Latency in the Uplink of massive MIMO CRAN with Packetized Fronthaul: Modeling and Analysis

Jay Kant Chaudhary, Jobin Francis, André Noll Barreto and Gerhard Fettweis

***Abstract*—With the emergence of cloud radio access network (C-RAN) architecture, latency in fronthaul (FH) network is a critical performance metric especially for ultra-reliable and low-latency communication applications. The stringent FH capacity and latency requirements of C-RAN can be relaxed by offloading some baseband functionalities to remote radio unit (RRU), referred to as functional splitting. This allows packetized FH network solutions such as ubiquitous Ethernet. In this paper, we calculate the FH latency in the uplink of a C-RAN system with massive MIMO-based RRUs and 3GPP functional Split 7, wherein MIMO equalization is done at the RRU. We derive tractable, closed-form expressions for the steady-state probabilities of queue length and sojourn time distribution at the output port of an Ethernet switch in the FH network. We first present these results for Poisson file arrivals from users in the network and exponential file size distribution. We then extend the results to general file size distribution. The numerical results show that the file size and spectral efficiency of the users are critical in determining the FH latency. Further, results show that switch speed can be decreased without incurring significant increase in FH latency revealing the possibility for statistical multiplexing gains.**

***Index Terms*—Cloud radio access network (C-RAN), Massive MIMO, Fronthaul, Functional split, Latency**

I. Introduction

A plethora of use cases and application scenarios, broadly classified as enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communication (URLLC) are envisioned to be enabled by the fifth generation (5G) cellular networks [1], [2]. These services place diverse and stringent requirements in terms of data rate, latency, and reliability. Massive multiple-input-multiple-output (MIMO) and cloud radio access network (C-RAN) are two promising technologies to meet these requirements.

In massive MIMO, the base station is equipped with a large number of antennas. This allows the base stations to generate highly directional beams to the users and serve multiple users at the same time via spatial multiplexing. This yields a significant improvement in spectral efficiency (SE) and energy efficiency (EE) compared with conventional MIMO [3], [4]. In C-RAN, all the baseband processing is centralized at the baseband unit (BBU) pool while the radio processing is performed at the remote radio unit (RRU). The BBU and RRU are connected by a high-speed and low-latency transport link known as fronthaul (FH). It transports the baseband in-phase/quadrature-phase (I/Q) samples between the BBU and RRU with the commonly used common public radio interface (CPRI) standard.

Despite the promising advantages of C-RAN [5], the key challenge in the deployment of C-RAN is the need for a FH with stringent requirements on latency, jitter, data rate, and reliability. In CPRI, the end-to-end latency is required to be less than 250 μs and the reliability target is 10^{-12} [5]. The required FH capacity is generally determined by the number of antennas at the RRU, sampling frequency, and resolution of the time-domain quantizer [6]. As an example, a three sector, 8×8, 20 MHz LTE system requires a FH capacity of nearly 30 Gbps. This requirement increases linearly with the number of antennas and sampling frequency. Hence, the CPRI standard is not suitable for massive MIMO RRUs in a C-RAN system.

In order to circumvent the problem of huge capacity and tight latency FH requirements of C-RAN, a hybrid base station architecture with different functional splits has been introduced [6]. Functional split refers to the division of baseband processing functionalities between the BBU and RRU. As more functionalities are offloaded to the RRU, the requirements on the FH reduce. However, the benefits from centralization and cloudification/virtualization decrease at the same time. 3GPP has identified eight functional splits and we focus on Split 7, also referred to as intra-PHY split, as this split is expected to be suitable for massive MIMO applications [7]. At this split, precoding in the downlink and equalization in the uplink are offloaded to the RRU. As a result, quantized IQ streams are carried by the FH and not the signal for each antenna. This significantly alleviates the FH capacity requirements.

The latency constraints on FH for Split 7 follow either from the hybrid automatic repeat request (HARQ) process or from the use case itself. In HARQ, the acknowledgment message for any received packet has to be sent within a pre-specified time, which is 3 ms in LTE. Thus, the latency on the FH must be less than this budget minus the time for baseband processing. In certain cases, the applications themselves place latency constraints such as in the Tactile Internet [8], where end-to-end latency between tactile devices must be in the order of milliseconds.

As the FH data rate and latency requirements are relaxed, Ethernet-based packetized transmission is considered for FH by the next generation fronthaul interface (NGFI) [9] and eCPRI [10] standardization bodies. Ethernet is widely deployed due to its cost effectiveness, flexibility, and ubiquity.

