# Packet Loss in Latency-constrained Ethernet-based Packetized C-RAN Fronthaul

Jay Kant Chaudhary*, Jobin Francis†, André Noll Barreto‡ and Gerhard Fettweis*
*Vodafone Chair Mobile Communications Systems
Technische Universität Dresden, Germany
Email: {jay_kant.chaudhary, gerhard.fettweis}@tu-dresden.de
† IIT Palakkad, Kerala, India
Email: jobinfrancis@iitpkd.ac.in
‡Barkhausen-Institut GmbH, Dresden, Germany
Email: andre.nollbarreto@barkhauseninstitut.org

***Abstract*—Latency is the one of the critical performance metrics for 5G and beyond mobile networks, particularly for ultra-reliable and low-latency communications (URLLC). In URLLC applications, it is required that the transmitted packets reach the destination within a certain time and the packets that are unable to meet this strict latency requirement will be discarded. In this paper, we compute the waiting time in the packetized fronthaul at the Ethernet switch and compute packet loss rate (PLR) incurred due to the inability of the transmitted packets to meet the FH latency threshold. In addition, we derive the tractable closed-form solution for the waiting time distribution and verify it with simulation results. Our results show that PLR is affected mainly by packet size, spectral efficiency, switch speed and arrival rate.**

## I. Introduction

Cloud radio access network (C-RAN) is considered as a one of the prime technologies in order to meet the diverse and stringent requirements of a plethora of new use cases and application scenarios envisioned by the 5G cellular networks. Contrary to the traditional baseband architecture, often called distributed radio access architecture (D-RAN), in C-RAN most of the baseband functionalities are offloaded from the BS and centralized to a common location, called baseband unit (BBU), while the remote radio units (RRUs) are located close to the antenna. The BBU and RRU are connected by a high bitrate, low latency and highly reliable transport link, known as fronthaul (FH). In C-RAN, the notion of functional split (refer to Section I-A ) is employed that determines how baseband signal processing is split between the BBU and the RRUs. In fully centralized C-RAN, FH transports raw in-phase/quadrature-phase (I/Q) samples between the BBU and RRU, and transportation of these I/Q samples requires a special protocol to encapsulate them, e.g., the widely used common public radio interface (CPRI) protocol. Although this C-RAN architecture is more energy efficient and offers centralization benefits, a fully centralized C-RAN with a classical functional split (Split 8 in Fig. 1) is quite challenging to implement, mainly because it requires strict latency, and a very high and static FH bandwidth, which in practice is not so economical for network operators and vendors to deploy. Thus, it is clear that the classical split is not an optimal and economically viable solution as it does not scale to 5G radio access technologies (RATs). Moreover, as the FH bitrate is very high with constant load and not associated with the actual user data rate, it provides no statistical multiplexing gain. Therefore, new functional splits towards packet-switched FH networks are required, which is explained in I-A.

### A. Packetized Fronthaul: Ethernet as new Fronthaul

In order to cope with C-RAN challenges, alternative RAN functional splits have been proposed e.g., by eCPRI and the IEEE 1914 next generation fronthaul interface (NGFI) Working Group enabling the use of packetized FH such as Ethernet. Ethernet offers several advantages: it is cost effective, flexible and widely used. In addition, it allows an efficient sharing of network infrastructure through standardized network function virtualization techniques and saves FH resources through statistical multiplexing. In addition, 3GPP [1] has identified eight functional split options with different sub-options for some of the splits. Offloading more functions to the RRU

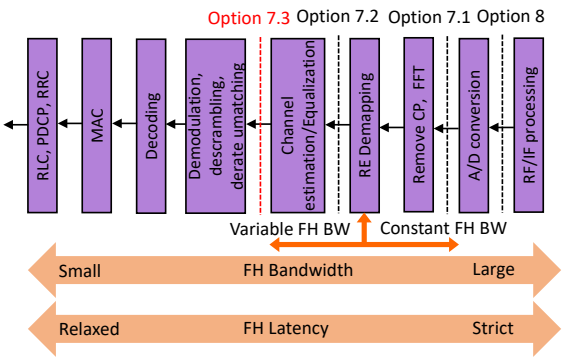


Fig. 1. C-RAN with selected functional splits, highlighting Split 7.3 as the considered split in this paper.

