Latency Bounds of Packet-Based Fronthaul for Cloud-RAN with Functionality Split

Ghizlane Mountaser, Maliheh Mahlouji, Toktam Mahmoodi
Centre for Telecommunications Research
Department of Informatics, King's College London
London WC2B 4BG, UK

{ghizlane.mountaser, maliheh.mahlouji, toktam.mahmoodi}@kcl.ac.uk

***Abstract*—The emerging Cloud-RAN architecture within the fifth generation (5G) of wireless networks plays a vital role in enabling higher flexibility and granularity. On the other hand, Cloud-RAN architecture introduces an additional link between the central, cloudified unit and the distributed radio unit, namely fronthaul (FH). Therefore, the foreseen reliability and latency for 5G services should also be provisioned over the FH link. In this paper, focusing on Ethernet as FH, we present a reliable packet-based FH communication and demonstrate the upper and lower bounds of latency that can be offered. These bounds yield insights into the trade-off between reliability and latency, and enable the architecture design through choice of splitting point, focusing on high layer split between PDCP and RLC and low layer split between MAC and PHY, under different FH bandwidth and traffic properties. Presented model is then analyzed both numerically and through simulation, with two classes of 5G services that are ultra reliable low latency (URLL) and enhanced mobile broadband (eMBB).**

***Index Terms*—Cloud-RAN; Fronthaul; Ethernet; Latency; Reliability; Upper bound; Lower bound.**

I. Introduction

One of the architecture enablers of fifth generation (5G) is Cloud-RAN that is supported through multiple technological advances in the network including softwarization, virtualization and cloudification [1]. Cloud-RAN despite bringing higher flexibility and granularity to the network architecture, introduces an additional communication link, i.e. fronthaul (FH). Hence, the ability for the FH to flexibly scale up with data rate has become critical to the success of Cloud-RAN. The need for flexibility in the FH has opened up the possibility of flexibly splitting Radio Access Network (RAN) functionalities between central unit (CU) and distributed unit (DU). The advantage to such an architectural approach is the use of different transport such as packet-based FH. Adopting packet-based FH in the Cloud-RAN architecture allows the use of widely deployed Ethernet-based network. At the same time, packet-based networks impose challenges in ensuring the high reliability and low-latency over the FH communication, which are the key performance indicators expected of 5G.

Reliability can be typically achieved by retransmission or redundancy. However, reliability is usually increased at the cost of latency which poses a major challenges when latency requirement is very stringent. For this reason, the design of Cloud-RAN solution involves one key design question, that is, which functional splits may be suitable from a reliability-latency point of view under the constraint required by 5G scenarios.

Given that each split comes with its own delay requirements, having the knowledge of latency bounds of FH will allow us to decide which split is the most appropriate in the Cloud-RAN architecture. On the other hand, 5G traffic classes eMBB (enhanced mobile broadband), URLLC (ultra reliable low latency communications) and Massive machine type communications, each come with varied requirements on reliability that should also be maintained on the FH link; improving reliability often results in increasing latency. To this end, focusing on reliable packet-based FH [2], the aim of this paper is to compute the lower bound and upper bound of latency analytically, using stochastic network calculus [3], [4], [5]. We simulate the reliable packet-based FH, demonstrating where and how delay bounds are achieved for two classes of traffic, which are eMBB and URLLC. Having these bounds, we further analyse where each functionality split can be the best architectural choice. This paper is an expended work to our previous work [2] where we investigated how to improve reliability and latency of packet based fronthauling by means of multi-path diversity and erasure coding; and [6], [7] which examined lower layer and higher layer splits through an experimental testbed considering an Ethernet-based FH.

The remainder of this paper is organized as follows. Section II gives a brief overview on Cloud-RAN and its different transport technologies and explores reliability on the FH. In section III, we elaborate system model of Cloud-RAN with multi-path FH using coding to analyse reliability-latency and we shed light on functional split requirements in term of latency. In section IV, we compute stochastic delay lower and upper bounds of the system model. The analytic and simulation results are studied in section V. Finally, conclusion and future research are presented in section VI.