J. K. Chaudhary, J. Francis, and G. Fettweis are with Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Germany. A. N. Barreto is with Barkhausen-Institut GmbH, Germany. Email: {jay_kant.chaudhary, jobin.francis, gerhard.fettweis}@tu-dresden.de, andre.nollbarreto@barkhauseninstitut.org

Further, it supports network virtualization and software defined networking, and takes advantage of statistical multiplexing. However, providing latency guarantees on Ethernet FH is difficult due to the randomness in latency caused by queuing in Ethernet switches. Developing an analytical model to characterize the delay in the FH for a massive MIMO CRAN system with Split 7 as the functional split is the focus of this paper.

A. Literature Review

The feasibility of functional Split 6 has been studied in [11] considering Ethernet-based FH with focus on latency and jitter over the link between PHY and MAC. This work is extended to two additional functional splits, Split 2 and Split 7 in [12] considering eMBB, mMTC and URLLC traffic. However, they consider a simple case with single RRU without massive MIMO. Further, they present the experimental results but lack the closed-form analysis. In [13], the impact of FH latency on the performance of automatic repeat request (ARQ) protocols is studied. The work in [14] proposed to improve the latency and reliability of packet-based FH network through multi-path diversity and erasure coding of media access control (MAC) frames. In addition, studies in [15]–[17] have been carried out towards FH transport network modeling and dimensioning considering, e.g., G/G/1 and D/G/1 queuing models. We note that none of the above contributions considered massive MIMO, which will cause an impact on the required bandwidth and latency of the FH segment.

B. Contributions

In this paper, we consider a practical massive MIMO scenario and calculate the latency in a packetized FH network for Split 7. We model the access link traffic generated by massive MIMO RRU, and map the queue at the switch as Poisson arrivals and service process as a hyperexponential (HE) distribution, leading to an M/HE/1 queuing model. Then, using the Pollaczek-Khinchin formula for M/G/1 queue, we derive tractable, closed-form expressions for the steady-state probabilities of queue length and sojourn time distribution at the output port of an Ethernet switch in the FH network for M/HE/1 queue. We first present these results for Poisson file arrivals with exponentially distributed file size, and later extend the results to general file size distribution. We show through numerical results that the file size and spectral efficiency of the users are critical in determining the FH latency. In addition, we show that speed of the switch can be reduced without causing significant increase in FH latency, which further reveals the benefits of possible statistical multiplexing.

C. Organization and Notations

The rest of the paper is organized as follows. In Section II, the system model is introduced. We present the queuing theoretic analysis in Section III. The numerical results are presented in Section IV and our conclusions in Section V.

Notation: We use uppercase bold face letters and lowercase bold face letters to denote vectors and matrices, respectively. I_M is an identity matrix of size $M \times M$. Further, $\mathcal{CN}(\cdot,\cdot)$ represents circularly symmetric complex Gaussian distribution, where $\mathcal{N}$ is normal distribution. For random variable (RV) X, let $\mathbb{E}[X]$ and $\Psi_X(s) = \mathbb{E}[\exp(-sX)]$ denote its expectation and moment generating function (MGF), respectively.

II. System Model

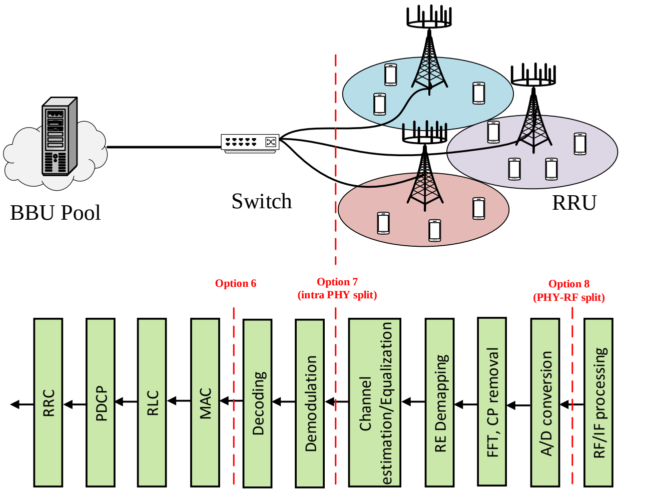

Fig. 1. C-RAN with an Ethernet switch considering functional Split 7 being implemented at the RRU.

A schematic diagram of the system model is shown in Fig.1. It consists of a massive MIMO access network, an Ethernet-based FH, and a BBU. They are described in detail below. Thereafter, the traffic model is presented. For such a system, we are interested in the SE of each user and consequently, the number of uplink channel uses needed to send files for each user.

