

Received May 7, 2019, accepted June 4, 2019, date of publication June 14, 2019, date of current version July 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923020

5G New Radio Fronthaul Network Design for eCPRI-IEEE 802.1CM and Extreme Latency Percentiles

GABRIEL OTERO PÉREZ^{ID}, DAVID LARRABEITI LÓPEZ, AND JOSÉ ALBERTO HERNÁNDEZ^{ID}

Department of Telematics Engineering, Universidad Carlos III de Madrid, 28911 Madrid, Spain

Corresponding author: Gabriel Otero Pérez (gaoterop@it.uc3m.es)

This work was supported in part by the Spanish National TEXEO Project under Grant TEC2016-80339-R and in part by the H2020 EU-Funded BlueSpace Project under Grant 762055. The work of G. Otero Pérez was supported by the Spanish Ministry of Education, Culture and Sport through the FPU Grant under Grant FPU16/01760.

ABSTRACT Packet-switched fronthaul networks are often designed following the rule that the worst-case network delay must be below a given target end-to-end network latency budget. However, the theoretical maximum delay can be too pessimistic in particular scenarios, where the latency budget needs to be a very small or there is a need to stretch the distance between the radio heads and the baseband units. In this paper, we propose to use a very high packet delay percentiles as an alternative to the maximum theoretical delay in order to stretch the range of the fronthaul links at the expense of a higher frame loss ratio (FLR), within the limits established by eCPRI and the IEEE 802.1CM. Several methods to estimate the percentiles for the I_U / II_D eCPRI functional splits are analyzed. Namely, G/G/1 and N*D/D/1 queueing models are tested and compared with simulation as dimensioning tools. The results support that the N*D/D/1 queue is able to model the behavior of a packet-switch fronthaul aggregator using the eCPRI standard for 5g New Radio (NR) Fronthaul streams and can be used as a tool to dimension the length of the links. The experiments show that the fronthaul links' lengths can be increased by 60% and 10% for 50- and 100-MHz NR channels, respectively, while keeping the latency budget and frame loss ratio within the IEEE 802.1CM limits.

INDEX TERMS 5G, C-RAN, delay percentiles, eCPRI, fronthaul networks, G/G/1, IEEE 802.1CM, N*D/D/1, new radio (NR), time-sensitive networking (TSN).

I. INTRODUCTION

The Cloud Radio Access Network (C-RAN) architecture proposed as an implementation option for 5G Mobile Networks introduces the concept of cloud-based processing of radio signals. In C-RAN, the radio signals received by the Remote Radio Heads (RRHs) are digitized and transported over the fronthaul (FH) network to a pool of shared remote Baseband Units (BBUs) where the processing takes place.

On the one hand, this scheme reduces the complexity of the base stations, enabling the sharing of signal processing capacity by several antennas. Since the signal processing is performed in a centralized facility, adding new features such as Coordinated Multi-Point (CoMP) [1], [2] has become easier. On the other hand, these advancements come at the expense of a higher bandwidth utilization in the

fronthaul network. Since a great deal of processing is offloaded from the base stations, the C-RAN architecture poses stringent delay and jitter requirements for the transport of FH data. Until recently, the Common Public Radio Interface (CPRI) [3] specification has been used as the most popular RRH-BBU interface.¹ However, CPRI requires very high-capacity and ultra-low latency links for the digitized RF signal. Therefore, more efficient schemes that rely on other functional splits of the radio processing chain are necessary to support 5G. In addition, the demand for a packet-switching-based fronthaul network [4] has led to an enhanced version of CPRI (eCPRI [5]), which is designed for packet networks, namely Ethernet and IP.

¹In CPRI terminology, the terms employed for RRH and BBU are RE (Radio Equipment) and REC (Radio Equipment Control) respectively; eRE and eREC if eCPRI is supported. Finally, 3GPP RAN architecture uses the terms DU (Distributed Unit) and Central Unit (CU).

The associate editor coordinating the review of this manuscript and approving it for publication was Guangdeng Zong.

With the aim of cost reduction, hardware reuse, and backwards compatibility, Ethernet-based packet-switch networks are being taken into account for the implementation of such FH networks. Given the potential of this solution to exploit the statistical multiplexing of variable-rate fronthaul and backhaul traffic, there exists an intense research and standardization effort in this field. Particularly, the IEEE 802.1CM standard published in 2018 [6] includes important recommendations for the configuration of Ethernet for the transport of fronthaul traffic and specifies relevant QoS targets for such transport. These parameters include the end-to-end latency budget and the maximum Frame Loss Ratio (FLR) for each type of fronthaul traffic, which are used as design targets in this article. We shall review the aspects of IEEE 802.1CM relevant to this paper in Section III-B.

Finally, all the above-mentioned aspects are affected by the planned data rate growth for 5G New Radio (NR). In December 2017, the numerology for the New Radio air interface for 5G was released by 3GPP in TS38.104 [8] as Release 15. This document defines two frequency ranges: FR1 (below 6 GHz) with component bandwidths ranging 5-100 MHz and sub-carrier spacings 15/30/60 KHz; and FR2 (24-86 GHz) with component bandwidths ranging 50-400 MHz and sub-carrier spacings 60/120 KHz. Additionally, eight possible functional split options are further defined in TR38.801 [9]. This leads to a wide range of very-high-rate fronthaul traffic patterns with different QoS requirements which require cost-efficient transport solutions given the size and economic impact of the access network in the telecommunication business. We study the optimization of the fronthaul network for transporting 5G NR signals in Section VI.

The remainder of this article is organized as follows. Section II makes a review of the related literature, including theoretical, simulated, and experimental works concerning the modeling of the FH network. Section III describes the C-RAN architecture and gives a short overview of eCPRI and its functional splits in order to identify the traffic patterns. Additionally, we highlight the transport requirements established by eCPRI and how IEEE 802.1CM proposes to implement them on an Ethernet switched network. The section includes a description of the main design parameter and sets the goal of this paper. Section IV identifies a number of options to compute the queuing delay percentiles for an eCPRI-driven 5G New Radio fronthaul, which are later compared in Section V. Finally, Section VI describes a concrete use case of application of N*D/D/1 using, as a target percentile, the one corresponding to the maximum FLR allowed for HPF packets according to 802.1CM. The practical gain, in terms of distance (link length), of using very high latency percentiles is assessed for the particular transport of 5G New Radio signals with eCPRI. Section VII concludes this paper, summarizing the findings and contributions of this work.

II. RELATED WORK IN FRONTHAUL MODELING

The fronthaul network appears to be a vital part of the future 5G C-RAN architecture. The performance of CPRI over

Ethernet has been evaluated in the past in several research efforts [10]. However, to the best of our knowledge, no previous standard-oriented works analyzing the tradeoff between delay and FLR have been published, and few contain 5G NR transport results.

Studies performed with several standards [11]–[13] propose frame preemption and traffic scheduling to alleviate end-to-end latency and jitter. In [14] and [15], the authors investigate the effects of different queuing regimes (weighted round robin and strict priority) on the mean and standard deviation of the frame inter-arrival delay of LTE traffic in the presence of background Ethernet traffic. The authors of [16] address the dimensioning problem of next generation fronthaul networks from a different perspective. They compare multiple fronthaul architectures in terms of bandwidth requirements, delay budgets, deployment costs, complexity of the RRHs, and the ability to support advanced wireless functions. In order to do so, they set up a mathematical framework to solve an optimization problem that takes into account deployment costs, distances, capacity, coverage, etc. Then, they give insights into the modeling of future optical transport networks.

Waqar *et al.* study in [17] the impact of jitter on the performance of CPRI over Ethernet and propose a fronthaul architecture with two algorithms that enforce constant inter-packet delay by transmitting the packets at pre-calculated timing values and use buffering to avoid rescheduling. Simulations confirm that the algorithms are able to maintain the jitter within reasonable limits. The authors of [18] examine the different packet switching mechanisms for Time Sensitive Networks proposed in standardization, focusing on solutions using inter-packet gap detection and scheduling.

Among simulation-based studies, it is worth mentioning the following: Chang *et al.* [19] evaluate different packetization strategies for a number functional splits and user densities. They provide insights on the feasibility of each combination via simulations and theoretical analysis assuming worst-case peak rates. Simulations in [20] further study the impact of packetization by computing the optimal payload sizes. The 95th percentile queueing delay is used as the dimensioning tool to decide the maximum number of supported RRHs. The analysis carried out in [21] presents a packet-based 5G transport network that includes a scheduler to exploit inter-packet gaps. The authors of [22] present an R package called *Simmer* for simulating 5G scenarios and show its applicability in [23]. Finally, [24] compares the throughput and cost of distributed versus centralized RAN via simulation.

Regarding our previous work, in [25], we study the delay constraints imposed by the CPRI protocol in ring-star topologies used by mobile operators. We derived the theoretical expressions for the propagation and queueing delays, adjusting a G/G/1 queueing model to our scenario. We showed that this estimation is an upper bound on the simulation output and is accurate under certain conditions. Also, based on these results, a packetization strategy is proposed to reduce the average aggregated queueing delay.