II. Background

Cloud-RAN is considered one of the key enablers of 5G architecture, given its desired properties [8]. Despite the attractive advantages of the conventional Cloud-RAN, the architecture whereby a standard common public radio interface (CPRI) is used to transport base band radio samples between

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

CU and DU faces several challenges. The first challenge is the user data is transmitted in the form of an IQ-data block which requires large bandwidth of 157.3 Gbps, considering a 100 MHz transmission bandwidth and 32 antenna ports [9]. Thereby, the high-throughput requirement poses challenges for the FH interface. The second challenge is that CPRI requires stringent latency and jitter requirements. It requires an end-to-end latency of around 250 μs [10]. Such critical requirements are making CPRI very challenging. To relax the excessive bandwidth and latency requirements, as well as to enhance the flexibility of the FH, functional split is introduced whereby a more flexible placement of baseband functionality between the DU and the CU is considered. In fact several options of functional splits have been standardized in 3GPP [9]. Moreover, fundamental simulations and experimentations are carried out by various academic studies [11], [7], [12]. Nonetheless, each split point has different requirements such as latency and bandwidth. These requirements have to be considered in order to select the appropriate functional split. In general, the lower the split point, the greater the level of centralization, the higher is the required interface data rate and the more stringent is the latency requirement.

In traditional Cloud-RAN architecture, fibre has been defined as an ideal attractive solution to meet the strict requirements of high bandwidth and low latency of CPRI. However, there are situations where deployment of fibre is difficult or not a good choice due to cost. In this end, packet-based FH can be considered as a promising alternative transport. This highly cost effective solution allows sharing and convergence with Ethernet-based fixed networks and offers great flexibility. However, packet-based FH imposes many challenges such as high latency and high jitter. Nevertheless, it can be used in functional split where the latency and jitter requirements are relaxed. For example, some of the authors of this work has demonstrated the feasibility of splitting between MAC and PHY in their past work [6]. Further studies have shown that the requirements of different 5G service classes, including the URLLC service can be accommodated using packet-based FH [7]. In [13] authors analysed impact of packetization on the Cloud-RAN and they analyzed different packet scheduling to increase the multiplexing gain.

In addition to low latency and excessive data rate, reliability is also an important metric in 5G [14], [15]. Under the Cloud-RAN architecture, FH needs to provide comparative reliability to enable adoption of Cloud-RAN. The most well used methods to improve reliability are retransmission, multi-path with packet duplication and multi-path with coding. Retransmission is a straight forward way to achieve reliability. However, retransmission can have significant impact on increasing the latency making it non-viable solution on the FH where delays cannot be afforded. By contrast, path diversity with duplication offers better latency in the expense of significant transmission overhead by duplicating packets over multiple interfaces increasing FH network congestion. Two important considerations in such approaches are latency and FH overhead. An alternative solution that provides trade-off between latency and FH overhead is channel coding which can add controlled redundancy to achieve desired reliability and splits the total amount of information to transmit across different paths. Additional reliability, using any technique, sacrifies latency and hence looking at the boundaries of latency that can be offered under certain reliability is of interest to the application in-need of both [14].

III. System Model

In this section, we first introduce the system model for the Cloud-RAN system with multi-path FH and then shed a light on functional split requirements in term of latency.

A. System Model

Our system model consists of Cloud-RAN with a single CU and a single DU connected with multiple FH paths (n different paths), where each path i has a capacity ψ_i. Packets of size B bits are arrived to the system with exponential inter-arrival periods with average $1/\lambda$ seconds (s). We assume that the FH links are identical. Each link is modelled as a single queue. We suppose that the service time of each queue follows an exponential distribution. The mean service time to transmit a packet of size B bits from CU to DU is $1/\mu = B/\psi$ s. The packets within each queue is served in a first in first out manner and the buffer length is assumed to be infinite. The focus of our model is on downlink (DL) direction. However, all arguments are valid in the reverse direction of communications.

