A Variable Rate Fronthaul Scheme for Cloud Radio Access Networks (C-RAN)

Sandip Das, *Member, IEEE*, and Marco Ruffini, *Senior Member, IEEE*
CONNECT, Trinity College Dublin, Ireland, Email: dassa@tcd.ie, marco.ruffini@scss.tcd.ie

Abstract—Cloud Radio Access Networks (C-RANs) are considered one of the most promising candidates for implementing 5G mobile communication systems. C-RAN enables centralisation of baseband processing, enabling advanced coordination between base stations, such as coordinated multi-point and inter-cell interference cancellation. In addition, it allows pooling of resources across several cells, providing statistical multiplexing gains of computing resources. However, the link between the remote radio unit and the baseband unit requires high transmission capacity, making the fronthaul link a potential bottleneck for future dense cell deployments. One of the current solutions to this issue is to compress the fronthaul transmission rate. A second under standardisation is to adopt a different functional split that can reduce the transmission capacity requirement. While this solution decreases the capacity requirements on the transport link, it decentralises some of the computational resources, requiring a more complex remote radio unit (e.g., compared with a CPRI type of solution), whose resources cannot be utilised by other cells when not in use. It is thus expected that in the future, multiple solutions (different functional splits and CPRI) will coexist.

In this paper, we introduce the concept of Variable Rate Fronthaul (VRF) for C-RAN. This scheme operates on a CPRI type of interface (e.g., one that transmits I/Q data samples) with the novelty of dynamically changing the cell bandwidth, and consequently the fronthaul data rates, depending on the cell load, with the support of a Software Defined Network (SDN) controller. This allows for a more efficient transport of C-RAN cells' data over a shared backhaul. We first propose a mathematical analysis of the VRF performance using a queuing theory approach based on the Markov model. We then provide the results of our simulation framework both for validation and in support of the mathematical analysis. Our results show that the proposed VRF scheme provides significantly lower blocking probability over a shared backhaul than standard CPRI.

Index Terms—Variable Rate Fronthaul, Variable Bandwidth Fronthaul, 5G, Cloud Radio Access Network (C-RAN), Passive Optical Network (PON), CPRI, Fronthaul Aggregator, Statistical Multiplexing, Functional Split.

I. INTRODUCTION

In the last decade, we have witnessed a massive growth in mobile data connectivity, which is expected to continue steadily for the foreseeable future [1], [2]. In order to support the demand for higher data rates, seamless mobile connectivity and machine type communication, Fifth Generation (5G) Radio Access Networks (RAN) is being developed with higher cell density and data rates to meet these requirements. Cloud-RAN is a promising framework for 5G, where baseband processing from multiple base stations is centralized (in baseband units - BBUs), while the antenna sites are highly simplified (the remote radio units - RRUs) [3]. The mobile protocol stack is thus split between the BBU and RRU, and the split point

can be chosen among several options (more than eight possible splits are currently being investigated [4]). One of these splits, called split-8, and the first one to be deployed in the field through a CPRI interface (Common Public Radio Interface) [5], locates all baseband processing in the centralised BBU, leaving only sampling and RF up/down conversion at the RRU. The advantages of the split-8 are those of a simpler RRU, fully centralized resource sharing and the ability to carry out advanced coordinated transmission across multiple cells, such as Coordinated Multi-Point (CoMP) communication [6]. However, it requires much higher data rates than the legacy distributed RAN (or than a functional split option equal or less than 4, i.e., above the MAC) [7].

CPRI supports transport of baseband samples at a fixed rate only and the per Antenna Carrier (AxC) fronthaul capacity demand is higher than 1Gbps (1.2288 Gbps) for a 20MHz LTE RRU [8]. When scaled to a future 100MHz bandwidth, 8-channel MIMO systems over three sectors, the required fronthaul capacity reaches 150 Gb/s. In order to overcome these bottlenecks, several solutions have been proposed in recent years, among which compressed CPRI [9] and different functional splits [10] are most popular. The compressed CPRI scheme applies compression in the I/Q baseband samples transported through fronthaul. Though these schemes have the advantage of reducing the fronthaul rate, they nonetheless transport data at a fixed rate, which is independent of the actual cell usage. Another possibility, currently under standardisation is to adopt a different functional split (i.e., below split-8). This reduces the fronthaul capacity demand and enables statistical multiplexing as the fronthaul capacity varies depending on the number of users the RRU is currently serving. However, it is achieved at the expense of a more complex, expensive and power-hungry RRU, as this needs to carry out some part of the physical layer or even MAC processing of the BBU. In addition, any processing resource installed in the RRU remains local to the cell and cannot be shared with other cells. In [11], the advantage of using centralized processing in C-RAN with a split-8 is analyzed through teletraffic theory and queuing systems. In this work, the authors analyze the improvement of blocking probability in RRU and BBU processing due to C-RAN architecture. Similar work is carried out for C-RAN employing functional split in [12].

The insight of our work is based on the observation that in future, the progressive densification of mobile cells will dramatically change their traffic patterns. Smaller cells will serve a smaller number of users, leading to much larger statistical fluctuation in cell traffic. Especially when using next generation high-capacity multi-media application, for example,

just a few users in one given cell could easily drive it to its maximum rate, while the nearby cells might have no users. If such small cells are multiplexed over a shared fronthaul transport system such as a Passive Optical Network (PON), statistical multiplexing can be used to reduce the overall backhaul requirement. However, today statistical multiplexing of C-RAN streams is only possible if functional decomposition is applied with a split-PHY level equal or lower than split-7 [4].

In this paper, we propose a Variable Rate Fronthaul (VRF) scheme that transports the raw baseband I/Q samples as the traditional CPRI transport mechanism (i.e., at split-8). The possibility for commercial TDM-PONs to support CPRI through the use of fixed upstream capacity allocation was experimentally demonstrated in [13]. However in this work the upstream capacity is fixed, independently of the actual cell usage. In our scheme instead, a Software Defined Network (SDN) controller monitors the cell usage, and dynamically adapts the cell wireless bandwidth to meet the traffic demand. A reduction in traffic demands will thus trigger a reduction in wireless spectrum usage, which in turn will decrease the I/Q sampling rate and consequently the fronthaul transport rate. The advantage is that the RRU remains simple while restoring statistical multiplexing even for a fronthaul transport system shared between multiple cells (e.g., such as a PON). It should be noted that split-7.1 [4], which puts the Inverse Fast Fourier Transform (IFFT) at the RRU, can also be used with our VRF scheme, since it produces a constant transmission rate (although at a lower rate than split-8). Using both mathematical analysis and event-driven simulation, we evaluated the performance of our proposed scheme for a typical cloud-RAN scenario. The results show that our VRF scheme can achieve significant reduction of traffic congestion in shared fronthaul medium reducing blocking probability of end-user services.

The rest of the paper is organized as follows. Section II provides the system model for the work considered here. Section III provides a detailed description of the mathematical analysis of the system based on queuing theory. The details of the simulator framework and the simulation parameters are discussed in Section IV. In Section V, we compare and discuss the results obtained from the analytical method with those obtained through simulation. Finally, in Section VI we conclude this article by briefly discussing the outcomes of this work.

II. System Model

Let us consider the system illustrated in Fig. 1. The mobile User Equipments (UEs) are connected to RRUs via an LTE network. Each RRU transmits data to its BBU through CPRI links for centralized processing. We refer to these as fronthaul links. A Fronthaul Aggregator (FHA) aggregates several such links from a cluster of RRUs and creates a Fronthaul Aggregated Link (FAL) which then connects to a centralized BBU pool, forming a tree-like network topology [14]. This abstract configuration could be implemented in practice with a Passive Optical Network (PON), by using a power splitter and Optical Line Terminal (OLT) as the fronthaul aggregator,

with each RRU having an Optical Networking Unit (ONU) for communicating to the OLT [15]. It should be noticed that in general, a shared backhaul can provide substantial cost savings in dense cell deployments, as it allows to take advantage of statistical multiplexing across base stations. However, statistical multiplexing cannot be exploited by traditional CPRI, which produces constant data rate over the fronthaul link. Our Variable Rate Fronthaul is a solution to this issue: by dynamically changing the wireless bandwidth of each cell and consequently the associated fronthaul rate, it enables both reuse of the wireless spectrum and statistical multiplexing over the shared fronthaul link.
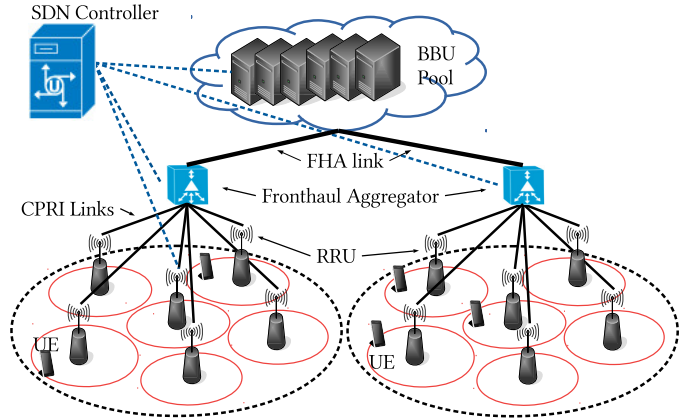

Fig. 1. System architecture of cloud-RAN

In our proposed architecture, the SDN controller interacts with the BBUs to constantly monitor the cell usage of each RRU. As a result, it coordinates operations with the BBU and RRU to dynamically adapt the wireless bandwidth and fronthaul rate. Recently, in [16], we have experimentally demonstrated the practical feasibility of our variable rate fronthaul concept using a software LTE BBU connected to its RRU (implemented using a USRP board [17]) via a fronthaul link operating over a PON.

