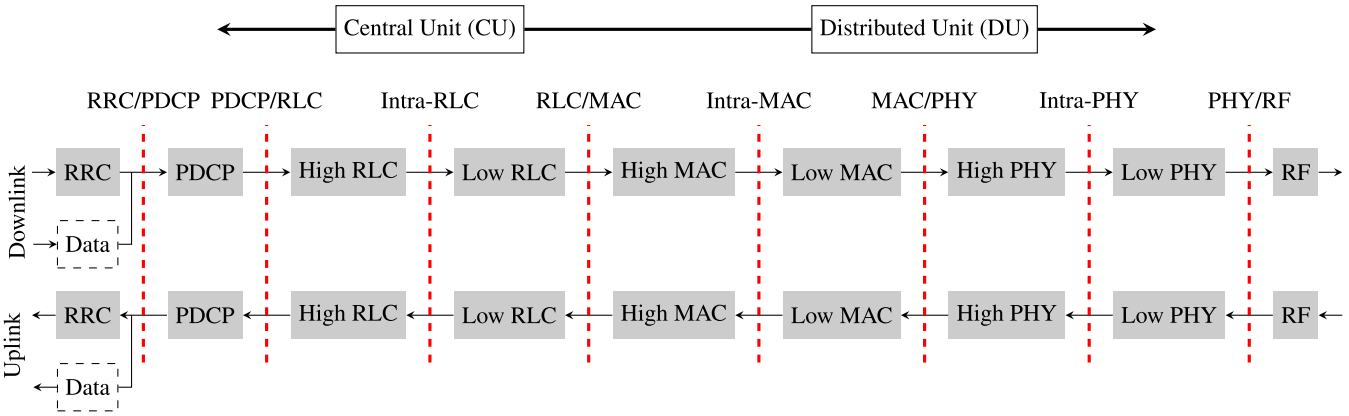


FIGURE 1. Possible functional splits between the CU and DU according to 3GPP. The vertical dashed red line establich the split, so that the function at left are placed in the CU and the functions at right are moved to the DU. [2].

TABLE 1. Overview of requirements, benefits and cons of each split according to [2] for a base station with 100 *MHz* channel bandwidth, 256 QAM modulation and 8 MIMO layers.

Split	One-way latency	DL/UL bandwidth	Benefits	Cons
RRC/PDCP	10 *ms*	$\sim 4/3$ *Gbps*	User plane separation. Potential benefits for edge computing.	Potential issues for security and aggregation.
PDCP/RLC	$1.5 - 10$ *ms*	4/3 *Gbps*	User plane separation and traffic aggregation.	Security configuration issues.
Intra-RLC	$1.5 - 10$ *ms*	$\sim 4/3$ *Gbps*	Traffic aggregation and better flow control. Potential handling of more connected mode UEs	Latency requirements and duplication of buffers.
RLC/MAC	~ 100 *us*	$\sim 4/3$ *Gbps*	No benefit with LTE protocol stack.	–
Intra-MAC	hundreds of *us*	$\sim 4/3$ *Gbps*	Traffic aggregation and better interference management.	Additional scheduling complexity.
MAC/PHY	250 *us*	$\sim 4/5$ *Gbps*	Traffic aggregation, COMP joint transmission, centralized scheduling.	Stringent timing between CU and DU.
Intra-PHY	250 *us*	$10.1 - 22.2/16.6 - 86.1$ *Gbps*	Implementation of advanced receivers	–
PHY/RF	250 *us*	157.3/157.3 *Gbps*	More efficient resource management. Improvement of RF/PHY scalability.	High requirement in fronthaul for latency and bandwidth.

characteristics, traffic features and service requirements. The traditional BS is now divided in three entities, so that the RRH is called the radio unit (RU) and the BS protocol stack is divided between a distributed unit (DU), which is located close to the RU, and a central unit (CU). Hence, CUs can be grouped together in virtualization pools or clusters. Figure 1 depicts the functional splits defined by the 3GPP [2] for LTE, which set the limit between the functions implemented in the CU (at left) and DU (at right). While the splits are defined based on the LTE protocol stack, the specification of the 5G radio access network assumes the split of the base station (i.e. gNB) [6]. Indeed, the protocol stack of LTE and 5G are almost identical in their high level descriptions, but for the new Service Data Adaptation Protocol (SDAP) included in 5G. As can be seen, most of the splits are defined in the protocols boundaries, while a few of them separate functionalities that belong to the same protocol. It is worth noting that the RF functionalities would be located at the RU, so that the PHY/RF split would boil down to a C-RAN solution. The applicability of the different splits is still being analyzed by standardization and industrial bodies like 3GPP and O-RAN [7], where special attention is being payed to low layer splits [8]. According to those analysis, Table 1 summarizes the requirements, benefits and limitations identified for each split by the 3GPP [2].

Altogether, the new RAN is made of physical and virtual entities, leading to the so called virtual RAN (vRAN) [9]. These architectural changes have also imposed new challenges on the fronthaul network that connects such entities, since it needs to provide quality of service (QoS) levels aligned with the configured split. Recent initiatives, such as the Next Generation Fronthaul Interface (NGFI) [10], envision a packet-switched fronthaul network divided in a segment connecting the RU and DU (NGFI-I), and another one, the so-called midhaul, which provides connectivity between the DU and CU (NGFI-II). Eventually, recent studies consider

the dynamic, or flexible, shifting of the selected functional split, to adapt the network to the varying environment [11] (i.e. new services, traffic changes). In this novel scenario, the network adaptation encompasses both the selection of the most appropriate functional split and the re-configuration of the underlying fronthaul network to provide the communication capacity demanded by the selected centralization level.

As an example, Figure 2 shows the functional diagram of a base station with PDCP/RLC split. As can be observed, the CU hosts both RRC and PDCP protocols and it is connected to the gateways of the operator core. On the other hand, the DU implements protocols below PDCP and it is connected to the RU which performs the RF tasks. Finally, NGFI-I and NGFI-II connect both the CU with the DU, and the DU with the RU.

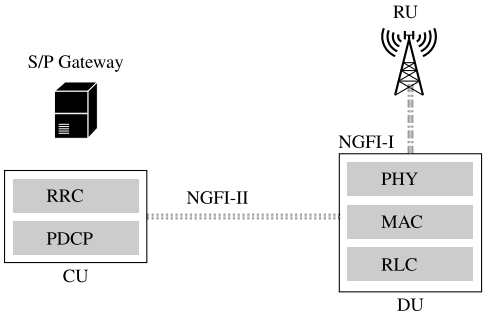


FIGURE 2. Example of PDCP/RLC functional split. The figure depicts the functionalities distribution and location of interfaces.

In this paper we consider a flexible functional split scenario, and we broaden the model that was originally presented in [12], which allows us to accurately predict the delay traversing CU or DU nodes. The main contributions are:

- We first extend the DU/CU model, by considering not only different service rates per split configuration, but also dwell times. In addition, this enhanced model also permits having different stand-by times after each split, and avoids the possibility of going to the same split after leaving it.
- We exploit Jackson Theory and the CU/DU model to yield the overall end-to-end delay, considering the fronthaul packet-switched network. To the best of our knowledge, this is the first work that exploits queuing theory to study the delay for vRAN architecture with Flexible Functional Split, embracing the fronthaul network.
- We exploit an event-driven simulator to assess the validity of the proposed model, which also allows us to study the variability of the observed behavior, and to carry out various complementary analysis.
- Finally, we also discuss the behavior exhibited by a real network configuration, where the particular characteristics might severely hinder the observed performance, due to rather long buffer lengths. We assess how the delay is distributed, and we discuss how this could also be used as a worst-case design parameter.

The proposed model can help to understand the expected behavior of different functional split policies, and to find reasonable limits on the maximum acceptable traffic load. In addition, the developed simulator could also help to derive appropriate queuing management policies.

The rest of the paper is structured as follows. In Section II we provide an in-depth literature review of flexible functional split, and we highlight the main contributions of our paper. Then, Section III introduces the queuing model that is used to characterize the delay of CUs and DUs, and how this can be extended to study the overall fronthaul delay, exploiting Jackson Open Networks Theory. In Section IV we validate the model, comparing its performance with that obtained after a thorough simulation-based study, over a simple scenario with various configurations. In order to assess the validity of the model with more complex scenarios, we use a realistic network deployment in Section V, which has an optimal split policy, and we again compare the results yielded by the proposed model with the behavior obtained by means of simulation-based experiments. We conclude the paper in Section VI, where we summarize the work and identify our future lines of research.

II. RELATED WORK

In recent years the research community has studied functional split solutions from different angles. From a general perspective, some works analyze the capabilities of each split [13], and their performance [12]. These studies have been complemented with other works that mostly focus on the constraints imposed by the fronthaul network [14]. Worth of attention are papers that consider the x-haul (networking resources shared between backhaul and fronthaul) [15]. Along with it, other lines of research have paid attention to the application of regular networking techniques in RAN, which exploit functional splits. Within this group, an overview of scheduling techniques suitable for dynamic functional splits is presented in [16]. In addition, and also from an overall perspective, Lagén *et al.* provide in [17] an overview of architectures supporting functional splits, and the fronthaul compression proposed by 3GPP and O-RAN.

Apart from the literature devoted to assessing the capabilities and requirements of functional splits, other research efforts aim at implementing dynamic functional split solutions. In this regard, a functional split prototype is introduced in [18] to evaluate the impact of low-level split options over the energy consumption, while the authors of [19] depict a 5G RAN implementation that allows dynamic split shifting. Also from this implementation perspective, some works have paid attention to the reconfiguration of the underlying network to support the split requirements. For instance, in [20], [21] the authors analyze and assess solutions to reconfigure the optical fronthaul network, while Chang *et al.* evaluate in [22] the implementation of a framework to enable flexible functional split over Ethernet-based fronthaul. A few works focus on the resources in the access elements, like [23], where the authors propose a new architecture, called F-RAN, to enable split selection considering availability of radio resources. From a different perspective, the authors of [24] propose

a technique to select the functional split, shifting baseband signal precoding, using compression-after-precoding (CAP) or data-sharing (DS) strategies, when it is implemented in the BBU or RRH.

As can be observed, the scope of the aforementioned works is different from ours in essence. The related literature analyzed so far aims to study the goodness and limitation of flexible functional split, and how it can be implemented. On the other hand, our goal is to propose a framework to analyze the system performance when a given split selection policy is applied, in particular focusing on the end-to-end delay. In this sense, some works have addressed the modeling, definition, development and assessment of such policies. For instance, in [25], [26] Harutyunyan *et al.* model the functional split selection as a Virtual Network Embedding (VNE) problem, which is formulated as a Integer Linear Program (ILP). Similarly, the authors of [27] propose an algorithm to allocate Customer Virtual Networks (CVN) considering functional split. Another proposal can be found in [28] where Rodriguez *et al.* suggest a split selection algorithm to ensure efficient utilization of the fronthaul network, while allowing a cooperation among BSs. In this same line, the authors of [29] propose adjusting the split selection to enable URLLC services, exploiting Coordinated Multi-Point techniques.

Furthermore, a number of split selection policies have been proposed to optimize different metrics, taking into account various scenario features. For instance, the data rate is maximized in [30], while end-to-end delay is considered in [31]. Differently, the mobility of users and interference level are taken into account in the split selection solution in [32] and [33], respectively. Worth of mentioning is the work of Temesgene *et al.* [34], where reinforcement learning is applied to select the split in a scenario with small cells. As we have mentioned before, the adaptation of the fronthaul network is a requirement to efficiently implement flexible split solutions. In this sense, Alameer and Sezgin proposed in [35] a solution for allocating functions that jointly tackles resource allocation and routing, using Alternating Direction Method of Multipliers (ADMM). A solution for functional split selection to optimize energy consumption, in scenarios comprising small cells, is proposed in [36], where the optimization problem is posed as a constraint Markov Decision Process (MDP).