In [26], we extended the previously mentioned work by studying the behavior of the Cloud Radio Access Network (C-RAN) architectures with eCPRI protocol. To that end, we derived the p -th percentile queueing delay expression based on the Kingman's Exponential Law of Congestion. Simulations revealed that it provides accurate estimates on such delays (90th, 99th percentiles) for the particular case of aggregating a number of eCPRI fronthaul flows, namely functional splits \mathbf{I}_U and \mathbf{II}_D . Nevertheless, meeting the extreme percentiles required in 802.1CM is not supported by this approach.

In the practical experimentation side, we should note several studies. In [27], the authors experimentally evaluate the fronthaul latency and how the virtualization affects the latency budget in an experimental 5G testbed [28]. They focus on the intra PHY split (Option 7-1 defined by 3GPP [9]). Their results suggest that virtualization further decreases the latency budget.

The authors of [29] study the Ethernet-based fronthaul as an alternative to the expensive deployment and use costs of CPRI. They present this option as a cost-efficient and more-easily reconfigurable alternative. They investigate the delay and jitter requirements from a more practical perspective by making use of FPGA-based Verilog experiments and further propose a scheduling policy, based on [11] and [13], to cope with the jitter introduced by encapsulation. In [30], both size- and time-based Ethernet encapsulations are considered. Results show that time-based encapsulation is preferable to avoid jitter upon CPRI line bit rate reconfiguration. In [31], they use the an OpenAirInterface (OAI) setup to characterize the traffic of different functional splits as a guide for choosing the appropriate transport network.

III. PRELIMINARIES AND STANDARDS

In this section, we give an overview of the C-RAN architecture envisioned for the fronthaul network. In addition, we review the standardization efforts that will shape the future FH, i.e., 5G New Radio, eCPRI, and IEEE 802.1CM, and their practical implications, while paying special attention to the constraints that they pose in terms of latency and bandwidth consumption.

A. 5G NEW RADIO AND eCPRI IN C-RAN

The C-RAN approach for cellular networks advocates for splitting the radio processing chain in order to simplify base stations and share baseband processors. The scenario addressed by this paper is the single-hop case, a frequent setting where a single Ethernet switch is employed to multiplex a number of fronthaul flows coming from different RRHs (see Fig. 1). These flows are then aggregated and forwarded to a centralized pool of baseband processing units over a fiber access network. In this scheme, a flexible distribution of the radio processing functions enables the network designer to trade off RRH complexity, fronthaul rate, and distance. A low-level functional split means simpler RRHs but higher fronthaul rates, ultra-low latency requirements, and, therefore,

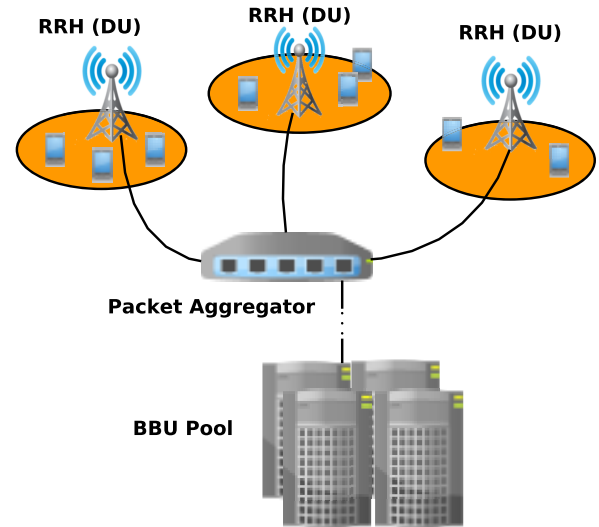


FIGURE 1. Target fronthaul network scenario.

distance limitations. Thus, it is important to properly identify the split, as it determines a particular traffic pattern and delay budget. Fig. 2 plots the envisioned functional splits in the eCPRI specification, similar to the ones in 3GPP 5G New Radio specification [8].

As noted in Fig. 2, Functional Split **E** is equivalent to the CPRI functional split. It consists of the quantization and digitalization of the down-converted radio waveform in the time domain [33]. Since no further processing is performed at the RRH side, information, such as the Cyclic Prefix (CP), is transmitted towards the BBU as overhead. In this case, complex processing devices are no longer needed at the RRH because all the functions required to decode the signal are centralized at the BBU. Since the fronthaul bitrate that has to be provisioned to give support to this functional split is too high (see Table 1), we focus our study in the next functional split, as suggested by eCPRI.

If we apply further processing to the radio signals by removing the cyclic prefix, performing the Fast Fourier Transform, removing guard band subcarriers and demapping the resource blocks, a large amount of overhead data is eliminated and, therefore, the bandwidth requirements are relaxed. At this point (Split \mathbf{I}_U), the generated data rate depends on the fraction of radio resource blocks that are being used (i.e., fronthaul data rate is proportional to cell load).

Let us analyze the shape of the generated traffic in a given RRH. Assuming a worst-case utilization scenario –that is, all the resource blocks are being utilized ($\eta = 1$) – the traffic at the output of the RRH can be expressed as

$$R_{\text{Split } \mathbf{I}_U} = N_{sc} \cdot 0.9 \cdot (T_s)^{-1} \cdot \eta \cdot 2 \cdot N_{bits} \cdot N_{ant} \quad (1)$$

where N_{sc} is the total number of subcarriers in the channel. Assume that 5% are used as guard bands [34] (10% for LTE [35]). T_s is the symbol duration, and N_{bits} and N_{ant} stand for the number of quantization bits and the number of antennas, respectively. Finally, the 2-factor accounts for the

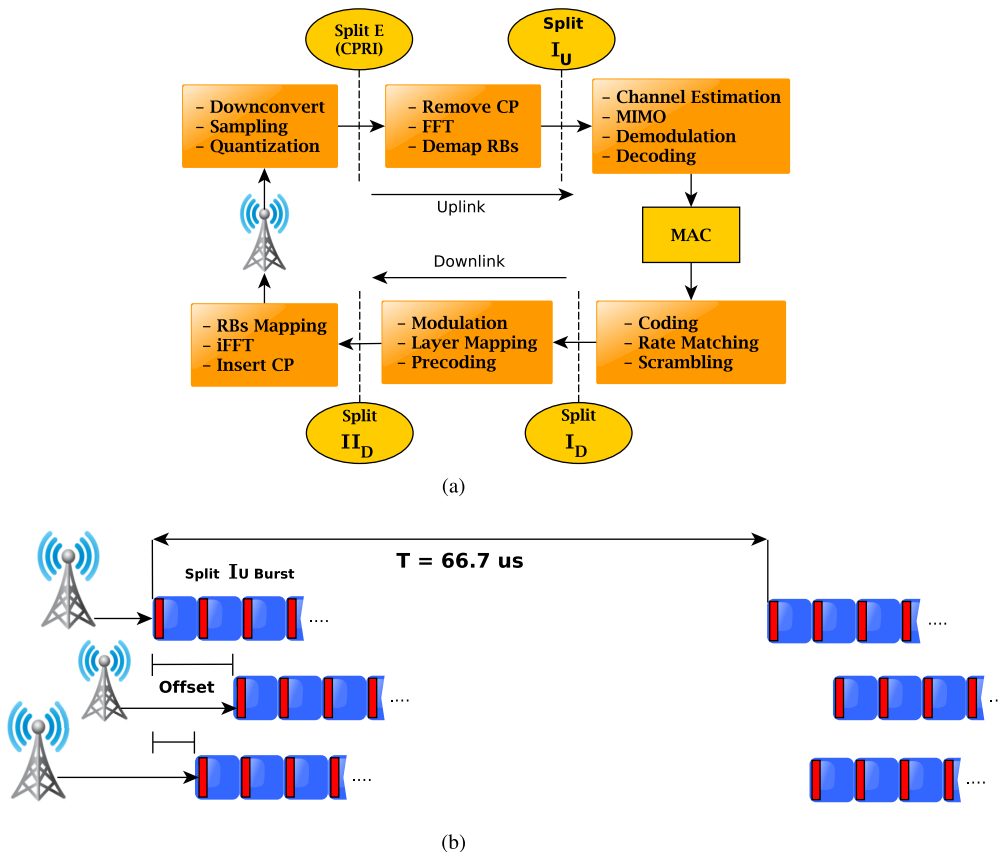


FIGURE 2. Fronthaul network and traffic pattern. (a) [eCPRI vision of 5G processing chain (see [5]). (b) Packetized Split I_U bursts: example for $T_s = 66.6 \mu s$.

TABLE 1. Functional splits traffic profiles for 5G New Radio user plane; $N_{ant} = 2$ MIMO, $N_{bit} = 15$ bit/sample, 5% guardband.

	Split E		Split I_U ($\eta = 1$)			
	50 MHz	100 MHz	50 MHz	100 MHz	200 MHz	400 MHz
Channel Bandwidth	50 MHz	100 MHz	50 MHz	100 MHz	200 MHz	400 MHz
Subcarrier Spacing	60 KHz	60 KHz	15 KHz	60 KHz	60 KHz	120 KHz
Burst Size [B]	120	240	23753	11880	23753	23753
Period [μs]	0.260416		66.6	16.6	16.6	8.3
Bitrate [Mb/s]	3686	7372	2851	5702	11401	22802

complex nature of signals, i.e., in-phase and quadrature (IQ) components.