In this work, we consider that the UEs are connected to the nearest RRU. Therefore, the cell coverage area per RRU can be estimated analytically using Voronoi diagrams as performed in [12]. However, unlike [12], where authors assume arbitrary transmission rates depending on user's behavior, and make use of a functional split architecture, we assume a fully centralized scheme with standard CPRI rates [5], listed in the Table I. Each line of the table indicates the wireless bandwidth supported, the size of the Fast Fourier Transform (FFT) required, the number of Physical Resource Blocks (PRBs) supported, the transmission rate at the CPRI interface and the maximum number of users that can be supported. In practice, the number of supported users depends on multiple factors such as traffic type, data rate requirement per user, per user service fairness index, etc. However, in this work, we assume for simplicity that each user is allocated two PRBs (or one Resource Group (RG)) when an application data stream (which we refer to as "call", following typical queuing theory terminology) is accepted for service. From this, we derive the maximum number of supported users per bandwidth, listed in Table I.

TABLE I
STANDARD CPRI DATA RATES FOR LTE

LTE bandwidth configuration (in MHz)	FFT size	Number of PRBs	CPRI data rate (in Mbps)	Max supported users
1.25	128	6	76.8	3
2.5	256	12	153.6	6
5	512	25	307.2	12
10	1024	50	614.4	25
15	1536	75	921.6	37
20	2048	100	1228.8	50

Fig. 2. State transition diagram for individual RRUs

In order to explain how the variable rate fronthaul system operates, let us consider a scenario where an RRU is serving already the maximum number of users for the allocated bandwidth. If a new user arrives, which cannot be handled within the current bandwidth, the SDN controller triggers a request to increase the allocated wireless spectrum. As a consequence, the fronthaul CPRI rate also increases to support the higher bandwidth configuration (as listed in table I). Similarly, when a call departs and the number of remaining users can be supported by the next lower bandwidth configuration, both wireless spectrum and fronthaul rate are decreased accordingly. In the following paragraph, we provide a mathematical interpretation of this model.

Let ϕ_i be the homogeneous Poisson random process with intensity Λ^i, modeling the call arrival process corresponding to the ith RRU. In this work, we consider the simple case where $\Lambda^i = \lambda$, $\forall i \in \{1, 2, \ldots, N\}$, where N denotes the number of RRUs connected to the same aggregator. Therefore, at any given time instant t, the capacity at the FHA link is represented as,

$$C(t) = \sum_{i=1}^{N} \alpha\{u_{i,t}\} \tag{1}$$

In (1), $u_{i,t} \in \phi_i$ represents the number of users the ith RRU is serving at time instant t. $\alpha : \phi_i \to D$, where $D = \{d_1, d_2, \ldots, d_M\}$ represents the data rates available for the fronthaul interface. In this work, we also consider the latency introduced by the SDN controller and BBU-RRU system to reconfigure the system on a different bandwidth, which, as shown in [16], can be estimated in a few hundred milliseconds.

This work aims to find the probability that a new customer call request is blocked, for a given network configuration. Here, we define the call request as the request from an UE to set up a connection with its nearest RRU for transmission and reception of data over one RG. Different network configurations are achieved by varying N, M, D and λ. Therefore, if B_C represents the capacity of the FHA link, then the steady state blocking probability at the aggregator is given by (2)

$$P_b = \lim_{t \to \infty} P\{C(t) \geq B_C\} \tag{2}$$

III. THEORETICAL ANALYSIS

In this section, we provide an analytical solution to the calculation of the blocking probability of a variable rate fronthaul system, based on queuing theory. The process can be divided in two phases. In the first phase, we find the probability for each RRU in a cluster to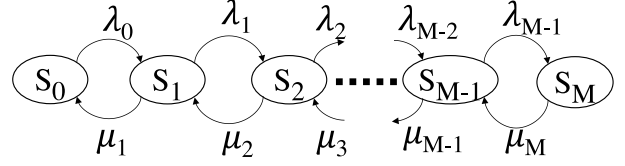 use a specific CPRI rate. In the next phase, we find the probability that the aggregator ends up in a blocking state, i.e., where an increase in CPRI rate cannot be supported as the FHA link is already at full capacity.

The queuing analysis for an RRU in a cluster using a given CPRI rate at any time instant requires the use of a threshold-based queuing system. Let F_i represent the forward threshold for an RRU to transition to a higher CPRI rate ($d_i \to d_{i+1}$) and R_i be the reverse threshold for an RRU to transition to a lower CPRI rate ($d_i \to d_{i-1}$). Let S_i denote the state of an RRU using CPRI rate d_i (or in bandwidth configuration B_i) with F_i and R_i being the forward and reverse thresholds, respectively. This means that if the RRU is in state S_i and currently serving F_i number of users, then any incoming call request to the RRU will trigger the adoption of the next higher CPRI rate $d_i \to d_{i+1}$ (i.e., the adoption of the next higher bandwidth configuration). Similarly, if the RRU is in state S_i, and currently it is serving R_i number of users, then any call departure from the RRU will trigger the adoption of the next lower CPRI rate $d_i \to d_{i-1}$. This state transition is represented in Fig. 2.

Our aim in the first phase of the analysis is to find the steady-state probabilities of S_i and the corresponding transition rates λ_i and μ_i. We can clearly notice that the steady-state probability of S_i, and the transition rates from S_i to S_{i+1} and S_{i-1} depend on the sub-states of S_i. Thus, in order to provide a solution for this type of system, we use the concept of stochastic complementation technique. This technique has been studied extensively in the field of computer science for adaptive processor allocation in computer servers [18], [19]. In the remainder of this section, we briefly introduce the stochastic complementation technique, before applying it to our model to obtain the steady-state probability and transition rates.

A. Background of Stochastic Complementation

The concept of stochastic complementation is based on the classic theory of decomposability of Markov chains which was introduced back in 1977 by P. J. Courtois [20]. The stochastic complementation technique originally introduced by C.D. Meyer in 1989 [21] provides a way of decoupling a Markov chain by means of partitioning. Let P be the transition probability matrix of a discrete space, discrete time Markov chain $\mathcal{M}$ with state space S. Let us partition this state space into two disjoint sets $\mathcal{L}$ and $\mathcal{R}$. The one step transition probability matrix of $\mathcal{M}$ becomes:

$$P = \begin{bmatrix} P_{\mathcal{L},\mathcal{L}} & P_{\mathcal{L},\mathcal{R}} \\ P_{\mathcal{R},\mathcal{L}} & P_{\mathcal{R},\mathcal{R}} \end{bmatrix}$$

