Uplink Latency in Massive MIMO-based C-RAN with Intra-PHY Functional Split

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

***Abstract*—Functional splitting and packetized fronthaul (FH) are two approaches to realize cloud radio access networks in a cost-effective manner. In the former, some baseband functionalities are offloaded to remote radio units (RRUs) instead of centralizing all of them at a baseband unit pool. This reduces the capacity and latency requirements on FH when massive multiple-input-massive-output RRUs are used. The latter approach aims to use the ubiquitous Ethernet networks for FH. However, this leads to random packet delays due to queuing at switching/aggregating gateways. In this paper, we present a novel analytical framework to characterize the distribution of queuing delays at an aggregation gateway in the uplink. This framework incorporates random user activity, the uplink spectral efficiencies of users, the slotted nature of uplink transmissions, and FH capacity. We study the impact of packet arrival rates, average packet sizes, and FH capacity on queue length and queuing delay distributions. We show that significant statistical multiplexing gains are possible by aggregating traffic from multiple RRUs.**

***Index Terms*—Cloud radio access network, Massive MIMO, Uplink latency, Packetized fronthaul, Functional split**

I. Introduction

Functional splitting is used to relax the capacity and latency requirements imposed on the fronthaul (FH) by the common public radio interface (CPRI) protocol in cloud radio access networks (C-RANs) [1]. It involves offloading some baseband functionalities from the baseband unit (BBU) pool to remote radio units (RRUs). The functional split determines the functionalities that are offloaded. For example, in *intra-PHY split* [1], the physical layer functionalities are split between the BBU pool and RRUs. In massive MIMO-based C-RAN, where RRUs are equipped with tens or even hundreds of antennas, precoding in the downlink and equalization in the uplink are moved to the RRUs as shown in Fig. 1. This lowers the required FH capacity as it scales with the number of spatial streams and not with the number of antennas, as in the case of the CPRI protocol. The latency constraint is also more relaxed, as it is determined by the hybrid automatic repeat request (HARQ) process and not by the CPRI protocol [2]. Radio-over-Ethernet [1] is considered for FH given its cost-effectiveness and widespread use in core networks. However, packetized transport over FH introduces latency concerns due to the possibility of queuing delays at aggregation/switching gateways such as Ethernet switches. Due to the random nature of these delays, they can exceed the delay budget for the FH – an event referred to as *outage*. An analytical expression for the queuing delay distribution is needed to study the probability of an outage, and developing it for a massive MIMO-based C-RAN with intra-PHY split is the focus of this paper.

Related Literature: The uplink queuing delay at a switching gateway is studied in [3] for a functional split with equalization at the BBU pool. Unlike ours, that work considers a different functional split, presents only approximate results, and does not model the links between users and RRUs. Bounds on the latency distribution with MAC-PHY and RLC-PDCP splits are presented in [4]. However, the model is different from ours as packets are fragmented and then forwarded to the BBU pool via parallel paths. The latency in FH for PHY-RF, intra-PHY, MAC-PHY splits are evaluated in [5], [6]. Also, the impact of packetization and scheduling policies on FH latency is studied in [2]. However, no analytical results are presented in [2], [5], [6]. A delay exponent approach is used to satisfy the delay constraints of different service classes in [7].

The uplink latency in a massive MIMO-based C-RAN is analyzed in [8], [9]. However, they do not model the slotted nature of uplink transmissions and allow uplink transmissions to occur at any arbitrary time, which is not the case in practice. This leads to a continuous-time queuing model, unlike ours, which is discrete in nature. Moreover, a worst-case spectral efficiency (SE), which does not account for the dynamic uplink interference due to random user activity, is used in [8], [9].

Contributions: We propose a novel discrete-time queuing model for an RRU gateway in the FH, which aggregates uplink traffic from multiple RRUs. This yields closed-form expressions for the generating functions of steady-state queue length and sojourn time probability mass functions (PMFs). The sojourn time measures the latency in the RRU gateway. The analytical results are verified via numerical simulations. The proposed model is then used to study the probability of an outage, which occurs when the sojourn time exceeds a delay budget. We see that the outage probability decreases as the FH capacity increases. Further, the FH capacity required per RRU to meet a delay constraint decreases when traffic is aggregated from a higher number of RRUs, due to statistical multiplexing.