Numerical example: Considering a MIMO system with 2 antennas, 50 MHz channels, and 15 KHz subcarrier spacing; $50 \text{ MHz}/15 \text{ KHz} = 3333.\bar{3}$ subcarriers are available inside that bandwidth. In order to maintain orthogonality, a symbol rate of $T_s = 66.6 \mu s$ is needed. Finally, assuming 15 bits to represent each IQ sample, we have a bit rate of 2, 851 Mbit/s, that is, a burst of $\approx 23,753$ bytes every $66.6 \mu s$ that each RRH periodically sends to the BBU. Applying the same methodology for different NR channel configurations, we obtain the numbers included in Table 1. Subcarrier spacings as well as the burst's periods and sizes are included.

Depending on the size of the packets used to transport the FH bursts, the performance may change, as studied in [25].

However, for the sake of backwards compatibility, we consider a payload size of 1, 500 bytes so as to meet the maximum payload length defined in the IEEE 802.3 Ethernet standard. This is compliant with IEEE 802.1CM (Standard Sections 8.1.1, 8.2.1). Accordingly, it would take a number of back-to-back frames to transport each burst from the RRH, as depicted in Fig. 2b. Regarding the frame overhead, we take into account an 8-byte preamble, a 14-byte Ethernet header, 4 bytes for the checksum, 12 bytes for the interpacket gap, and, finally, a 4-byte eCPRI header, adding up to 1, 542 bytes per burst packet. Each RRH periodically sends bursts of packets that contain the digitized IQ for OFDM symbols.

It is rather important to observe that the different FH flows may overlap in different ways at the aggregation point, leaving silence periods or causing important queuing delays.

In the following, we assume that the offset of each flow with respect to the first one (reference flow) follows a uniform distribution between 0 and the burst period, $U(0, T_s)$.

In addition to these numbers, the eCPRI standard defines the *Real-Time Control Information* messages. These are sent before the transmission of the user data bursts to inform the remote node about how to process the data contained in them. This message type includes information for control, configuration, and measurement. Nevertheless, at the time of writing, eCPRI does not provide a way to compute the generated Real-Time Control data rate. The reason is that the payload included in these messages is vendor-specific and depends on the particular functional split and implementation [5]. This traffic may be transported as IEEE 802.1CM Medium Priority Fronthaul (see Table 2). In this case, this traffic does not alter the delay calculations of this paper.

TABLE 2. Per-flow transport requirements for splits E, I_D, II_D, and I_U.

802.1cm Fronthaul Class	Data Type	802.1cm Strict Priority Configuration	Maximum One-way Delay	Maximum One-way FLR
HPF	Class 1 IQ data (CPRI) and Class 2 (eCPRI) User Plane fast data	Highest	100 μ s	10^{-7}
MPF	Class 2 (eCPRI) User Plane slow data and C&M fast data	Second Highest	1 ms	10^{-7}
LPF	Classes 1, 2 C&M data	Third Highest	1 ms	10^{-6}

B. LATENCY BUDGETING: ECPRI AND IEEE 802.1CM

The IEEE Standard for local and metropolitan area networks, IEEE 802.1CM *Time-Sensitive Networking for Fronthaul* [6], defines a set of profiles usable to configure Ethernet networks to transport time-sensitive fronthaul streams. The standard covers two classes of fronthaul interfaces:

- **Class 1** refers to interfaces in which the functional decomposition of an E-UTRA base station is done according to CPRI V7.0; also present in eCPRI as split option E.
- **Class 2** refers to eCPRI interfaces in which the functional decomposition of an E-UTRA base station is intraphy, i.e., Splits I_U/II_D [5]. E-UTRA splits above PHY do not have such stringent QoS constraints and are not addressed by neither IEEE 802.1CM nor eCPRI.

In addition, IEEE 802.1CM suggests different timing distribution schemes to fulfill the synchronization requirements of the four timing categories identified in [7] to implement 3GPP features (handovers, MIMO, COMP, etc.). Table 2 summarizes the per-flow transport requirements integrating information from eCPRI v1.1 [7] and IEEE 802.1CM. In summary, the three types of fronthaul flows identified in CPRI and eCPRI are as follows:

- High Priority Fronthaul (HPF):** includes Class 1 IQ data and Class 2 User Plane data, both with 100 μ s maximum end-to-end one-way latency.
- Medium Priority Fronthaul (MPF):** includes Class 2 User Plane slow data and Class 2 Control & Management (C&M) fast data, with 1 ms of one-way latency budget.
- Low Priority Fronthaul (LPF):** carries Class 1 and Class 2 C&M data.

The standard defines such profiles to configure a bridged network for the transport of fronthaul traffic. Furthermore, different profiles are defined with the aim of handling each fronthaul flow properly. Profile A makes use of strict priority queueing and recommends setting the highest possible priority to HPF traffic. Subsequent lower priorities should be assigned to MPF and LPF, in this order. Profile B extends Profile A with frame preemption in order to reduce the impact of background traffic on jitter [23]. However, the extra latency saving of this profile is limited given the high interface data rates (and, hence, small frame transmission times) required to transport fronthaul traffic, especially with the advent of 5G NR.

The network latency in IEEE 802.1CM and IEEE 802.1 standards is defined as the time elapsed between the reception of a frame’s first bit at the ingress switch and the moment that the last bit leaves the egress switch of the access network. Fig. 3a is a graphical representation of the network latency definition.

As reviewed in [6], the network latency t_{network} comprises a number of different components. In general, these can be considered either as fixed (or bounded) or variable. On the one hand, the **variable** terms that we consider in our design problem are the *self-queueing delay* (queueing time due to flows of the same HPF class competing for the same output ports) and the *propagation delay*. On the other hand, the terms that can be regarded as **fixed** are as follows: (a) the *frame transmission time*; (b) the waiting time until the transmission of the current packet is finished (if Profile A, i.e., no preemption is used) or, alternatively, the “*preemption time*”, if Profile B is used and we need to seize the output link. This is named as *queueing delay* in 802.1 CM; (c) the switch *store-and-forward latency*. Latency terms (a) and (b) are bounded by the transmission time of the maximum frame size allowed by IEEE 802.1CM at the ingress (2, 000 bytes). Finally, (c) is bounded by the switch’s hardware characteristics, which is constant and in the order of a few microseconds, typically between 200 ns and 5 μ s [6], [40]. Consequently, excluding the small fixed delay terms from t_{network} , we consider the overall latency budget t'_{network} that includes the variable latency terms used for network planning, as:

$$t'_{\text{network}} = t_{\text{self-queueing}}^{\text{Worst-case}} + t_{\text{propagation}} \quad (2)$$

C. TARGET FRONTHAUL DESIGN PARAMETERS

In single-hop architectures as the one shown in Fig. 3b only one switch is required to aggregate traffic from many distant RRHs and distribute it among the pool of BBUs allocated in

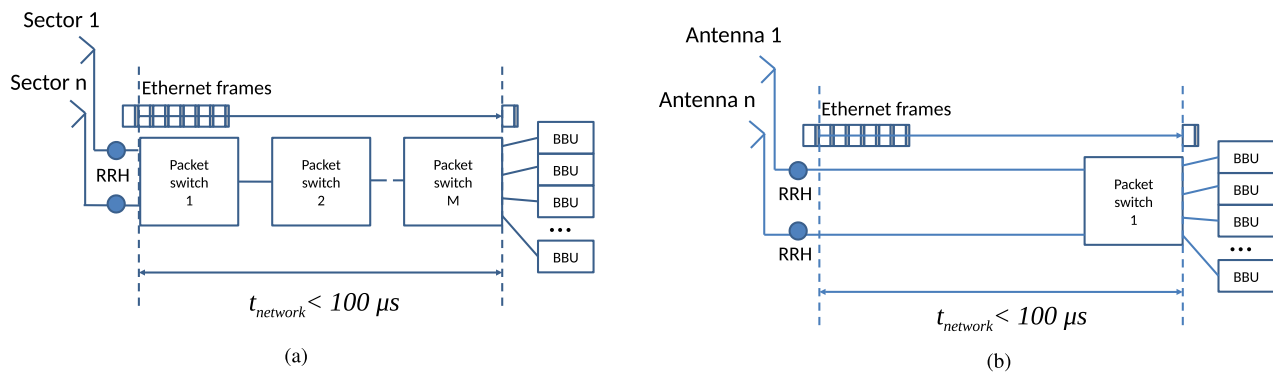


FIGURE 3. Maximum network latency budget for high-priority fronthaul traffic. (a) eCPRI and IEEE 802.1CM [6]. (b) eCPRI latency budget in the one-hop scenario.

the cloud. This is particularly suitable for network virtualization scenarios, where virtual BBUs and technologies like Edge Computing [38] coexist. The high data rate intended for 5G as well as the above-mentioned link capacities needed to support 5G New Radio generally demand the use of shorter transmission distances. In this context, a fronthaul network featuring a reduced number of hops also seems to be a wise choice [10], [39]. Ideally, in a C-RAN architecture, the BBUs are dynamically allocated and shared by a large number of RRHs that may or may not be generating traffic at a given time. In all these contexts, very low latency and few hops to reach the final user are paramount to ensure a proper system operation.