We analyze the performance of the system by considering coexistence of both eMBB and URLLC traffics over orthogonal and non-orthogonal sharing of FH resources as described in [2] using multi-path FH with coding (MPC).

In this solution (Fig. 1), packets arrive to the CU with exponential inter-arrival periods with average $1/\lambda$ s. Each packet goes through the four steps below,

- *Fragmentation* block fragments the arrival packet into k equal blocks, each with size B/k.
- *Encoder* block encodes the blocks into n encoded blocks with size B/k. Each block is then forked into n paths and serviced in parallel. The service time of each path follows an exponential distribution with service rate $\mu_{\text{MPC}} = \frac{k\psi}{B}$.
- At the receiver, the original packet can be retrieved if any k out of n are received successfully. Thereby, once k blocks are received, they are passed into *decoder* to start decoding them without waiting for the remaining ones.
- The k decoded blocks are passed to *concatenation* block to be merged into one packet.

Latency in this solution is determined from the time packet is transmitted over the FH until k blocks are successfully received.

B. Functional Split and Latency Requirement

Different functional split points have different latency and bandwidth requirements on the FH [9]. These requirements should be considered to support 5G scenarios since each scenario requires different end-to-end requirements in term of latency and reliability as shown in Table I [16]. Hence, a split

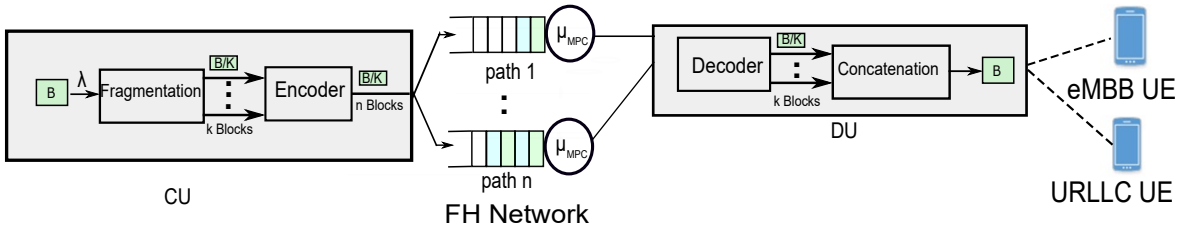


Fig. 1. Multi-path FH with erasure coding (MPC) for downlink communication.

TABLE I
REQUIREMENTS FOR 5G SCENARIOS

Scenario	End-to-end latency	Reliability	Payload size
Tactile interaction	0.5 ms	99.999%	Small
Electricity distribution (high voltage)	5 ms	99.9999%	Small
Electricity distribution (medium voltage)	25 ms	99.9%	Small to big
Discrete automation	10 ms	99.99%	Small to big
Intelligent transport systems	10 ms	99.9999%	Small to big

TABLE II
BANDWIDTH AND LATENCY REQUIREMENTS FOR DIFFERENT SPLIT POINTS

Split Point	One-way Latency	DL Bandwidth	UL Bandwidth
PDCP-RLC	$1.5-10$ ms	4016 Mbps	3024 Mbps
MAC-PHY	250 μs	4133 Mbps	5640 Mbps

point that can sustain scenario requirements can be considered appropriate.

Among all available splits, we will focus our attention on PDCP-RLC and MAC-PHY splits (option 2 and option 6 respectively according to [9]). The expected latency requirements as estimated by 3GPP [9] for each split is listed in Table II.