Manuscript received August 20, 2019; revised December 6, 2019; accepted January 2, 2020. This work was supported by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 762057 (5G-PICTURE) and the Saxon State government out of the State budget approved by the Saxon State Parliament. The associate editor coordinating the review of this letter and approving it for publication was G. Zheng.

J. Francis is with Indian Institute of Technology Palakkad, India. J. K. Chaudhary and G. Fettweis are with Technische Universität Dresden, Germany. A. N. Barreto is with Barkhausen-Institut gGmbH, Germany. Email: jobinfrancis@iitpkd.ac.in, {jay_kant.chaudhary, gerhard.fettweis}@tu-dresden.de, andre.nollbarreto@barkhauseninstitut.org

Digital Object Identifier

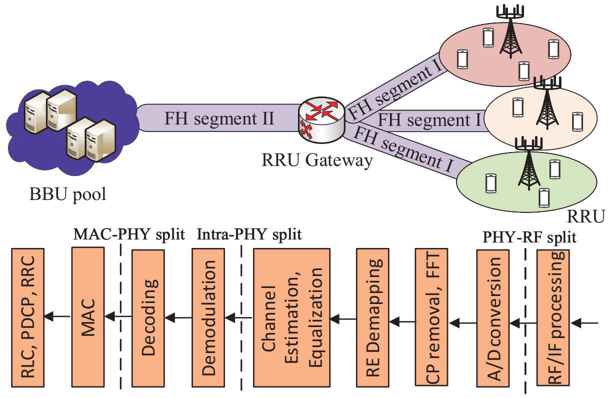

Fig. 1. C-RAN with Ethernet-based FH and intra-PHY split.

II. System Model

The C-RAN topology consists of L cells, K users in each cell, a two-hop Ethernet-based FH with an RRU gateway, and a BBU pool. A massive MIMO-aided RRU with M antennas is located at the cell center. The network is illustrated in Fig. 1.

Uplink Transmissions: The uplink transmissions from the K users in a cell are spatially multiplexed onto the same time-frequency resource, referred to as resource element (RE) as in LTE. Let $\mathbf{h}_{ik,l}$ denote the complex baseband channel gain vector from user k in cell i to RRU l. We assume spatially uncorrelated Rayleigh fading [10], i.e., $\mathbf{h}_{ik,l} \sim \mathcal{CN}(\mathbf{0}, \beta_{ik,l}\mathbf{I}_M)$, where $\beta_{ik,l}$ denotes the large-scale fading coefficient.

The channel gains are estimated via uplink training, which is repeated every coherence interval of τ_c REs. Orthogonal pilot sequences of length τ_p are used with pilot reuse, which results in pilot contamination. Let $\mathcal{B}_{ik}$ denote the set of users that reuse the pilot sequence used by user k in cell i. Each RRU generates the minimum mean square error (MMSE) estimates of the uplink channel gain vectors of users in its cell. That is, RRU l estimates $\mathbf{h}_{lk,l}$, for $k = 1, \ldots, K$. These are then used for matched-filter equalization at the RRU.

Note that only users with data to transmit are active in a transmission slot and cause interference to other users. Let p_{lk} and α_{lk} denote the transmit power and activity factor of user k in cell l, respectively. The activity factor is the probability of a user being active in a slot. It is the same in each slot and captures the temporal behavior of a user. The maximum uplink transmit power is P_{ue}. Then, the signal-to-inference-plus-noise ratio (SINR) of user k in cell l when active is given by

$$\gamma_{lk} = \frac{M\eta_{lk}p_{lk}\beta_{lk,l}^2}{\sigma^2 + p_{lk}\beta_{lk,l} + \sum_{it\in\mathcal{K}_{lk}}\alpha_{it}p_{it}\beta_{it,l} + M\eta_{lk}\sum_{it\in\mathcal{K}_{lk}}\alpha_{it}p_{it}\beta_{it,l}^2},$$

where $\eta_{lk} = \left(\sum_{uv\in\mathcal{B}_{lk}}\beta_{uv,l} + \sigma^2/\left(P_{\text{ue}}\tau_p\right)\right)^{-1}$, $\mathcal{K}_{lk} = \{(it) \neq (lk) : i \in \{1, \cdots, L\}, t \in \{1, \cdots, K\}\}$ is the set of all users in the network except user k in cell l, and σ^2 is the additive white Gaussian noise variance. The SINR expression is derived in a manner similar to that in [10]. The uplink SE R_{lk} is then given by $R_{lk} = \nu_{(\text{ul})}\left(1 - \tau_{\text{p}}/\tau_{\text{c}}\right)\log_2\left(1 + \gamma_{lk}\right)$ bits/RE, where $\nu_{(\text{ul})}$ is the fraction of REs allocated for uplink data transmission. Note that this SE expression accounts for the dynamic nature of interference owing to the random user activity. A min-max uplink power optimization is carried out to ensure a minimum SE of R_{min} for all the users [11].