reduces the required FH bandwidth and increases the flexibility of the FH yielding FH bandwidths that actually scale with user load. [2]. However, low and variable FH bitrate comes with a price: it reduces the centralization and virtualization gains, and increases the complexity at the RRU. In this paper, we focus on Split 7.3 as shown in Fig. 1, whereby the precoding in the downlink and channel equalization in the uplink are performed at the RRU. We concentrate on the uplink protocol stack throughout this paper. The notion behind using this split is that it is more suitable for future RATs employing massive MIMO because the required FH bandwidth is significantly reduced as the required FH bandwidth now scales with the number of spatial streams rather than with number of antenna elements as in the classical CPRI (Split 8). Moreover, Fig. 1 also highlights that locating the resource demapper at the RRU makes the FH data rate variable, thus enabling the exploitation of statistical multiplexing gain, which is one of the main inherent advantages of Ethernet. However, one of the problems of Ethernet as the packetized FH solution is traffic is quite likely to queue, which could lead to unwanted delay and jitter in a mobile network [2]. In this paper, we focus on latency constraints of the FH and investigate suitability of a packet-switched FH to meet such constraints.

### B. Literature Review and Contributions

Often literatures on latency analysis of C-RAN [2], [3] have either ignored the waiting time at the switch or assumed, for simplicity, some deterministic value for delays at the switch. A few works [4], [5] have considered a very simple scenario with a single RRU. However, waiting time will play a significant role, particularly for heavily loaded system such as massive MIMO RRUs, where the switch has several arrivals from different users in the network with varying requirements. Hence, it is important to model and analyze the effects of waiting time in real scenarios. In this paper, we attempt to do exactly that. We extended the model presented in [6] for waiting time calculation and thereby compute the waiting time distribution at the switch for random packet arrivals from the users in the network considering massive MIMO RRUs. Our main contributions in this paper can be briefly summarized as:

1) We simulate and compute the waiting time at the switch in a packet-switched FH employing Split 7.3 for massive MIMO-aided RRUs.
2) Using the Pollaczek-Khinchin formula, we derive a tractable, closed-form expression for the waiting time distribution at the Ethernet switch.
3) We demonstrate the validity of our analytical results by means of simulation.
4) We discuss the impact of file size, arrival rate, switch speed and spectral efficiency on waiting time, and provide insights for network dimensioning, particularly in terms of packet loss rate (PLR) for a latency-constrained FH.

## II. Latency in Packetized Fronthaul

In this section, we analyze the main latency components in the FH. We consider end-to-end (e2e) latency requirements between the BBU and RRU. The e2e latency of the FH network mainly consists of three parts: delays in the access link, delays in the FH network and delays in the backhaul (BH) network. These latency components can be broken down, for simplicity, into transmission delay, propagation delay, processing delay, serialization delay, fabric delay and queuing delay, and hence, the total round trip latency can be computed as:

$$T_{\text{tot}} = 2 \cdot T_{\text{trans}} + T_{\text{Proc}} + 2 \cdot T_{\text{P}} + 2 \cdot N(T_{\text{q}} + T_{\text{f}} + T_{\text{se}}), \quad (1)$$

where, $T_{\text{trans}}$ = packet size/FH bitrate[1] is the transmission delay, $T_{\text{Proc}}$ the net processing delay at BBU and RRU, $T_{\text{se}}$ = packet size/switch speed the serialization delay, $T_{\text{f}}$ the fabric delay, $T_{\text{q}}$ the queuing delay and $N$ the number of switches.