Among the works that propose split selection policies, a group of them pay special attention to the optical fronthaul. In this sense, a novel mobile fronthaul architecture to reduce latency in functional split enabled networks is proposed in [37], while the authors of [38] introduce different techniques with the goal of minimizing the delay. Similarly, the available capacity of the optical network is considered in [39], [40] as a constraint in the split selection procedure, while wavelength usage and transponder cost of optical networks is addressed in [41]. Other works focus on the orchestration and split reconfiguration for the optical fronthaul network [42], and on the development of simulation tools to study the impact of flexible functional split solutions over the underlying optical network [43], [44]. As can be observed, although these works develop split selection policies, they model and analyze the system performance according to particular indicators and features. Opposed to that, our paper does not propose a split policy, but it aims to model the system in a generic way, regardless of the specific split selection policy, to understand its behavior and fairly compare different strategies.

Another group of works seek to optimize energy consumption in flexible functional split scenarios. Under this category, the authors of [45] discuss the optimal centralization level, considering energy consumption and midhaul bandwidth. Temesgene *et al.* [46] suggest the use of Q-learning and SARSA algorithms to optimize the placement of functions in terms of energy. In a similar way, an online solution for flexible functional split selection considering energy is proposed in [47], where the problem is formulated as a MDP. The energy consumption, together with functional split, is also studied in [48], using a real implementation based on Open Air Interface (OAI). Energy is also considered in [49], although in this case the flexible functional split is applied to optimize the energy consumption, including baseband processing, in scenarios with Unmanned Aerial Vehicles (UAVs)

Finally, it is also worth mentioning some works that consider the interplay of functional split with techniques used for service deployment and provisioning in cellular networks. In this regard, functional split is considered in [50] as part of network slicing to ensure certain QoS. Following a similar approach, Papa *et al.* study the combination of functional split and network slicing in [51]. From a more generic perspective, the authors of [52] propose a framework to handle heterogeneous RAN, functional split selection, and network slicing for multiple services. Finally, the authors of [53] consider split selection together with task offloading, analyzing the interplay of functional split and fog/cloud services. Once again, these works differ from ours in their scope, since they do not model the behavior of the fronthaul network with functional split.

All in all, after this thorough literature review, we can conclude that the modeling of vRAN has not been sufficiently addressed before. The theoretical model described and evaluated in this paper aims to shed light on the end-to-end performance of the vRAN, in terms of delay, and for any arbitrary split selection policy. In this sense, it is worth remarking that the propose model can be configured to consider any split strategy, and so it would provide an arena to fairly compare them. To our best knowledge, this is the first paper proposing a theoretical model that yields the overall end-to-end delay in vRAN networks, embracing the use of flexible functional split, as well as the impact of the fronthaul network.

III. FRONTHAUL QUEUING MODEL

In this section we discuss the proposed queuing-based model. It encompasses two types of nodes: (1) the one used to reflect the behavior of both CU and DU; and (2) the nodes that are

used, in the packet-switched fronthaul network, to connect CUs and DUs. While the latter will be based on the legacy M/M/1, CUs and DUs, which might implement different splits, require a more complex approach. Hence, we first present the model for CU and DU entities. Afterwards, we exploit Open Jackson Network theory to establish the average end-to-end delay.

Table 2 enumerates all the variables and symbols that are used in the proposed model, including those that are used to solve it.

TABLE 2. Variables and symbols.

	CU/DU nodes
s	Number of slit configurations
λ	Frame arrival rate
μ_j	Service rate of the j^{th} split
$\alpha_{j,k}$	Probability of shifting from j^{th} to k^{th} split $\sum_{k=1}^{s} \alpha_{j,k} = 1,\ \alpha_{j,j} = 0$
γ_j	Change rate for the j^{th} split
ξ_j	Inverse of time at stand-by after leaving j^{th} split
$\pi_i(t)$	Probability of state (i,t) There are i frames in the node when: (1) t odd, using split $j:\ j = \frac{t+1}{2},\ (i,j)$ (2) t even, standby after split $j:\ j = \frac{t}{2},\ (i,\widetilde{j})$
π_i	Column vector: $[\pi_i(1) \ldots \pi_i(t) \ldots \pi_i(2s)]$
$\mathcal{Q}$	Infinitesimal matrix of the QBD process
F	Forward transition matrix
B	Backward transition matrix
L, L_0	State transition matrices within the same level (i.e. having i frames at the node)
	Fronthaul network
λ	Frame arrival rate at a node (switch/link)
μ	Service rate for a node (switch/link)
ρ	Load/occupancy for a node (switch/link)
Λ	Vector with the incoming frame rates for all nodes
Γ	Vector with the external traffic $\gamma_i \neq 0$ only for CU nodes
$\mathcal{R}$	Routing matrix of the fronthaul network

A. CU AND DU MODEL

As has been mentioned earlier, the proposed model for the CU and DU is an extension of the one that was presented in [12]. We focus on downlink communication, although the same reasoning can be applied to uplink. We assume that frames arrive at the CU following a Poisson process of rate λ pkt/ms. The CU/DU might be configured in s different functional split configurations, each of them characterized by a certain service time, which we assume exponentially distributed, with mean μ_j^{-1} *ms*, for the jth split.

We assume that such nodes might change their current split, after a time that we also assume exponentially distributed, with mean γ_j^{-1} *ms* for the jth split. Before moving to the next configuration, a standby situation happens, which captures the time devoted to the reconfiguration tasks required in the CU/DU. The time that the node spends in this standby situation (where it does not process frames) is also modeled with an exponential random variable, with average ξ_j^{-1} *ms*, for the jth split. A change in the functional split is modeled as a random event, with probability α_{jk} of going from split jth to kth. We impose that α_{jj} equals 0, ensuring that whenever there is a functional split change, the node does actually modify the configuration.

The main improvements from the model that was originally presented in [12] are:

- We can consider different sojourn times for the various functional splits.
- The time spent at the standby status is also different for each split.
- When changing a particular configuration, we ensure that the next functional split is different from the previous one.

We can thus capture a more realistic behavior, having a greater number of knobs to tune the node configuration. The CU/DU node can be modeled with the 3-dimensional Markov chain that is shown in Figure 3.

We define two types of states by means of the tuples (i,j) and $(i,\widetilde{j})$, respectively. The first one denotes a state of normal operation, where i corresponds to the current number of frames at the node, and j is the index of the current functional split. The second tuple represents a standby state reached upon leaving the j^{th} split. The proposed Markov chain has s horizontal planes, each of them representing a particular split.

If the CU/DU node is active and working at a particular split j, anytime a frame arrives there is a rightwards transition (rate λ), and when a frame finishes its processing and exits the node, we can see a leftwards transition (rate μ_j). At any time the node might shift to a standby situation, modeled with a transition to the corresponding state $(i,\widetilde{j})$, in the same plane, with rate γ_j. Once in this standby situation, frames might keep arriving, reflected by rightwards transitions, but the node would not be able to process them, and there is not any leftwards transition, as can be seen in Figure 3.

When in the standby state, the node will eventually go to another split configuration. The corresponding sojourn time is modeled with an exponential random variable, and so the overall rate towards other splits is ξ_j. Once the standby status is over, the next functional split is selected with probability α_{jk}, with $k \in \{1,\ldots,s\},\ k \neq j$, and so the transition rate between $(i,\widetilde{j})$ and (i,k) is $\alpha_{jk} \cdot \xi_j$. Although the CU/DU node cannot process frames during this standby situation, we assume that the node has enough buffer capacity to keep incoming frames until they can be eventually processed, provided it works in a stable regime of operation.

The underlying model boils down to a quasi-birth-death (QBD) process, where each level corresponds to all states having the same number of frames: (i,j) and $(i,\widetilde{j})$, for $j,\widetilde{j} \in \{1,\ldots,s\}$. Therefore, we use the Matrix Geometric method to find the average delay of processing a frame in this node. The reader can refer to the seminal works from Neuts [54] and Hajek [55] for a thorough treatment of this theoretical framework.

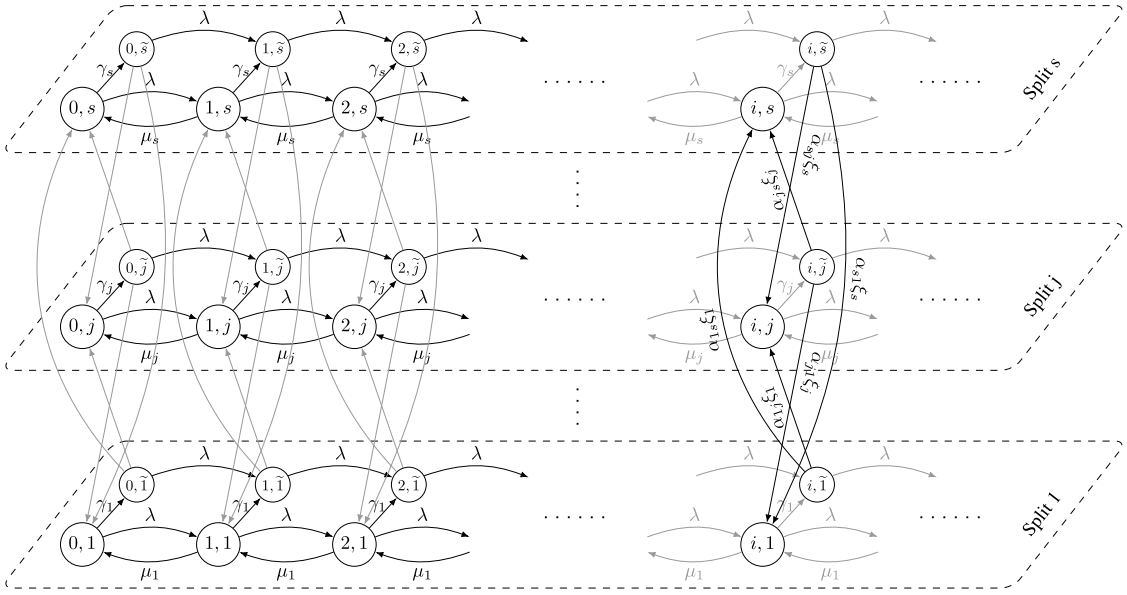


FIGURE 3. Markov chain for CU and DU nodes.

The infinitesimal matrix characterizing the QBD process is defined as follows:

$$\mathcal{Q} = \begin{bmatrix} L_0 & F & 0 & 0 & \cdots \\ B & L & F & 0 & \cdots \\ 0 & B & L & F & \cdots \\ \vdots & & \ddots & & \ddots \end{bmatrix} \tag{1}$$

where $L_0, B, L, F \in \mathbb{R}^{2s\times 2s}$. Matrices B, F are given in equation (2), while L is given in (3), as shown at the bottom of the page. On the other hand, $L_0 = L + B$.

$$F = \begin{bmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{bmatrix},$$

$$B = \begin{bmatrix} \mu_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \mu_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mu_s & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{2}$$

We denote the stationary distribution of the QBD process as $\Pi = [\pi_0, \pi_1, \pi_2, \ldots]$, where π_i is a column vector of length $2s$, and $\pi_i(t)$, $t \in \{1, \ldots, 2s\}$ is the probability of having i frames at the node when: (1) for odd t, the node is working at the jth split, and $j = \frac{t+1}{2}$, (2) for even t, the node is at standby, after split jth, $j = \frac{t}{2}$.

If the node is working at a stable operation regime, then a stationary solution for Π exists and there is a constant matrix R that fulfills the following relation [56, Theorem 3.1.1]:

$$R^2 \cdot B + R \cdot L + F = 0, \tag{4}$$

where $R \in \mathbb{R}^{2s\times 2s}$. Although there is not a closed solution for the quadratic equation (4), an iterative method can be used instead to find R. In addition, there exists a unique positive solution to the finite system of equations, from which we can obtain vector π_0:

$$\begin{aligned} \pi_0^{\mathsf{T}} (L_0 + R \cdot B) &= \mathbf{0}^{\mathsf{T}}, \\ \pi_0^{\mathsf{T}} (I - R)^{-1} \mathbf{1} &= 1, \end{aligned} \tag{5}$$

where $\mathbf{0}, \mathbf{1}$ are all-zeros and all-ones column vectors of length $2s$, respectively.