User Traffic Model: The uplink data is generated as follows. The user k in cell l generates packets for uplink transmission as a Poisson point process with an arrival rate of λ_{lk} packets/s. Hence, the inter-arrival time between two consecutive packet arrivals is exponentially distributed with rate λ_{lk}. We define slot as the basic unit of time. It could be the transmission time interval (TTI), which is $T = 0.125/0.25/0.5/1$ ms in 5G NR, or the OFDM symbol duration, which is $T = 66.7/33.3/16.7/8.3/4.2\,\mu$s in 5G NR [12]. A user is active if there is at least one packet arrival in a slot. Hence, the activity factor with Poisson arrivals is $\alpha_{lk} = 1 - \exp(-\lambda_{lk}T)$.

Let F_{lk} denote the packet size for user k in cell l. It is modeled as an exponential random variable (RV) with mean $\overline{F}$. There could be multiple packet arrivals in a slot, and in such cases, the packets are transmitted together to the RRU in the next slot, as REs are assumed to be sufficient in number. While packets can arrive at any time due to Poisson arrivals, the transmissions start only at the slot boundaries. This models the slotted nature of RE grants and transmission in practical systems such as LTE and 5G NR. The user's SE determines the number of REs needed to transmit the arrived packets.

At the RRU, the received symbols at different antennas are equalized to recover the spatially multiplexed symbols on an RE. Each of these symbols is quantized to $2N_q$ bits and are then encapsulated in an Ethernet frame for transport over FH.

Ethernet-based FH: We consider a two-hop FH network as shown in Fig. 1. The hops are referred to as FH segment I and FH segment II. While the former is dedicated to each RRU, the latter is shared by the RRUs in the network. The Ethernet frames from RRUs are aggregated at the RRU gateway, which does the switching of traffic between the RRUs and BBU pool. The Ethernet frames from RRUs are stored in a first-in-first-out queue in a random order until FH segment II is available. This models the behavior of an Ethernet switch. Since the capacity C_{FH} of FH segment II is finite, queuing delays may occur. This can result in outages if these delays exceed the budget D for the FH. This budget is computed by deducting from the one-way HARQ trip time, which is 3 ms in LTE, the fixed delays involved in RRU and BBU pool processing, packetization, and propagation [2]. The FH delay budget for LTE is generally in the order of hundreds of microseconds.

Simplifications and Discussion: We note that while practical systems such as LTE and Ethernet-based FH motivate several aspects of the paper, not all the aspects are modeled. These simplifications have been made to arrive at a model that is practically relevant, yet analytically tractable. They include an idealized FH segment I with large enough capacity, so that the queuing delays at the RRUs are negligible and less than a slot duration, ignoring retransmissions arising from transmission failures, and availability of sufficient number of REs to send the packet arrivals in a slot to the RRU.

III. Queue Modeling and Steady-state Analysis

We first develop a queuing model for the RRU gateway. Then, the queue length and sojourn time PMFs are derived.

A. Queue Model

In order to study the queuing dynamics at the RRU gateway, the arrival and service processes of the queue need to be characterized. This is done below.

Arrival Process: The Ethernet frames arrive at the RRU gateway only at slot boundaries since the uplink transmissions from users last for a slot duration. The digitized received symbols are encapsulated in an Ethernet frame at the end of the slot. No frame arrival from an RRU occurs if no user in its cell transmits in the previous slot. Thus, the arrivals from an RRU l follow a Bernoulli process with probability of no arrival $p_l = \prod_{k=1}^{K} \exp(-\lambda_{lk} T) = \exp\left(-\Lambda_l T\right)$, where $\Lambda_l = \sum_{k=1}^{K} \lambda_{lk}$. It is the probability of no packet arrival in a slot duration T from the K users in cell l. Note that simultaneous arrival of frames from different RRUs is possible. Hence, the frame arrival process is a batch arrival process and the batch size A is the sum of L Bernoulli RVs, which indicate the frame arrivals from different RRUs. Let $\mathcal{L} = \{1, \ldots, L\}$. Then, the probability $p_A(i)$ that the batch size equals i is

$$p_A(i) = \sum_{E \in \mathcal{L}, |E| = n} \prod_{l \in E} p_l \prod_{l \in E^c} (1 - p_l), \text{ for } i = 0, \ldots, L.$$