After computing all the involved delay components, the maximum allowable FH latency can be calculated as:

$$\Delta T_{\text{FH, threshold}} = T_{\text{max, delay}} - T_{\text{tot}}, \quad (2)$$

where $T_{\text{max, delay}}$ is the maximum e2e delay. Note that $T_{\text{tot}}$ is a random quantity since the queuing delay is random. The allowable latency budget in the FH limits the distance between the BBU and RRU. Thus, knowing the FH latency, the maximum (one way) distance between the BBU and RRU can be computed using

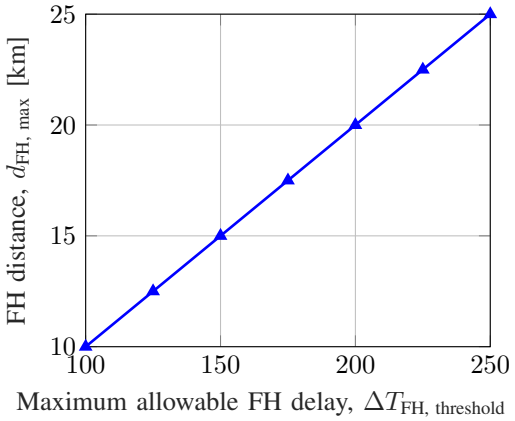


Fig. 2. FH distance, $d_{\text{FH, max}}$ for round trip fronthaul delay, $\Delta T_{\text{FH, threshold}}$

$$d_{\text{FH, max}} = \Delta T_{\text{FH, threshold}} / \Delta T_{\text{P}}, \quad (3)$$

where, $\Delta T_{\text{P}}$ is the propagation delay per km, which is 10 $\mu$s/km for fiber-based FH. Fig. 2 shows the FH distance

[1]The fronthaul bitrate value is different for each split.

(one way) corresponding to the maximum allowable e2e fronthaul delay. In the remainder of this paper, we focus on the queueing delay at the switch, which is explained in detail in Section IV.

The latency constraint in the FH originates either from the timing requirement of the hybrid automatic repeat request (HARQ) or from use cases such as Tactile Internet, autonomous driving or augmented and/or virtual reality. In the LTE MAC, HARQ process is co-located[2] with a scheduler and it requires the acknowledgement signal to be sent within a pre-defined time denoted as $T_{\text{max, delay}}$. Most of the round trip time $T_{\text{max, delay}}$ is spent at the BBU and RRU for baseband signal and RF processing, respectively, and the remaining time $\Delta T_{\text{FH, threshold}}$ is left for the FH transport. In general, the latency budget left for the FH with the HARQ process located at the BBU is a few hundreds of microseconds, typically $\Delta T_{\text{FH, threshold}} \leq 250$ $\mu$s [1], [7].

## III. System Model

The traffic from the users is likely to experience some waiting time in the queue at the switch. Hence, we first need to model user traffic. For this purpose, we consider a massive MIMO system model that follows from the approach in [6].

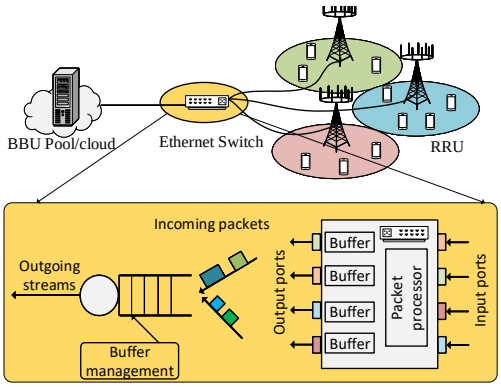


Fig. 3. A packetized Fronthaul C-RAN with simplified Ethernet switch Structure.

The system model is shown in Fig. 3 consisting of a massive MIMO access network, an Ethernet-based FH network, and a BBU. Further, the FH network consists of two FH segments: FH Segment I and FH Segment II and an Ethernet switch. The switch consists of input-output ports, a packet processor and buffer elements. FH Segment I connects the RRUs to the input ports of the switch and FH Segment II connects the output port of the switch to the BBU pool. The switch is configured as a multiplexer, and the traffic from the users in the access network is multiplexed at the switch and forwarded to the BBU for further processing. The switch has source and destination MAC addresses. The packet processor routes the incoming packets to an appropriate output port by looking at the destination address of the packet. Thus, the packet is queued at the switch before it is transmitted. We assume that buffer length at the switch is sufficiently large enough so that packet dropping at the switch can be ignored.