As reviewed in Table 2, IEEE 802.1CM allocates a network latency budget of $100 \mu s$ for HPF traffic so as to be aligned with the requirements established by eCPRI [36], [37]. Unlike IEEE 802.1CM, eCPRI [7] includes the ingress and egress links in the latency budget. Thus, the most restrictive definition (i.e., eCPRI) is taken into account. Having this in mind, a maximum propagation distance of 20 Km could be supported (assuming $5 \mu s/\text{Km}$). In a typical scenario, roughly half of the delay budget ($50 \mu s$) is allocated for propagation delay, which allows a target RRH-BBU distance of up to 10 Km . The remaining $50 \mu s$ should be enough to cope with the rest of the switching delay, that is, processing and queueing delays. However, given the stringent delay requirements of C-RAN, the dimensioning usually considers the maximum theoretical queueing delay (see (2) and 802.1CM).

Since our design objective is to stretch the range of the fronthaul links as much as possible, we propose to loose this requirement and apply the maximum (one-way) Frame Loss Ratio (FLR) defined in 802.1CM for eCPRI HPF (see Table 2) instead of the worst-case delay. The FLR criteria includes all causes of frame drops (transmission errors, congestion, etc.) and frames experiencing late delivery. The IEEE 802.1CM standard (see Section 6.2.3.2) explicitly excludes service unavailability factor in the definition of FLR. Provided that the buffers for IQ data have been dimensioned to the worst-case situation –all IQ bursts arriving at the same time to the switch– the congestion effect on FLR can be considered null.

Additionally, FLR caused by transmissions errors, assuming 1, 500 bytes packets and a link BER of 10^{-12} , is $1.2 \cdot 10^{-8}$ [41], which is almost an order of magnitude smaller than the maximum FLR.

Therefore, assuming that most of the latency budget can be spent on late delivered packets, it is acceptable to use FLR as a design rule. The fronthaul links are dimensioned such that only one out of every 10^7 packet is lost due to late delivery, i.e., $\text{FLR} = 10^{-7}$. This implies finding the right configuration of link lengths and FH network such that the 99.999999th network delay percentile remains below the HPF latency limit of $100 \mu s$. Thus, contrary to what we would do for the worst-case, we use the 99.999999th percentile, and (2) may be expressed as

$$t'_{network} = t_{self-queueing}^{99.999999th} + t_{propagation}. \quad (3)$$

This allows us to maximize the propagation delay budget and, hence, the total fiber length and service coverage. Next, we focus on finding the appropriate tools to model the behavior of the aggregation of eCPRI HPF flows in IEEE 802.1p. For the sake of simplicity, *self-queueing* delay will be referred to as *queueing delay* in the remainder of the paper. With the aim of modeling the 99.999999th queueing delay percentile, we consider both analytical and simulation solutions and assess its suitability for different 5G New Radio settings.

IV. QUEUING LATENCY MODELING AND SIMULATION OPTIONS FOR HIGH PRIORITY FRONTHAUL TRAFFIC

In this section, we review the different options we identified to model and compute the values of the queueing delay percentiles for HPF. In all cases, we shall assume that RRHs and BBUs send OFDM symbols as bursts of back-to-back frames, as specified in Section III.

A. THE G/G/1 MODEL

Firstly, we studied the applicability of a generalized queueing model G/G/1 by adapting it to model the target scenario. The reason for choosing G/G/1 is the following. Markovian models, like the well-known M/M/1 and M/G/1 models, are widely used due to the existence of closed-form expressions

for the mean waiting time in queue. However, the assumption of exponentially distributed time between arrivals in this scenario is not realistic, as shown in [25], [26]. Therefore, the G/G/1 model based on the Allen-Cunneen approximation [42] was chosen. Unfortunately, contrary to the previous one, this model does not provide closed-form formulas; instead, it offers an upper-bound (see (4)) for the mean waiting time in queue that depends on the load of the system (ρ) and the mean service time ($E[S]$). Additionally, it can be particularized for our scenario via the squared coefficient of variation of the arrivals ($C^2[T]$) and service times ($C^2[S]$).

$$E[W_q] \leq E[S] \cdot \frac{\rho}{1-\rho} \cdot \frac{C^2[T] + C^2[S]}{2} \quad (4)$$

Note that, for exponentially distributed time between arrivals and service times, equation (4) can be simplified to the M/M/1 model since $C^2[T] = 1 = C^2[S]$.

Since we are interested in the 99.999999th percentile, we may derive an expression for the p th percentile, as shown in [26], where p represents the percentile of interest:

$$W_q^{(p)} = \max \left\{ 0, E[S] \frac{1}{1-\rho} \frac{C^2[T] + C^2[S]}{2} \ln \left(\frac{\rho}{1-p} \right) \right\}. \quad (5)$$

B. DISCRETE EVENT SIMULATION OF EXTREME PERCENTILES

In addition to validating the accuracy of the G/G/1 queueing model via simulation [26], we evaluate the potential of a simulation tool to find the value of the extreme percentiles that we are interested in as a complementary part to this theoretical models. To this end, we make use of a custom discrete-event simulator specifically programmed to simulate the aggregation of the fronthaul flows in a packet switch, that is, to emulate the arrival of overlapping Split I_U bursts coming from many RRHs. In each repetition, we choose a random alignment of the flows (see *Offset* in Fig. 2b) ranging from 0 to T .

C. N*D/D/1 QUEUEING MODEL

Our third option is the N*D/D/1 analytical model developed in [43]. It models the delay in a first-in/first-out queue and single-resource system. We assume that the arrivals are a superposition of N streams (RRHs) with a reference flow ($N + 1$ flows) following a deterministic pattern and a constant service time in the system's resource (the packet aggregator), i.e., N flows following a D/D/1 profile. Clearly, the worst-case scenario occurs when all the RRHs' bursts are synchronized and arrive at the packet aggregator at the same time. Nevertheless, such scenario is very unlikely and a more accurate estimation can be made by using this model. Assuming RRH's burst lengths of M bits and an output link rate of C bits/s, the service time is deterministic for each burst and equal to $\tau = M/C$. Consequently, for the system to be stable, (6) must hold

$$(N + 1) \cdot \tau < T, \quad (6)$$

where T is the burst period of each RRH –that is, the time elapsed between consecutive RRH's OFDM symbols shipments. The main result of the n*D/D/1 model is the CCDF of the waiting time in queue, $F(x) = P(W_q > x)$. For any waiting time in queue $x \geq 0$, this function is of the form

$$F(x) = T^{-N} P_N(T, x). \quad (7)$$

In addition, for fixed $x \geq 0$ and $n \geq 0$, the function $P_n(t, x)$ is a polynomial of degree $(n - 1)$ in t .

$$P_n(t, x) = \sum_{l=0}^{n-1} q_{n,l}(x)(t - n\tau + x)^l. \quad (8)$$

This means to treat N and T as variables to obtain the distribution of the exact delay for the system. The polynomial's coefficients can be computed by starting at $q_{0,l}(x) = 0$ and continuing with

$$\begin{cases} q_{n,0}(x) = [(n\tau - x)^+]^n \\ q_{n,k}(x) = \frac{n}{k} \sum_{l=k-1}^{n-2} \binom{l}{k-1} \tau^{l-k+1} q_{n-1,l}(x). \end{cases} \quad (9)$$

Let us consider a generic C-RAN architecture that is using a packet switch in order to aggregate N fronthaul flows. The system can be characterized in terms of the load $\rho = \frac{(N+1) \cdot \tau}{T}$, as shown in [43]. Normalizing the problem by using $T = 1$, the survivor function for the waiting time in queue can be computed by solving (7), (8) and (9). Fig. 4 shows the CCDF of the waiting time in queue in the packet switch that aggregates all the fronthaul flows coming from the RRHs, according to the network topology shown in Fig. 1, for different number of remote radio heads (5, 10, 15, 20).

Using these survivor functions, we may compute the percentile values that we desire for any number of merging RRHs at the packet aggregation point. As shown in the following sections, this can be a very useful parameter in order to properly dimension the capacity and size of the fronthaul network.

D. WORST-CASE DELAY DIMENSIONING MODEL

For the purpose of clarification, we present the worst-case dimensioning of the fronthaul network; however, note that this is the option we try to avoid and improve. In this approach, the general functioning and dimensioning of the fronthaul network is conditioned by the worst-case queueing delay –that is, all IQ bursts arriving aligned and at the same time to the switch. Among other things, this includes providing buffers of the appropriate length as well as setting a maximum range for the fronthaul optical links so that the total delay budget remains uncompromised in any situation, including the worst case.