Service Process: The service time S is independent across frame arrivals since the number of packet arrivals in a slot and their packet sizes are independent across slots and users. Thus, only the marginal distribution of S conditioned on an arrival is needed. Towards this, we first compute the number of bits B_l in a frame arriving from RRU l as follows. Let N_{lk} denote the number of packet arrivals in a slot for user k in cell l. It is a Poisson RV with rate $\lambda_{lk} T$ since the packet arrival process is Poisson. Let $F_{lk}^{(n)}$ denote the packet size in bits for the n^{th} arrival. Thus, the total number of bits to be transmitted in the uplink is $\sum_{n=1}^{N_{lk}} F_{lk}^{(n)}$. The number of REs needed to transmit these bits is $\sum_{n=1}^{N_{lk}} F_{lk}^{(n)} / R_{lk}$. Note that it is also the number of symbols at RRU l from user k after equalization. Thus, the total number of received symbols at RRU l is computed by summing over K users in cell l and is $\sum_{k=1}^{K} \sum_{n=1}^{N_{lk}} F_{lk}^{(n)} / R_{lk}$. Since each received symbol is quantized into $2N_q$ bits, the frame size $B_l = 2N_q \sum_{k=1}^{K} \sum_{n=1}^{N_{lk}} F_{lk}^{(n)} / R_{lk}$.

Let $G_{lk}^{(n)} = 2N_q F_{lk}^{(n)} / R_{lk}$. It is exponentially distributed with mean $\mu_{lk} = 2N_q \overline{F} / R_{lk}$. Therefore, $\sum_{n=1}^{N_{lk}} G_{lk}^{(n)}$ is Erlang distributed with shape parameter N_{lk} and scale parameter μ_{lk}. Hence, B_l is the sum of K independent, but non-identical, Erlang RVs. The following result gives its distribution.

Result 1. *The cumulative distribution function (CDF) $F_{B_l}(x)$ of B_l conditioned on the event that there is an arrival from cell l, i.e., $N_l = \sum_{k=1}^{K} N_{lk} > 0$, is given by*

$$F_{B_l}(x) = \sum_{m=1}^{\infty} \frac{T^m \exp(-\Lambda_l T)}{1 - \exp(-\Lambda_l T)} \sum_{\substack{n_1, \ldots, n_K \geq 0 \\ \sum_{k=1}^{K} n_k = m}} \left[\prod_{k=1}^{K} \frac{\lambda_{lk}^{n_k}}{n_k!} \right] \times \left(1 - \boldsymbol{e}_1^T \exp\left(x \boldsymbol{M}\right) \mathbf{1}\right), \quad (1)$$

where $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^T$ and $\mathbf{1} = [1, \ldots, 1]^T$ are $m \times 1$ vectors, and $\exp(x\boldsymbol{M})$ is the matrix exponential of $x\boldsymbol{M}$. The $m \times m$ block-diagonal matrix $\boldsymbol{M}$ has entries $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K$, where $\boldsymbol{M}_k$ is an $n_k \times n_k$ matrix with $-1/\mu_{lk}$ in the main diagonal, $1/\mu_{lk}$ in the super diagonal, and zero elsewhere. ■

Proof: The proof is relegated to Appendix A.

The service time S_l in number of slots needed for the RRU gateway to forward the B_l bits to segment II is $\lceil B_l / (C_{\text{FH}} T) \rceil$, where $\lceil \cdot \rceil$ denotes the ceil operation[1]. Using (1), PMF $p_{S_l}(i)$ of S_l, for $i = 1, \ldots, \infty$, is given by

$$p_{S_l}(i) = \mathbb{P}(S_l = i) = F_{B_l}\left(i C_{\text{FH}} T\right) - F_{B_l}\left((i-1) C_{\text{FH}} T\right).$$

Since an arriving frame can be from any of the L RRUs, the service time $S = S_l$, for $l = 1, \ldots, L$, with probability q_l. Here, q_l is the probability that a frame arriving at the RRU gateway is from RRU l and is given by

$$q_l = \sum_{n=1}^{L} \frac{p_l}{n \left(1 - \prod_{k=1}^{L} p_k\right)} \sum_{\substack{E \in \mathcal{L} - \{l\}, \\ |E| = n-1}} \prod_{k \in E} p_k \prod_{k \in E^c} (1 - p_k).$$

This is because, in a batch of size n containing an arrival from RRU l, the frame from RRU l is chosen with probability $1/n$. Thus, the PMF of service time S conditioned on an arrival is $p_S(i) = \sum_{l=1}^{L} q_l p_{S_l}(i)$, for $i = 1, \ldots, \infty$.