Then, the complete stationary distribution $\mathbf{\Pi} = [\pi_0, \pi_1, \ldots]$ can be obtained as:

$$\pi_i^{\mathsf{T}} = \pi_0^{\mathsf{T}} \cdot R^i. \tag{6}$$

$$L = \begin{bmatrix} -(\lambda+\mu_1+\gamma_1) & \gamma_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -(\lambda+\xi_1) & \alpha_{12}\cdot\xi_1 & 0 & \alpha_{13}\cdot\xi_1 & \cdots & \alpha_{1s}\cdot\xi_1 & 0 \\ 0 & 0 & -(\lambda+\mu_2+\gamma_2) & \gamma_2 & 0 & \cdots & 0 & 0 \\ \alpha_{21}\cdot\xi_2 & 0 & 0 & -(\lambda+\xi_2) & \alpha_{23}\cdot\xi_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -(\lambda+\mu_s+\gamma_s) & \gamma_s \\ \alpha_{s1}\cdot\xi_s & 0 & \alpha_{s2}\cdot\xi_s & 0 & \alpha_{s3}\cdot\xi_s & \cdots & 0 & -(\lambda+\xi_s) \end{bmatrix} \tag{3}$$

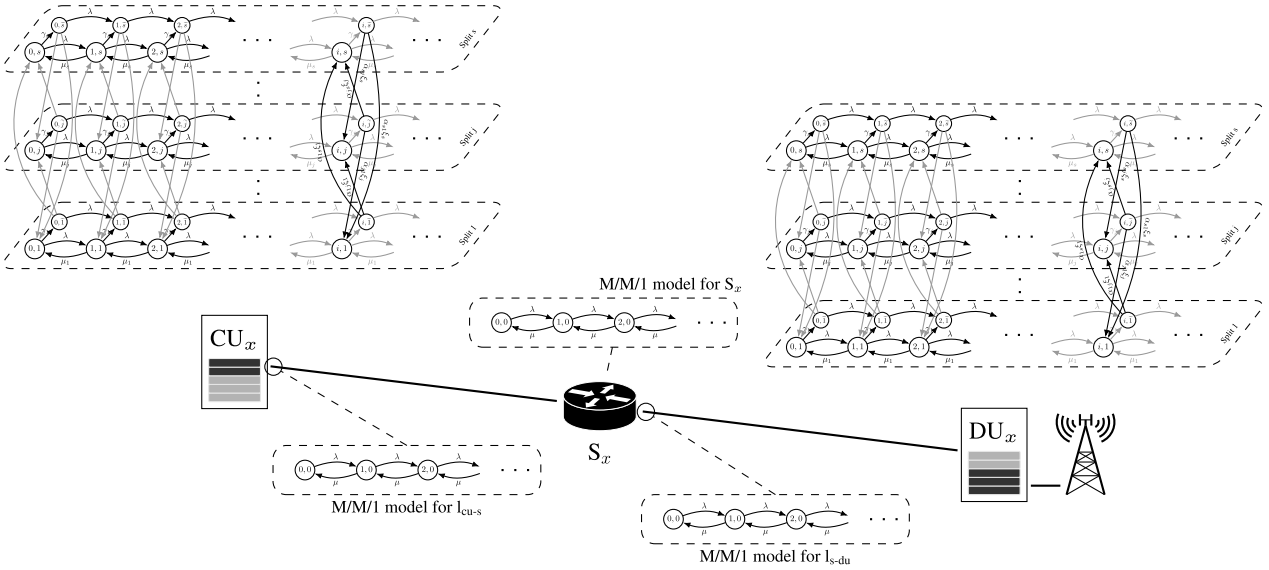


FIGURE 4. Markov chain for CU and DU nodes.

From the stationary probability distribution $\mathbf{\Pi}$, we can straightforwardly obtain the average number of frames in the node $\overline{N_{cu/du}}$:

$$\overline{N_{cu/du}} = \left\| \frac{\pi_1}{(I-R)^2} \right\|_1 = \left\| \frac{\pi_0^{\mathsf{T}} \cdot R}{(I-R)^2} \right\|_1 \tag{7}$$

where $\|\cdot\|_1$ is the 1-norm.

Finally, applying Little's Law, we can find the average delay per frame $\tau_{\text{cu/du}}$, which encompasses both the waiting and processing times:

$$\tau_{\text{cu/du}} = \frac{\overline{N_{\text{cu/du}}}}{\lambda} \tag{8}$$

As mentioned before, the stationary distribution is only guaranteed if the average service rate of the node is higher than the incoming data rate. We can thus establish the maximum packet rate $\lambda_{\max}$ that ensures system stability:

$$\lambda_{\max} = \sum_{i=1}^{s} \theta_i \cdot \mu_i \tag{9}$$

where θ_i is the probability that the CU/DU node works at a particular functional split. The value of each θ_i can be obtained by solving:

$$\Theta^{\mathsf{T}} \cdot M = \mathbf{0}^{\mathsf{T}}; \quad \Theta^{\mathsf{T}} \cdot \mathbf{1} = 1 \tag{10}$$

where Θ is a column vector of length $2s$, with the probability of working at a particular split (and the corresponding standby configuration): $\Theta = [\theta_1, \widetilde{\theta_1}, \theta_2, \widetilde{\theta_2}, \ldots \theta_s, \widetilde{\theta_s}]$; $M = L + B + F$; and $\mathbf{0}$ and $\mathbf{1}$ are all-zeros and all-ones column vectors of length $2s$, respectively.

B. FRONTHAUL END-TO-END DELAY

Once we have established the delay in both CU and DU nodes, we are now interested in finding the end-to-end delay between a CU and its corresponding DU. As mentioned earlier, we assume they are connected through a packet-switched fronthaul network, comprising switches as well as links connecting them, which might be of different technologies. In the most generic case, we consider that both switches and links can be modeled as legacy M/M/1 queuing systems and so exploit Open Jackson Networks theory to find the end-to-end delay.

We model the network topology as a directed graph $\mathcal{G} = (\mathbb{V}, \mathbb{E})$, where $\mathbb{V}$ is the set containing all network nodes and $\mathbb{E}$ is the set of all links. If we assume that there are c CUs, d DUs, n switches, and l links, then we can define $V \triangleq |\mathbb{V}| = c + d + n + l$. Based on the network topology and the particular routing strategy we can establish the routing matrix, $\mathcal{R}$, of size $V \times V$, which defines how frames travel from the CU to its corresponding DU. In Section IV we provide an illustrative example of such matrix. Figure 4 shows a typical fronthaul connection, where CU_x and DU_x are connected through a single switch, S_x, and the corresponding two links. As can be seen, the QBD model that was introduced previously is used to capture the performance of both CU and DU nodes, while M/M/1 systems are used for both the switch and the corresponding two links.

In a packet-switched network, provided that the conditions established by Burke and Jackson's Theorems [57]–[60] are met, we can establish the end-to-end delay as the sum of the individual contributions from each of the considered nodes within the path. These conditions impose that the output process is statistically identical to the one at the input, for a given node. As for the M/M/1 node, the delay can be calculated as:

$$\tau_{\text{mm1}} = \frac{1}{\mu - \lambda} \tag{11}$$

where μ and λ are such node's service and incoming rates, respectively. In order to guarantee a stable regime of operation, it is required that $\mu > \lambda$.

Since both the switches and the corresponding links can be shared by different traffic flows, the incoming data rate at each of them might be different. We assume that only CUs receive external traffic, and that the aforementioned routing matrix, $\mathcal{R}$, captures how the flows traverse the fronthaul network. We define Λ as a row vector, where each component λ_v corresponds the arrival rate at each of the $v \in \mathbb{V}$ nodes [59], [60]:

$$\Lambda = \Phi \cdot (\mathcal{I} - \mathcal{R})^{-1} \tag{12}$$

where Φ is another row-vector containing the external traffic in the network, i.e., $\phi_v = 0$ for all switches, links, DUs, and $\phi_v \neq 0$, for all CUs. Hence, based on the routing matrix $\mathcal{R}$ and the incoming traffic at all CUs we can obtain the incoming traffic rate at each node and so yield the corresponding delays.

Once we have the delay of all nodes in the network, we could use the following expression to establish the end-to-end delay for any particular flow $f \in \mathbb{F}$, where $\mathbb{F}$ is the set of all flows, as:

$$\overline{\tau_f} = \sum_{v \in \mathcal{P}(f)} \tau_v; \quad \mathcal{P}(f) : \mathbb{F} \longrightarrow \mathbb{V} \tag{13}$$

where $\mathcal{P}(f)$ returns the nodes traversed by flow f.

We can also obtain the overall average delay (for all considered flows), by applying Little's Law to the whole network:

$$\overline{\tau} = \frac{\sum_{v \in \mathbb{V}} n_v}{\lambda_0} \tag{14}$$

where λ_0 is the overall external traffic in the network: $\lambda_0 = \sum_{v \in \mathbb{V}} \phi_v$, and n_v is the average number of frames at node v. For CUs and DUs this value is provided by equation (7), and for switches and links (M/M/1) it can be obtained as:

$$n_v = \frac{\rho}{1 - \rho} \tag{15}$$

being ρ the corresponding node occupancy, which can be calculated as: $\rho = \frac{\lambda}{\mu}$.

As will be discussed later, the output process of CUs is not strictly *Poisson* and this would actually hinder the possibility of applying the Open Jackson framework. We will discuss that, under mild conditions (i.e. short standby times), the results are still valid, and close to real performances.

IV. MODEL VALIDATION AND DISCUSSION

In this section we validate the previously described model, comparing the theoretical results with those obtained from extensive simulation-based experiments. For that, we exploit an event-driven simulator, which was developed from scratch in C++. In a nutshell, it implements the two types of nodes used in the model: M/M/1 for links and switches, and the QBD for the CU and DU, and it considers four types of events. First, for all nodes, we implement two event types: (1) arrival of an external frame; and (2) end of frame processing. In addition, in the case of the QBD nodes (CU/DU), two additional event types are taken into account: (3) change of functional split; and (4) end of stand-by situation. Several flows can be configured, and a routing matrix is used to establish the node that needs to process a frame, when it first enters the system, or whenever it finishes its processing by any other node.[1]

Table 3 shows the configuration parameters that we use in all scenarios. We consider four functional splits ($s = 4$), with service rates $\mu_{1,2,3,4} = \{1, 1.5, 2, 4\}$ pkt/ms. These values are chosen for illustrative purposes, and they reflect the different processing delays featured by each functional split option. In addition, the average time at each of the functional splits is given by the corresponding rates: $\gamma_{1,2,3,4} = \{\frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \frac{4}{100}\}$ ms^{-1}. As can be seen, these rates are flipped for the DU, since processing is divided between the two entities (i.e. when split 1 is used in the CU for a particular frame, the DU should use split 4, and so appropriately complete the processing). Furthermore, we assume that ξ_j is constant for all possible configurations, $\xi_j = \xi \ \forall j$, and we modify the corresponding standby time to evaluate its impact. Matrix A establishes the probabilities of selecting the next functional split, upon a change from this particular configuration. In this sense, $\alpha_{j,k}$ corresponds to the probability of going to split k from j, with $\alpha_{j,j} = 0$, and $\sum_{t=1^s} \alpha_{t,k} = 1$. As can be observed in Table 3, the corresponding matrix for the DU is the flipped version of the CU one, to reflect that a frame processed with a certain split in the CU requires a particular one in the DU.

TABLE 3. Scenario configuration.