- PDCP-RLC split: For this split, RRC and PDCP are centralized whereas RLC, MAC, PHY and RF are distributed. From a latency point of view, PDCP-RLC split has a relaxed latency requirement on the FH. It tolerates high latency as PDCP doesn't require a strict lower layer synchronization. The maximum tolerable one way latency should be in maximum 10 ms.
- MAC-PHY split: For this option, the split is between MAC and PHY wherein only PHY and RF are in DU. The split offers a high level of centralization and pooling gain compared to PDCP-RLC split. In this split, the HARQ process and other timing critical functions are located in CU which results in tighter latency constraints on the FH. This split can support 250 μs latency in maximum.

IV. STOCHASTIC DELAY BOUNDS FOR (N,K) FORK-JOIN SYSTEM

The fronthaul delay for the MPC method described in section III can be computed by analysing (n,k) fork-join system. Although (n,n) fork-join system, also known as basic fork-join system, has been thoroughly studied, there are many open problems in analysing its generalization, i.e. (n,k) fork-join system. Mean value analysis for (n,k) fork-join system has been done in [4] and [5]. Authors in [3] used stochastic network calculus to define a stochastic upper bound for distribution of delay in (n,k) fork-join system in a general case. Nevertheless, there is no reasonable way to use that formula without knowing the joint distribution of parallel queues (in case of dependent queues). In this paper, we compute an upper bound and lower bound for (n,k) fork-join system delay using the concept of independency and full dependency between parallel links.

It is worth mentioning that compared to the work presented in [5], we make an additional assumption of non-purging scenario i.e. after k out of n blocks exit the queuing system, the other $n-k$ remaining blocks are not removed from the queues and they will continue being processed. This assumption is more realistic in this context given dispatched packets can not be removed from the links and switches. As it has been discussed in [5], split-merge system (which is a variation of (n,k) fork-join system that blocks processing of the next packets until k out of n blocks of the current packet finish being processed) provides an upper bound for delay in purging scenario, however, it is not an upper bound in non-purging scenario.

In this section, our objective is to compute stochastic delay bounds of (n,k) fork-join system. As detailed in section III, in the MPC method, CU encodes the packet into n equal length blocks and sends those blocks into n parallel links. Hence assumption of independence between n links is not valid. Moreover, without knowing the dependency between the links, e.g. their joint distribution, computation of delay bounds are not tractable. Therefore, in this paper we calculate stochastic lower bound and upper bound for delay distribution

under certain assumptions.

A. Stochastic Lower Bound for Delay

In a homogeneous (n, k) fork-join system, the smallest delay stochastically occurs when all links are independent, since in that case if some links are highly congested, other links might be less congested with larger probability. Therefore, here we will find the stochastic distribution of delay in the case that all links are independent. Authors in [3] computed delay bound for independent links, using stochastic network calculus for a general case. However, we will derive this bound for the case in which each parallel link is an $M/M/1$ queue by applying classical queuing theory.

For an $M/M/1$ queue with the iid Poisson arrival process, with mean λ and iid exponentially distributed service times with mean $1/\mu$, we have

$$P\{d > \tau\} = e^{-(\mu-\lambda)\tau} =: p_0 \ , \tag{1}$$

where d is the block delay in an $M/M/1$ queue which includes waiting time of the block in the queue plus its own service time. Let us assume (n, k) fork-join system, which consists of n parallel homogeneous $M/M/1$ queues. In this system, delay of a packet is defined as the time between execution of the n encoded blocks into the n parallel links until the first k out of n blocks have been processed in the queues. Note that Equation (1) can be viewed as a Bernoulli process in which, the number of successes in n independent trials has Binomial distribution. Therefore, (n, k) fork-join delay, denoted by D, would be,

$$P\{D > \tau\} = \sum_{j=0}^{k-1} \binom{n}{j} (1-p_0)^j p_0^{n-j}. \tag{2}$$

which is the probability that more than $n - k$ links have greater delay than τ.

B. Stochastic Upper Bound for Delay

Similarly in a homogeneous (n, k) fork-join system, the largest delay stochastically occurs when all links are fully dependant, such that all queues have the same length; in this case congestion happens at the same time in all links. Therefore, we use the "equal queue length" assumption to compute the worst case of dependency, instead of looking for joint distribution of the queues, and find a stochastic upper bound for (n, k) fork-join system delay.