B. Steady-state Analysis

We now present results for stability, queue length PMF, and sojourn time PMF, which follow from [13, Chap. 4].

1) Stability: The load ρ is defined as the product of the average number of frame arrivals and the average service time. Thus, the load $\rho = \left[\sum_{i=1}^{\infty} i p_A(i)\right] \left[\sum_{i=1}^{\infty} i p_S(i)\right]$. The stability of the queue is ensured when the load $\rho < 1$.

2) Queue Length: The generating function $Q(z)$ of queue length can be expressed in terms of the generating functions $A(z)$ and $S(z)$ of RVs A and S, respectively. Here, $A(z) = \sum_{i=0}^{\infty} p_A(i) z^i$ and $S(z) = \sum_{i=1}^{\infty} p_S(i) z^i$. Then, $Q(z)$ is

$$Q(z) = \frac{(1-\rho)(1-z) S(A(z))}{S(A(z)) - z}, \quad (2)$$

where $S(A(z))$ is the generating function of the number of arrivals during a service time. The queue length PMF is obtained by taking the inverse Z-transform of $Q(z)$.

3) Sojourn Time: The sojourn time T of a frame is the sum of the waiting time W_b of the batch and the sojourn time T_x measured from the start of service of the batch to which the frame belongs to. Therefore, $T = W_b + T_x$. The generating function $W_b(z)$ of W_b is given by $W_b(z) = (1-\rho_b)(1-z)/(1 - p_{\text{batch}} - z - S_b(z))$, where $p_{\text{batch}} = 1 - \prod_{l=1}^{L} p_l$ is the probability of a batch arrival and $S_b(z)$ is the generating function of batch service time S_b conditioned on a batch arrival. Lastly, $\rho_b = p_{\text{batch}} \mathbb{E}[S_b]$, where $\mathbb{E}[\cdot]$ denotes the expectation operator. Since S_b is the sum of service times of the frames in the batch, the generating function $S_b(z)$ is

$$S_b(z) = \frac{1}{p_{\text{batch}}} \sum_{E \in \mathcal{L}, |E| > 0} \prod_{l \in E} \left[p_l S_l(z)\right] \prod_{l \in E^c} (1 - p_l), \quad (3)$$

where $S_l(z) = \sum_{i=0}^{\infty} p_{S_l}(i) z^i$ is the generating function of S_l.

[1] For analytical tractability, we assume that the RRU gateway and uplink transmissions have slot as the same basic unit of time.

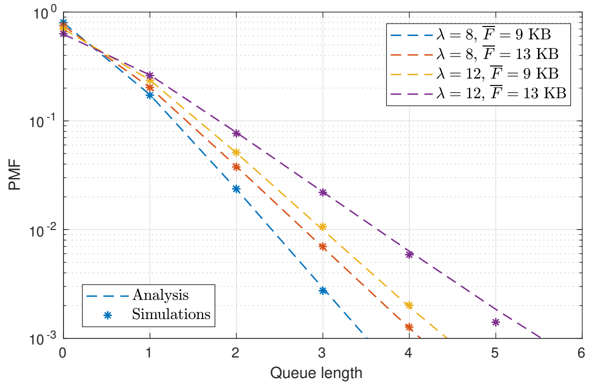

Fig. 2. Queue length PMF for different arrival rates λ and average packet sizes $\overline{F}$ when $C_{\text{FH}} = 1$ Gbps.

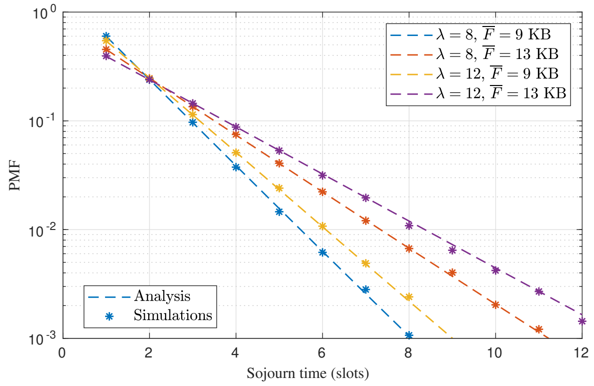

Fig. 3. Sojourn time PMF for different arrival rates λ and average packet sizes $\overline{F}$ when $C_{\text{FH}} = 1$ Gbps.