For such a system, we need to model user arrival traffic. However, before modeling the user traffic, we calculate the spectral efficiency of each user and consequently, the number of uplink channel uses required to send files for each user.

### A. Massive MIMO Access Network

We consider massive MIMO RRUs equipped with $M$ antennas that serve $K$ single antenna users. The users are spatially multiplexed onto the time-frequency resource grid. Let there be a total of $L$ cells and each RRU is located at the center of the cell. We assume the network operates in time division duplex mode and the channel is reciprocal. Thus, the RRU obtains the channel information from the uplink pilots and later the RRU uses them for downlink data transmission. Further, we assume that channel between the users and RRUs is frequency-flat in a coherence interval, $\tau_{\text{c}} = B_{\text{coh}}\tau_{\text{coh}}$ symbols, where $B_{\text{coh}}$ is the coherence bandwidth and $\tau_{\text{coh}}$ is the coherence time.

Let us consider, we need $\tau_{\text{p}}$ OFDM symbols which are used for pilot signalling. Hence, the remaining $\tau_{\text{c}} - \tau_{\text{p}}$ OFDM symbols will be used for data transmission. To be precise, let $\zeta^{(\text{ul})}(1-\frac{\tau_{\text{p}}}{\tau_{\text{c}}})$ and $\zeta^{(\text{dl})}(1-\frac{\tau_{\text{p}}}{\tau_{\text{c}}})$ symbols be used for uplink and downlink data transmission, respectively, where $\zeta^{(\text{ul})} + \zeta^{(\text{dl})} = 1$ and $1 \leq \tau_{\text{p}} < \tau_{\text{c}}$. We assume full pilot reuse, which will cause pilot contamination, and that the pilots reused in every cell are assigned randomly to the users. Further, we assume that there is no pilot power and data power control, and all the users transmit with their peak power $P_{\text{UE}}$. Let $B_{it}$ denote the set of users that use the same pilot sequence as user $t$ in cell $i$.

Assuming the matched filtering at the RRU, the uplink spectral efficiency $R_{lk}^{l}$ (in bits/s/Hz) of transmission can be obtained as [6], [8]:

$$R_{lk}^{l} = \zeta^{(\text{ul})}\left(1 - \frac{\tau_{\text{p}}}{\tau_{\text{c}}}\right)\log_2\left(1+\gamma_{lk}^{l}\right), \tag{4}$$

where $\gamma_{lk}^{l}$ is the received signal-to-interference-plus-noise-ratio (SINR), that is given by [6], [8]:

$$\gamma_{lk}^{l} = \frac{\left(\frac{M(\beta_{lk}^{l})^2}{\sum_{(it)\in B_{lk}}\beta_{it}^{l}+\frac{\sigma^2}{\tau_{\text{p}}P_{\text{UE}}}}\right)}{\frac{\sigma^2}{P_{\text{UE}}} + \sum_{(it)\in\mathcal{S}}\beta_{it}^{l} + \left(\frac{M\sum_{(it)\in B_{lk}\setminus(l,k)}\left(\beta_{it}^{l}\right)^2}{\sum_{(it)\in B_{lk}}\beta_{it}^{l}+\frac{\sigma^2}{\tau_{\text{p}}P_{\text{UE}}}}\right)}, \tag{5}$$

[2]The HARQ timing requirement is very critical if HARQ is located at the BBU, however, the timing requirement is much relaxed if the process is located at the RRU [1].

where, $\beta_{it}^{l}$ is the large-scale fading coefficient. Note that the numerator in (5) is the received signal power whereas the three terms in denominators are the noise power, multiuser interference, and interference due to pilot contamination, respectively.