For an $M/M/1$ queue, the queue length, denoted by L_q, would be equal to l with the following probability,

$$P\{L_q = l\} = (1-\rho)\rho^l \tag{3}$$

$$\rho := \frac{\lambda}{\mu}. \tag{4}$$

Also, delay profile for an $M/M/1$ queue with length l is as follows,

$$P\{d > \tau\} = \sum_{m=0}^{l} \frac{(\mu\tau)^m}{m!} e^{-\mu\tau} := p_1. \tag{5}$$

Similar to the previous computation, to find the delay distribution of (n, k) fork-join system consisting of n homogeneous $M/M/1$ queues, we should compute the probability that more than $n - k$ queues have the delay greater than τ. Therefore, in the case of dependant parallel queues, the (n, k) fork-join system delay will be as follows,

$$P\{D > \tau\} = \sum_{j=0}^{k-1} \binom{n}{j} \times$$
$$P\{d_1 < \tau, ..., d_j < \tau, d_{j+1} > \tau, ..., d_n > \tau\}, \tag{6}$$

where d_i, $i = 1, ..., n$ denotes the block delay in the i^{th} queue (i.e. i^{th} link). In this analysis, we assume all parallel links have the same queue length equal to L_q. We further assume that solely this property, i.e. equal queue length, defines the dependency between links, while the queues are assumed to be independent. Hence, joint probability in Eq. (6) can be computed using Bayes' law, as follows,

$$P\{D > \tau\} = \sum_{j=0}^{k-1} \binom{n}{j} \times$$
$$\sum_{l=0}^{\infty} P\{d_1 < \tau, ..., d_n > \tau | L_q = l\} P\{L_q = l\} = \sum_{j=0}^{k-1} \binom{n}{j} \times$$
$$\sum_{l=0}^{\infty} P\{d_1 < \tau | L_q = l\} ... P\{d_n > \tau | L_q = l\} P\{L_q = l\}$$
$$= \sum_{j=0}^{k-1} \binom{n}{j} \sum_{l=0}^{\infty} (1-p_1)^j p_1^{n-j} (1-\rho)\rho^l \ . \tag{7}$$

V. Simulation Results

In this section, we develop a simulation model in MATLAB to validate our analysis in the presence of coexisting eMBB and URLLC services. Characterization of the two services are shown in Table III. We assume there are $n = 10$ independent FH paths, where each path i has a capacity of 100 Mbps, i.e. $\psi_i = 100$ Mbps, $\forall i$.

We initially plot the non-orthogonal sharing of FH resources with orthogonal FH transmission schemes that can allocate a different amount of resources to URLLC. In these first plots, the aim is to determine the allocations that improve the probability of error for a given latency for the URLLC services in orthogonal as compared to non-orthogonal FH shared resources.

In Fig. 2, we plot the error probability for URLLC using orthogonal bandwidth allocation on the FH with different URLLC bandwidth fractions; in each case bw_u fraction of the available path bandwidth, ψ_i, is allocated to URLLC. The plot shows that the choice $bw_u \geq 1/2 \ \psi_i$ can reduce the latency as compared to shared FH transport.

TABLE III
SYSTEM PARAMETERS FOR THE 5G SERVICES

Type of traffic	eMBB	URLLC
Packet Size (Bytes)	1500	500
λ (packet/ms)	4	8

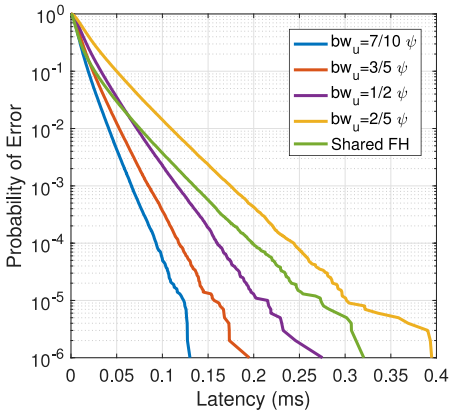


Fig. 2. Achievable probability of error Vs. latency under orthogonal FH bandwidth with different bandwidth fractions for URLLC.