The generating function $T_x(z)$ of T_x is given in [13]. In order to simplify the computations, we approximate T_x with the batch service time S_b. Specifically, S_b is an upper bound on T_x and equality happens if the frame under consideration is serviced last in a batch or if the batch itself is of size one. Thus, $T_x(z) \approx S_b(z)$. Then, the generating function $T(z)$ of RV T is

$$T(z) = \frac{(1-\rho_b)(1-z)S_b(z)}{1 - p_{\text{batch}} - z - S_b(z)}. \tag{4}$$

The sojourn time PMF is obtained by taking the inverse Z-transform of $T(z)$.

The mean sojourn time $\overline{T}$ is computed using (2) and Little's law [13]. It is given by $\overline{T} = \mathbb{E}[S] + (\Lambda\mathbb{E}[S^2] - \rho)/(2(1-\rho))$.

4) Efficient Computation of Inverse Z-transform: We use the long-division method [14] to efficiently compute the inverse Z-transform. This involves expressing the Z-transform as a ratio of two polynomials. Note that $[p_{S_b}(0), \ldots, p_{S_b}(s_{\max})]$ are the coefficients of the polynomial $S_b(z)$, where $s_{\max}$ is the maximum service time beyond which the PMF values are negligible. Then, the coefficients for the numerator polynomial of $T(z)$ are $(1-\rho_b)[p_{S_b}(0), p_{S_b}(1) - p_{S_b}(0), \ldots, p_{S_b}(s_{\max}) - p_{S_b}(s_{\max}-1), p_{S_b}(s_{\max})]$ and for the denominator polynomial of $T(z)$ are $[1 - p_{\text{batch}} + p_{S_b}(0), p_{S_b}(1) - 1, \ldots, p_{S_b}(s_{\max})]$. The long-division method is used to find the quotient polynomial, whose coefficients yield the sojourn time PMF values.

This procedure can be repeated to evaluate the queue length PMF. However, evaluating the coefficients of $S(A(z)) = \sum_{i=0}^{s_{\max}} p_S(i)(A(z))^i$ is slightly more involved. This can be done efficiently by using repeated convolution to compute the coefficients of $(A(z))^i$ and then appropriately summing the coefficients for different i.

IV. Numerical Results

We consider a hexagonal cellular layout with $L = 7$ cells and a cell radius of 500 m. In each cell, $K = 10$ users are randomly dropped. The RRUs are equipped with $M = 200$ antennas. We set $P_{\text{ue}} = 23$ dBm, $\sigma^2 = -174$ dBm, and $\tau_c = 200$ REs. Pilot sequences are uniquely assigned to the users. Hence, $\tau_p = 70$. The large-scale fading coefficient in dB is [8] $\beta_{ik,l} = -128.1 + 37.6\log_{10}(d_{ik,l}/d_0) + \Psi_{\text{shad}}$, where $d_{ik,l}$ is the distance between user k in cell i and RRU l, and $d_0 = 100$ m. Here, Ψ_{shad} is a Gaussian RV with zero mean and standard deviation of 8 dB, which models lognormal shadowing. Minimum user SE is set to $R_{\min} = 1$ bit/symbol. The packet arrival rate λ is the same for all users. The slot duration $T = 0.25$ ms, which is one of the possible TTIs in 5G. The average packet size $\overline{F}$ is set to be in the range of possible transport block sizes in 5G NR [12]. For these simulation parameters, we have observed that the first 3 terms of the series in (1) are sufficient to ensure numerical accuracy.

Fig. 2 plots the queue length PMF for different values of arrival rate and average packet size for a random realization of the large-scale fading coefficients. We see an excellent match between the analysis and simulation curves, which validates our analytical results. The PMF value at zero queue length for $\lambda = 8$ is higher than that for $\lambda = 12$. This is expected as the queue becomes empty more often at lower arrival rates. The PMF values at larger queue lengths are lower for $\lambda = 8$ compared to $\lambda = 12$. This is because large queue lengths occur less often at lower arrival rates. Similar trends are observed when $\overline{F}$ increases. The PMF value at zero queue length is higher for $\overline{F} = 9$ KB when compared to that for $\overline{F} = 13$ KB. The reverse is true at higher queue lengths.