A. Massive MIMO Access Network

The access network involves L cells with massive MIMO RRUs equipped with M antennas located at the cell center. There are K single antenna users in each cell, which are spatially multiplexed onto the same time-frequency resource. We assume that the network operates in time division duplex mode such that the RRU obtains the channel state information from uplink pilots. The RRU exploits them for downlink data transmission assuming that the channel is reciprocal. Further, we assume that the channel between the users and RRUs is time-invariant and frequency-flat in a coherence interval of $\tau_{\mathrm{c}} = B_{\mathrm{coh}}\tau_{\mathrm{coh}}$ symbols, where B_{coh} is the coherence bandwidth and τ_{coh} is the coherence time.

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

1) Uplink Pilot Training: In a coherence interval τ_{c}, τ_{p} OFDM symbols are utilized for uplink pilot signaling, $\zeta^{(\mathrm{ul})}(1-\frac{\tau_{\mathrm{p}}}{\tau_{\mathrm{c}}})$ symbols for uplink data transmission and $\zeta^{(\mathrm{dl})}(1-\frac{\tau_{\mathrm{p}}}{\tau_{\mathrm{c}}})$ symbols for downlink data transmission. Here, $\zeta^{(\mathrm{ul})} + \zeta^{(\mathrm{dl})} = 1$ and $1 \leq \tau_{\mathrm{p}} < \tau_{\mathrm{c}}$. We assume full pilot reuse and random pilot assignment to the users. Hence, pilots are reused in every cell and are assigned randomly to the users in a cell. Full pilot reuse results in pilot contamination. Let B_{it} denote the set of users that use the same pilot sequence as user t in cell i. We assume that there is no pilot power control and all the users transmit at the maximum power P_{UE}.

2) Uplink Data Transmission: Let x_{it} denote the signal transmitted by user t in cell i. This user's complex channel gain vector to RRU l is denoted by $\mathbf{h}_{it}^{l}$. It is distributed as $\mathbf{h}_{it}^{l} \sim \mathcal{CN}(\mathbf{0}, \beta_{it}^{l}\mathbf{I}_M)$, where β_{it}^{l} is the large-scale fading coefficient. As in the uplink pilot training, there is no power control for uplink data transmission and users transmit at full power P_{UE}. Then, the received signal $\mathbf{y}_l$ at the RRU l is obtained by the superposition of the transmitted signals from all the users in the network $\mathcal{S} = \{(it) : i \in \{1, \cdots, L\}, t \in \{1, \cdots, K\}\}$ and is given by

$$\mathbf{y}_l = \sum_{(it)\in\mathcal{S}} \sqrt{P_{\text{UE}}}\mathbf{h}_{it}^{l} x_{it} + \mathbf{n}_l, \tag{1}$$

where $\mathbf{n}_l \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_M)$ is the additive Gaussian noise.

We assume matched filtering at the RRU. That is, $\left(\mathbf{h}_{lk}^{l}\right)^H \mathbf{y}_l$ is used to recover signal x_{lk}. Then, the signal-to-interference-plus-noise-ratio (SINR) γ_{lk}^{l} can be obtained as [18]:

$$\gamma_{lk}^{l} = \frac{\left(\dfrac{M(\beta_{lk}^{l})^2}{\sum\limits_{(it)\in B_{lk}} \beta_{it}^{l} + \frac{\sigma^2}{\tau_p P_{\text{UE}}}}\right)}{\frac{\sigma^2}{P_{\text{UE}}} + \sum\limits_{(it)\in\mathcal{S}} \beta_{it}^{l} + \left(\dfrac{M \sum\limits_{(it)\in B_{lk}\setminus(l,k)} \left(\beta_{it}^{l}\right)^2}{\sum\limits_{(it)\in B_{lk}} \beta_{it}^{l} + \frac{\sigma^2}{\tau_p P_{\text{UE}}}}\right)}. \tag{2}$$

The numerator in (2) is the received signal power. The first, second, and third terms in the denominator can be identified as noise power, multiuser interference, and interference due to pilot contamination, respectively. Note that the pilot contamination term persists even if the number of antennas grows to infinity. This shows that pilot contamination becomes the limiting factor when the number of antennas is large. The uplink SE R_{lk}^{l} (in bits/s/Hz) of transmission is then given by

$$R_{lk}^{l} = \zeta^{(\text{ul})} \left(1 - \frac{\tau_{\text{p}}}{\tau_{\text{c}}}\right) \log_2\left(1 + \gamma_{lk}^{l}\right). \tag{3}$$