CU and DU nodes	
Service rates	$\mu = \{1, 1.5, 2, 4\}$ (pkt/ms)
Split change rates	$\gamma_{\text{cu}} = \{\frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \frac{4}{100}\}$ (ms^{-1})
	$\gamma_{\text{du}} = \{\frac{4}{100}, \frac{3}{100}, \frac{2}{100}, \frac{1}{100}\}$ (ms^{-1})
Standby duration	$\xi^{-1} = 1, 5, 10, 20, 50$ (ms)
Split transition probs.	$A_{\text{cu}} = \begin{pmatrix} 0 & 0.6 & 0.2 & 0.2 \\ 0.1 & 0 & 0.3 & 0.6 \\ 0.3 & 0.3 & 0 & 0.4 \\ 0.2 & 0.3 & 0.5 & 0 \end{pmatrix}$
	$A_{\text{du}} = \begin{pmatrix} 0 & 0.5 & 0.3 & 0.2 \\ 0.4 & 0 & 0.3 & 0.3 \\ 0.6 & 0.3 & 0 & 0.1 \\ 0.2 & 0.2 & 0.6 & 0 \end{pmatrix}$
Fronthaul network	
Switches service rates	$\mu_n = 5, 3$ (pkt/ms)
Optical fiber service rate	$\mu_{\text{of}} = 8$ (pkt/ms)
mmWave service rate	$\mu_{\text{mmw}} = 1, 2, 4$ (pkt/ms)
Routing matrix (cf. Fig. 7)	$\mathcal{R} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_1}{\phi_1+\phi_2} & \frac{\phi_2}{\phi_1+\phi_2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\phi_3}{\phi_2+\phi_3} & 0 & 0 & 0 & \frac{\phi_2}{\phi_2+\phi_3} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

[1]The simulator will be made publicly available in an open GitHub repository upon paper acceptance.

As was mentioned earlier, M/M/1 nodes are used to model the behavior of both links and switches on the fronthaul network. In particular, we will use various service rates to reflect different situations. As a starting point, the service rates of the switches will be $\mu_n = 5$ pkt/ms (we will reduce it to 3 pkt/ms in the last scenario), while for the fronthaul links we consider two underlying technologies: optical fiber, $\mu_{\text{of}} = 8$ pkt/ms, and milimeter waves, whose service rate, μ_{mmw}, will be varied (1, 2, 4 pkt/ms) to assess its impact. Table 3 also depicts the routing matrix that corresponds to the scenario shown in Figure 7. The indexes for both rows and columns are: $cu_1, cu_2, cu_3, du_1, du_2, du_3, s_1, s_2, s_3, s_4$; and we assume that there are three different flows in the network: $\phi_i,\ i = 1, 2, 3$, from cu_i to du_i. The corresponding routes are depicted in Figure 7. As we mentioned before, this section is intended to validate both the proposed model and the simulator implementation. To this end, we have selected configuration parameters that permit us exemplifying different situations, but which, while sensible, are synthetic. In the next section, the proposed model will be used to analyze a realistic setup.

A. SINGLE NODE CU/DU

In the first set of experiments, we validate the model for a single CU/DU node. In Figure 5 we show the average sojourn time at the CU (upper figure) and DU (lower figure) when using the configuration depicted in Table 3, as we increase the incoming frame rate. We repeat the experiment for different values of the average standby time. The results that are obtained with the theoretical model are shown with solid lines, while the markers correspond to the values yielded by the simulator. In this case, 100 independent simulations, comprising the transmission of 10^6 frames, were carried out for each configuration (λ and ξ^{-1} combination), to ensure statistically tight results. First, we can observe an almost perfect match between the two approaches, thus validating both the proposed model for the CU/DU nodes, as well as the

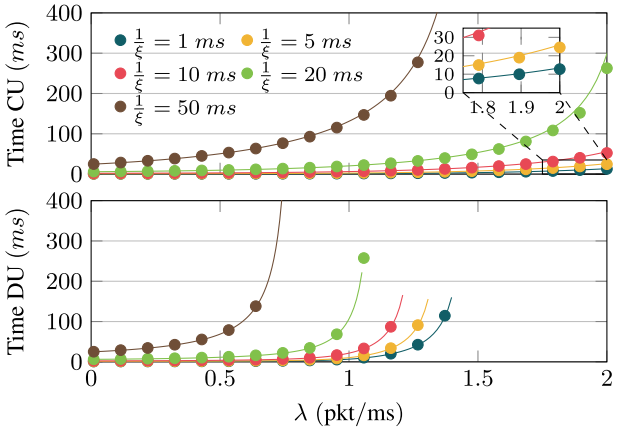


FIGURE 5. Sojourn time at a CU/DU node Vs. incoming frame rate (λ) for different standby times. Simulation and theoretical results are shown with markers and solid lines respectively.

TABLE 4. Maximum admissible λ for CU/DU nodes.

ξ^{-1} (ms)	CU $\lambda_{\max}$ (pkt/ms)	DU $\lambda_{\max}$ (pkt/ms)
1	2.8617	1.4713
5	2.6752	1.3755
10	2.4737	1.2719
20	2.1498	1.1053
50	1.5435	0.7936

simulator. On the other hand, the Figure 5 also shows the great impact of the standby duration, since the average sojourn time heavily increases when ξ^{-1} gets higher. It is worth mentioning that, in real systems, it is quite likely that the time required at the standby configuration is much shorter than those characterizing the different functional splits, as was reported in [19]. We can see that the DU yields longer times than the CU, since the probability of working at the quicker split configurations is lower. The figure also reflects the maximum admissible frame rate to ensure system stability (asymptotically increase of the delay for a certain λ). The corresponding values, which can be obtained using (9), are summarized in Table 4. Although the DU seems to be more restrictive than the CU, it is worth recalling that the stability of both nodes needs to be guaranteed, and so the maximum allowable rate for a particular flow shall be the lowest one.

In order to characterize the overall end-to-end delay, along the complete fronthaul network, we exploit, as was previously discussed, the Jackson Theory, which requires that all nodes comply with the Burke's Theorem, so that the output process at every node is statistically identical to the one at the input [58], [60]. Hence, for the CU nodes, we need the output process to be *Poisson*, which implies that the inter-departure times follow an exponential random variable. Even if the incoming frame rate ensures system stability, as established by (9), there might be circumstances that hinder the aforementioned requirement. In this sense, in order to guarantee that the corresponding Jackson Theory conditions hold, we need that: (i) the incoming frame rate is lower than the slowest function split service rate; and (ii) the time at the stand-by situation can be neglected and the node thus moves instantaneously from one split to the next one.

In order to assess whether these two aspects need to be strictly respected or not, we use the simulator to study the inter-departure times at the CU. Figure 6 represents the corresponding relative standard deviation (RSD) of such times, which is defined as the ratio between the standard deviation and the average value. If the output of the CU node were a pure Poisson process, the corresponding RSD would equal 1.

We observe that the RSD of the output process is substantially greater than 1 when the average standby time is large, and thus the output process could not be modeled as a Poisson process in this situation. Conversely, when the value of the standby time is much smaller than the split times, the RSD barely differs from 1. Hence, we can conclude that,

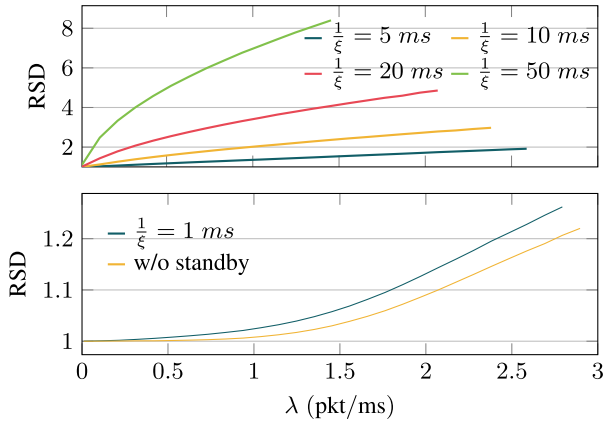


FIGURE 6. Relative standard deviation of the inter-departure times at the CU Vs. incoming rate for different standby times.

under realistic conditions (i.e. standby-times much shorter than split times), the output of the CU node mostly corresponds to a Poisson process, even if the incoming frame rate is slightly higher than the slowest functional split rate (1 pkt/ms). According to that, the use of Jackson Theory (as was discussed in Section III) to analyze the end-to-end delay in the fronthaul network is valid.

B. FRONTHAUL END-TO-END DELAY

After validating the model that we have introduced for the CU/DU nodes, and studying whether it can be exploited to assess the overall end-to-end delay, we now focus on studying such parameter. We consider the scenario shown in Figure 7, which comprises three CU/DU pairs, and four switches that interconnect them. A flow is established between each CU/DU pair, and the corresponding paths are as follows (see Figure 7): (i) $CU_1 \rightarrow S_1 \rightarrow S_2 \rightarrow DU_1$; (ii) $CU_2 \rightarrow S_1 \rightarrow S_3 \rightarrow S_4 \rightarrow DU_2$; (iii) $CU_3 \rightarrow S_3 \rightarrow DU_3$.

We first assume that all links are of high capacity (optical) and they are not bottleneck, so that they do not impact the overall delay. Under this assumption, the links are not included in the evaluation, but only the CU/DU nodes, as well as the four switches, are considered. We increase the frame rate for all flows and we study the average end-to-end delay

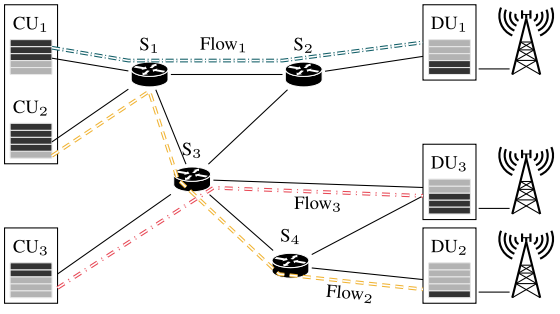


FIGURE 7. Fronthaul network to validate the proposed model.

for the three of them, as well as the overall delay. We use (14) and (15) to obtain the analytical delays, which are compared with the values yielded by the simulator. We execute 100 independent simulations per configuration, each of them comprising the transmission of 10^6 frames per flow (it is worth recalling that all of them are using the same rate, and so the time required to transmit such number of frames would be alike). Figure 8 shows the average delay. As can be seen, there is again an almost perfect match between the delays obtained with the analytical model and those yielded by the simulator. As the frame rate increases, the end-to-end delay gets higher. More interestingly, the graph also shows that the proposed model yields delays rather close to the ones obtained in the experiments, even when the requirements to apply Jackson Theory do not completely hold, i.e. when the traffic rate is higher than the slowest service rate (1 pkt/ms). In fact, there is not a relevant difference with the delays obtained with the simulator, even when getting closer to the maximum λ ensuring system stability, which is (when the stand-by time could be neglected) $\approx$1.4974 pkt/ms. The results show that for low rates the maximum delay is below 10 ms, while

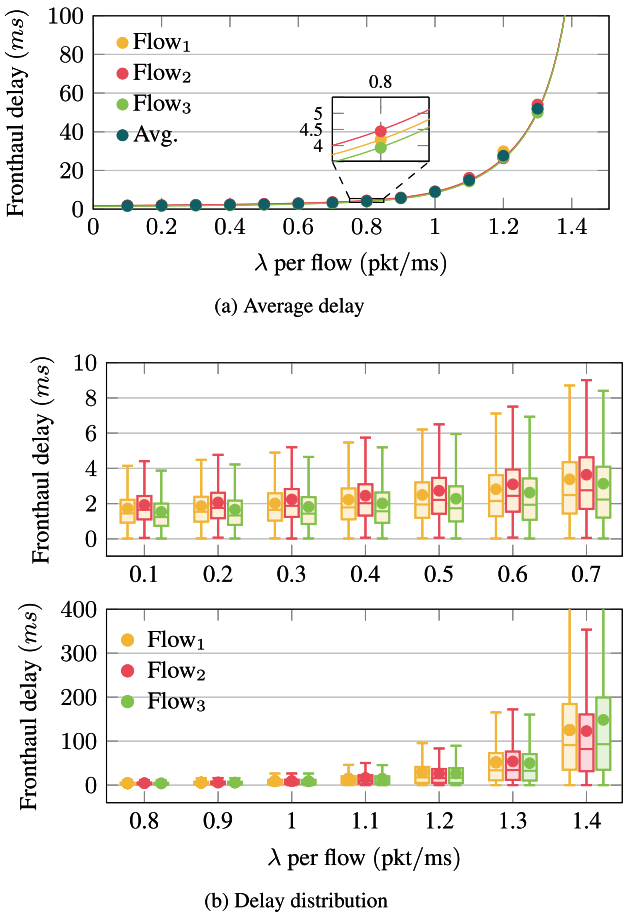


FIGURE 8. End-to-end fronthaul delay as λ (per flow) increases. Upper figure shows average delays (simulation and theoretical results are shown with markers and solid lines respectively), while the two bottom graphs represent the variability of the observed results.

it sharply increases when the incoming rate surpasses the slowest service rate (1 pkt/ms).