The steady-state probability vector of $\mathcal{M}$ is $\pi = [\pi_{\mathcal{L}}, \pi_{\mathcal{R}}]$, where $\pi_{\mathcal{L}}$ and $\pi_{\mathcal{R}}$ are the steady-state probability vectors of
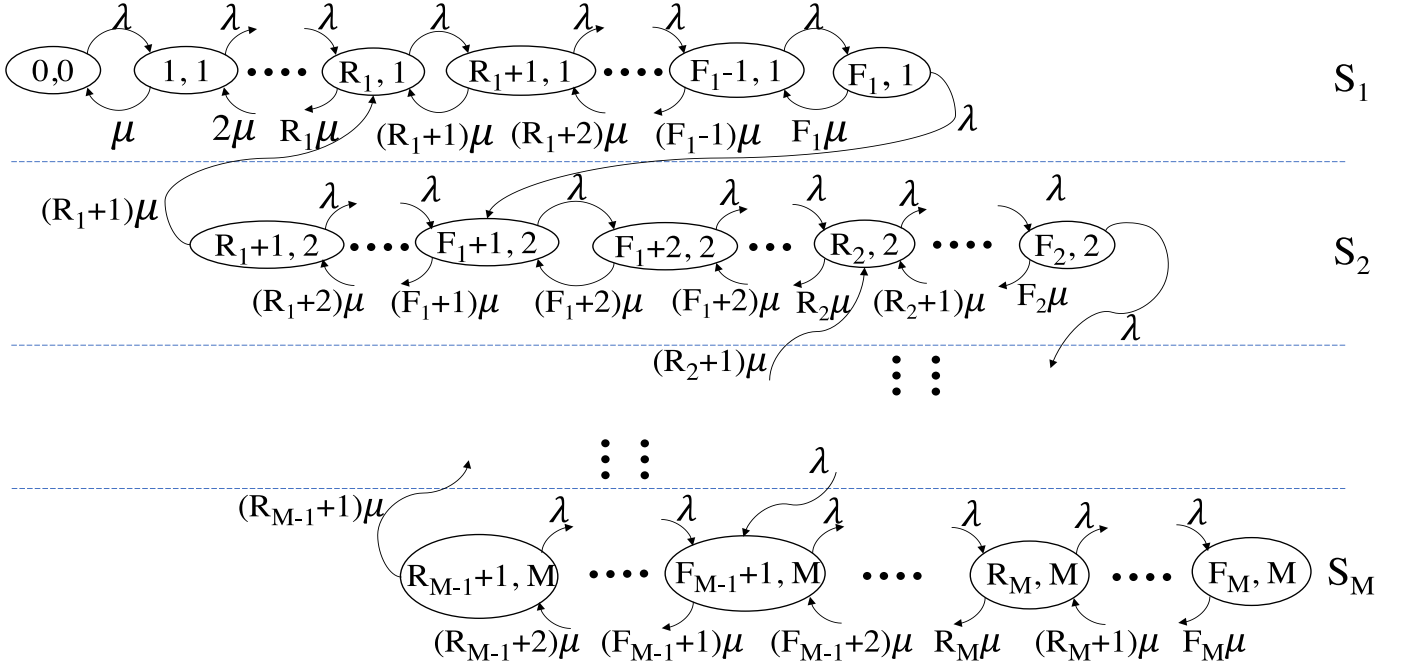
Fig. 3. State transition diagram of the RRU with partition

$\mathcal{L}$ and $\mathcal{R}$, respectively. The stochastic complement of $P_{\mathcal{L},\mathcal{L}}$ which is denoted by $C_{\mathcal{L},\mathcal{L}}$, is given by:

$$C_{\mathcal{L},\mathcal{L}} = P_{\mathcal{L},\mathcal{L}} + P_{\mathcal{L},\mathcal{R}}[I - P_{\mathcal{R},\mathcal{R}}]^{-1}P_{\mathcal{R},\mathcal{L}} \quad (3)$$

Let $\pi_{|\mathcal{L}}$ denote the steady-state probability vector corresponding to the states of $C_{\mathcal{L},\mathcal{L}}$. We can write $\pi_{|\mathcal{L}} = \frac{\pi_{\mathcal{L}}}{\pi_{\mathcal{L}}.e}$, where e is the column vector with all entries equal to 1. $\pi_{|\mathcal{L}}$ can be interpreted as the conditional state probabilities of the associated states of the original Markov chain $\mathcal{M}$.

We can rewrite (3) as

$$C_{\mathcal{L},\mathcal{L}} = P_{\mathcal{L},\mathcal{L}} + diag(P_{\mathcal{L},\mathcal{R}}e)Z \quad (4)$$

In (4), $diag(v)$ is a diagonal matrix whose ith diagonal element is the ith element of vector v and $Z = P^n_{\mathcal{L},\mathcal{R}}[I - P_{\mathcal{R},\mathcal{R}}]P_{\mathcal{R},\mathcal{L}}$. $P^n_{\mathcal{L},\mathcal{R}}$ is essentially the matrix $P_{\mathcal{L},\mathcal{R}}$ with all rows normalized. If r_i is the ith element of $P_{\mathcal{L},\mathcal{R}}e$, and z_i is the ith row of Z, then (4) can be re-written as:

$$C_{\mathcal{L},\mathcal{L}} = P_{\mathcal{L},\mathcal{L}} + \begin{bmatrix} r_1 z_1 \\ r_2 z_2 \\ \vdots \\ r_n z_n \end{bmatrix} \quad (5)$$

Expression in (5) can be interpreted as follows. Due to the partitioning of the original Markov process to the sub-processes $\mathcal{L}$ and $\mathcal{R}$, any transition from $\mathcal{L}$ to $\mathcal{R}$ in the original Markov process becomes a transition to some states in $\mathcal{L}$ instead (i.e., it folds back to itself). This process is well known as decoupling of Markov chain. Finding Z can be computationally intensive, although some special cases exist where Z can be easily computed. The following describes one such special case, which we use in our analysis.

Theorem (1): Let Q be the transition rate matrix of a given irreducible Markov process with state space S. If we partition the state space into two disjoint sets $\mathcal{L}$ and $\mathcal{R}$, then we can write

$$Q = \begin{bmatrix} Q_{\mathcal{L},\mathcal{L}} & Q_{\mathcal{L},\mathcal{R}} \\ Q_{R,\mathcal{L}} & Q_{\mathcal{R},\mathcal{R}} \end{bmatrix}$$

In the representation above, $Q_{j,k}$ is the transition rate sub-matrix corresponding to the transitions from partition-j to partition-k. If $Q_{\mathcal{R},\mathcal{L}}$ has all zero entries except for some non-zero entries in the ith column, then the conditional steady-state probability vector (corresponding to the states in $\mathcal{L}$), given that the system is in partition $\mathcal{L}$, is denoted by $\pi_{|\mathcal{L}}$ and is the solution to the following system of linear equations.

$$\pi_{|\mathcal{L}}[Q_{\mathcal{L},\mathcal{L}} + Q_{\mathcal{L},\mathcal{R}}ee_i^T] = \mathbf{0}, \qquad \pi_{|\mathcal{L}}e = 1$$

where e_i^T is a row vector with a 0 in each component except for a 1 in the ith component.

Proof: The proof of this theorem can be obtained by following the arguments of the stochastic complementation. We need to apply the simple transformation between a continuous time Markov chain with rate matrix Q and a discrete time Markov chain with probability matrix P which can be obtained via uniformization [22] as provided in (6). In this equation, $\zeta = \max\{|q_{ii}|\}$, where q_{ii} is the ith diagonal element of Q. For a more detailed proof of this theorem, the reader can refer to [18], [19].

$$P = I + Q/\zeta \quad (6)$$

B. Steady-state Probability Analysis of the RRU using Stochastic Complementation

In our work, we adopt the stochastic complementation method mentioned above to analyze the probability of each RRU using a particular CPRI rate and subsequently its transition rates. We consider the case where all RRUs belonging to the same cluster are characterized by similar arrival rate λ and service rate μ. When a UE call request is accepted, a user session is created at the BBU by allocating an RG to

the corresponding UE, over which it transmits and receives data. From this point, we will refer to this user session as 'server', following the terminology used in queuing theory. Therefore, the maximum number of servers '$\mathcal{K}$' per RRU is constant and is determined by the maximum number of RGs in the highest bandwidth configuration. Increments and decrements of the CPRI rates or bandwidth configuration are governed by the forward and reverse threshold vectors $F = [F_1, F_2, \ldots, F_{M-1}]$ and $R = [R_1, R_2, \ldots, R_{M-1}]$, respectively, where $R_i \leq F_i$.

We consider a homogeneous threshold-based queuing system with hysteresis. Let us construct a Markov process $\mathcal{M}$ with the following state space $\mathcal{S}$ according to our system model.

$$\mathcal{S} = \{(N_u, s) \mid N_u \geq 0, (s \mid d_s \in D)\}$$

Where N_u represents the number of users currently served by the given RRU and d_s is the CPRI rate associated with the current bandwidth configuration. Fig. 3 illustrates the transition diagram of the Markov process of each RRU. The horizontal lines in the figure are used to partition the whole state space $\mathcal{M}$ into different CPRI spaces. All states (i, j) in any partition S_j are states where the RRU uses the same CPRI rate d_j. Horizontal transitions, within the same space S_j are those where the arrival of a call does not trigger a transition to the next higher CPRI rate. This occurs until the number of users reaches the F_j value, after which a further call arrival triggers the transition to the next CPRI space S_{j+1}. A similar process occurs for the transition to lower rates, with the difference that these occur when the number of users, for a given CPRI state S_j, decreases below the value R_j.