B. User Traffic Model and Ethernet-based FH Network

We now describe the dynamics of the uplink data traffic from different users. The file arrival process from a user is a Poisson point process. Let λ_{lk} denote the arrival rate for user k in cell l. The file size F_{lk} for user k in cell l is a RV and follows an exponential distribution with mean $\overline{F}$. The extension to the case of general file size distribution is discussed in Section III-C. We note that the SE expression in (3) has not accounted for the dynamic interference resulting from dynamic user traffic. This simplification is made to make the latency analysis tractable as it avoids the coupling between the arrival and service processes of different users. We also note that the above SE expression is a lower bound on SE with dynamic traffic. The SE of a user determines the number of I/Q symbols needed to send its file. Each received I/Q symbol after equalization is quantized to N_q bits at the RRU. The quantized bits corresponding to a user file is encapsulated in an Ethernet packet and is sent over the FH network.

The FH network in Fig. 1 consists of two FH segments and an Ethernet switch. In the first FH segment, FH links connect RRUs to the input ports of the switch and Ethernet packets from the RRUs. The second FH segment is the link that connects the output port of the switch to the BBU. The switch is configured as a multiplexer that aggregates the packets from different RRUs to the single output port. A schematic diagram of a switch is shown in Fig. 2. It involves a packet processor and output queues at each output port. The packet processor looks at the destination address of the packet and directs it to the appropriate output port. There, the packet is queued before it is transmitted. We assume that the switch speed is matched to the capacity C_{FH} of the second FH segment. Hence, the packet is pushed out of the queue as fast as possible. Further, we assume that the queue buffer is sufficiently large so that packet dropping at the switch is ignored.

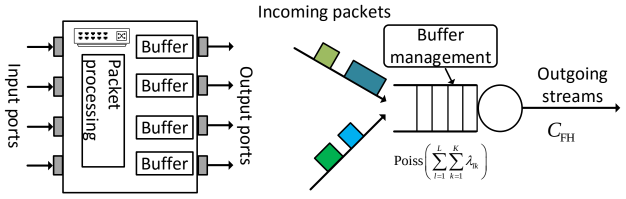

Fig. 2. Simplified structure of an Ethernet switch/aggregator (left) and its output port structure (right).

III. Queuing Theoretic Modeling and Steady-state Analysis

In this section, we first model the output queue at the Ethernet switch and then derive closed-form expressions for its steady-state probabilities and sojourn time distribution.

A. Queue Model

For modeling the queuing dynamics of the Ethernet switch, we need to understand the arrival and service processes of the queue. They are modeled as follows.

Arrival Process: Recall that the I/Q streams of the users are recovered after equalization at the RRU. Since the users' I/Q streams are generated from their Poisson file arrival process, the I/Q streams at the RRU for each user are also Poisson process[1]. The aggregate arrival process from an RRU to the switch is also Poisson as it is the sum of K independent Poisson processes. That is, the arrival process from RRU l is Poisson with arrival rate $\sum_{k=1}^{K} \lambda_{lk}$. Then, the overall arrival process at the queue is the sum of independent Poisson arrival processes from different RRUs. It is a Poisson process with arrival rate $\Lambda = \sum_{l=1}^{L} \sum_{k=1}^{K} \lambda_{lk}$.

[1] We ignore the slotted nature of uplink transmission in the access network since the symbol duration is small. For example, in LTE, it is 66.7 μs (without cyclic prefix), which is an order of magnitude lower than the time scale of interest (ms).

Service Process: The process time of a file depends on its size. Since file sizes are independent across different arrivals and users, service processes are independent and identically distributed. The marginal service time distribution is computed as follows. For user k in cell l, the number of subcarriers needed to send a file in the uplink is F_{lk}/R_{lk}^{l}. This is also the number of I/Q symbols as each subcarrier carries one I/Q symbol. At the RRU, as mentioned before, each I/Q symbol is quantized to $2N_q$ bits before being sent over the FH[2]. Then, the number of FH bits corresponding to file F_{lk} is $2N_qF_{lk}/R_{lk}^{l}$. These bits are from the packet. The time required by the switch to forward this packet is the number of bits divided by the switch speed, as the switch is operating at a constant speed. Hence, the service time S_{lk} for the packet corresponding to file F_{lk} is $S_{lk} = 2N_qF_{lk}/(R_{lk}^{l}C_{\text{FH}})$. Since F_{lk} is exponentially distributed with mean $\overline{F}$, the service time is also exponentially distributed but with mean $2N_q\overline{F}/(R_{lk}^{l}C_{\text{FH}})$. Note that the mean of the service time distribution is different for different users and depends on their SE.