### B. User Traffic Model

Let us consider a user $k$ in cell $l$ with an arrival rate $\lambda_{lk}$ of requests with the file size $F_{lk}$. The file size $F_{lk}$ is a random variable and we assume, for simplicity, that it is exponentially distributed with mean $\overline{F}$. Hence, its probability distribution function (PDF) is $f_{F_{lk}}(F_{lk} = x, \overline{F}) = (1/\overline{F})\ \exp(-x/\overline{F})$.

## IV. Waiting Time Analysis

In this section, we model the queue at the switch, which requires us to have information about the arrival process and service process. In addition, we have to ensure that the stability condition of the switch is met. Later, we derive the closed-form expression for the waiting time distribution and verify the analytical results with simulation results.

### A. Queue Model

File requests by any user is modelled as a Poisson point process. Corresponding to the user $k$ in $l$ cell with file size $F_{lk}$, arrival rate $\lambda_{lk}$ and the spectral efficiency $R_{lk}$, the quantized bit steams at the output of equalizer that will be transported over the FH are $N_{\text{bitstreams},lk} = 2N_qF_{lk}/R_{lk}^{l}$. Hence, the service time required to process the packet corresponding to file $F_{lk}$ can be obtained by

$$S_{lk} = N_{\text{bitstreams},lk}/C_{\text{FH}} = 2N_qF_{lk}/(R_{lk}^{l}C_{\text{FH}}), \quad (6)$$

where $C_{\text{FH}}$ is the speed of the switch operating at a constant speed. As $F_{lk}$ is exponentially distributed with mean $\overline{F}$, the service time $S_{lk}$ is also exponentially distributed but with mean $\mu_{lk} = \mathbb{E}[S_{lk}] = 2N_q\overline{F}/(R_{lk}^{l}C_{\text{FH}})$. Therefore, the PDF of $S_{lk}$ $f_{S_{lk}}(S_{lk} = x, \mu_{lk}) = (1/\mu_{lk})\ \exp(-x/\mu_{lk})$.

Following the discussion in [6], the queue model at the switch is M/HE/1, where M indicates that the arrival process is Poisson and the service process has a hyperexponential (HE) distribution. The arrival process at the switch is Poisson with aggregated arrival rate $\Lambda = \sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_{lk}$ because it is the sum of $LK$ independent Poisson processes.

Any random packet arriving at the switch could be from one of the $LK$ possible users with probability (w.p.) $p_{lk} = \lambda_{lk}/\Lambda$ [9]. Hence, the service time RV $S$ has mixed density, known as hyper exponential (HE) distribution, with the PDF, $f_S(x)$ and the mean, $\mathbb{E}[S]$ given by

$$f_S(S = x) = \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk} f_{S_{lk}}(x), \quad (7)$$

$$\mathbb{E}[S] = (2N_q\overline{F}/C_{\text{FH}})\sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}/R_{lk}^{l}. \quad (8)$$

An M/HE/1 queue needs to fulfil the stability condition $\rho = \Lambda\mathbb{E}[S] < 1$. Thus, we get

$$\rho = \frac{2N_q\overline{F}}{C_{\text{FH}}}\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\lambda_{lk}}{R_{lk}^{l}} < 1. \quad (9)$$

(9) shows the involved parameters that affect the switch stability. For our results, we choose these parameters such that stability of queue at the switch is always ensured.

### B. Waiting Time Distribution

Let $T$ and $W$ denote the sojourn time and waiting time, respectively. $S$ is the service time, defined previously. Sojourn time, $T = W + S$, is the time spent in the switch. Assuming that the waiting time and the service time are independent, the PDF of sojourn time is obtained by convolving the PDF of the waiting time with the PDF of the service time as $f_T(x) = f_W(x) * f_S(x)$. Next, in order to compute the waiting time distribution, we employ the Pollaczek-Khinchin formula [10] for M/G/1 queue and derive the relation to our M/HE/1 queue model. The Pollaczek-Khinchin formula expresses moment generating function (MGF) of sojourn time in terms of MGF of waiting time and MGF of service time.