The sojourn time PMFs for different arrival rates and average packet sizes are plotted in Fig. 3. Note that the PMF value at zero sojourn time is equal to zero because a minimum of one slot is needed to service an arrival. We again observe an excellent match between analysis and simulation results. The trends exhibited by the curves for $\lambda = 8$, 12 and $\overline{F} = 9$ KB, 13 KB are similar to those of the queue length PMFs in Fig. 2.

Fig. 4 plots the outage probability as a function of delay budget for different values of C_{FH}. These curves are the complementary CDFs of sojourn time averaged over large-scale fading. These results can be used to appropriately dimension the FH; for example, $C_{\text{FH}} = 10$ Gbps ensures that the outage probability is less than 10^{-4} when $\lambda = 8$ packets/s. We see that the outage probability is lower for higher C_{FH}. This is expected, as the service time decreases as C_{FH} increases.

We now study the statistical multiplexing gains possible by aggregating traffic from RRUs at the RRU gateway. This is done as follows [15]. First, the FH capacity needed to ensure that the average sojourn time $\overline{T}$ is below a threshold D is

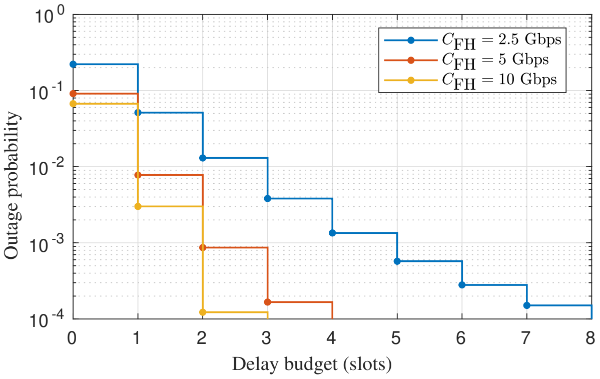

Fig. 4. Outage probability for different values of FH capacity C_{FH} when $\lambda = 8$ packets/s.

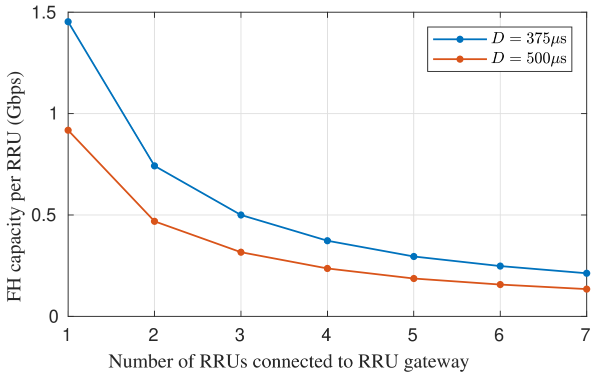

Fig. 5. FH capacity per cell as a function of the number of cells connected to the RRU gateway for $\lambda = 8$ packets/s.

computed. It is then divided by the number of RRUs connected to the RRU gateway to determine the required FH capacity per RRU. This is repeated for different number of RRUs connected to the RRU gateway. These results are plotted in Fig. 5 for two different values of D. We see that the required FH capacity per RRU decreases as the number of RRUs connected to the RRU gateway increases. This saving in the FH capacity is the statistical multiplexing gain. It is 85% for $D = 500$ μs when the number of connected RRUs is increased from one to seven. We also see that the required FH capacity per RRU is lower when D is higher, i.e., for a more relaxed delay requirement.

V. Conclusions

We proposed a novel queuing model for the RRU gateway with uplink traffic from users to RRUs and then to a BBU pool through an Ethernet-based FH. We derived closed-form expressions for the steady-state queue length and sojourn time distributions. The analysis took into account the user activity factors, time-slotted uplink transmissions, users' SEs, and FH capacity. The analytical results were validated through simulations. The impact of FH capacity on outage probability was then studied. Lastly, we evaluated the FH capacity savings possible via statistical multiplexing. In the investigated scenario, the statistical multiplexing gains were as high as 85%.

Some interesting avenues for future research are incorporating queuing delays at the RRUs, exploring the relation between the odds of simultaneous frame arrivals and outage probability, optimizing the network for delay savings, accounting for retransmissions, and considering other functional splits.