The simulator does not only allow us to validate the proposed model, but it can also be exploited to broaden the analysis. One particular aspect of interest is the dispersion of the delay, since not only its average value, but its variability as well might jeopardize the behavior of services with time-stringent requirements. Since the model can only be used to ascertain the average value, we use the simulator to look at the delay variability. Figure 8b uses whisker plots to represent such variability for various λ (per flow) values. Each whisker plot includes the median (0.5-percentile) as an horizontal line within each box, as well as the 0.25- and 0.75-percentiles, which correspond to the box lower and upper limits, respectively. In addition, the 0.05- and 0.95-percentiles are also represented, as the lower and upper limits of the vertical lines. In addition, we have added, as a circular marker, the corresponding average delay. As can be observed, not only the delay grows with the incoming traffic rate, but the variability gets also higher. For instance, for a packet rate of 1.2 pkt/ms, the average delay is roughly around 2 ms, but the 95% confidence interval might be as large as 100 ms (5 times the average value).

We now use a different configuration. We keep the frame rate for flows 1 and 3 at 0.8 pkt/ms, and we increase the traffic for flow 2. As can be seen in Figure 7, f_2 traverses S_1 (which is also used by f_1), and S_3, shared with f_3. Figure 9 shows the end-to-end delays. We use solid lines to represent the analytical values, while markers correspond to the delays obtained with the simulator. We also carried out 100 independent experiments per configuration. In this case, for each run we ensure that the flow having the lowest rate generates 10^6 frames, and that the other two flows are always active (the number of transmitted packets is adapted to guarantee they are active during the whole experiment), so as to ensure the validity of the results. Again, the analytical and simulation-based results show almost no difference between them. We can see that when the fronthaul switches are not heavily loaded the impact of the increased traffic over the delay perceived by the other flows is not very relevant, and they stay almost constant for all λ_{f_2}. Hence, we can also conclude that under these particular circumstances, the increased end-to-end delay for flow 2 is mostly due to the time spent at both the CU and DU nodes.

C. IMPACT OF BACKGROUND TRAFFIC

In order to complement the previous results, we synthetically increase the load of a number of the fronthaul switches, to assess how the end-to-end delay for the flows of interest is affected. We fix the rates for all flows to 0.8 pkt/ms, which ensures that the conditions to apply Jackson Theory hold. Then, we add some background traffic, with rate λ_{bg} in S_1 and S_3, and we study the end-to-end delay for the three flows. In order to add the background traffic in the model we just need to include an external flow at a particular switch (vector Φ), and accordingly adapt the routing matrix ($\mathcal{R}$). We represent the results in Figure 10, where again solid lines correspond to analytical results, while markers are the values yielded by the simulator. The duration of each experiment is established by sending 10^6 for the slowest flow (including the background traffic), while we ensure that all the others are sending packets during the whole time. Once again, we can see an almost perfect match between simulation results and

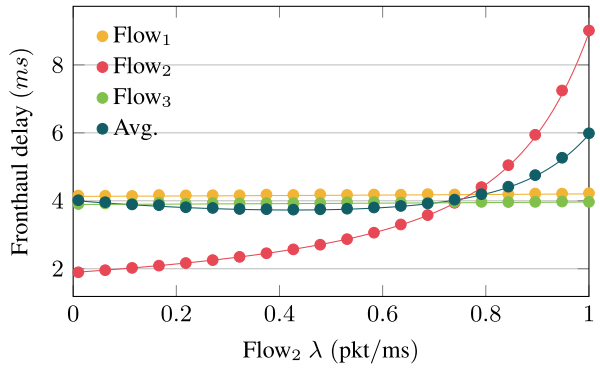
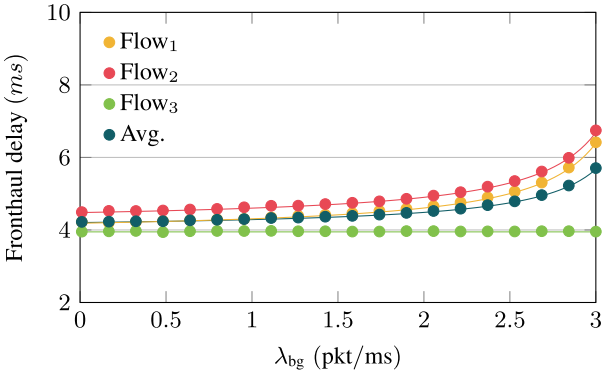
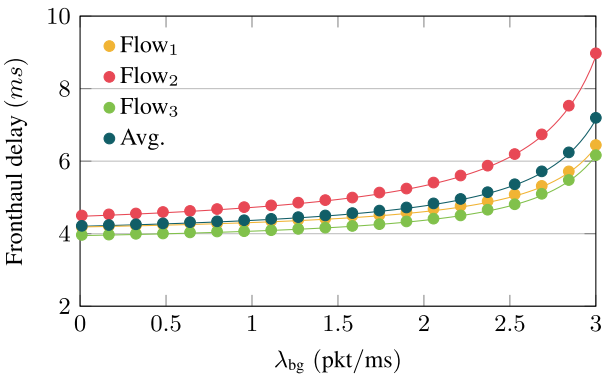


FIGURE 9. End-to-end fronthaul delay as λ_{f_2} increases. Simulation and theoretical results are shown with markers and solid lines respectively.



(a) Background traffic at S_1



(b) Background traffic at S_1 and S_3

FIGURE 10. Impact of background traffic in the fronthaul network over the end-to-end delay. Simulation and theoretical results are shown with markers and solid lines respectively.

analytical values. In Figure 10a, we add the background traffic only at S_1, while in Figure 10b, it affects both S_1 and S_3. We can observe that when background traffic is larger, and so the load of the corresponding nodes gets higher, there is a clear increase on the end-to-end delay for the affected flows. In this sense, when the background traffic only affects S_1, the end-to-end delay for flow 3 remains constant, since this flow does not traverse such node. On the other hand, when both S_1 and S_3 have some background traffic, the end-to-end delay for flow 2 is more heavily affected, since it goes through both nodes, while the other two flows only use one of them. It is worth recalling that the service rate of the switches is $\mu_{\text{mm1}} = 5$ pkt/ms and when $\lambda_{\text{bg}} = 3$ pkt/ms, the load of S_3 would reach 4.6 pkt/ms (the sum of $\lambda_{\text{bg}}, \lambda_{f_2}, \lambda_{f_3}$), so at that point the network is fairly congested.

D. IMPACT OF HETEROGENEOUS LINKS AND ROUTING STRATEGY

The last experiment that was run over this validation scenario (cf. Figure 7), aims to evaluate the impact of the links characteristics over the network performance. We also assess how the routing strategy might yield lower delays. We assume that all links in the fronthaul network are of high capacity ($\mu_{\text{fo}} = 8$ pkt/ms), but the one connecting S_1 and S_2, which is of lower capacity μ_{mmw}, reflecting the use of a different technology (for instance, millimeter wave). Under these conditions, we also vary the routing strategy at S_1 for flow f_1, so that with probability φ frames use the shortest path (i.e., traversing the link between S_1 and S_2), and with probability $1 - \varphi$ they will use the path: $\text{CU}_1 \rightarrow S_1 \rightarrow S_3 \rightarrow S_2 \rightarrow \text{DU}_1$. Furthermore, in this setup we decrease the processing capacity of the four switches, to $\mu_n = 3$ pkt/ms, so that their impact over the overall delay is comparable with that of the millimeter wave link. Figure 11 shows how the overall average delay (considering all flows) varies as φ is modified. The rates for the three flows were 0.8 pkt/ms. Analytical results are represented with solid lines, while the markers are the values obtained with the simulator, again averaging the output of 100 independent runs, in each of them transmitting 10^6 frames per flow. There is again a good match between analytical and simulation-based results. The results show that the routing strategy has an impact over the network performance. More interestingly, we can actually see that there might exist optimum operation points, where the overall delay is minimum, which could be found by using the proposed theoretical model. In the scenario we are considering, when the service rate of the link between S_1 and S_2 is 1 pkt/ms, this optimum value is seen for $\varphi \approx 0.6$.

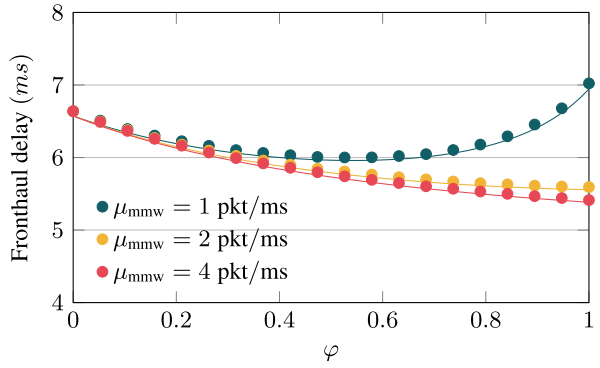


FIGURE 11. Fronthaul end-to-end delay with heterogeneous underlying technologies and different routing policies. Simulation and theoretical results are shown with markers and solid lines respectively.

V. PERFORMANCE OF A REALISTIC TOPOLOGY

In order to illustrate the applicability of the proposed model, in this section we use it to analyze the performance of a split selection policy over a realistic network deployment. We will first describe the system under analysis, and the particular split selection policy applied. Later on, we will use an outcome of that policy to feed the model and evaluate the expected system behavior.

A. SYSTEM DESCRIPTION

As mentioned before, the network consists of a set of links, switches, and base stations, each of which is divided into a CU and a DU. We now assume that all CUs are deployed into a single data center, located at a convenient position for the operator. Conversely, DUs are deployed close to the radio equipment. One fourth of the base stations correspond to macro cell, whereas the rest are small cells [61]. The geographical location of the base stations is that of a dense urban scenario, in which macro cells are distributed over a triangular grid, with an inter-site distance of 200 m, and small cells are randomly distributed over the covered area. Altogether, the scenario comprises G base stations.

DUs and the data center containing the CUs are connected by means of a packet-switched fronthaul network, which, in addition to DUs and the data center, consists of layer-3 or layer-2 switches and high-speed links (1 Tbps). We assume that there is, on average, one network switch per 10 DUs and they are connected to the data center via a minimum spanning tree plus additional links from a Waxman model [62], until an average node degree of 3.5 is achieved [63]. The number of UEs is modeled with variable U, which corresponds to $U = 10 \times G$ [61]. UEs can be either uniformly distributed over the covered area, or concentrated into clusters.

Under this scenario, we assume that the functional split is dynamically chosen with the goal of optimizing user-perceived performance. Namely, the network operator aims at instantaneously maximizing user data rates in a proportionally fair manner. This can be accomplished by selecting the functional split such that the sum of the logarithm of the user spectral efficiencies is maximized. In particular, we adopt the model described in [33], where the functional split selection is modeled as an optimization problem, as follows:

$$\max_{\mathbf{x},\mathbf{f}} \sum_{u=1}^{U} \log\left(\log_2\left(1 + \frac{p_u}{\varsigma + \sum_{g=1}^{G} i_{u,g} c(\min(x_g, x_{h_u})}\right)\right) \tag{16}$$

$$\text{subject to:} \sum_{e\in\mathbb{E}^+(n)} \phi_e^g - \sum_{e\in\mathbb{E}^-(n)} \phi_e^g = \begin{cases} 0 & n \text{ is a switch} \\ r(x_g) & n \text{ is a CU} \\ -r(x_g) & n \text{ is a DU} \end{cases} \quad \forall g \in \mathbb{G}, \tag{17}$$

$$\sum_{g=1}^{G} \phi_e^g \le \Phi_e \quad \forall e \in \mathbb{E}, \tag{18}$$

$$\phi_e^g \ge 0 \quad \forall e \in \mathbb{E},\ \forall g \in \mathbb{G}, \tag{19}$$

$$x_g \in \{1, \ldots, s\} \quad \forall g \in 1, \ldots, G, \tag{20}$$

where p_u is the signal power received by UE u from its serving base station, ς is thermal noise, $i_{u,g}$ is the interference power received by UE u from cell g. The serving base station of u is denoted by the index h_u, $c(x_g)$ is the maximum interference cancellation factor that can be applied in base station g given its current functional split x_g, $\mathbb{G} \triangleq \{1, \ldots, G\}$, where we recall that G is the number of base stations. As for traffic, ϕ_e^g is the flow proasduced by base station g on link e, Φ_e is the capacity of link e, $\mathbb{E}$ is the set of all links, $\mathbb{E}^+(n)$ is the set of links leaving node n, $\mathbb{E}^-(n)$ is the set of edges entering node n and $r(x)$ is the capacity required by split x. Notice that in (16) the cancellation factor that multiplies i_{u_g} is determined by the lowest functional split of the interfering and serving base stations.