Following the numerical example developed in Section III, in which a 2-antenna, 40 MHz channel MIMO system was studied, assume that the same system is operating in Split I_U mode with a 100% occupancy of radio resources ($\eta = 1$). These numbers led to a Split I_U burst size of 18, 000 bytes

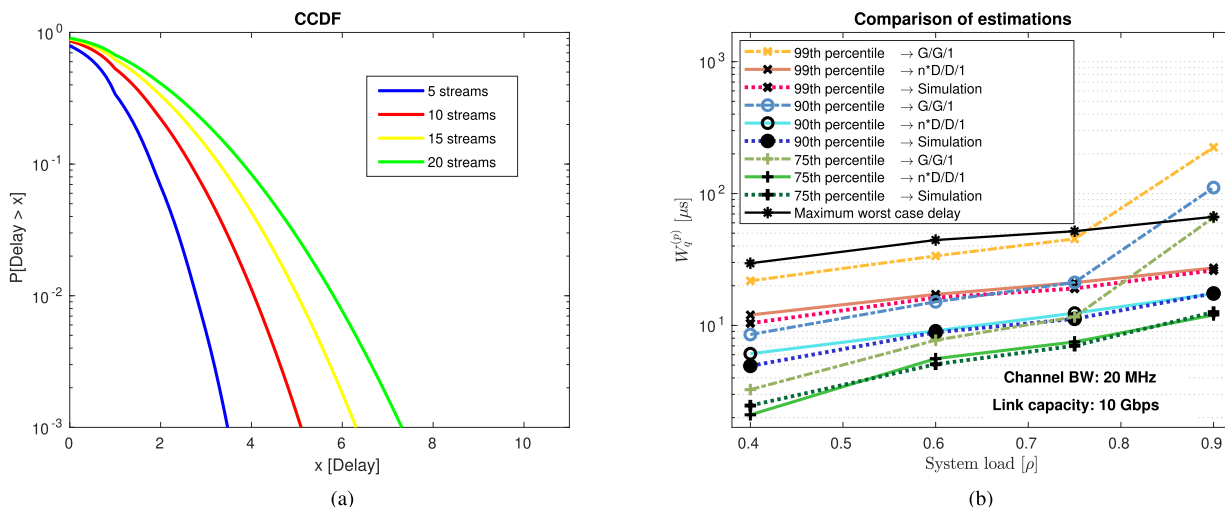


FIGURE 4. N*D/D/1 delay distribution and comparison of estimators. (a) CCDF of the waiting time in queue for the N*D/D/1 example. (b) Percentiles of the waiting time in queue for different system loads.

every $66.\bar{6} \mu s$. Considering a payload size of 1,500 bytes and an overhead of 42 bytes per packet, the burst can be sent using 12 packets. This adds up to a total of $\approx 18,500$ bytes per burst and $R_{RRH} = 2,220$ Mbit/s. Then, assuming a link capacity of $C = 10$ Gb/s and targeting a system load of $\rho = 0.8$, the maximum number of RRHs (Max_{RRHs}) that can be aggregated in the same link is given by,

$$\text{Max}_{RRHs} = \left\lfloor \rho \cdot \frac{C}{R_{RRHs}} \right\rfloor = 3. \quad (10)$$

This means that a maximum of three RRHs can be multiplexed using this optical link. Taking into account the value of Max_{RRHs} , we can compute the maximum theoretical worst-case queueing delay suffered by any burst that has to wait for the others, as $t_{\text{queueing}}^{\text{Worst-case}} = (\text{Max}_{RRHs} - 1) \cdot \frac{\text{Burst Size}}{C} \approx 14.8 \mu s$, which is unrealistic from the practical point of view and can be significantly improved, as we show in the upcoming sections.

V. EXPERIMENTS: COMPARISON OF ESTIMATORS

In this section, we evaluate the accuracy of each of the estimation approaches presented above. We do this for several operation regimes and different percentile values. In addition, we assess the precision of each method for different values of the link load (ρ).

A. GENERAL OVERVIEW: SMALL PERCENTILES

For the sake of the example, consider that the output link's capacity of the aggregation packet switch is $C = 10$ Gb/s. Additionally, assume 20 MHz channels and 1,500 bytes packets to transport the RRHs' Split I_U bursts. This means a 9,000-byte burst transmission every $66.\bar{6} \mu s$, plus overheads. Fig. 4b plots the obtained results for different percentile values of the waiting time in queue while increasing the load of the aggregation point by multiplexing more RRHs

(4, 6, 7, and 9, respectively). In view of the figure, it is worth highlighting several facts.

First, note that the estimation based on the G/G/1 model is, in general, a good upper bound to the percentile of the waiting time in queue. The lower the percentile, the better the estimation. However, it is clear that it overestimates the delay for high-load scenarios, sometimes even exceeding the maximum theoretical delay (see the line with star markers) computed using (6). Having in mind that we aim at even higher percentiles, the G/G/1 queue is not a precise enough tool for our target despite the fact that it is useful for more conservative percentiles [26], as shown in Fig. 4b.

Secondly, we look into the possibility of using a custom discrete event simulator to evaluate the state of the queue. Close inspection of Fig. 4b reveals that simulations give promising results that remain under the maximum theoretical queueing values, which will occur whenever all bursts arrive synchronously to the aggregation point. This simulator has already been validated in [25] and [26]. Nevertheless, the higher the percentile we are looking for in a simulation process, the more trials (and time) we need to achieve results with the appropriate significance values. During these simulations, 99% confidence intervals were computed. They are sufficiently small to be considered negligible and, therefore, are not plotted.

Finally, the N*D/D/1 results (see dotted lines in Fig. 4b) match almost perfectly the simulation outputs for all system loads, even for heavy load scenarios. It is important to note that computing the polynomial together with its coefficients using (8) and (9) is a recursive task that can potentially be very time-consuming for high values of N —that is, for a large number of RRHs being aggregated at the packet switch.

B. EXTREME PERCENTILES

Figures 5a and 5b justify the final model election chosen to dimension the fronthaul network. Particularly, Fig. 5a depicts

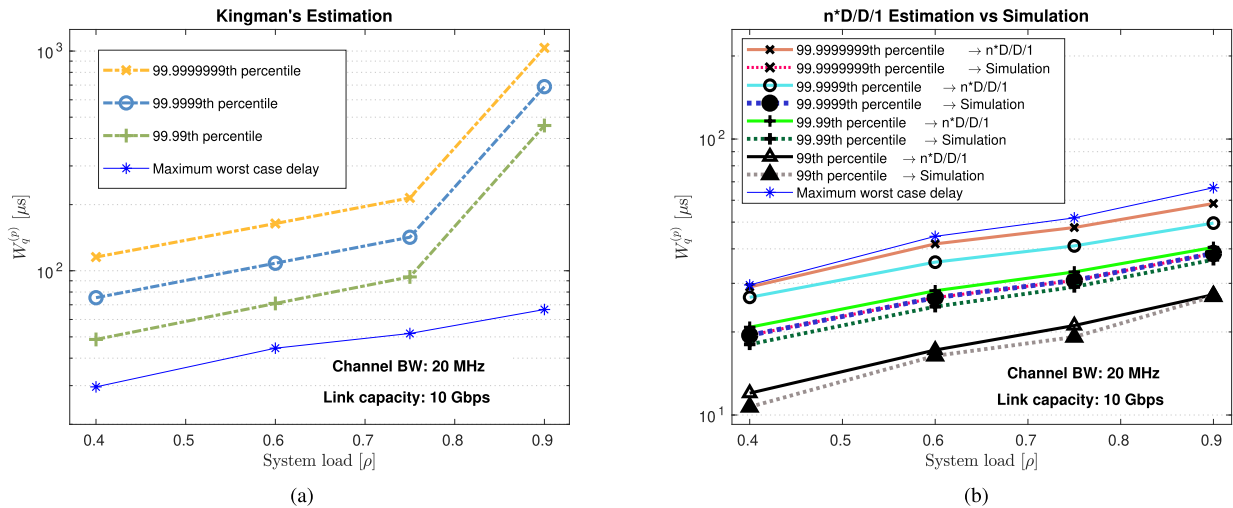


FIGURE 5. Comparison of estimation approaches. (a) Behavior of the percentile estimation using the $G/G/1$ model. (b) Simulation saturation for high percentile values.

the behavior of the $G/G/1$ estimation in the context of extreme percentiles. Namely, 99.99th, 99.9999th, and 99.999999th percentiles are shown. Note that the $G/G/1$ bound for these percentiles is always above the theoretical maximum value for all system loads, which makes it a useless tool for extreme percentiles.

On the other hand, Fig. 5b compares $N^*D/D/1$ results against the **simulation outputs** for extreme percentiles. We include the 99th queueing delay percentile as a reference point where both approaches produce similar results. However, as we look for more extreme percentiles, this does no longer hold. We observe that simulation estimations saturate at a certain point and converge to a certain value (see overlapping red and blue dotted lines that represent the simulated 99.999999th and 99.9999th percentiles, respectively).

On the contrary, $N^*D/D/1$ estimation keeps approaching the maximum theoretical value as the percentile grows. This means that the number of simulations is not enough to grasp the desired percentile value, and even a narrow confidence interval can be deceptive.

Increasing the number of repetitions in the simulation is unaffordable from a practical point of view. Given a 99.999999th percentile value x , it is clear that, on average, only one out of every 1,000,000,000 packets would suffer from a queueing delay higher than x . For the sake of statistical significance, let us seek 100 occurrences of that event, which would mean, on average, to simulate $100 \cdot 10^9$ packets.

Considering the aggregation of 9 RRHs, each one transmitting 9,000 bytes bursts and using 1,500 bytes packets, $6 \cdot 9 = 54$ packets are generated per period. Averaging 2,000 simulations, which is the value used to obtain the above figures, we are able to see $2,000 \text{ simulations} \cdot 54 \text{ packets/simulation} = 108 \cdot 10^3$ packets. Hence, we need ≈ 1000 times more packets. Having in mind that each batch of 2,000 simulations takes, on average, 12 seconds to

complete,² the whole process would take approximately $11 \cdot 10^6$ seconds. This represents around 128 days of computation time, making the simulation a slow tool for the most extreme percentiles. However, note that the simulation outputs and $N^*D/D/1$ estimations match for 99.99th, 99th, 90th, 75th or lower percentiles, as shown in Fig. 4b.