Fig. 3 shows the error probability for URLLC using orthogonal path allocation on the FH with different number of paths allocated to URLLC, i.e. using n_u of the available path, $n = 10$. The plot shows the latency to achieve the error probability of better than 10^{-1} can be reduced as compared to shared FH transport by choosing $n_u \geq 5$. For example the error probability of 10^{-6} obtained using orthogonal paths is improved by approximately 160% as compared to that obtained by non-orthogonal sharing of FH resources.

Focusing on orthogonal bandwidth allocation on the FH with $bw_u = 1/2\ \psi_i$, Fig. 4 shows that simulation results are bounded by the lower and upper bounds computed from Equations (2) and (7). For URLLC (Fig. 4(a)), to achieve a reliability of 99.9999% the latency ranges from 0.1 ms to 0.4 ms. As for eMBB (Fig. 4(b)), the latency range is wider varying from 0.53 ms to 1.92 ms. From Fig. 5 we can observe the performance of both URLLC and eMBB are enhanced as compared to orthogonal bandwidth allocation (Fig. 4).

Focusing on Fig. 5, we use the lower and upper bounds obtained in this figure to choose appropriate functional split that offers the required reliability for a given scenario. For example using MAC-PHY split with the URLLC traffic (Fig. 5(a)), the upper bound can provide a reliability of 99.9999% at latency of 0.167 ms. Therefore, this setup can be used for low latency applications which requires a reliability as high as 99.9999%. Considering the requirements listed in Table I, this setup can, for example, be used for all scenarios. As for eMBB (Fig. 5(b)), MPC with MAC-PHY split can offer a reliability less than 99.9% which is not suitable for any scenario listed in Table I whereby the reliability requirements are of at least 99.9%. In such a case MPC with PDCP-RLC split is the only

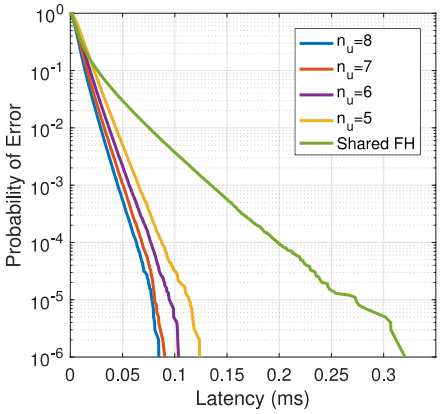


Fig. 3. Achievable probability of error Vs. latency under orthogonal FH path with different number of paths for URLLC.

choice available since the upper bound can provide a reliability of 99.9999% at latency of 0.6 ms.

To summarise, MAC-PHY split is the most appropriate split for scenarios using URLLC traffic considering the system model and traffic patterns in Table III, since it meets their latency and reliability requirements. Whereas, PDCP-RLC split is more suitable for scenarios using eMBB traffic.

VI. CONCLUSION AND FUTURE RESEARCH

In this paper, we presented a Cloud-RAN model based on multi-path FH with coding solution for enhancing the reliability of the FH. The paper aims at providing an upper and a lower bounds of reliability-latency function on the FH under orthogonal FH allocation.

We first derived lower and upper bounds analytically. Then we simulated the Cloud-RAN model to demonstrate the effectiveness of the analytic by showing the simulation results are bounded by lower and upper bounds. Finally, based on this result, we discussed the recommendations for split point focussing on MAC-PHY and PDCP-RLC splits for different scenarios to meet their latency and reliability requirements.

Future work will be focusing on analyzing multi-path FH with multi-hops.