Problem (16) can be approximated by a Mixed Integer Linear Program (MILP), as shown in [33], which allows for timely solving, even for relatively large networks. Our setup considers $G = 300$ base stations and 4 possible functional splits: PDCP-RLC, MAC-PHY; Intra-PHY and C-RAN. The fronthaul protocols used for these splits are CPRI or eCPRI for C-RAN and Intra-PHY protocols [64], the nFAPI protocol for MAC-PHY [65], and the F1 application protocol for PDCP-RLC [66], as described in 3GPP recommendations. The use of these protocols produces a signaling overhead that can be comparable or even greater than the actual user throughput [61]. Nonetheless, these additional throughput requirements are already considered in the model, since each functional split is described separately. For this scenario the equation (16) can be solved in less than 500 *ms* using operator grade equipment [33].

Our simulated time spans 12 hours and the optimal functional split is computed every second. Since there are strategies in the state of the art that can change the functional split in the millisecond range, we consider that changing the split every second is a feasible option, if required. Users are randomly distributed over the covered area, and move according to the mobility parameters proposed in TS38.913 [61]: 20% are vehicles moving at 30 km/h and the remaining 80% are pedestrians walking at 3 km/h. Their movement is mainly confined to street and squares without special preference for any specific point. Nonetheless, small clusters do form randomly, which influences the optimal functional split selection.

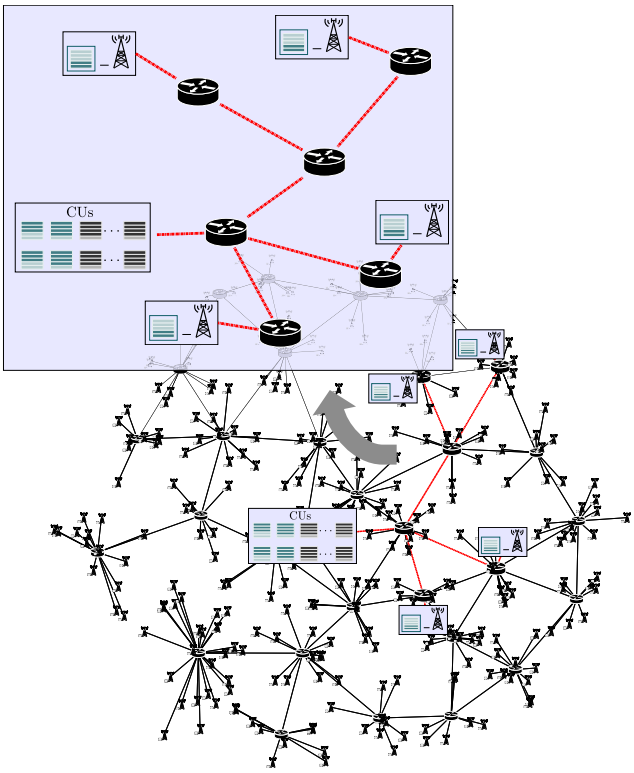


FIGURE 12. Real network scenario, the sub-network used for the evaluation is highlighted and zoomed in.

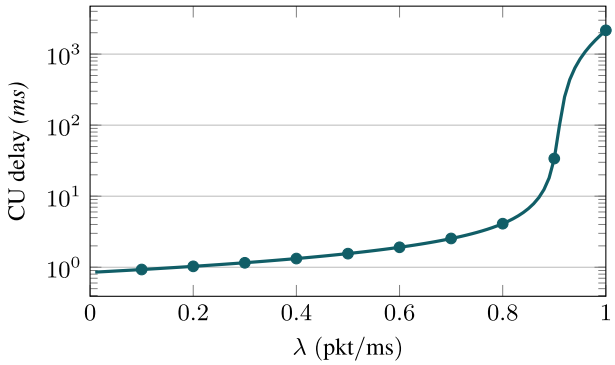


FIGURE 13. Delay at CU for different traffic rates. λ_{max} to ensure system stability is ≈1.22 pkt/ms. Simulation and theoretical results are shown with markers and solid lines respectively.

B. SYSTEM PERFORMANCE

By considering the features that were previously discussed to establish the optimal functional split policy, we then exploit the model introduced in Section III, as well as the event-driven simulator, to assess the performance of a realistic network scenario. In particular, the network topology under consideration is shown in Figure 12. The picture shows the overall network that has been used during the simulation and from which we have obtained the statistics needed for the model. Then, we have chosen a sub-network, which is highlighted in Figure 12, to carry out the analysis.

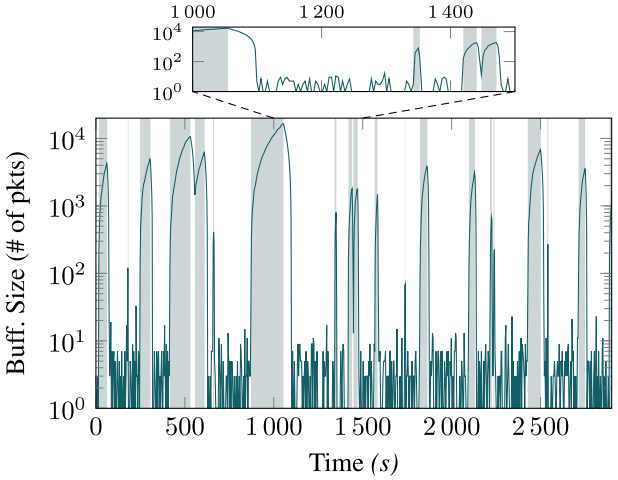
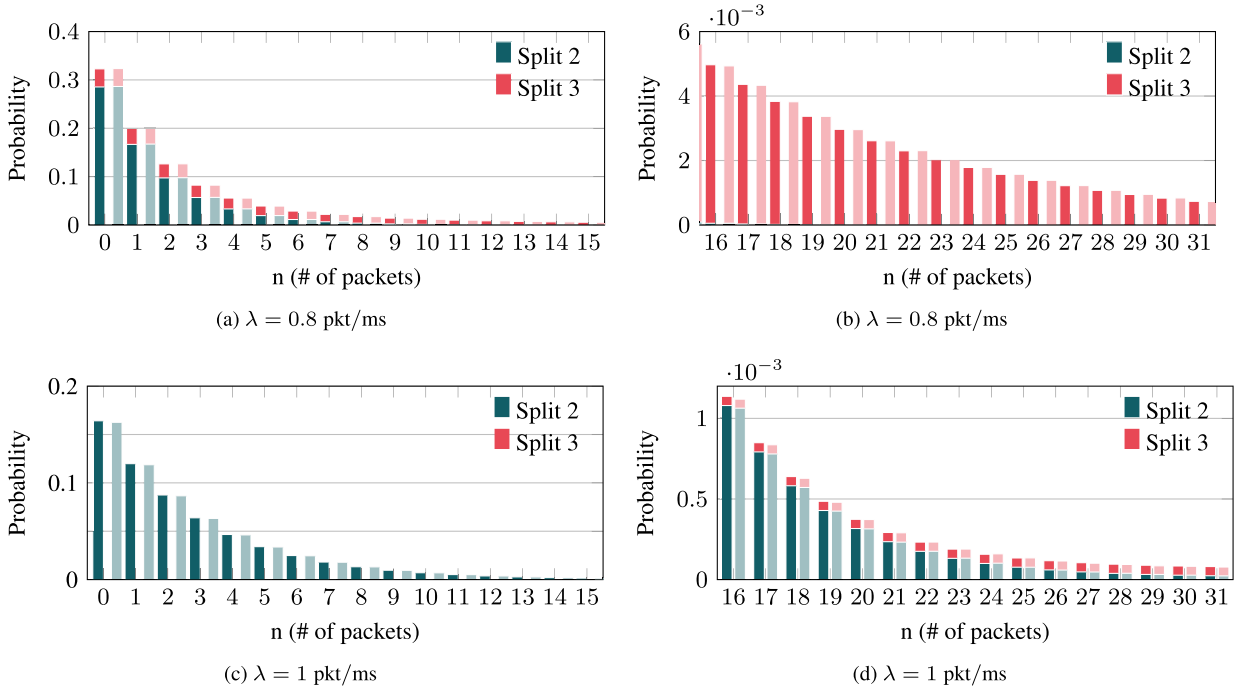


FIGURE 14. Buffer evolution at CU with $\lambda = 1$ pkt/ms. Low rate split is highlighted with shaded areas.

Taking the outcome of the 12-hour simulation that was described above, we use the split change probabilities to obtain the corresponding rates (γ). To estimate the service rates, we assume that the network is limited by the computational resources in the data center, which is a sensible assumption, since all CUs are deployed in a single facility. This way, we use the computational complexity in the CU associated to each split [67], so that the C-RAN configuration would yield the lowest service rate (highest computational needs at the CU). We also assume that the service rate for this C-RAN configuration would correspond to an average channel quality. For that, we use mean transport block size (TBS) of a base station using 15 physical resource blocks (PRBs), which is 695 packets per seconds (packets are assumed to have 1500 bytes). From that value, we estimate the service rate for others splits by scaling that of the C-RAN scheme by the ratio of computational complexities (i.e. using a linear relationship). We reckon that different assumptions could be taken, rather than the computational limitation, and we leave other network configurations for our future work. Nevertheless, as will be seen below, the described configuration illustrates the applicability of the proposed model on a realistic setup. Furthermore, we neglect the propagation delay, since it is much lower than the overall delay. For a global distance of 3 *km* in the fronthaul network, the propagation delay would be more than 100 times lower than the values that were observed for the lowest traffic rate.

All in all, we select four CU-DU pairs, which are connected by means of the fronthaul network comprising 6 switches, as shown in Figure 12. From the 12-hours simulation selecting the optimal split, we obtained the model parameters which are summarized in Table 5. As can be observed, the outcome of the split selection policy described above, and the corresponding solution of (16), does not embrace all the splits in a single base station. On the contrary, for the first base station, the optimal policy shifts between PDCP-RLC and MAC-PHY splits, while for the others it selects the highest centralization options. It is worth pointing out that different statistics would be obtained with other policies, network setups or users deployments, but our primary goal is to assess the validity of the proposed model for realistic configuration



(a) $\lambda = 0.8$ pkt/ms (b) $\lambda = 0.8$ pkt/ms

(c) $\lambda = 1$ pkt/ms (d) $\lambda = 1$ pkt/ms

FIGURE 15. Probability of having *n* packets when being in each split in the CU for different traffic rates. Model and simulation results are shown in with solid and shaded colors respectively.

TABLE 5. Scenario configuration.