The transition structure of the Markov process $\mathcal{M}$ can be mathematically expressed as follows:

$$
\begin{aligned}
(0,0) &\to (1,1) & &\lambda \\
(i,j) &\to (i+1,j) & &\lambda \cdot \mathbf{1}\Big\{ j \in \{1,2,\ldots,M\} \wedge \big((i \notin F) \\
& & &\qquad \vee \big((i = F_z \in F) \wedge (j \neq z)\big)\big)\Big\} \\
(i,j) &\to (i+1,j+1) & &\lambda \cdot \mathbf{1}\Big\{ j \in \{1,2,\ldots,M\} \\
& & &\qquad \wedge (i = F_z \in F) \wedge (j = z)\Big\} \\
(i,j) &\to (i-1,j) & &\mu \cdot \mathbf{1}\Big\{ (i \geq 1) \wedge \big((i,j) \neq (1,1)\big) \wedge \\
& & &\qquad (j \in \{1,\ldots,M\}) \wedge \big((i-1) \notin R\big) \\
& & &\qquad \vee \big((i-1 = R_z \in R) \wedge (j \neq z+1)\big)\Big\} \\
(i,j) &\to (i-1,j-1) & &\mu \cdot \mathbf{1}\Big\{ (j \in \{1,\ldots,M\}) \wedge \\
& & &\qquad \big((i-1) = R_z \in R\big) \wedge (j = z+1)\Big\} \\
(1,1) &\to (0,0) & &\mu
\end{aligned}
$$

$$(7)$$

In (7), $\mathbf{1}\{x\}$ is an indicator function such that $\mathbf{1}\{x\} = 1$ if the condition x is true and 0 if the condition is false. The operators $\wedge$ and $\vee$ represent logical "AND" and "OR", respectively.

Let us partition the state space $\mathcal{S}$ into M disjoint sets $\mathcal{S}_l$ $(l \in \{1, 2, \ldots M\})$, where

$$\mathcal{S}_l = \{(i,j) \mid (i,j) \in \mathcal{S} \text{ and } j = l\}$$

Therefore, $\mathcal{S}_l$ represents the state where the RRU is using CPRI rate d_l (or in bandwidth configuration B_l). For $2 \leq l \leq M-1$ we can order the states in $\mathcal{S}_l$ as follows:

$$\{(R_{l-1}+1, l), \ldots, (F_{l-1}+1, l), \ldots, (R_l, l), \ldots, (F_l, l)\}$$

We define another Markov process $\mathcal{M}_l$ for $l \in \{2, \ldots, M-1\}$ which corresponds to the state space $\mathcal{S}_l$. The transition structure of $\mathcal{M}_l$ resembles $\mathcal{M}$ for the states in $\mathcal{S}_l$ with the following adjustments (shown in Fig. 4).

1) A transition from $(R_{l-1}+1, l)$ to $(R_{l-1}, l-1)$ in the original process $\mathcal{M}$ is replaced by a transition from $(R_{l-1}+1, l)$ to $(F_{l-1}+1, l)$.
2) A transition from (F_l, l) to $(F_l+1, l+1)$ in the original process $\mathcal{M}$ is replaced by a transition from (F_l, l) to (R_l, l).

For $l = 1$, adjustment (1) simply does not apply (see Fig. 5) as there are no states to transition to from the state $(0,0)$ in the original Markov chain $\mathcal{M}$, because it is the initial state. Similarly, for $l = M$, adjustment (2) does not apply (see Fig. 6) as (F_M, M) is the terminal state of the original Markov chain $\mathcal{M}$.
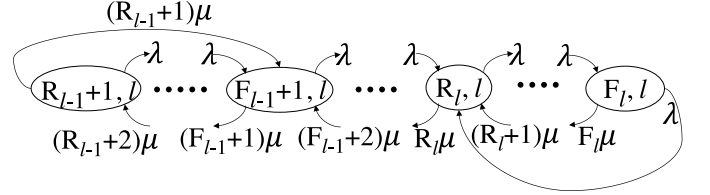

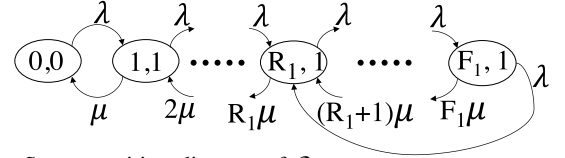Fig. 4. State transition diagram of $\mathcal{S}_l$ for $l \in \{2, \ldots, M-1\}$

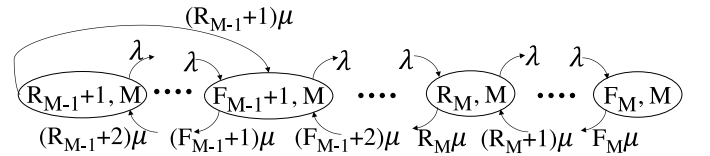Fig. 5. State transition diagram of $\mathcal{S}_1$

Fig. 6. State transition diagram of $\mathcal{S}_M$

Given that we construct the partitioned Markov process $\mathcal{M}_l$ according to the procedures mentioned above, we can prove the following:

The steady-state probabilities derived from Markov process $\mathcal{M}_l$ are the conditional steady-state probabilities for the states in $\mathcal{S}_l$ of the original Markov process $\mathcal{M}$, provided that the system is in partition S_l.

This statement can be proved using the argument of stochastic complementation together with the help of *Theorem (1)*. The reader can refer to [18] and [19] for a detailed proof.

Analysis of $\mathcal{M}_l$:

Let us now derive the steady-state probability vector for the states in $\mathcal{S}_l$, namely $\pi_l(n)$, where $l \in \{1, \ldots, M-1\}$. Based on the flow balance equation of $\mathcal{M}_l$ for $l \in \{2, \ldots, M-1\}$, we can express the steady-state probabilities $\pi_l(i)$ for $(R_{l-1}+1) \leq i \leq F_l$ in terms of $\pi_l(R_{l-1}+1)$ in the following form :

$$\pi_l(i) = C_i^l \pi_l(R_{i-1}+1) \quad \text{for} \quad l = 2, 3, \ldots, M \quad (8)$$

The unified expression of C_i^l for $(R_{l-1}+1) \le i \le F_l$ is provided in (9), where $\rho = \lambda/\mu$. This can be obtained as follows:

Considering the flow balance equation of $\mathcal{M}_l$ for $(R_{l-1}+1) \le i \le (F_{l-1}+1)$, we express $\pi_l(i)$ in terms of $\pi_l(R_{l-1}+1)$. Therefore, we obtain the expression of C_i^l as provided in (9a), for the corresponding sub-states of $\mathcal{M}_l$.

Next, we consider the flow balance equation of $\mathcal{M}_l$ for $(F_{l-1}+2) \le i \le R_l$ and express $\pi_l(i)$ in terms of $\pi_l(R_{l-1}+1)$. Thus, obtaining the expression of C_i^l for the corresponding sub-states of $\mathcal{M}_l$ as shown in (9b).

Similarly, based on the flow balance equation of $\mathcal{M}_l$ for $(R_l+1) \le i \le F_l - 1$, we get the expression of $\pi_l(i)$ in terms of $\pi_l(R_{l-1}+1)$ and $\pi_l(F_l)$. Therefore, we obtain the expression of C_i^l for the corresponding sub-states of $\mathcal{M}_l$ as provided in (9c). In (9c), the expression of $C_{F_l}^l$ can be found from the from the steady-state probability of equation: $\pi_l(F_l) = C_{F_l}^l \pi_l(R_{l-1}+1)$

Finally, considering the flow balance of $\mathcal{M}_l$ for $i = F_l$, we obtain the expression of $\pi_l(F_l)$ in terms of $\pi_l(R_{l-1}+1)$. Therefore, we obtain the expression of $C_{F_l}^l$ as shown in (9d).

We follow a similar process and apply it to Markov chains $\mathcal{M}_1$ and $\mathcal{M}_M$ to derive the steady-state probabilities for π_1 and π_M, shown in (10) and (11), respectively. The unified expression of C_i^l ($\forall l \in \{1, 2, \ldots, M\}$) is given in equations (9) - (11).