A packet arriving at the switch can be from any one of the LK users in the network. Since the arrival process at the switch is the superposition of LK independent Poisson processes, as discussed above, a packet arriving at the switch is from user k in cell l with probability (w. p.) $p_{lk} = \lambda_{lk}/\Lambda$ [19]. Hence, the service time RV S is given by

$$S = \begin{cases} S_{11}, & \text{w. p.} \quad p_{11}, \\ \vdots & \\ S_{LK}, & \text{w. p.} \quad p_{LK}. \end{cases} \tag{4}$$

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

$$f_S(x) = \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk} f_{S_{lk}}(x), \tag{5}$$

where $f_{S_{lk}}(x)$ is the PDF of S_{lk}. Because S_{lk} is exponentially distributed, the distribution of S is known as hyper exponential (HE) distribution. The mean service time is given by $\mathbb{E}[S] = (2N_q\overline{F}/C_{\text{FH}})\sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}/R_{lk}^{l}$.

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

B. Steady-state Analysis

We now use the results available in the literature for M/G/1 queue to obtain closed-form results for the steady-state queue length and sojourn time distributions of an M/HE/1 queue [20]. These results follow from the embedded Markov chain at instances when a packet leaves the queue.

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

$$\rho = \frac{2N_q\overline{F}}{C_{\text{FH}}}\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\lambda_{lk}}{R_{lk}^{l}} < 1. \tag{6}$$

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

2) Steady-state Queue Length Probabilities: Let π_i denote the steady state probability of the queue length being equal to i, for $i = 0, 1, \dots \infty$. As shown in [20], these steady-state probabilities satisfy the following recursion:

$$\pi_i = \frac{1}{k_0}\left(a_{i-1}\pi_0 + \sum_{j=1}^{i-1} a_{i-j}\pi_j\right). \tag{7}$$

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

$$k_i = \int_0^{\infty}\frac{(\lambda x)^i}{i!}\exp(-\lambda x) f_S(x)dx. \tag{8}$$

When S has HE distribution, evaluating k_i yields

$$k_i = \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{\Lambda}{\Lambda + {\mu_{lk}}^{-1}}\right)^i\left(\frac{{\mu_{lk}}^{-1}}{\Lambda + {\mu_{lk}}^{-1}}\right), \tag{9}$$

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

3) Sojourn Time Distribution: For the sojourn time distribution, we employ the Pollaczek-Khinchin formula [20], which expresses the MGF of the sojourn time RV T in terms of the MGF of the service time RV S. Let $\Psi_T(s)$ and $\Psi_S(s)$ denote the MGF of RVs T and S, respectively. Then, $\Psi_T(s)$ is

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

For the HE service time distribution, $\Psi_S(s)$ is given by

$$\Psi_S(s) = \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{{\mu_{lk}}^{-1}}{s + {\mu_{lk}}^{-1}}\right). \tag{11}$$

Substituting (11) in (10) yields the MGF of sojourn time. The PDF of sojourn time can be obtained by taking the inverse Laplace transform of $\Psi_T(s)$. However, it cannot be evaluated in closed-form and numerical techniques are used to evaluate the inverse Laplace transform.

C. Extension to General File Size Distribution

Notice that the above analysis is not specific to the exponential file size distribution. Hence, it can easily be extended to any general file size distribution. Accordingly, we will have different expressions for $f_S(x)$, k_i and $\Psi_S(s)$ from (5), (9) and (11), respectively.

To demonstrate the generality, we consider the case when the file size is gamma distributed. Therefore, $F_{lk} \sim \Gamma(a, b)$, where a and b are the shape and scale parameters, respectively. Then, its mean is $\mathbb{E}[F_{lk}] = \overline{F} = ab$ and the pdf