The MGF of a RV $x$ is in fact the Laplace transform of its PDF. Hence, taking the Laplace transform of $f_T(x) = f_W(x) * f_S(x)$, we get $\Psi_T(s) = \Psi_W(s)\cdot\Psi_S(s)$, where $\Psi_T(s)$, $\Psi_W(s)$ and $\Psi_S(s)$ denote MGF of the sojourn time, waiting time and service time, respectively. Employing the Pollaczek-Khinchin formula to our M/HE/1 queue model, we obtain the waiting time MGF as [10]

$$\Psi_W(s) = \frac{s\,(1-\rho)}{s - \Lambda + \Lambda\Psi_S(s)}. \quad (10)$$

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

$$\begin{aligned}\Psi_S(s) = \mathcal{L}\left\{f_S(x)\right\} &= \int_{-\infty}^{+\infty}\exp(-sx)\left\{f_S(x)\right\}dx \\ &= \int_{0}^{+\infty}\exp(-sx)\left\{\sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk} f_{S_{lk}}(x)\right\}dx \\ &= \int_{0}^{+\infty}\sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\mu_{lk}^{-1}\ \exp\left(-(s+\mu_{lk}^{-1})x\right)dx \\ \Rightarrow \Psi_S(s) &= \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{{\mu_{lk}}^{-1}}{s+{\mu_{lk}}^{-1}}\right). \end{aligned} \quad (11)$$

Substituting $\Psi_S(s)$ and $\rho$ in (10), we get the final expression of the MGF of the waiting time as

$$\Psi_W(s) = \frac{s\left(1 - \frac{2N_q\overline{F}}{C_{\text{FH}}}\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\lambda_{lk}}{R_{lk}^{l}}\right)}{s - \Lambda + \Lambda\sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{{\mu_{lk}}^{-1}}{s+{\mu_{lk}}^{-1}}\right)}. \tag{12}$$

(12) shows how the different parameters will impact the waiting time. Lastly, we take the inverse Laplace transform of $\Psi_W(s)$ to get the distribution of the waiting time, which we later evaluate against the simulation results.

### C. Packet Loss Rate

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

$$\text{PLR} = P_{\text{r}}(W > \Delta T_{\text{FH, threshold}}), \tag{13}$$

where, $\Delta T_{\text{FH, threshold}}$ is the FH latency threshold obtained from (2).

## V. Results

In order to calculate the channel usage for a given file size, we first compute the spectral efficiency of each user. Let us consider there are $L = 7$ cells and we drop $K = 10$ users in each cell while guaranteeing that users are located at a distance $\geq 35$ m from the center of the cell. Let both the pilot and data transmission powers be $P_{\text{UE}} = 23$ dBm.

The large-scale fading coefficient $\beta_{it}^{l}$ in dB can be obtained using [11]

$$\beta_{it}^{l} = -148.1 - 37.6\log_{10}(d_{it}^{l}) + X_{\sigma,it}^{l} \text{ dB}, \tag{14}$$

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

TABLE I: Simulation parameters.

| Parameters | Symbol | Value |
|---|---|---|
| No. of antennas/cell | $M$ | 300 |
| Intersite distance (ISD) | $d_{\text{ISD}}$ | 500 m |
| Number of pilots | $\tau_{\text{p}}$ | 10 |
| Channel bandwidth | $B$ | 20 MHz |
| Coherence interval | $\tau_{\text{c}}$ | 200 symbols |
| Noise power | $\sigma^2$ | -96 dBm |
| Average file size | $\overline{F}$ | 0.5 MB |
| Quantizer resolution | $N_{\text{q}}$ | 8 bit |

Let each user transmits a maximum $N = 100000$ packets and let the latency budget for the FH be $\Delta T_{\text{FH, threshold}} = 250\mu$s. Before we discuss PLR, first we evaluate simulated waiting time distribution with analytical results.