Using (8), together with the newly obtained equations for π_1 and π_M, we can express $\pi_l(i)$ for $(R_{l-1}+1) \le i \le F_l$, $\forall l \in \{1, 2, 3, \ldots, M\}$ ($R_0 = 0$) in the following form:

$$\pi_l(i) = C_i^l \pi_l(R_{i-1}+1) \quad \text{for} \quad l = 2, 3, \ldots, M \quad \text{and,}$$
$$\pi_1(i) = C_i^1 \pi_1(0) \tag{12}$$

where the expression for $\pi_l(R_{i-1}+1)$ can be obtained by summing over the state probabilities of Markov chain $\mathcal{M}_l$ and set it equal to 1.

$$\pi_l(R_{l-1}+1) = \Big[\sum_{i=0}^{F_l} C_i^l \Big]^{-1} \quad \text{for} \quad l = 1, 2, \ldots, M \tag{13}$$

Now, let us consider the aggregated process to calculate the transition rate between the CPRI states S_l. Fig. 2 shows the transition diagram of the resulting process. These are the transition rates of the RRU across the different CPRI rates. They can be computed as follows:

$$\lambda_i = \lambda \pi_i(F_i) \quad \text{for} \quad i = 1, 2, \ldots, M-1 \quad \text{and,}$$
$$\lambda_0 = \lambda \tag{14}$$

$$\mu_i = (R_{i-1}+1)\mu\pi_i(R_{i-1}+1) \quad \text{for} \quad i = 1, 2, \ldots, M \tag{15}$$

C. Steady-state Analysis of the Fronthaul Aggregator using Multidimensional Queuing Model

After we complete the computation of the state transition rates between different CPRI configurations, we need to analyze the steady-state probabilities at the aggregator process ($\mathcal{M}_A$). Let us consider a system with a cluster of N RRUs connected to an aggregator, where each RRU adopts a set of M CPRI rate configurations $(d_1, \ldots, d_M)$. The maximum

number of RRU that can be supported with this configuration is $N_{\text{RRU}}^{\max} = \lfloor B_c/d_1 \rfloor$. The reason is that even for a single active user in the RRU, a CPRI rate of d_1 will be adopted. This causes the aggregated capacity to surpass the FHA link capacity, beyond a cluster size of $N_{\text{RRU}}^{\max}$ RRUs. Therefore, we will always get blocking probability $P_b \approx 1$. Let k_m represents the fact that RRU k is using the CPRI rate configuration d_m. Therefore we can consider an M-dimensional vector $\mathbf{k} = (k_1, \ldots, k_m, \ldots, k_M)$ to represent any steady-state of the aggregator.

Depending on call arrival and departure at the RRU in the corresponding cluster, the following events can occur at the aggregator:

1) An RRU transitions to a higher CPRI rate due to a call arrival, thus creating a transition from $d_m \to d_{m+1}$ for that RRU with a transition rate λ_m (λ_m is the transition rate computed in Section III-B). This creates an aggregator event for a state transition from $(k_1, \ldots, k_m, k_{m+1}, \ldots, k_M)$ to $(k_1, \ldots, k_m-1, k_{m+1}+1, \ldots, k_M)$. A similar event occurs for a call departure, triggering an event at the aggregator with rate μ_m.

2) A call request arrives to an RRU belonging to a cluster which had no previous user, thus creating a transition from $(d_0 = 0) \to d_1$. This is also a wake-up RRU transition with rate λ. This creates an aggregator event for a state transition from $(k_1, \ldots, k_M)$ to $(k_1+1, \ldots, k_M)$. A similar event occurs for a call departure, triggering an event at the aggregator with a rate μ_1.

An incoming call request triggering any of these events is accepted if the new aggregated capacity, which accounts for the increase of the CPRI rate, is less than or equal to the FHA link capacity. Therefore, we get the following possible states for the aggregator:

$$\mathbb{K} = \{\mathbf{k} \mid k_1, k_2, \ldots k_M \ge 0, \sum_{i=1}^{M} k_i \le N \text{ and } \sum_{i=1}^{M} d_i k_i \le B_c\} \tag{16}$$

As there can only be one possible event at any instant of time, either due to call arrival or departure at an RRU, only a single entry of $\mathbf{k}$ can change at any epoch. Therefore, we can consider the aggregator as a multi-dimensional birth and death process. In the following text, we provide the mathematical description of the transition from the ith to the jth state of the aggregator process $\mathcal{M}_A$.

$$Q_{\mathbf{k}^{(i)}\mathbf{k}^{(j)}}$$
$$= \mathbf{k}_m^{(i)} \lambda_m \qquad \text{if} \quad \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = e_m^{(1)}, \ \mathbf{k}_m^{(i)} \ne 0$$
$$= \Big(N - \sum_{m=1}^{M} \mathbf{k}_m^{(i)}\Big)\lambda \quad \text{if} \ \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = e_1^{(2)}, \ \sum_{m=1}^{M} \mathbf{k}_m < N,$$
$$\text{and} \ N \le N_{\text{RRU}}^{\max}$$
$$= \mathbf{k}_m^{(i)} \mu_m \qquad \text{if} \quad \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = -e_m^{(1)}, \ \mathbf{k}_m^{(i)} \ne 0$$
$$= \mathbf{k}_1^{(i)} \mu_1 \qquad \text{if} \quad \mathbf{k}^{(j)} - \mathbf{k}^{(i)} = -e_1^{(2)}, \ \mathbf{k}_1^{(i)} \ne 0$$
$$= 0 \qquad \qquad \text{Otherwise} \tag{17}$$

$$C_i^l = \left[\frac{(R_{l-1}+1)!}{i!}\right]^{i-(R_{l-1}+1)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{R_{l-1}!}\right\} \qquad \text{for } (R_{l-1}+1) \le i \le (F_{l-1}+1) \tag{9a}$$

$$= \left[\frac{(R_{l-1}+1)!}{i!}\right]^{i-(R_{l-1}+1)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{R_{l-1}!}\right\} - (R_{l-1}+1)\left[\frac{(F_{l-1}+1)!}{i!}\right]^{i-(F_{l-1}+2)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{(F_{l-1}+1)!}\right\}$$
$$\text{for } (F_{l-1}+2) \le i \le R_l \tag{9b}$$

$$= \left[\frac{(R_{l-1}+1)!}{i!}\right]^{i-(R_{l-1}+1)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{R_{l-1}!}\right\} - (R_{l-1}+1)\left[\frac{(F_{l-1}+1)!}{i!}\right]^{i-(F_{l-1}+2)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{(F_{l-1}+1)!}\right\}$$
$$- \left[\frac{R_l!}{i!}\right] \sum_{j=1}^{i-R_l} \rho^j \left\{\frac{(i-j)!}{R_l!}\right\} C_{F_l}^l \qquad \text{for } (R_l+1) \le i \le (F_l-1) \tag{9c}$$

where,

$$C_{F_l}^l = \left[\left\{1+\frac{F_l}{\rho}\right\} + \frac{R_l!}{(F_l-1)!}\sum_{j=1}^{(F_l-1)-R_l} \rho^j \left\{\frac{((F_l-1)-j)!}{R_l!}\right\}\right]^{-1} \left[\frac{(R_{l-1}+1)!}{(F_l-1)!}\right]^{(F_l-1)-(R_{l-1}+1)} \sum_{j=0} \rho^j \left\{\frac{((F_l-1)-j-1)!}{R_{l-1}!}\right\} -$$
$$(R_{l-1}+1)\left[\frac{(F_{l-1}+1)!}{(F_l-1)!}\right]^{(F_l-1)-(F_{l-1}+2)} \sum_{j=0} \rho^j \left\{\frac{((F_l-1)-j-1)!}{(F_{l-1}+1)!}\right\}\right] \tag{9d}$$

$$\forall \quad l \in \{2,3,\ldots,M-1\}$$

$$C_i^1 = \frac{\rho^i}{i!} \qquad \text{for } 1 \le i \le R_1 \tag{10a}$$

$$= \frac{\rho^i}{i!} - \left[\frac{R_1!}{i!}\right] \sum_{j=1}^{i-R_1} \rho^j \left\{\frac{(i-j)!}{R_1!}\right\} C_{F_1}^1 \qquad \text{for } (R_1+1) \le i \le (F_1-1) \tag{10b}$$

$$\text{where,} \quad C_{F_1}^1 = \left[\left\{1+\frac{F_1}{\rho}\right\} + \frac{R_1!}{(F_1-1)!}\sum_{j=1}^{(F_1-1)-R_1} \rho^j \left\{\frac{((F_1-1)-j)!}{R_1!}\right\}\right]^{-1} \left[\frac{\rho^{F_1-1}}{(F_1-1)!}\right] \tag{10c}$$

$$C_i^M = \left[\frac{(R_{M-1}+1)!}{i!}\right]^{i-(R_{M-1}+1)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{R_{M-1}!}\right\} \qquad \text{for } (R_{M-1}+1) \le i \le (F_{M-1}+1) \tag{11a}$$

$$= \left[\frac{(R_{M-1}+1)!}{i!}\right]^{i-(R_{M-1}+1)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{R_{M-1}!}\right\} - (R_{M-1}+1)\left[\frac{(F_{M-1}+1)!}{i!}\right]^{i-(F_{M-1}+2)} \sum_{j=0} \rho^j \left\{\frac{(i-j-1)!}{(F_{M-1}+1)!}\right\}$$
$$\text{for } (F_{M-1}+2) \le i \le F_M \tag{11b}$$

where states $\mathbf{k}^{(i)}, \mathbf{k}^{(j)} \in \mathbb{K}$, $\mathbf{k}_m^{(i)}$ is the mth entry of $\mathbf{k}^{(i)}$.

$$e_m^{(1)^T} = \{0,0,\ldots,\overset{m^{\text{th}}}{-1},\overset{(m+1)^{\text{th}}}{1},\ldots,0,0\} \qquad \text{and}$$
$$e_1^{(2)^T} = \{1,0,\ldots,0,0\}$$

An example of the aggregator process is shown in Fig. 7.