ACKNOWLEDGEMENT

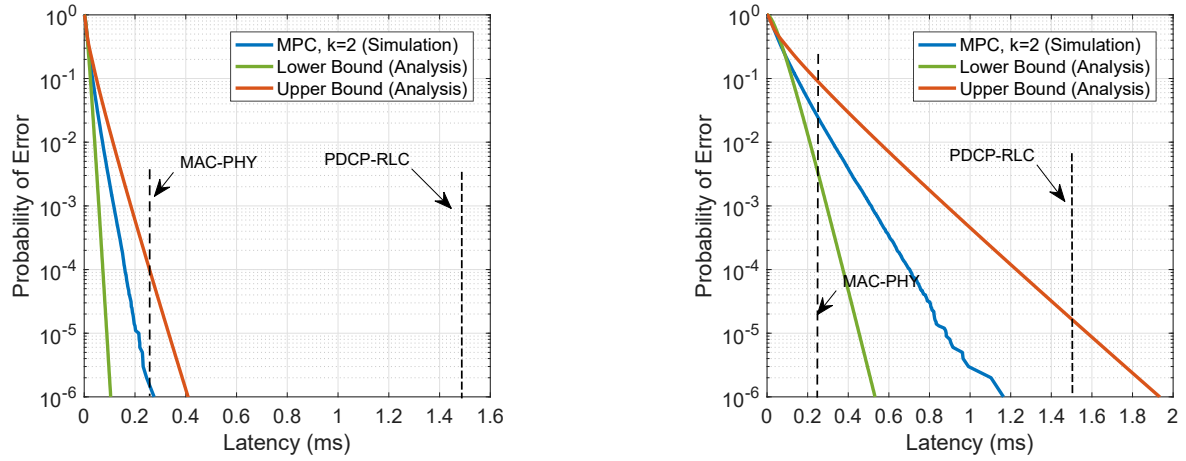
This work has been supported by The Engineering and Physical Sciences Research Council (EPSRC) industrial Cooperative Awards in Science & Technology (iCASE) award and by the British Telecom (BT). Additional support is received from EU H2020 5GCAR.

REFERENCES

[1] M. Condoluci and T. Mahmoodi, "Softwarization and virtualization in 5g mobile networks: Benefits, trends and challenges," *Computer Networks*, vol. 146, pp. 65 – 84, 2018.

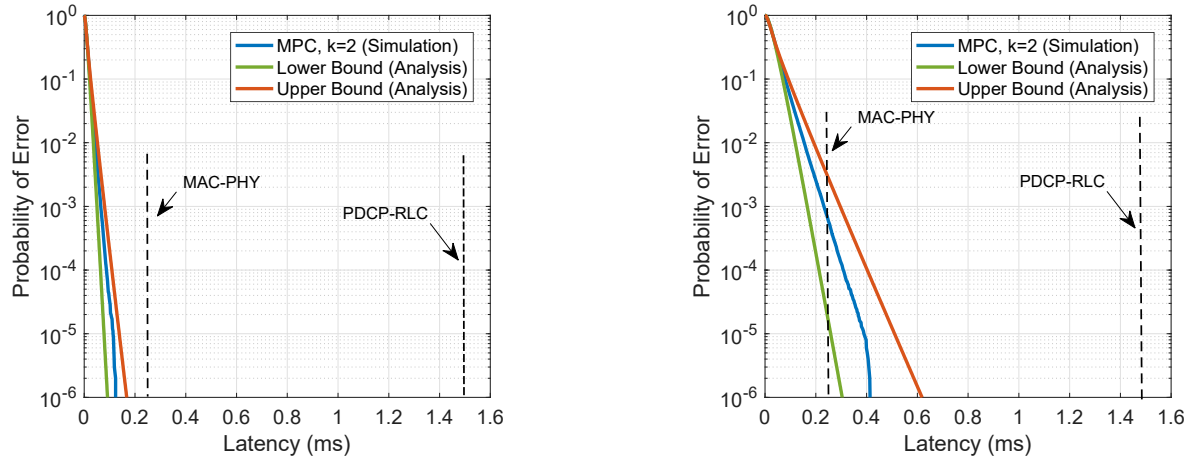
[2] G. Mountaser, T. Mahmoodi, and O. Simeone, "Reliable and Low-Latency Fronthaul for Tactile Internet Applications," *IEEE JSAC*, vol. 36, pp. 2455–2463, Nov 2018.