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

is $f_{F_{lk}}(x,a,b) = x^{a-1}e^{-x/b}/(b^a\Gamma(a))$. Following the discussion in Section III-A, the service time S_{lk} of user lk, $S_{lk} = 2N_q F_{lk}/(R_{lk}^l C_{\text{FH}})$ is also gamma distributed, i.e., $S_{lk} \sim \Gamma(a, c_{lk})$, where $c_{lk} = 2N_q b/(R_{lk}^l C_{\text{FH}})$. Hence, the mean of S_{lk} is $2N_q\overline{F}/(R_{lk}^l C_{\text{FH}})$. The PDF of S_{lk} is given by $f_{S_{lk}}(x,a,c_{lk}) = x^{a-1}e^{-x/c_{lk}}/(c_{lk}^a\Gamma(a))$. Using this and (5), the PDF of the service time RV S can be obtained. Since $\mathbb{E}[S_lk] = 2N_q\overline{F}/(R_{lk}^l C_{\text{FH}})$ is same as before, $\mathbb{E}[S]$ and the queue stability condition in (6) are also the same. Substituting the $f_S(x)$ in (8), k_i can be evaluated as

$$k_i = \frac{\Gamma(a+1)}{i!\Gamma(a)} \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{\Lambda}{\Lambda + {c_{lk}}^{-1}}\right)^i \left(\frac{{c_{lk}}^{-1}}{\Lambda + {c_{lk}}^{-1}}\right)^a . \tag{12}$$

Further, the MGF of the RV S is given by

$$\Psi_S(s) = \sum_{l=1}^{L}\sum_{k=1}^{K} p_{lk}\left(\frac{{c_{lk}}^{-1}}{s + {c_{lk}}^{-1}}\right)^a . \tag{13}$$

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

IV. Numerical Results

A. Access Link Throughput

We consider a C-RAN system with massive MIMO RRUs employing $M = 300$ antennas in each cell. The cellular layout is 7-cell hexagonal with wrap around implementation. We drop $K = 10$ users in each cell such that no user lies within a distance of $d_{\text{min}} = 35$ m from the center of the cell. The pilot and data transmission powers are set to $P_{\text{UE}} = 23$ dBm. The remaining simulation parameters are listed in Table I. Using the 3GPP LTE model [21], we compute the large-scale fading coefficient β_{it}^l in dB as

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

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.

Fig. 3 shows that the cumulative distribution function (CDF) of SE (in b/s/Hz) for three values of M. The plot shows that the SE increases with M.

TABLE I: Simulation parameters.

Parameters	Symbol	Value
Intersite distance (ISD)	d_{ISD}	500 m
Number of pilots	τ_{p}	10
Channel bandwidth	B	20 MHz
Coherence interval	τ_{c}	200 symbols
Noise power	σ^2	-96 dBm
Average file size	$\overline{F}$	0.5 MB
Quantizer resolution	N_{q}	8 bit

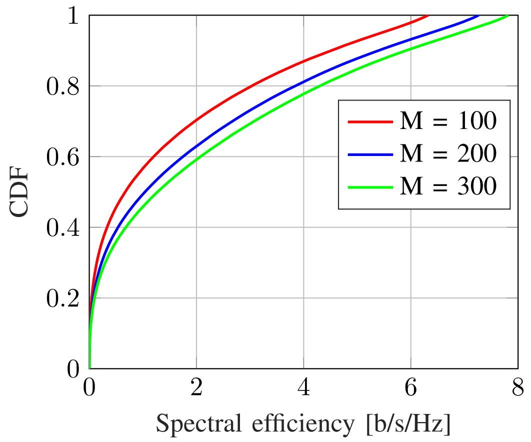

Fig. 3. CDF plot of spectral efficiency, $K = 10$, $\tau_p = 10$.

B. Sojourn Time and Queue Length

Some users, especially at the cell edges, might experience low data rates, which can occur due to bad channel conditions or due to severe multiuser interference and pilot contamination. We ensure 5 Mbps for each user in order to guarantee that the load is less than one, thereby ensuring the stability of the queue. This choice is justified as more than 75% of the users had SE higher than this value for all user drops.

Now, to compare the simulation result with the analytical solution, we take the inverse Laplace transform of (10) using a built-in MATLAB function. We compare the results with varying file sizes and different arrival rates. Fig. 4 shows simulation and analytical results of sojourn time distribution for $\lambda = 1$ and $\lambda = 5$. As we see, both the simulation and analytical results match quite well. However, we have some mismatch around zero. This occurs due to MATLAB's precision in handling the inverse Laplace transform at the vicinity of zero. Further, notice that the higher value of arrival rates, stretches the curve reducing the PDF peaks. Hence, latency increases with the higher values of arrival rates.