Analysis of $\mathcal{M}_A$:

It can be shown that $\mathbf{k}$ satisfies the reversibility property by following the method described in [23]. We use this property to derive the steady-state probabilities at the aggregator and the obtained expression is given in (18).
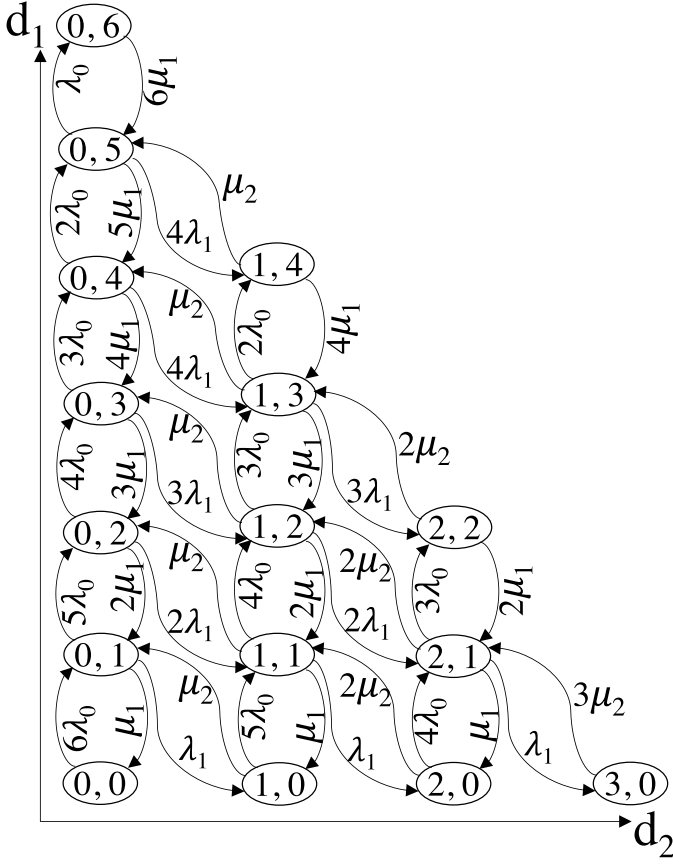
$$P(k_1,k_2,\ldots,k_M) =$$

$$P(0,0,\ldots,0)\left[\binom{N_{\text{RRU}}}{K_s}\frac{K_s!}{\prod_{i=1}^M k_i!}\prod_{i=1}^M \left(\frac{\lambda_{i-1}}{\mu_i}\right)^{\sum_{j=i}^M k_j}\right] \tag{18}$$

In (18), $N_{\text{RRU}} = N$, for the cluster size $N \le N_{\text{RRU}}^{\text{max}}$ and $N_{\text{RRU}} = N_{\text{RRU}}^{\text{max}}$, for the cluster size $N > N_{\text{RRU}}^{\text{max}}$. The reader is invited to see APPENDIX A for the derivation of (18). We can derive the expression for $P(0,0,\ldots,0)$ as

$$P(0,0,\ldots,0) =$$

$$\left[\sum_{\mathbf{k}\in\mathbb{K}}\left[\binom{N_{\text{RRU}}}{K_s}\frac{K_s!}{\prod_{i=1}^M k_i!}\prod_{i=1}^M \left(\frac{\lambda_{i-1}}{\mu_i}\right)^{\sum_{j=i}^M k_j}\right]\right]^{-1} \tag{19}$$

Fig. 7. An example of the aggregator process with each RRU using two CPRI rate configurations (d_1 and d_2), assuming $d_2 = 2d_1$, FHA link capacity $= 6d_1$ and a N-RRU cluster connected to the aggregator ($N = 6$ for this example).

by using the fact that

$$\sum_{\mathbf{k} \in \mathbb{K}} P(k_1, k_2, \ldots, k_m, \ldots, k_M) = 1$$

Therefore, the steady state probabilities of the aggregator can be obtained from (18). In this expression, the transition rates between different CPRI configurations λ_i and μ_i can be obtained from (12) - (15).

D. Blocking Probabilities at the Fronthaul Aggregator

With the expression of steady-state probabilities derived in the previous section, we can evaluate the blocking probability at the aggregator. In order to accomplish this, we can decompose the calculation of blocking probability into M-parts with M being the number of CPRI rates (or bandwidth configurations) available at the RRUs belonging to a cluster. Let $P_B^{\lambda_m}$ be the probability that the aggregator goes into a blocking state due to a transition of CPRI rate from d_m to d_{m+1}. We take $d_0 = 0$ and $\lambda_0 = \lambda$ as we have described

previously. Then $P_B^{\lambda_m}$ can be determined as follows:

$$P_B^{\lambda_m} = \frac{k_m \lambda_m \sum_{\mathbf{k} \in \mathbb{K}^{\lambda_m}} P(\mathbf{k})}{\sum_{i=0}^{M-1} \lambda_m} \qquad \text{for } i = 1, 2, \ldots, M-1$$

and

$$P_B^{\lambda_0} = \frac{\left(N_{\text{RRU}} - \sum_{i=1}^{M} k_i\right) \lambda \sum_{\mathbf{k} \in \mathbb{K}^{\lambda_0}} P(\mathbf{k})}{\sum_{i=0}^{M-1} \lambda_m}$$

where,

$$\mathbb{K}^{\lambda_m} = \{\mathbf{k} \mid \quad \mathbf{k} \in \mathbb{K}, k_m \neq 0, \text{and} \sum_{i=1}^{M}(k_i + e_{m\_i}^{(1)})d_i > B_c\},$$

$$\mathbb{K}^{\lambda_0} = \{\mathbf{k} \mid \quad \mathbf{k} \in \mathbb{K}, \sum_{i=1}^{M} k_i < N, \text{and} \sum_{i=1}^{M}(k_i + e_{1\_i}^{(2)})d_i\},$$

$e_{m\_i}^{(1)}$ is the ith entry of $e_m^{(1)}$ and $e_{1\_i}^{(2)}$ is the ith entry of $e_1^{(2)}$.

Finally, the overall blocking probability at the aggregator can be found as

$$P_B = \sum_{m=0}^{M-1} P_B^{\lambda_m} \qquad (21)$$

IV. SIMULATION MODEL

In addition to providing an analytical expression for the blocking probability at the fronthaul aggregator, we have also simulated the above model in Matlab using the discrete-event-blocks called 'Sim Events'. The model is shown in Fig. 8. For each RRU belonging to an RRU-cluster, the call arrival and service is simulated as an $M/M/S(0)$ loss system queuing process. In this work, we have taken the number of servers S as 50, as shown in table I, which corresponds to the number of RGs available at the 20 MHz LTE bandwidth configuration. We use an event-based function call module to translate the current state of the RRU to its corresponding fronthaul rate. In this simulation, an incoming call is dropped whenever the sum of all fronthaul rates at the aggregator is higher than the aggregator capacity. In the case of an XG-PON system, this is equivalent to 10 Gbps. The events in the aggregator are triggered by the events in any RRU. Therefore, an incoming call at any RRU can be blocked for two reasons:

1) If there is no idle server available at the call arrival instance to serve the incoming request, i.e., all 50 servers are busy. This means no RG is available to serve the call, which is thus blocked and dropped by the RRU.
2) If due to a call arrival, a higher bandwidth configuration needs to be adopted but there is not enough capacity available in the FHA link. Therefore the call request is blocked due to unavailable bandwidth at the FHA link.