### A. Waiting Time Distribution

We take the inverse Laplace transform of (12) using a built-in MATLAB function to get the analytical result for the waiting time distribution and compare it with the simulation result.

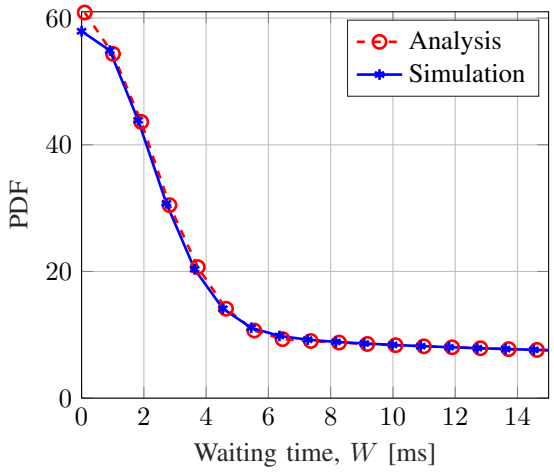


Fig. 4. Waiting time distribution, $\overline{F} = 0.5$ MB, $\lambda = 1$, $C_{\text{FH}} = 10$ Gbps

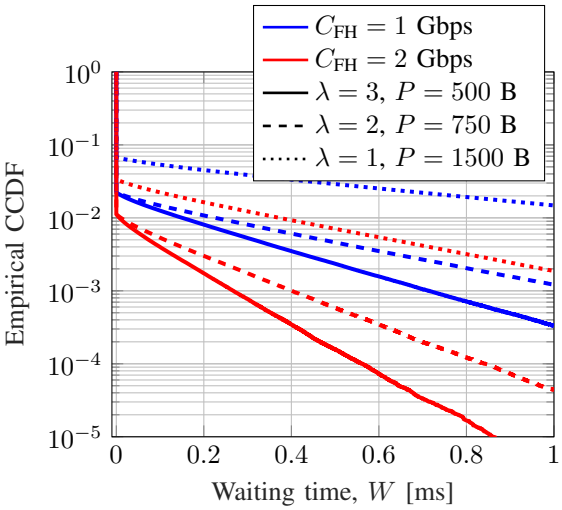


Fig. 5. CCDF plot of waiting time, $W$ .

Fig. 4 plots the waiting distribution for simulated and analytical results. We consider an average file size, $\overline{F} = 0.5$, mean arrival rate, $\lambda = 1$ and switch speed of 10 Gbps. We observe that both the simulated and analytical results match well. A slight deviation of the analytical solution occurs in the vicinity of zero, which is due to

Matlab's precision for inverse Laplace transform at the vicinity of zero. Waiting time is impacted mainly by the file size and switch speed. A bigger file size takes more resources and hence, more time to process for a given switch speed. On the other hand, even a bigger file size could be processed much quicker if the switch is operating at faster speeds. Note that frequency of arrival of a big file size will be less compared to smaller file sizes.

Fig. 5 plots the simulation results for the empirical complementary cumulative distribution function (CCDF) of the waiting time. Practically, the waiting time at the switch will be much smaller. Hence, we compare three smaller but fixed packet sizes ($P$) of 500 Byte, 750 Byte and 1500 Byte with corresponding mean arrival rates $\lambda = 3$, $\lambda = 2$ and $\lambda = 1$ while keeping the load at the switch constant.

### B. Packet Loss Rate

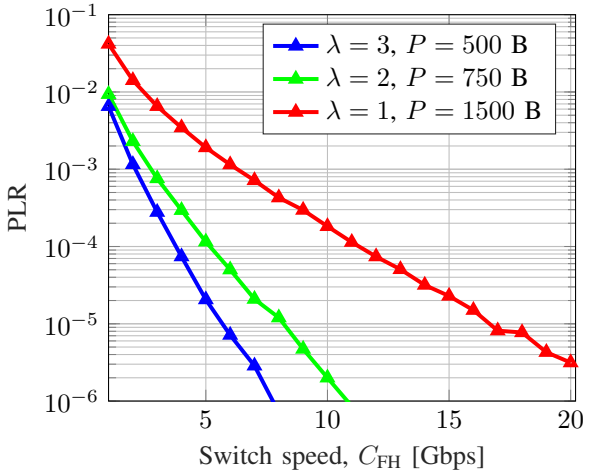


Fig. 6. Packet loss rate for varying switch speeds for different packet sizes and arrival rates.