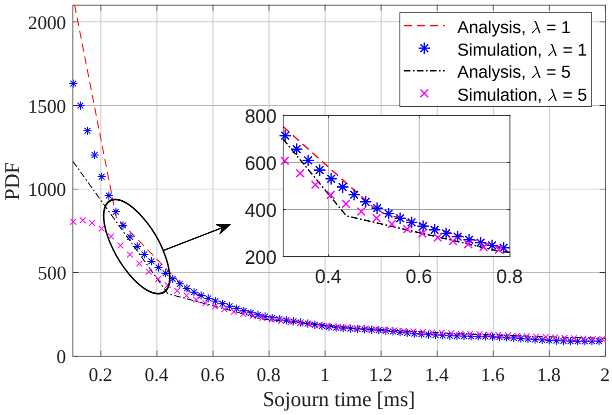

Fig. 4. Sojourn time distribution, $\overline{F} = 0.5$ MB, $C_{\text{FH}} = 100$ Gbps.

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

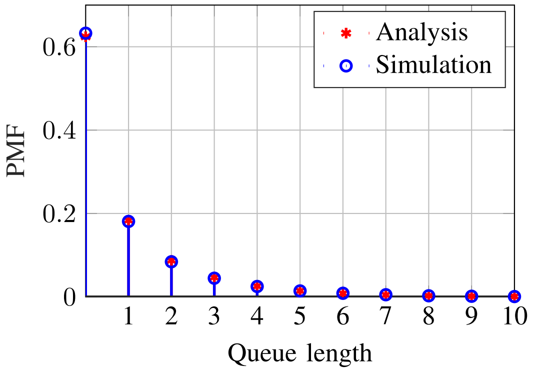

Fig. 5. Queue length distribution, $\lambda = 5, \overline{F} = 0.5$ MB, $C_{\text{FH}} =$ 100 Gbps.

Next, in order to illustrate that the presented model works for any general file size distribution, we consider that the file size is gamma distributed with the values of a and b fulfilling the stability condition in (6). Fig. 6 shows the results for two values of scale parameter, $a = 2$ and $a = 3$ for fixed b. Depending upon different values of the shape parameter a [3], the shape of the distribution will have different forms for given b. This is especially apparent in comparison to Fig. 4, where $a =$ is set to 1. Contrary to a, which changes shape of the distribution, the scale parameter b for a given a has the effect of stretching or shrinking the distribution shape. Fig. 7 shows the results for different values of b while keeping a fixed. Notice that the peak value of the distribution curve in Fig. 7 decreases when the value of b increases. As illustrated in Figs. 6 and 7, both results also match each other when the file size has general distribution.

C. Packet Size Impact

Now, we are interested to know the lowest achievable latency for different packet sizes at different percentiles. Depending upon the use cases and application, the FH will have its own latency threshold. This value can be as low as some hundreds of μs, typically it is assumed to be 250 μs. In order to guarantee such a low FH latency requirement, we assume the file size is small such that it contains only a single packet. According to [22], we assume a packet size of 500 B for URLLC and 1500 B for eMBB. Fig. 8 illustrates the 99th, 90th and 50th percentiles of the sojourn time for 500 B and 1500 B packet size. The following observations can be made from Fig 8. First, the sojourn time increases significantly with the increase in packet size as it requires more resources to process it. Second, with the faster switch speed, sojourn time decreases. For the slower switch speed, sojourn time

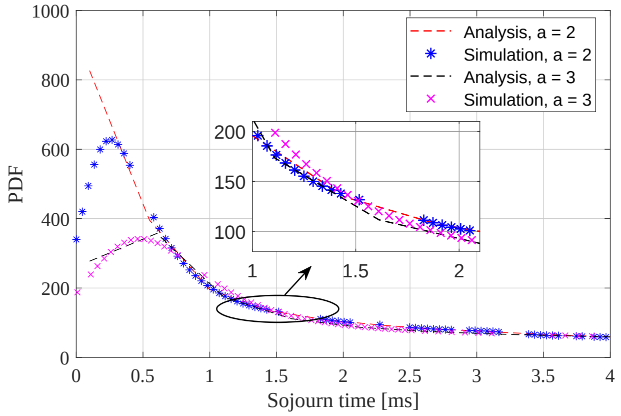

Fig. 6. Sojourn time distribution for gamma distributed file size, $\lambda = 1, b = 0.5$ MB, $C_{\text{FH}} = 100$ Gbps.

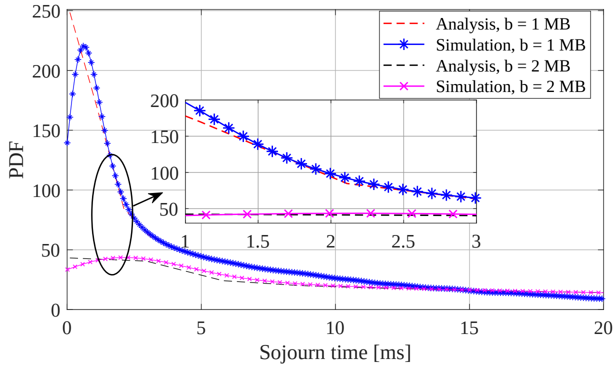

Fig. 7. Sojourn time distribution for gamma distributed file size, $\lambda = 1$, $a = 2$, $C_{\text{FH}} = 100$ Gbps.