VI. APPLICATION: $N^*D/D/1$ DIMENSIONING FOR 5G NR

Once that we have weighted the pros and cons of each approach, the $N^*D/D/1$ queueing model is selected as the tool to dimension the total length of the fronthaul links. The FLR specified by the 802.1CM standard can be met under the assumptions explained in Section IV by considering the 99.999999th queueing delay percentile ($1 - \text{FLR}$) instead of the worst-case delay.

Let us consider the same scenario but, this time, with the numbers of the new air interface developed for the next generation mobile networks: 5G New Radio (NR) [8]. Since future services are envisioned to be data-intensive (e.g., video streaming, immersive applications, virtual reality), there is a growing need for high end-user data rates. Consequently, the capabilities of the fronthaul network must scale accordingly. In Table 1, we showed the numerology for the 5G NR interface regarding different functional splits and channel bandwidths. Particularly, we focus our study on three use cases: 50 MHz, 100 MHz, and 200 MHz channels, so as to show the pattern that arises as we increase the demand in the fronthaul network.

A. SCENARIO I: 5G NEW RADIO (50 MHz CHANNEL)

Consider the aggregation of Split I_U fronthaul flows coming from a number of RRHs. Assume that these support **50 MHz channels** with 15 KHz subcarrier spacing [8]. The aggregator

²Intel Xeon Processor E3-1505M v6, 8M Cache, 3.00 GHz

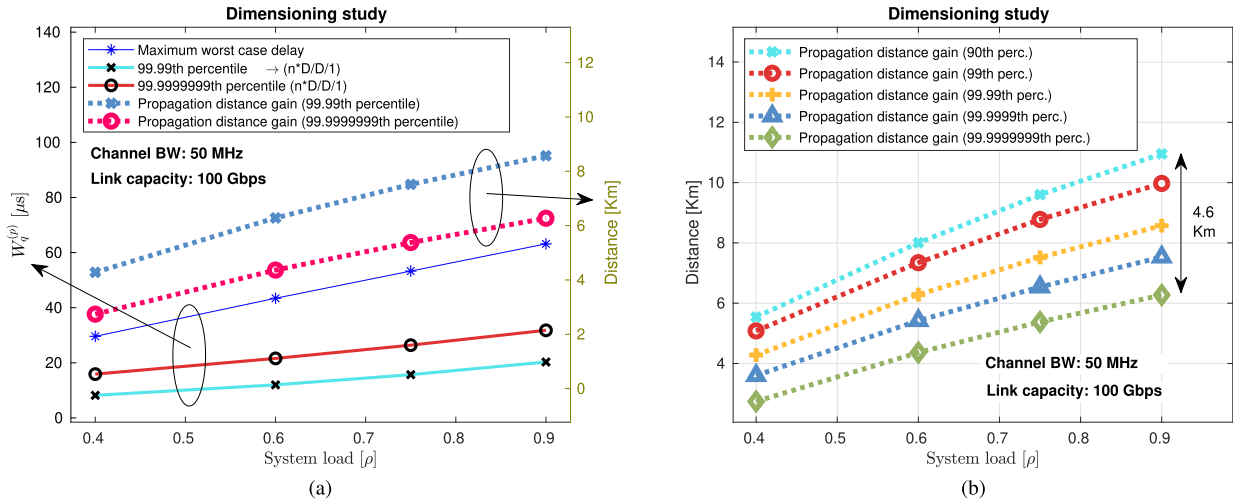


FIGURE 6. Distance gained due to the use of (1-FLR)% packet latency percentile instead of worst-case delay. (a) Distance gain computation (50 MHz channels). (b) Distance gain for 50 MHz channels and different dimensioning percentiles.

output is driven by a 100 Gb/s optical transceiver. Also, we make use of 1,500 bytes Ethernet packets in order to transport Split I_U flows. Next, we compute the value of the 99.999999th queuing delay percentile, by modeling our system as a $N^*D/D/1$ queue and compare the results with the maximum theoretical queuing delay. By taking the difference between these two values, we obtain the extra delay budget available that is gained by using the 99.999999th percentile instead of the worst case as the dimensioning reference. This extra time can be spent at our discretion –either to increase the reach of the fronthaul links or to aggregate more RRHs.

Assuming a $5 \mu s/km$ propagation delay in the FH’s optic fiber, we can compute the extra propagation distance that is gained if the additionally available delay budget is spent on propagation. Fig. 6 illustrates the aforementioned comparison for different number of aggregated RRHs, i.e., for various aggregator’s load conditions. Namely, we plot the results for system loads equal to $\rho = 0.4, 0.6, 0.75$ and 0.9 . The number of RRHs, i.e., Split I_U flows that we are able to multiplex is given by (10), reworked as

$$\text{Num}_{RRHs} = \left\lfloor \frac{\rho \cdot C}{\text{Bitrate}_{\text{Split } I_U}} \right\rfloor, \quad (11)$$

which, for 100 Gb/s links, means aggregating roughly 14, 21, 26, and 31 eCPRI Split I_U flows, respectively. In view of the results, there are two interesting facts that are worth highlighting:

- 1) **The higher the load** of the system, the bigger the gap between the maximum theoretical queuing delay and the $N^*D/D/1$ solution.
- 2) Consequently, **the more extra latency budget** can be spent on propagation, with respect to the worst-case solution.

For instance, under heavy load conditions ($\rho = 0.9$), including overheads (eCPRI and Ethernet) and using (10),

we get that the maximum theoretical queuing delay is $t_{\text{queuing}}^{\text{Worst-case}} \approx 63.16 \mu s$ (see the solid line with star markers in Fig. 6). However, if we choose to dimension according to the 99.999999th queuing delay percentile:

- a) $W_q^{99.999999} \approx 31.8 \mu s$ (see the solid line with circle markers in Fig. 6).
- b) We save an extra: $t_{\text{queuing}}^{\text{Worst-case}} - W_q^{99.999999} \approx 31.36 \mu s$, which represents a $\approx 30\%$ of the $100 \mu s$ delay budget.

Additionally, this translates into an **additional propagation distance budget**, i.e., these extra $31.36 \mu s$ enable us to extend the link up to nearly 6.3 Km . This gain is achieved by using the 99.999999th queuing delay percentile given by the $N^*D/D/1$ model. It would represent an approximately 60% increase in the maximum distance with respect to the 10 Km -baseline mentioned in Section III-B under heavy load conditions ($\rho = 0.9$). Alternatively, we may choose to relax the 99.999999th percentile rule and use lower percentiles. Note, in Fig. 6, that the 99.99th percentile is obviously farther from the maximum queuing delay than the 99.999999th percentile. This enables us to take advantage of additional propagation delay budget. Now, for the same system load ($\rho = 0.9$), a supplementary $\approx 8.6 \text{ Km}$ would be available if we choose to dimension using the 99.99th percentile. By doing so, we achieve larger propagation distances at the expense of higher FLR. This could be tackled by making use of different techniques, such as Forward Error Correction (FEC) protocols, network coding, etc. However, its impact on the final delay [44] should be studied in detail in future work.

Following the same reasoning for other percentiles, Fig. 6b shows the propagation distance gains achieved for different system loads, depending on which percentile we use in the dimensioning process. It is worth highlighting that the delay budget savings for $\rho = 0.9$ span from around 6.3 Km using the 99.999999th queuing delay percentile to roughly 10.95 Km the 99th percentile is used, which represents a difference of nearly 4.6 Km .

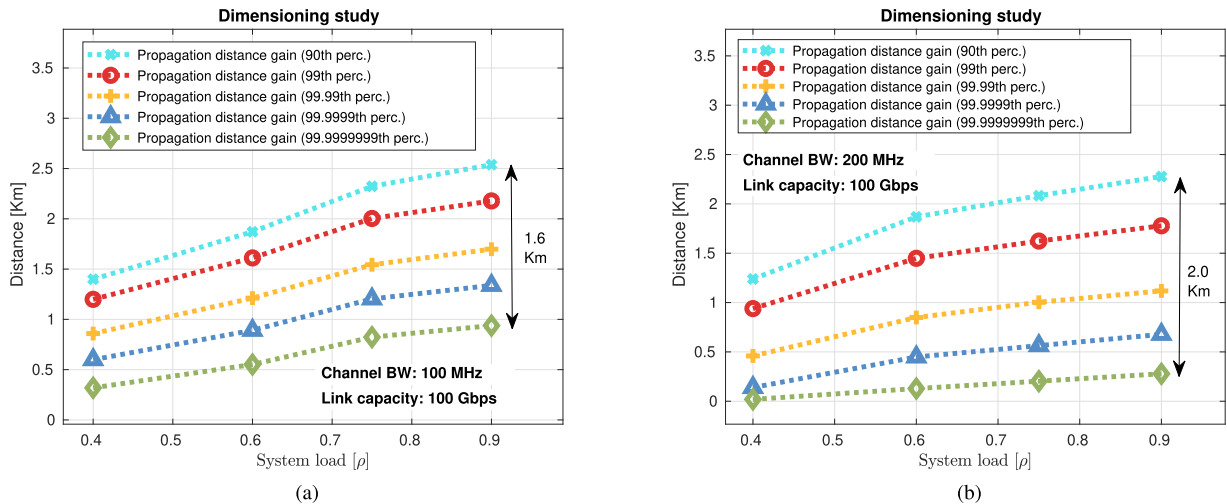


FIGURE 7. Distance gains for the 90th, 99th, 99.99th, 99.9999th, 99.999999th percentiles; 100 MHz and 200 MHz channels. (a) Propagation distance gains (100 MHz channels). (b) Propagation distance gains (200 MHz channels).