Appendix

A. Derivation of $F_{B_l}(x)$

The CDF $F_{B_l}(x)$ of B_l conditioned on an arrival event $\{N_l > 0\}$ is the probability $F_{B_l}(x) = \mathbb{P}(B_l < x | N_l > 0) = \mathbb{P}(B_l < x, N_l > 0) / \mathbb{P}(N_l > 0)$. The probability of an arrival from RRU l is $\mathbb{P}(N_l > 0) = 1 - \exp(-\Lambda_l T)$. To evaluate $\mathbb{P}(B_l < x, N_l > 0)$, we use the law of total probability to get $\mathbb{P}(B_l < x, N_l > 0) = \sum_{m=1}^{\infty} \mathbb{P}(B_l < x, N_l = m)$. Enumerating over the possibilities of m arrivals at K users, we get

$$\mathbb{P}(B_l < x, N_l > 0) = \sum_{m=1}^{\infty} \sum_{\substack{n_1,\ldots,n_K \geq 0 \\ \sum_{k=1}^{K} n_k = m}} \mathbb{P}\left(\sum_{k=1}^{K} \sum_{n=1}^{n_k} G_{lk}^{(n)} < x\right)$$
$$\times \mathbb{P}\big(N_{l1} = n_1, \ldots, N_{lK} = n_K | N_l = m\big) \mathbb{P}\big(N_l = m\big).$$

The first probability term in the summation is the CDF of a sum of Erlang RVs and is given by $1 - \boldsymbol{e}_1 \exp(x\boldsymbol{M})\mathbf{1}$ [16]. The second term is the probability of partitioning m arrivals among K users. It is given by $\binom{m}{n_1 \cdots n_K} \prod_{k=1}^{K} (\lambda_{lk}/\Lambda_l)^{n_k}$. Lastly, $\mathbb{P}\big(N_l = m\big) = (\Lambda_l T)^m \exp(-\Lambda_l T)/m!$. Putting everything together, we get $F_{B_l}(x)$ in (1).

References

[1] 3GPP, "Study on new radio access technology: Radio access architecture and interfaces," NTT Docomo, Inc., Tech. Rep. 38.801, Aug. 2016.

[2] C. Chang, N. Nikaein, and T. Spyropoulos, "Impact of packetization and scheduling on C-RAN fronthaul performance," in *Proc. Globecom*, Dec. 2016, pp. 1–7.

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

[4] G. Mountaser, M. Mahlouji, and T. Mahmoodi, "Latency bounds of packet-based fronthaul for cloud-RAN with functionality split," 2019.

[5] G. Mountaser, M. L. Rosas, T. Mahmoodi, and M. Dohler, "On the feasibility of MAC and PHY split in Cloud RAN," in *Proc. WCNC*, Mar. 2017, pp. 1–6.

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

[7] H. Ren, N. Liu, C. Pan, M. Elkashlan, A. Nallanathan, X. You, and L. Hanzo, "Low-latency C-RAN: An next-generation wireless approach," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 48–56, Jun. 2018.

[8] J. K. Chaudhary, J. Francis, A. N. Barreto, and G. Fettweis, "Latency in the uplink of massive MIMO CRAN with packetized fronthaul: Modeling and analysis," in *Proc. WCNC*, Apr. 2019.

[9] ——, "Packet loss in latency-constrained Ethernet-based packetized C-RAN fronthaul," *Proc. PIMRC*, Sep. 2019.

[10] E. Bjornson and E. G. Larsson, "Three practical aspects of massive MIMO: Intermittent user activity, pilot synchronism, and asymmetric deployment," in *Proc. Globecom Workshops*, Dec. 2015, pp. 1–6.

[11] T. V. Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," vol. 15, no. 9, pp. 6384–6399, Sep. 2016.

[12] 3GPP, "Ts 38.214 v15.3.0 - physical layer procedures for data," Tech. Rep. 38.214, 2018.

[13] S. K. Bose, *An Introduction to Queueing Systems*. Kluwer, 2002.

[14] S. Barnard and J. M. Child, *Higher Algebra*. Macmillan, 1959.

[15] R. R. Mazumdar, "Notes on statistical multiplexing," https://ece.uwaterloo.ca/~mazum/ECE610/statmux.pdf.

[16] B. Legros and O. Jouini, "A linear algebraic approach for the computation of sums of erlang random variables," *Applied Mathematical Modelling*, vol. 39, no. 16, pp. 4971 – 4977, 2015.