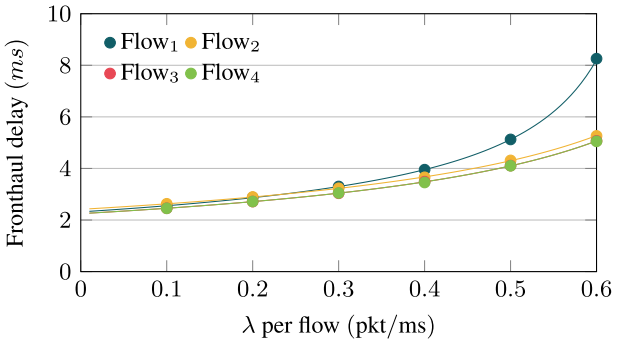
CU and DU nodes	
Service rates	$\mu = \{5.378, 1.37, 0.908, 0.695\}$ (pkt/ms)
Split change rates CUs	$\gamma_{cu1} = \{0.0714, 0.0652, -, -\}$ (s^{-1})
	$\gamma_{cu2} = \{-, 0.013, 0.0188, -\}$ (s^{-1})
	$\gamma_{cu3} = \{-, 0.0087, 0.019, -\}$ (s^{-1})
	$\gamma_{cu3} = \{-, 0.0147, 0.0097, -\}$ (s^{-1})
Split change rates DUs	$\gamma_{du1} = \{-, -, 0.0652, 0.0714\}$ (s^{-1})
	$\gamma_{du2} = \{-, 0.0188, 0.013, -\}$ (s^{-1})
	$\gamma_{du3} = \{-, 0.019, 0.0087, -\}$ (s^{-1})
	$\gamma_{du3} = \{-, 0.0097, 0.0147, -\}$ (s^{-1})

characteristics. On the other hand, the service rate of the switches is set high enough so that we can neglect their impact over the overall delay. Since all CUs are collocated, the corresponding routing matrix becomes trivial and it is therefore not included in Table 5.

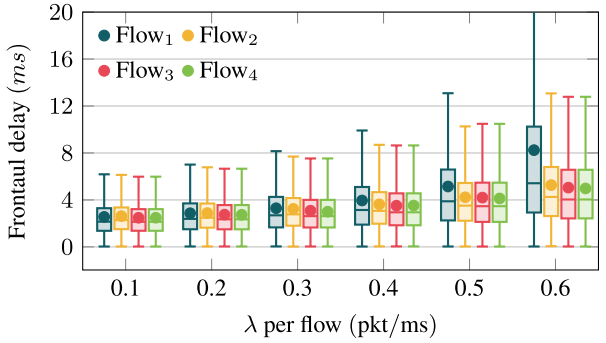
First, we focus on the CU behavior. Since the standby-times are rather short compared to the average dwell time at each functional split, it is sensible neglecting them. Hence, by using Eq. (9) we can find the traffic rate that ensures system stability. We take CU_3 configuration, which just uses splits 2 and 3, with inverse of dwell time given by γ_{cu3} in Table 5. The maximum traffic rate to ensure system stability is, for this particular configuration, $\approx$1.22 pkt/ms. We then conduct an experiment in which we increase the incoming rate until 1 pkt/ms (roughly 80% of the maximum admissible rate), thus ensuring system stability and we analyze the delay in the CU. Figure 13 shows the results. The solid line corresponds to the delay yielded by the proposed model, while the markers reflect the results that were obtained after simulations encompassing the transmission of 10^8 packets. We can again see that the model is able to perfectly match the expected behavior, regardless of the particular CU configuration. More interestingly, results show that the delay can significantly increase, even if we keep the load well below λ_{max}. The reason is that when the incoming packet rate is higher than the service rate of any split configuration, frames start to be kept at the buffer, until the CU moves to a quicker configuration. Since the dwell times at the various splits in real networks might be much longer than the ones used in the scenario that was studied in Section IV, the buffer length may strongly increase, and so the delay, which might become unacceptable.

In order to better highlight this behavior, Figure 14 and 15 show the evolution of the buffer lengths for a particular experiment. First, Figure 14 illustrates the instantaneous variation of the buffer length within a time interval of 3000 seconds and for an incoming traffic rate of 1 pkt/ms. We use gray areas to reflect the time intervals in which the CU was working at the third split configuration, with a service rate lower than the incoming traffic rate. As can be seen, the buffer length increases very sharply, reaching rather high values. On the other hand, when the CU moves to a quicker configuration (split #2), the graph evinces that the buffer occupancy also reduces at a very quick pace (the service rate in this split is 37% faster than the incoming traffic). As can be seen, the buffer length remains mostly below 10 frames, but when the slowest split configuration is active. Hence, the system is stable, but the average delay remarkably increases, up to unacceptable values, as was shown in Figure 13, showing in addition a very large variability.

Then, Figure 15 shows the probability density function (*pdf*) of the node occupancy (number of frames at the CU). We can first observe that the theoretical results match again the values obtained with the simulator. In accordance to what was seen in Figure 14, the results evince that lower buffer lengths are more likely with the fast split (#2), while longer buffer sizes are mostly caused by moving to split #3. On the other hand, the results, obtained for two different values of λ, also show the strong impact of the slowest service rate. When the traffic rate is slower than such value (upper figures), longer buffer lengths are not likely with split #2, but this is not the case when λ equals 1 pkt/ms, where the quickest split is needed to transmit the frames that were buffered when split #3 was active. In addition, the *pdf* is this case has a much longer tail (see the lower values of the y-axis in the figures), reflecting larger buffer lengths. Finally, Figure 16 depicts the end-to-end delay seen by the four flows, as we increase the traffic load for each of them. First, Figure 16a



(a) Average delay



(b) Delay distribution

FIGURE 16. End-to-end delay as λ (per flow) increases. Upper figure shows average delays (simulation and theoretical results are shown with markers and solid lines respectively), while the two bottom graphs represent the variability of the observed results.

shows the average delay comparing analytical and simulated results. As can be seen, the theoretical model yields again the same performance than the experiments carried out with the simulator. The results show that the base stations using splits #2 and #3, MAC-PHY and Intra-PHY respectively, are less impacted by the increase in the traffic rate. On the contrary, the first base station, which uses the fastest splits in the CU, shows higher end-to-end delay as we increase λ, due to the DU behavior, whose service rates are slower, as could be seen in Table 5. The results also evince that, provided the traffic load is below the slowest split configuration, the end-to-end delays remains within reasonable levels. Then, in Figure 16b we show the delay variability, obtained using the simulator. We use again whisker plots, and we add as well the corresponding average delay, represented with circular markers. For low data rates all the flows present similar distributions. However, as we increase the incoming traffic rate, we can see that the delay at the base station with fastest splits in the CU (Flow_1) does not only has a higher average value, but also far larger variability.

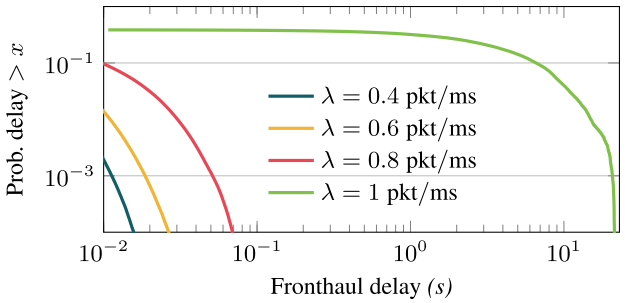


FIGURE 17. Complementary CDF of the end-to-end fronthaul delay.

In order to complement the previous discussion, Figure 17 shows the complementary *cdf* for the end-to-end delay, using the scenario depicted in Figure 12. As can be seen, for low values of λ, the probability of having long delays is very low. However, when λ gets higher, surpassing the service rate of a particular split (this mimics the configuration that was used to obtain Figure 14), the probability of suffering rather long delays is quite high. This can be used as another design parameter, considering the performance under a worst-case scenario.

All in all, we can conclude that the proposed model yields accurate results, even when using realistic configuration setups.

VI. CONCLUSION

In this paper we have introduced a novel model, based on queuing theory, which can be used to study the performance of CU/DU nodes in vRAN architectures. We can consider different service rates for the various functional splits, as well as dwell times for each of them and their corresponding standby times. The delay in traversing such nodes can be obtained with the matrix-geometric method. We have also studied the circumstances under which we could exploit this model, together with Open Networks Jackson Theory, to assess the end-to-end delay over a fronthaul network. We have shown that under sensible operation regimes, there is a very good match between the values yielded by the theoretical model and the real behavior.

We have also conducted a thorough study of the expected performance over a fronthaul network. An event-driven simulator was used not only to assess the validity of the proposed model, but also to broaden the analysis, by looking at the variability of the observed performance. In all cases, the match between the values obtained by the simulator and the proposed model is almost perfect.

Last, we have used a more realistic configuration, to assess the impact that the use of flexible functional split might have over the end-to-end delay over the fronthaul. The features of the scenario were selected from a realistic network setup, where the average sojourn times at each functional split might be much longer. We first confirmed that the proposed model still yields accurate results. On the other hand, we also saw that even if the stability of the CU/DU nodes was guaranteed, the buffer lengths, and so the delay, might strongly increase when the traffic load becomes higher than the service rate of a particular split configuration. In this sense, we saw that when the incoming traffic rate was higher than the slowest split processing rate, even if system stability was ensured, the delay could increase by a factor of $>100\times$. This can be used to carry out an analysis based on a worst-case performance, which might hinder the behavior of certain services, especially those having strict delay requirements.

There are two different lines of work that we will pursue in our future research. On the one hand we will study how the use of finite buffers and different traffic characteristics, as well as split selection policies, impact the system performance, using the developed simulator. On the other hand, we will also exploit the model to facilitate the design and planning of vRAN topologies, proposing sensible split selection policies. We will also exploit the developed simulator to propose and study different buffer management schemes.

ACKNOWLEDGMENT

The authors alone are responsible for the content of the paper.

REFERENCES

[1] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 10, no. 6, pp. 573–581, Jun. 2018.
[2] 3GPP, *Study on New Radio Access Technology: Radio Access Architecture and Interfaces*, document 3GPP, (TR) 38.801, version 14.0.0, Apr. 2017.
[3] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.
[4] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.
[5] C. L. I, Y. Yuan, J. Huang, S. Ma, C. Cui, and R. Duan, "Rethink fronthaul for soft RAN," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 82–88, Sep. 2015.
[6] *5G; NG-RAN; Architecture Description*, 3rd Generation Partnership Project (3GPP), Technical Specification (TS) document 38.401, 2021, version 16.6.0.

[7] C.-L. I. S. Kuklinski, T. Chen, and L. Ladid, "A perspective of O-RAN integration with MEC, SON, and network slicing in the 5G era," *IEEE Netw.*, vol. 34, no. 6, pp. 3–4, Nov. 2020.

[8] 3GPP, *Technical Specification Group Radio Access Network;Study on CU-DU Lower Layer Split for NR*, document 3GPP, TR 38.816, version 1.0.0, Dec. 2017.

[9] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[10] IEEE 1914Working Group. (Accessed: Nov. 2021). *Next Generation Fronthaul Interface*. [Online]. Available: https://sagroups.ieee.org/1914/

[11] P. Arnold, N. Bayer, J. Belschner, and G. Zimmermann, "5G radio access network architecture based on flexible functional control / user plane splits," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

[12] L. Diez, C. Hervella, and R. Aguero, "Understanding the performance of flexible functional split in 5G vRAN controllers: A Markov chain-based model," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 1, pp. 456–468, Mar. 2021.

[13] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 146–172, 1st Quart., 2019.

[14] A. Maeder, M. Lalam, A. De Domenico, E. Pateromichelakis, D. Wubben, J. Bartelt, R. Fritzsche, and P. Rost, "Towards a flexible functional split for cloud-RAN networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2014, pp. 1–5.

[15] A. F. Ocampo, M.-R. Fida, A. Elmokashfi, and H. Bryhni, "Evaluating the cloud-RAN architecture: Functional splitting and switched Ethernet xhaul," in *Proc. 16th Int. Conf. Netw. Service Manage. (CNSM)*, Nov. 2020, pp. 1–5.

[16] Z. Becvar, P. Mach, M. Elfiky, and M. Sakamoto, "Hierarchical scheduling for suppression of fronthaul delay in C-RAN with dynamic functional split," *IEEE Commun. Mag.*, vol. 59, no. 4, pp. 95–101, Apr. 2021.

[17] S. Lagen, L. Giupponi, A. Hansson, and X. Gelabert, "Modulation compression in next generation RAN: Air interface and fronthaul trade-offs," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 89–95, Jan. 2021.

[18] N. Bartzoudis, O. Font-Bach, M. Miozzo, C. Donato, P. Harbanau, M. Requena, D. Lopez, I. Ucar, A. A. Salona, P. Serrano, J. Mangues, and M. Payaro, "Energy footprint reduction in 5G reconfigurable hotspots via function partitioning and bandwidth adaptation," in *Proc. 5th Int. Workshop Cloud Technol. Energy Efficiency Mobile Commun. Netw. (CLEEN)*, Jun. 2017, pp. 1–6.