[3] M. Fidler and Y. Jiang, "Non-asymptotic delay bounds for (k, l) fork-join systems and multi-stage fork-join networks," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, Apr 2016.



(a) URLLC. (b) eMBB.

Fig. 4. Achievable probability of error Vs. latency at the upper and lower latency bounds, under bandwidth orthogonal FH with URLLC bandwidth fraction $1/2$ ($bw_e = 1/2\psi_i$ and $bw_u = 1/2\psi_i$). *The dashed line represents the maximum latency supported by functional split*



(a) URLLC. (b) eMBB.

Fig. 5. Achievable probability of error Vs. latency at the upper and lower latency bounds, under path orthogonal FH with URLLC path fraction $1/2$ ($n_e = 5$ and $n_u = 5$). *The dashed line represents the maximum latency supported by functional split*

[4] H. Wang, J. Li, Z. Shen, and Y. Zhou, "Approximations and Bounds for (n, k) Fork-Join Queues: A Linear Transformation Approach," in *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, vol. 00, pp. 422–431, May 2018.

[5] G. Joshi, Y. Liu, and E. Soljanin, "Coding for fast content download," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 326–333, Oct 2012.

[6] G. Mountaser, M. L. Rosas, T. Mahmoodi, and M. Dohler, "On the Feasibility of MAC and PHY Split in Cloud RAN," in *IEEE WCNC*, March 2017.

[7] G. Mountaser, M. Condoluci, T. Mahmoodi, M. Dohler, and I. Mings, "Cloud-RAN in Support of URLLC," in *2017 IEEE GLOBECOM Wkshps*, Dec 2017.

[8] China Mobile, "C-RAN: the Road Towards Green RAN," *White Paper*, vol. 2, 2011.

[9] 3GPP, "Study on New Radio Access Technology: Radio Access Architecture and Interface (Release 14)," Tech. Rep. 38.801, 2017.

[10] NGMN, "Further Studies on Critical Cloud RAN Technologies." White Paper, March 2015.

[11] A. S. Thyagaturu, Z. Alharbi, and M. Reisslein, "R-fft: Function split at ifft/fft in unified lte cran and cable access network," *IEEE Transactions on Broadcasting*, vol. PP, no. 99, pp. 1–18, 2018.

[12] J. Baranda, J. Mangues-Bafalluy, I. Pascual, J. Nunez-Martinez, J. L. D. l. Cruz, R. Casellas, R. Vilalta, J. X. Salvat, and C. Turyagyenda, "Orchestration of End-to-End Network Services in the 5G-Crosshaul Multi-Domain Multi-Technology Transport Network," *IEEE Communications Magazine*, vol. 56, pp. 184–191, July 2018.

[13] C. Chang, N. Nikaein, and T. Spyropoulos, "Impact of Packetization and Scheduling on C-RAN Fronthaul Performance," in *2016 IEEE GLOBECOM*, pp. 1–7, Dec 2016.

[14] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G Radio Network Design for Ultra-Reliable Low-Latency Communication," *IEEE Network*, vol. 32, pp. 24–31, March 2018.

[15] N. I. Sulieman, E. Balevi, K. Davaslioglu, and R. D. Gitlin, "Diversity and network coded 5G fronthaul wireless networks for ultra reliable and low latency communications," in *IEEE PIMRC*, Oct. 2017.

[16] 3GPP, "Technical Specification Group Services and System Aspects; Service requirements for the 5G system; Stage 1 (Release 16)," Tech. Rep. 22.261, Jun 2018.