Tables II and III provide the parameters used in our simulations. If the holding time of a service session (or call duration of an accepted call) is τ time units, the service rate
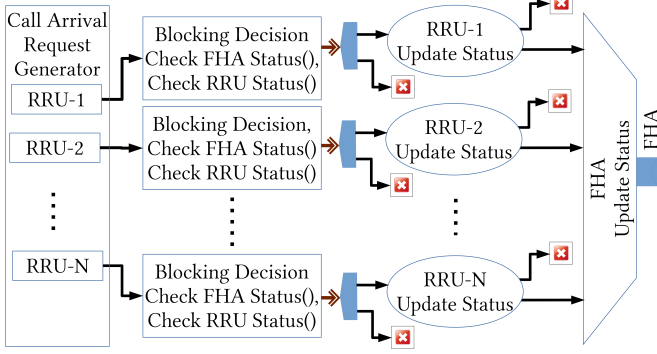
(20)

Fig. 8. Model of the variable rate fronthaul simulator implemented in MATLAB

(μ) corresponding to each service session of a particular RRU is: $\frac{1}{\tau}$ (calls/time units). If the maximum number of service sessions per RRU is '$\mathcal{K}$' then the condition for the call arrival rate (λ) in each RRU so that the system maintains a steady state blocking probability is $\frac{\lambda}{\mathcal{K}\mu} < 1$ [24]. The traffic load (in Erlang) for each RRU can be represented as $\rho = \lambda/\mu$. Hence, we can say the traffic load can essentially go up to $\mathcal{K}$ Erlang. Therefore, we denote $a = \frac{\lambda}{\mathcal{K}\mu}$ in order to obtain a normalized traffic load across RRUs. We use different values of 'a' to control the traffic load per RRU.

V. RESULTS

Here we report and compare the results of both our theoretical and simulation analysis of our variable rate fronthaul system, showing the blocking probability for a number of different configurations. We run the event-driven simulations for approximately 10^7 events for each specific system configuration (i.e., for a combination of M, N and a) and capture the blocking probability.

Fig. 9 to 11 illustrate the blocking probability (P_b) w.r.t the RRU cluster size for traffic intensity ('a') of 0.2, 0.3, and 0.5, respectively. In these figures, N_d represents the number of different data rates that can be used for the fronthaul transport (i.e., the parameter M, as defined in the theoretical analysis section). Therefore, $N_d = 1$ corresponds to the traditional fronthaul scheme. These figures also demonstrate the improvement of blocking probability when our proposed variable rate fronthaul scheme is used over traditional fronthaul (i.e., for $N_d = 1$). In these results, we have taken the forward thresholds (F_l) as the maximum number of supported users corresponding to the LTE bandwidth / CPRI rate configuration (as listed in Table I) and the reverse thresholds (R_l) as one step less than the F_l. Therefore, the difference between the forward and reverse threshold is equal to 1.

Firstly, we can see a close match between the analytical and simulated results, which corroborates the validity of our analysis. We do notice a slight discrepancy at the low end of the blocking probability ($\approx 10^{-4} - 10^{-5}$), which we believe is a statistical deviation due to the low number of instances where blocking occurred. In Fig. 9, it can be observed that for the traditional fronthaul scheme the blocking probability is zero up to a cluster of size 8. After that, it sharply increases

to 1. This is because, in traditional fronthaul ($N_d = 1$), each of the RRU adopts a static rate of 1228.8 Mbps even for a single active user in the system, causing the aggregated rate to surpass FHA link capacity of 10 Gbps. In addition, for a cluster size of 9, we notice that P_b is slightly less than 1 ($\approx$ 0.9). The reason for this is that at moderate traffic intensity ($a = 0.2$), it is relatively likely that there are no users in at least one RRU at any time instant. In that case, we have assumed that the RRU does not send any data over the fronthaul link. Thus, only 8 RRUs would be active, which is a non-blocking situation.
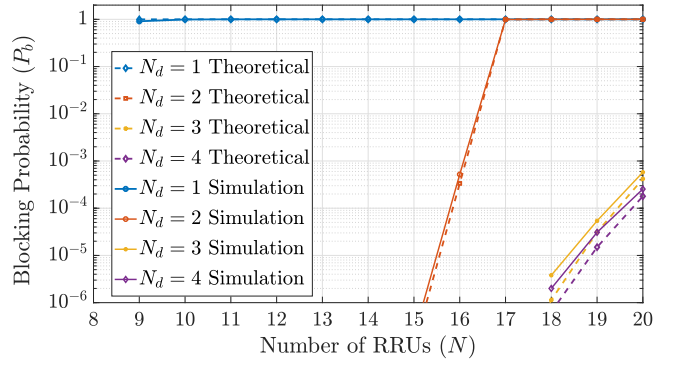

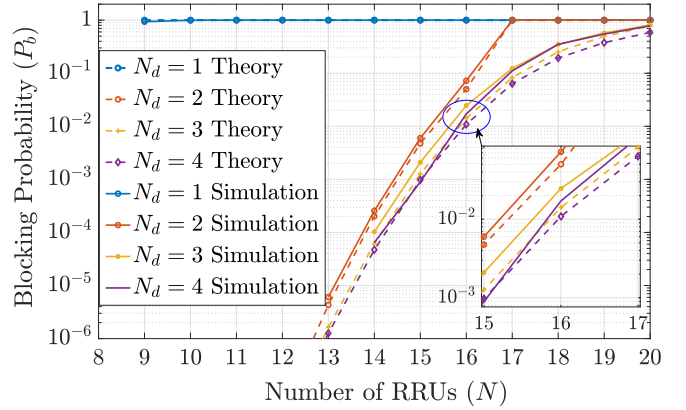Fig. 9. Blocking probability (P_b) vs. number of RRUs for $a = 0.2$

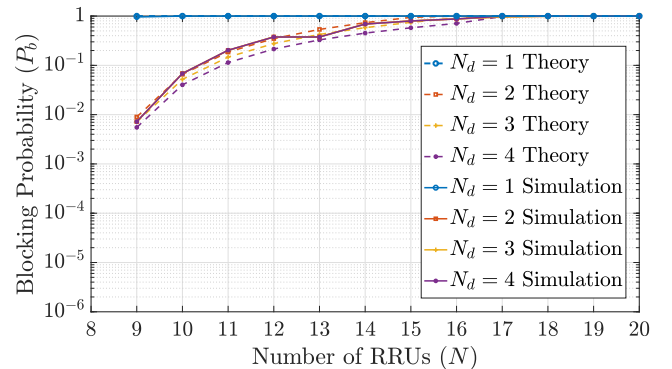Fig. 10. Blocking probability (P_b) vs. number of RRUs for $a = 0.3$

Fig. 11. Blocking probability (P_b) vs. number of RRUs for $a = 0.5$

TABLE II
FRONTHAUL RATES USED FOR DIFFERENT VBF CONFIGURATION

$N_d = 1$	$N_d = 2$	$N_d = 3$	$N_d = 4$
1228.8 Mbps	{614.4, 1228.8} Mbps	{307.2, 614.4, 1228.8} Mbps	{153.6, 307.2, 614.4, 1228.8} Mbps

TABLE III
SIMULATION PARAMETERS

Service Rate (per service session)	0.5
Max no. of service sessions per RRU	50
No. of available fronthaul rates (N_d)	1,2,3,4
Capacity of FHA link	10 Gbps
Max no. of RRU per Aggregator	20
Number of servers in Aggregator	20

Fig. 12. Blocking probability (P_b) *vs.* number of data rates for $a = 0.25$

Secondly, it is clear from the figures that our proposed variable rate fronthaul scheme achieves noticeable lower blocking probability compared to the static CPRI case. Indeed for values of $a = 0.2$, we can see that a blocking probability below 10^{-4} can be achieved even when 15 cells are aggregated (i.e., almost the double of using traditional fronthaul). The number of cells can be increased to 18 if four different rates are employed. However, it should also be noticed from Fig. 10 that we obtain almost the same performance for $N_d = 3$ and $N_d = 4$ for $a = 0.3$ (the difference is highlighted in the small square). This is due to the fact that in our system the different CPRI rates are considered in descending order and, for example, the $N_d = 4$ configuration only adds the low 153.6 Mbps data rate with respect to $N_d = 3$, which has a small impact on the aggregated rate. Furthermore, the maximum number of end users supported at the 153.6 Mbps rate is only 6, meaning that if more than six calls arrive at the RRU, this will move to a higher data rate. Thus, the probability for the RRU to remain in this lowest rate is significantly small. The difference can only be noticed when either the call arrival rate is very low or the size of the RRU cluster is large. As we increase the traffic intensity, for $a = 0.5$ (and for rates above $a = 0.5$, which are not shown here), there is very little difference in performance for $N_d = 2$, $N_d = 3$ and $N_d = 4$. This is illustrated in Fig. 11 and happens because at this high call arrival rate, higher rate configurations (614 Mbps and 1228.8 Mbps) are adopted most of the time.