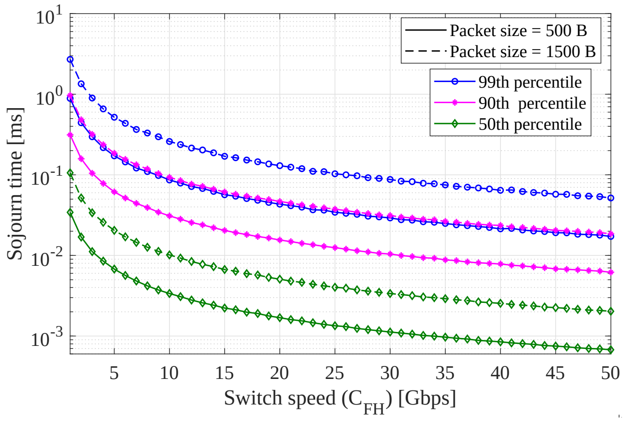

Fig. 8. %tile of sojourn time for 500 B and 1500 B packet sizes, $\lambda = 1$.

[3] The computation time of the sojourn time increases with a because the second term of the MGH of S in (13) is raised to exponent a.

grows abruptly, and given a 250 μs latency budget cannot be guaranteed. Hence, in order to meet the URLLC performance metric, one needs to have smaller packet sizes and the switch needs to operate at reasonably higher speeds. Third, the switch speed can be decreased without increasing the FH latency significantly, which means we can benefit from a statistical multiplexing gain as well.

V. Conclusion

The stringent latency requirement for FH traffic can be relaxed with an alternative functional split. In this paper, we presented an analytical framework to calculate the latency in the uplink of C-RAN massive MIMO system with functional Split 7. We considered both the access and FH networks in the analysis. We showed that the output port of an Ethernet switch can be modeled as an M/HE/1 queue when the file arrival process is Poisson and the file sizes are exponentially distributed. This allowed us to derive the tractable, closed-form expressions for sojourn time and queue length distribution. The simulation results corroborated the correctness of our analytical results. We showed that the analysis presented in this paper applies to any general file size distribution and we illustrated this by presenting the results for the gamma distribution. Our analysis also revealed the impact of different parameters such as average file size, arrival rate for the users, spectral efficiency of the users, and switch speed on the FH latency. We saw that the average file size, arrival rate, and spectral efficiency played a critical role. Furthermore, we observed that the switch speed can be reduced without incurring a notable increase in FH latency, which enables to exploit the benefits of statistical multiplexing. Future work will focus on extending the analysis to other functional splits, dimensioning FH traffic based on heterogeneous traffic and evaluating statistical multiplexing benefits.

Acknowledgment

The research leading to this work was supported from 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] NGMN Alliance, "NGMN white paper," March 2015.

[2] *3GPP TS 36.212: LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding* , 3GPP Std., Apr. 2013.

[3] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan 2013.

[4] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[5] China Mobile Research Institute, "C-RAN - The road towards green RAN," *White Paper*, Oct. 2011.

[6] U. Dötsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, June 2013.

[7] 3GPP, "3GPP TR 38.801 v14.0.0 (2017-03): Study on new radio access technology: Radio access architecture and interfaces (Release 14)," 2017.

[8] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, March 2016.

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

[10] eCPRI, "Common Public Radio Interface (CPRI); eCPRI Interface Specification (V1.0)," Tech. Rep., Aug. 2017.

[11] 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.

[12] 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.

[13] Q. Han, C. Wang, M. Levorato, and O. Simeone, "On the effect of fronthaul latency on ARQ in C-RAN systems," *CoRR*, vol. abs/1510.07176, 2015. [Online]. Available: http://arxiv.org/abs/1510.07176

[14] G. Mountaser, T. Mahmoodi, and O. Simeone, "Reliable and low-latency fronthaul for tactile internet applications," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2018.

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

[16] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 6, pp. 573–581, Jun. 2018.

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

[18] 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.

[19] 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.

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

[21] 3GPP, "3GPP TR 36.814 v9.0.0 (2010-03): Further advancements for e-utra physical layer aspects (release 9)," 2010.

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