B. SCENARIO II: 5G NEW RADIO (100 MHz CHANNEL)

Consider, again, the same output rate for the aggregation point, that is, 100 Gb/s. This time, each RRH supports 100 MHz channels with 60 KHz subcarrier spacing. Again, 1, 500 bytes packets are used to transport the bursts. For 100 Gb/s links, aggregating 7, 10, 13, and 15 eCPRI Split I_U flows leads to system loads of $\rho = 0.4, 0.6, 0.75,$ and $0.9,$ respectively (apply (10) with the numbers provided in Table 1).

By repeating the same procedure, we obtain the extra propagation delay available budget by taking the difference between the maximum theoretical queuing delay and the 99.999999th percentile computed with the $N^*D/D/1$ estimation. Close inspection of Fig. 7 reveals that the overall gain is worse than that obtained for 50 MHz channels. Under heavy load conditions –that is, when $\rho = 0.9$ – the distance gain using the 99.999999th queuing delay percentile as the dimensioning reference is near 1 Km. This gain is smaller in comparison to what we obtained for smaller channel bandwidths, but it is not negligible since it represents an approximate 10% length gain in the fronthaul links.

Again, relaxing the reference percentile, we may obtain roughly 1.3 Km, 1.7 Km, 2.1 Km, and 2.5 Km length gains for the 99.9999th, 99.99th, 99th, and 90th percentiles, respectively, assuming heavy load conditions.

C. SCENARIO III: 5G NEW RADIO (200 MHz CHANNEL)

In this third experiment, we assess the dimensioning results for high-bandwidth-demanding 5G NR channels. Namely, 200 MHz channels with 60 KHz subcarrier spacing. This represents eCPRI Split I_U bursts of $\approx 23, 753$ bytes every $16.67 \mu s,$ which requires a transport capacity of up to ≈ 11.4 Gb/s. These bursts are then split and packetized into 1, 500 bytes Ethernet packets.

In order to cope with this traffic load, we assume 100 Gb/s links that are carrying 3, 5, 6, and 7 RRH flows simultaneously and, therefore, achieving link occupancies of $\rho = 0.4,$

0.6, 0.75, and 0.9, respectively. Fig. 7b confirms that the distance gain keeps decreasing as we increase the channel bandwidth. In this case, only ≈ 0.2 Km are gained for the 99.999999th percentile at $\rho = 0.9.$ As for the rest of percentiles (99.9999th, 99.99th, 99th, and 90th), ≈ 0.6 Km, ≈ 1.1 Km, ≈ 1.8 Km, and ≈ 2.3 Km are achieved, respectively.

It is worth highlighting that the burst size in this example (200 MHz; 60 KHz spacing) is the same as that of Subsection VI-A (50 MHz; 15 KHz spacing). The main reason why the distance gain is hampered in this case is that the $1/Q$ symbol period is four times smaller, which makes a favorable alignment of the flows more unlikely, as there is less room for them to avoid overlapping in the aggregation point.

VII. SUMMARY AND CONCLUSIONS

In this work, we propose the use of extreme latency percentiles as a design parameter rather than maximum end-to-end one-way delay. The aim of this approach is to stretch the RRH-BBU distance by using the available FLR budget for scenarios in which such range extension is necessary.

To this end, we compared several options for computing extreme queuing delay percentiles for fronthaul traffic. Namely, we assessed the suitability of discrete-event simulations, $G/G/1$ and $N^*D/D/1$ queueing models. We concluded that, while the $G/G/1$ model and simulations can produce satisfactory results for moderate percentiles, both saturate in the context of high system loads and extreme percentile values. Only the $N^*D/D/1$ queue is appropriate for the extreme percentiles.

A better modeling of these percentiles enables us to comply with the defined FLR in IEEE 802.1CM. We may interpret the gap between this estimation and maximum worst-case delay as an extra delay budget. This extra budget becomes relevant at high loads. Experiments revealed that additional propagation can be gained at 100 Gb/s under the appropriate conditions, as discussed in Section VI. The rule of thumb is that the higher the load, the more extra latency budget

we can obtain, proportionally to the maximum worst-case delay, since the gap between the percentiles and the maximum theoretical queueing delay becomes wider.

Taking into consideration that the envisioned distance for the fronthaul links is up to 10 Km, we find that we are able to extend the fronthaul links up to around 60% for 50 MHz channels, 10% for 100 MHz channels, and 2% for 200 MHz channels. Alternatively, this extra budget could be used to aggregate more RRHs at the same aggregation point, or we could even think about dynamically switching to more resource-demanding functional splits on certain RRHs, if needed.

ACKNOWLEDGMENT

The work of G. Otero Pérez was supported by the Spanish Ministry of Education, Culture and Sport by means of the FPU under Grant FPU16/01760.

REFERENCES

- [1] 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.
- [2] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen, "Evaluating C-RAN fronthaul functional splits in terms of network level energy and cost savings," *J. Commun. Netw.*, vol. 18, no. 2, pp. 162–172, Apr. 2016.
- [3] Common Public Radio Interface. *Interface Specification v7.0*. Accessed: Mar. 15, 2019. [Online]. Available: <http://www.cpri.info/spec.html>
- [4] F. Cavaliere, P. Iovanna, J. Manges-Bafalluy, J. Baranda, J. Núñez-Martínez, K.-Y. Lin, H.-W. Chang, P. Chanclou, P. Farkas, J. Gomes, L. Cominardi, A. Mourad, A. De La Oliva, J. A. Hernández, D. Larrabeiti, A. Di Giglio, A. Paolicelli, and P. Ödling, "Towards a unified fronthaul-backhaul data plane for 5G The 5G-Crosshaul project approach," *Comput. Standards Interfaces*, vol. 51, pp. 56–62, Mar. 2017.
- [5] *Common Public Radio Interface: ECPRI Interface Specification v1.2*. Accessed: Jun. 25, 2018. [Online]. Available: <http://www.cpri.info/spec.html>
- [6] *IEEE Standard for Local and Metropolitan Area Networks-IEEE Time-Sensitive Networking for Fronthaul*, IEEE Standard 802.1cm, 2018. [Online]. Available: <http://www.ieee802.org/1/pages/802.1cm.html>
- [7] *Common Public Radio Interface: Requirements for the eCPRI Transport Network VI.1*. Accessed: Oct. 24, 2017. [Online]. Available: http://www.cpri.info/downloads/Requirements_for_the_eCPRI_Transport_Network_V1_1_2018_01_10.pdf
- [8] *NR; Base Station (BS) Radio Transmission and Reception*, document 3GPP TS 38.104 version 15.2.0 Release 15, 2018.
- [9] *Study on New Radio Access Technology; Radio Access Architecture and Interfaces*, document 3GPP TR 38.912 version 14.0.0 Release 14, Jun. 2017.
- [10] N. J. Gomes, P. Chanclou, P. Turnbull, A. Magee, and V. Jungnickel, "Fronthaul evolution: From CPRI to Ethernet," *Opt. Fiber Technol.*, vol. 26, pp. 50–58, Dec. 2015.
- [11] T. Wan and P. Ashwood-Smith, "A performance study of CPRI over Ethernet with IEEE 802.1Qbu and 802.1Qbv enhancements," in *Proc. IEEE Globecom*, Dec. 2015, pp. 1–6.
- [12] *IEEE Draft Standard for Local and Metropolitan Area Networks-Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks Amendment: Frame Preemption*, IEEE Standard P802.1Qbu/03.0, 2015.
- [13] *IEEE Approved Draft Standard for Local and Metropolitan Area Networks—Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks Amendment: Enhancements for Scheduled Traffic*, IEEE Standard P802.1Qbv/D3.1, 2015.
- [14] M. K. Al-Hares, P. Assimakopoulos, S. Hill, and N. J. Gomes, "The effect of different queuing regimes on a switched Ethernet fronthaul," in *Proc. 18th Int. Conf. Transparent Opt. Netw. (ICTON)*, Trento, Italy, Jul. 2016, pp. 1–4. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7550324&isnumber=7550246>
- [15] A. Gowda, J. A. Hernández, D. Larrabeiti, and L. Kazovsky, "Delay analysis of mixed fronthaul and backhaul traffic under strict priority queueing discipline in a 5G packet transport network," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 6, 2017, Art. no. e3168.
- [16] C. Ranaweera, E. Wong, A. Nirmalathas, C. Jayasundara, and C. Lim, "5G C-RAN with optical fronthaul: An analysis from a deployment perspective," *J. Lightw. Technol.*, vol. 36, no. 11, pp. 2059–2068, Jun. 1, 2018. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8194738&isnumber=8307216>. doi: 10.1109/JLT.2017.2782822.
- [17] M. Waqar, A. Kim, and P. K. Cho, "A transport scheme for reducing delays and jitter in Ethernet-based 5G fronthaul networks," *IEEE Access*, vol. 6, pp. 46110–46121, 2018. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8430513&isnumber=8274985>. doi: 10.1109/ACCESS.2018.2864248.
- [18] S. Bjørnstad, D. Chen, and R. Veisllari, "Handling delay in 5G Ethernet mobile fronthaul networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Ljubljana, Slovenia, Jun. 2018, pp. 1–9. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8442755&isnumber=8442432>. doi: 10.1109/EuCNC.2018.8442755.
- [19] C.-Y. Chang, R. Schiavi, N. Nikaiein, T. Spyropoulos, and C. Bonnet, "Impact of packetization and functional split on C-RAN fronthaul performance," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7511579&isnumber=7510595>. doi: 10.1109/ICC.2016.7511579.
- [20] C.-Y. Chang, N. Nikaiein, and T. Spyropoulos, "Impact of packetization and scheduling on C-RAN fronthaul performance," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–7. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7841885&isnumber=7841475>. doi: 10.1109/GLOBECOM.2016.7841885.
- [21] R. M. Rao, M. Fontaine, and R. Veisllari, "A reconfigurable architecture for packet based 5G transport networks," in *Proc. IEEE 5G World Forum (5GWF)*, Silicon Valley, CA, USA, Jul. 2018, pp. 474–477. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8517039&isnumber=8516707>. doi: 10.1109/5GWF.2018.8517039.
- [22] I. Ucar, B. Smeets, and A. Azcorra, "Simmer: Discrete-event simulation for R," 2017, *arXiv:1705.09746*. [Online]. Available: <https://arxiv.org/abs/1705.09746>
- [23] I. Ucar, J. A. Hernández, P. Serrano, and A. Azcorra, "Design and analysis of 5G scenarios with simmer: An R package for fast DES prototyping," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 145–151, Nov. 2018. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8423801&isnumber=8539002>. doi: 10.1109/MCOM.2018.1700960.
- [24] M. Jaber, D. Owens, M. A. Imran, R. Tafazolli, and A. Tukmanov, "A joint backhaul and RAN perspective on the benefits of centralised RAN functions," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 226–231. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7503792&isnumber=7503749>. doi: 10.1109/ICC.2016.7503792.
- [25] G. O. Pérez, J. A. Hernández, and D. L. López, "Delay analysis of fronthaul traffic in 5G transport networks," in *Proc. IEEE 17th Int. Conf. Ubiquitous Wireless Broadband (ICUWB)*, Salamanca, Spain, Sep. 2017, pp. 1–5. doi: 10.1109/ICUWB.2017.8250956.
- [26] 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. doi: 10.1364/JOCN.10.000573.
- [27] F. Giannone, H. Gupta, K. Kondepudi, D. Manicone, A. Franklin, P. Castoldi, and L. Valcarengi, "Impact of RAN virtualization on fronthaul latency budget: An experimental evaluation," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec 2017, pp. 1–5.
- [28] *ARNO-5G Testbed*. Accessed: Mar. 15, 2019. [Online]. Available: <http://arnotestbed.santannapisa.it>
- [29] D. Chitimalla, K. Kondepudi, L. Valcarengi, M. Tornatore, and B. Mukherjee, "5G fronthaul-latency and jitter studies of CPRI over Ethernet," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 9, no. 2, pp. 172–182, Feb. 2017.
- [30] L. Valcarengi, K. Kondepudi, and P. Castoldi, "Time-versus size-based CPRI in Ethernet encapsulation for next generation reconfigurable fronthaul," *J. Opt. Commun. Netw.*, vol. 9, no. 9, pp. D64–D73, Sep. 2017.