Fig. 6 plots simulated PLR for different switch speeds for given packet sizes. As an example, the probability of a waiting time of 0.25 ms when the switch operates at 2 Gbps is 2%, 0.2% and 0.1% for $P = 1500$ Byte, $P = 750$ Byte and $P = 500$ Byte packet sizes, respectively. This shows that PLR increases if larger packet sizes are used. Their corresponding PLRs are much lower if the switch operates at faster speeds. For a fixed packet size, we can also infer that PLR increases with the higher values of mean arrival rates. This occurs because higher values of arrival rates increases the waiting time at the switch for a given switch speed. Generally, the Ethernet switch are over provisioned to operate at faster switch speeds compared to the incoming traffic from the RRUs such that PLR is extremely low or no PLR. This is because packets are processed much quicker and hence, their waiting times in the queue are much smaller.

## VI. Conclusion

Waiting time at the Ethernet switch plays a crucial role for latency-constrained packetized fronthaul. In this paper, we derived a the tractable closed-form solution of the waiting time distribution for M/HE/1 queues at the FH network switch. We showed that the simulated and analytical results match well. The inability of the transmitted packets to reach the destination within a certain time causes packet loss, which we presented in our results. We can get a reasonable low packet loss rate by deploying faster Ethernet switches. Moreover, we showed that main factors affecting the waiting time at the switch are file size, spectral efficiency, switch speed and arrival rate. In the future, we aim to provide latency and multiplexing gain trade-off analysis.

## Acknowledgment

The research leading to this work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 762057 (5G-PICTURE). Neither the European Union nor its agencies are responsible for the contents of this paper; its contents reflect the views of the authors only.

## References

[1] 3GPP, "3GPP TR 38.801 v14.0.0 (2017-03): Study on new radio access technology: Radio access architecture and interfaces (Release 14)," 2017.

[2] M. P. Larsen, M. S. Berger, and H. L. Christiansen, "Fronthaul for Cloud-RAN enabling network slicing in 5G mobile networks," *Wireless Communications and Mobile Computing*, vol. 2018, no. 3, 2018.

[3] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5g mobile crosshaul networks," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 146–172, Firstquarter 2019.

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

[5] G. Mountaser, M. Condoluci, T. Mahmoodi, M. Dohler, and I. Mings, "Cloud-RAN in support of URLLC," in *2017 IEEE Globecom Workshops (GC Wkshps)*, Dec 2017, pp. 1–6.

[6] J. K. Chaudhary, J. Francis, A. N. Barreto, and G. P. Fettweis, "Latency in the uplink of massive MIMO C-RAN with packetized fronthaul: Modeling and analysis," in *IEEE Wireless Communications and Networking Conference, accepted*, Marrakech, Morocco, 15-19 April 2019.

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

[8] T. V. Chien, E. Björnson, and E. G. Larsson, "Joint pilot design and uplink power allocation in multi-cell massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2000–2015, March 2018.

[9] L. N. Singh and G. R. Dattatreya, "Estimation of the hyperexponential density with applications in sensor networks," *International Journal of Distributed Sensor Networks*, vol. 3, no. 3, pp. 311–330, 2007.

[10] J. Virtamo, "38.3143 queueing theory/the M/G/1 queue." [Online]. Available: http://www.netlab.tkk.fi/opetus/s383143/kalvot/english.shtml

[11] 3GPP, "3GPP TR 36.814 v9.0.0 (2010-03): Further advancements for e-utra physical layer aspects (release 9)," 2010.