[19] A. M. Alba, J. H. G. Velasquez, and W. Kellerer, "An adaptive functional split in 5G networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2019, pp. 410–416.

[20] Y. Alfadhli, M. Xu, S. Liu, F. Lu, P.-C. Peng, and G.-K. Chang, "Real-time demonstration of adaptive functional split in 5G flexible mobile fronthaul networks," in *Proc. Opt. Fiber Commun. Conf. Expo. (OFC)*, 2018, pp. 1–3.

[21] P. Monti, Y. Li, J. Martensson, M. Fiorani, B. Skubic, Z. Ghebretensae, and L. Wosinska, "A flexible 5G RAN architecture with dynamic baseband split distribution and configurable optical transport," in *Proc. 19th Int. Conf. Transparent Opt. Netw. (ICTON)*, Jul. 2017, p. 1.

[22] C.-Y. Chang, N. Nikaein, R. Knopp, T. Spyropoulos, and S. S. Kumar, "FlexCRAN: A flexible functional split framework over Ethernet fronthaul in cloud-RAN," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[23] Y. Li, J. Martensson, B. Skubic, Y. Zhao, J. Zhang, L. Wosinska, and P. Monti, "Flexible RAN: Combining dynamic baseband split selection and reconfigurable optical transport to optimize RAN performance," *IEEE Netw.*, vol. 34, no. 4, pp. 180–187, Jul. 2020.

[24] Y. Zhou, J. Li, Y. Shi, and V. W. S. Wong, "Flexible functional split design for downlink C-RAN with capacity-constrained fronthaul," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6050–6063, Jun. 2019.

[25] D. Harutyunyan and R. Riggio, "Flex5G: Flexible functional split in 5G networks," *IEEE Trans. Netw. Service Manag.*, vol. 15, no. 3, pp. 961–975, Sep. 2018.

[26] D. Harutyunyan and R. Riggio, "Flexible functional split in 5G networks," in *Proc. 13rd Int. Conf. Netw. Service Manage. (CNSM)*, Nov. 2017, pp. 1–9.

[27] A. Asensio, M. Ruiz, L. M. Contreras, and L. Velasco, "Dynamic virtual network connectivity services to support C-RAN backhauling," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 8, no. 12, pp. B93–B103, Dec. 2016.

[28] V. Q. Rodriguez, F. Guillemin, A. Ferrieux, and L. Thomas, "Cloud-RAN functional split for an efficient fronthaul network," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 245–250.

[29] Y.-T. Huang, C.-H. Fang, L.-H. Shen, and K.-T. Feng, "Optimal functional split for processing sharing based CoMP for mixed eMBB and uRLLC traffic," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[30] A. Martinez Alba and W. Kellerer, "A dynamic functional split in 5G radio access networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[31] A. Alabbasi, M. Berg, and C. Cavdar, "Delay constrained hybrid CRAN: A functional split optimization framework," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–7.

[32] A. M. Alba, S. Janardhanan, and W. Kellerer, "Dynamics of the flexible functional split selection in 5G networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[33] A. M. Alba, S. Janardhanan, and W. Kellerer, "Enabling dynamically centralized RAN architectures in 5G and beyond," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 3, pp. 3509–3526, Sep. 2021.

[34] D. A. Temesgene, M. Miozzo, D. Gunduz, and P. Dini, "Distributed deep reinforcement learning for functional split control in energy harvesting virtualized small cells," *IEEE Trans. Sustain. Comput.*, early access, Sep. 18, 2020, doi: 10.1109/TSUSC.2020.3025139.

[35] A. Alameer and A. Sezgin, "Optimization framework for baseband functionality splitting in C-RAN," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2017, pp. 1–5.

[36] H. Ko and S. Pack, "Energy-efficient mode switching mechanism with flexible functional splitting in energy harvesting cloud radio access networks," *IEEE Access*, vol. 6, pp. 65078–65087, 2018.

[37] S. Zhou, X. Liu, F. Effenberger, and J. Chao, "Mobile-PON: A high-efficiency low-latency mobile fronthaul based on functional split and TDM-PON with a unified scheduler," in *Proc. Opt. Fiber Commun. Conf. Exhib. (OFC)*, 2017, pp. 1–3.

[38] S. Das, F. Slyne, A. Kaszubowska, and M. Ruffini, "Virtualized EAST–WEST PON architecture supporting low-latency communication for mobile functional split based on multiaccess edge computing," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 12, no. 10, pp. D109–D119, Oct. 2020.

[39] A. Marotta, D. Cassioli, K. Kondepu, C. Antonelli, and L. Valcarenghi, "Efficient management of flexible functional split through software defined 5G converged access," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[40] A. Marotta, D. Cassioli, K. Kondepu, and C. Antonelli, "Exploiting flexible functional split in converged software defined access networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 11, no. 11, pp. 536–546, Nov. 2019.

[41] Y. Li, J. Martensson, M. Fiorani, B. Skubic, Z. Ghebretensae, Y. Zhao, J. Zhang, L. Wosinska, and P. Monti, "Flexible RAN: A radio access network concept with flexible functional splits and a programmable optical transport," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Sep. 2017, pp. 1–3.

[42] K. Kondepu, A. Sgambelluri, N. Sambo, F. Giannone, P. Castoldi, and L. Valcarenghi, "Orchestrating lightpath recovery and flexible functional split to preserve virtualized RAN connectivity," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 10, no. 11, pp. 843–851, Nov. 2018.

[43] M. P. Amaral, J. Gomes, H. R. O. Rocha, J. A. L. Silva, and M. E. V. Segatto, "Processing resource allocation in 5G fronthaul," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Nov. 2019, pp. 1–3.

[44] J. Yu, Y. Li, M. Bhopalwala, S. Das, M. Ruffini, and D. C. Kilper, "Midhaul transmission using edge data centers with split PHY processing and wavelength reassignment for 5G wireless networks," in *Proc. Int. Conf. Opt. Netw. Design Modeling (ONDM)*, May 2018, pp. 178–183.

[45] T. Ismail and H. H. M. Mahmoud, "Optimum functional splits for optimizing energy consumption in V-RAN," *IEEE Access*, vol. 8, pp. 194333–194341, 2020.

[46] D. A. Temesgene, M. Miozzo, and P. Dini, "Dynamic functional split selection in energy harvesting virtual small cells using temporal difference learning," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1813–1819.

[47] L. Wang and S. Zhou, "Flexible functional split and power control for energy harvesting cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1535–1548, Mar. 2020.

[48] H. Gupta, M. Sharma, A. Franklin A., and B. R. Tamma, "Apt-RAN: A flexible split-based 5G RAN to minimize energy consumption and handovers," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 1, pp. 473–487, Mar. 2020.

[49] L. Wang and S. Zhou, "Energy-efficient UAV deployment with flexible functional split selection," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.
[50] S. Matoussi, I. Fajjari, N. Aitsaadi, and R. Langar, "User slicing scheme with functional split selection in 5G cloud-RAN," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–8.
[51] A. Papa, M. Klugel, L. Goratti, T. Rasheed, and W. Kellerer, "Optimizing dynamic RAN slicing in programmable 5G networks," in *Proc. ICC - IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.
[52] R. Schmidt, C.-Y. Chang, and N. Nikaein, "FlexVRAN: A flexible controller for virtualized RAN over heterogeneous deployments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.
[53] Z. Cheng, Y. Tang, and H. Wu, "Joint task offloading and flexible functional split in 5G radio access network," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2019, pp. 114–119.
[54] M. F. Neuts, "Markov chains with applications in queueing theory, which have a matrix-geometric invariant probability vector," *Adv. Appl. Probab.*, vol. 10, no. 1, pp. 185–212, Mar. 1978.
[55] B. Hajek, "Birth-and-death processes on the integers with phases and general boundaries," *J. Appl. Probab.*, vol. 19, no. 3, pp. 488–499, Sep. 1982.
[56] M. Neuts, *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1981.
[57] F. P. Kelly, "Networks of queues," *Adv. Appl. Probab.*, vol. 8, no. 2, pp. 416–432, 1976.
[58] P. J. Burke, "The output of a queuing system," *Oper. Res.*, vol. 4, no. 6, pp. 699–704, Dec. 1956.
[59] J. R. Jackson, "Jobshop-like queueing systems," *Manage. Sci.*, vol. 10, no. 1, pp. 131–142, Oct. 1963.
[60] L. Kleinrock, *Queueing Systems: Theory*, vol. 1. Hoboken, NJ, USA: Wiley, 1975.
[61] *Study on Scenarios and Requirements for Next Generation Access Technologies*, 3rd Generation Partnership Project (3GPP), Technical Report (TR) document 38.913, Jul. 2018, version 15.0.0.
[62] B. M. Waxman, "Routing of multipoint connections," *IEEE J. Sel. Areas Commun.*, vol. I-6, no. 9, pp. 1617–1622, Dec. 1988.
[63] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith, and G. Iosifidis, "FluidRAN: Optimized vRAN/MEC orchestration," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 2366–2374.
[64] Ericsson AB, Huawei Technologies Co. Ltd, NEC Corporation and Nokia, *Common Public Radio Interface: eCPRI Interface Specification*, Common Public Radio Interface (CPRI), Interface Specification, May 2019, version 2.0.
[65] SCF, *5G nFAPI Specifications*, Small Cell Forum (SCF), document 225.2.0, May 2021.
[66] *NG-RAN; F1 Application Protocol (F1AP)*, 3rd Generation Partnership Project (3GPP), Technical Specification (TS) document 38.473, Aug. 2021, version 16.6.0. [Online]. Available: http://www.3gpp.org/DynaReport/38473.htm
[67] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–7.

LUIS DIEZ received the M.Sc. and Ph.D. degrees from the University of Cantabria, in 2013 and 2018, respectively. He is currently an Assistant Professor with the Communications Engineering Department, University of Cantabria. As for teaching, he has supervised 15 B.Sc. and M.Sc. thesis, and he teaches in courses related to cellular networks, network dimensioning, and service management. He has been involved in different international and industrial research projects. His research interests include future network architectures, resource management in wireless heterogeneous networks, and the IoT solutions and services. He has published more than 40 scientific and technical articles in those areas. He has served as a TPC member and a reviewer in a number of international conferences and journals.

ALBERTO MARTÍNEZ ALBA (Graduate Student Member, IEEE) received the bachelor's and master's degrees in telecommunication engineering from the Technical University of Madrid, Spain. He is currently pursuing the Ph.D. degree with the Chair of Communication Networks, Technical University of Munich, Germany. His current research interests include the design, optimization, and implementation of flexible next-generation mobile networks, adaptive radio access networks, and software-defined mobile networks.

WOLFGANG KELLERER (Senior Member, IEEE) is currently a Full Professor with the Technical University of Munich (TUM), where he is also heading the Chair of Communication Networks, Department of Electrical and Computer Engineering. Before, he was for over ten years with NTT DOCOMO's European Research Laboratories. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and as an Area Editor for Network Virtualization for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.

RAMÓN AGÜERO (Senior Member, IEEE) received the M.Sc. degree (Hons.) in telecommunications engineering from the University of Cantabria, in 2001, and the Ph.D. degree (Hons.), in 2008. Since 2016, he has been the Head of the IT Area (Deputy CIO) at the University of Cantabria. He is currently an Associate Professor with the Communications Engineering Department, University of Cantabria. He has supervised five Ph.D. and more than 70 B.Sc. and M.Sc. thesis. He is the main instructor in courses dealing with networks, and traffic modeling, both at B.Sc. and M.Sc. levels. His research interests include future network architectures, especially regarding the (wireless) access part of the network and its management. He is also interested on multi-hop (mesh) networks, and network coding. He has published more than 200 scientific articles in such areas. He is a regular TPC member and a reviewer on various related conferences and journals. He serves with the Editorial Board for the IEEE COMMUNICATION LETTERS (Senior Editor, since 2019), the IEEE OPEN ACCESS JOURNAL OF THE COMMUNICATIONS SOCIETY, *Wireless Networks* (Springer), and *Mobile Information Systems* (Hindawi).