- [31] F. Civerchia, L. Kondepue, F. Giannone, S. Doddikrinda, P. Castoldi, and L. Valcarenghi, "Encapsulation techniques and traffic characterisation of an Ethernet-based 5G fronthaul," in *Proc. 20th Int. Conf. Transparent Opt. Netw. (ICTON)*, Bucharest, Romania, Jul. 2018, pp. 1–5. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8473737&isnumber=8473576>. doi: [10.1109/ICTON.2018.8473737](https://doi.org/10.1109/ICTON.2018.8473737).
- [32] Z. Ghebretensae, K. Laraqui, S. Dahlfors, F. Ponzini, L. Giorgi, S. Stracca, J. Chen, Y. Li, J. Hansryd, and A. R. Pratt, "Transmission solutions and architectures for heterogeneous networks built as C-RANs," in *Proc. 7th Int. ICST Conf. Commun. Netw. China (CHINACOM)*, Aug. 2012, pp. 748–752. doi: [10.1109/ChinaCom.2012.6417583](https://doi.org/10.1109/ChinaCom.2012.6417583).
- [33] A. de la Oliva, J. A. Hernández, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 152–159, Feb. 2016.
- [34] *User Equipment (UE) Radio Transmission and Reception*, document 3GPP TS 38.101-1 version 15.2.0 Release 15, Jul. 2018. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/138100_138199/13810101/15.02.00_60/ts_13810101v150200p.pdf
- [35] (2011). *LTE 1800 MHz: Introducing LTE With Maximum Reuse of GSM Assets*. [Online]. Available: <https://www.gsma.com/spectrum/wp-content/uploads/2012/03/lte1800mhzwhitepaper0.9.pdf>
- [36] CPRI. (2016). *Functional Decomposition Requirements*. [Online]. Available: <http://www.ieee802.org/1/files/public/docs2016/cm-CPRI-functional-decomposition-requirements-0516-v01.pdf>
- [37] G. Garner and S. Bao. (2016). *Comments on 802.1cm Synchronization*. [Online]. Available: <http://www.ieee802.org/1/files/public/docs2016/cm-baosh-synchronization-comments-on-D0-4-0916-v02.pdf>
- [38] J. Pan and J. McElhannon, "Future edge cloud and edge computing for Internet of Things applications," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 439–449, Feb. 2018.
- [39] T. Chekolet, E. Larsen, and K. Mahmood, "Performance analysis of cloud radio access network," M.S. thesis, Dept. Inf. Secur. Commun. Technol., Norwegian Univ. Sci. Technol., Aug. 2017. [Online]. Available: <http://hdl.handle.net/11250/2461339>
- [40] Cisco, "Cisco nexus 3548 switch performance validation," Spirent, Crawley, U.K., Dec. 2012. [Online]. Available: https://www.cisco.com/c/dam/en/us/products/collateral/switches/nexus-3548-switch/white_paper_c11-716751.pdf
- [41] B. Varga and J. Farkas, *Packet/Frame Loss Considerations for CPRI Over Ethernet*, Standard IEEE 802.1 Interim, 2016. [Online]. Available: <http://www.ieee802.org/1/files/public/docs2016/cm-varga-CPRI-packetloss-considerations-0116-v02.pdf>
- [42] J. F. C. Kingman, "The single server queue in heavy traffic," *Math. Proc. Cambridge Phil. Soc.*, vol. 57, no. 4, pp. 902–904, 1961. doi: [10.1017/S0305004100036094](https://doi.org/10.1017/S0305004100036094).
- [43] A. Eckberg, "The single server queue with periodic arrival process and deterministic service times," *IEEE Trans. Commun.*, vol. 27, no. 3, pp. 556–562, Mar. 1979. doi: [10.1109/TCOM.1979.1094425](https://doi.org/10.1109/TCOM.1979.1094425).
- [44] W. Lautenschlaeger, L. Dembeck, and U. Gebhard, "Prototyping optical Ethernet—A network for distributed data centers in the edge cloud," *J. Opt. Commun. Netw.*, vol. 10, no. 12, pp. 1005–1014, 2018.



GABRIEL OTERO PÉREZ received the B.S. and M.S. degrees in telecommunications engineering from the University of Vigo, Spain, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree in telematics engineering with the Universidad Carlos III de Madrid (UC3M), Spain.

He joined the Research Group on Advanced Switching and Communication Technologies, UC3M, where he is currently involved in traffic modeling, access/network protocol modeling, and testing. He is also as an Assistant Professor. He has a first-author journal publication, two first-author conference publications, and coauthored another conference publication.

Mr. Otero Pérez has received the 22 Spanish Government Ph.D. Scholarships in the field of electrical engineering and telecommunications.



DAVID LARRABEITI LÓPEZ received the degree and the Ph.D. degree in telecommunications engineering from Universidad Politécnica de Madrid. He has been with the Telematics Engineering Department, Universidad Carlos III de Madrid (UC3M), since 1998, where he is currently a Professor of switching and optical networks.

He is currently involved in a number of EU research projects on new optical networking paradigms, including PASSION, BlueSPACE, and Metro-haul. He has also served on the TPC for ECOC, HPSR, GLOBECOM, and other conferences. His current research interests include fronthaul traffic transport, ultra-low latency switching, and the tactile Internet.



JOSÉ ALBERTO HERNÁNDEZ received the degree in telecommunications engineering from UC3M, in 2002, and the Ph.D. degree in computer science from Loughborough University, Leicester, U.K., in 2005.

He has been a Senior Lecturer with the Department of Telematics Engineering, Universidad Carlos III de Madrid, since 2010, where he combines teaching and research in the areas of optical WDM networks, next-generation access networks, metro Ethernet, energy efficiency, and hybrid optical-wireless technologies. He has published more than 100 articles in both journals and conference proceedings on his research topics. He is the coauthor of the book *Probabilistic Models for Computer Networks: Tools and Solved Problems*.

• • •