In Fig. 12, we use our results to determine the maximum cluster size for which a given blocking probability can be achieved, for $a = 0.25$. For example, a blocking probability of 10^{-3} can be attained with a cluster size of up to 17 RRUs, if we use a VRF configuration with three CPRI rates. However, if we only use two CPRI rates, then the maximum cluster size is reduced to 15.

The results shown up to this point have considered a difference between the forward (F_l) and reverse threshold (R_l) of 1. Fig. 13 shows how the blocking probability increases when we increase the difference between these thresholds. This result considers normalized traffic load as $a = 0.2$, where each RRU employs three different CPRI rates ($N_d = 3$). Indeed, as we increase such difference ($F_l - R_l$), a given CPRI state is retained for a longer period, till a higher number of call departs from the system (i.e., with respect to the case where $F_l - R_l = 1$). This result i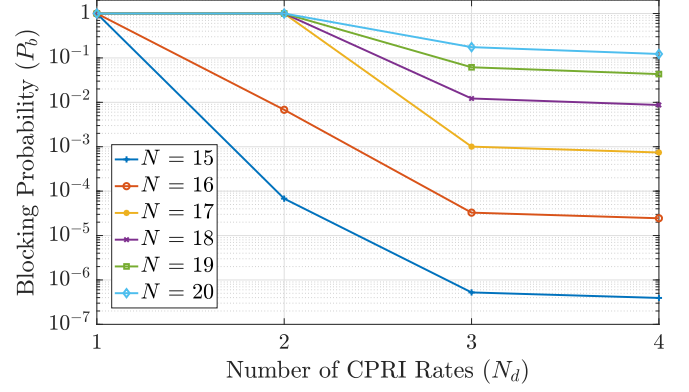s important when considering the finite amount of time required for a mobile system to transition between different CPRI rates.

The difference between the forward and reverse threshold helps to prevent hysteresis i.e., to prevent looping between two CPRI configurations. Furthermore, it also takes care of the latency encountered by the SDN controller to configure between different CPRI rates. For example, if the latency for reconfiguration between different CPRI rates is somewhere between 500ms (this is a value we experimented in our dynamic-bandwidth SDR testbed implementation [16]) to 5seconds (this is the average value taken to completely reboot the SDR BBU) and if we consider normalized traffic load (for example $a = 0.2$) per RRU then we can write,

$a = \lambda / \mathcal{K}\mu = 0.2$
$\implies \lambda = 0.2 \times \mathcal{K} \times \mu = 10\mu$ ($\mathcal{K} = 50 \rightarrow$ maximum number of users supported)

Now, if we consider a moderate traffic scenario, with $\lambda = 10$ calls/min (e.g., in terms of number of users joining the cell) and $\mu = 1$ calls/min, this means that one call, two calls and three calls are expected to arrive, within the reconfiguration window (0.5 seconds), respectively, with probabilities $(p) = 0.0767, 0.0032$ and 8.8739×10^{-5}. Therefore, if we choose $F_l - R_l < 2$ (say for example $F_l - R_l = 1$), then while the SDN controller is triggering a transition to a lower CPRI rate, it is highly probable that one or more calls might arrive. This immediately triggers a transition to a higher CPRI rate which makes the system unstable. Therefore $F_l - R_l \geq 2$ is a good choice to efficiently address the hysteresis (keeps it within 1 percent) and SDN controller configuration timings. However, if instead of implementing a dynamic reconfiguration mechanism, the BBU needs a reboot, with a reconfiguration timing window of 5 seconds, then using the same argument we see that for $a = 0.2$, $\lambda = 10$ and $\mu = 1$, the probability of arrival of two, three and four call requests is $0.1509, 0.0419$ and 0.0087 respectively. Therefore a choice of $F_l - R_l \geq 3$ efficiently keeps the hysteresis within 1 percent for this case.

It should be noted that the choice of the parameter $F_l - R_l$ only depends on the arrival and departure rates considered in the system. Furthermore, given an average traffic load with moderate intensity, if we let the reconfiguration window to be higher, then the system will be more prone to stay in the higher CPRI rate and the lower rates will be used fewer times, thereby reducing the advantage of using more CPRI configurations (N_d) in VRF. However, as the traffic load increases, the lower rates would be used less frequently in any case (as shown in Fig. 9-11), therefore reducing the effect of the reconfiguration window on the system performance.

Finally, while in the future we expect C-RAN software-based BBUs to all have similar reconfiguration times to the values above, today some hardware-based BBU might have notably longer reconfiguration times, in which case the rate should be modified less frequently, based on predicted average traffic.
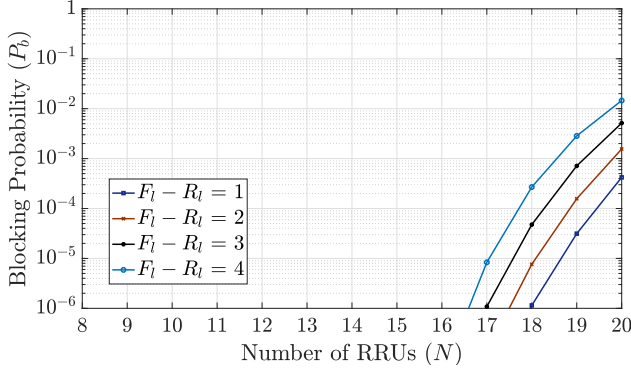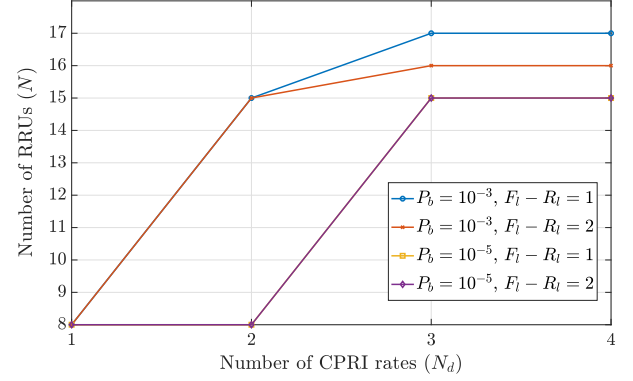

Fig. 14. Maximum number of RRUs that can be aggregated for different VRF configuration under certain Grade of Service ($P_b = 10^{-3}, 10^{-5}$) requirement, normalized traffic load ($a = 0.25$), and different choice of forward and reverse threshold difference ($F_l - R_l = 1, 2$).
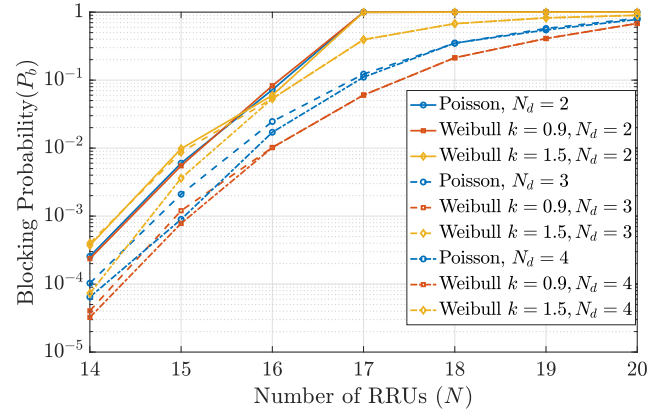

Fig. 13. Blocking probability (P_b) vs. number of RRUs for various differences between forward and reverse thresholds ($F_l - R_l$ =1,2,3,4), for $a = 0.2$ and $N_d = 3$

Fig. 15. Performance comparison between Poisson and Weibull Distribution for $a = 0.3$, $N_d = 2, 3, 4$, $F_l - R_l = 1$

Fig. 14 illustrates a typical deployment scenario that could be obtained summarizing the results from Fig. 9 to Fig. 13. This result provides the number of RRUs that can be aggregated for different VRF configurations ($N_d = 1, 2, 3, 4$) under the requirement of a certain Grade of Service (GoS) (or blocking probability ($P_b = 10^{-3}, 10^{-5}$)) while considering a given normalized traffic load ($a = 0.25$) and a choice of the forward and reverse threshold difference ($F_l - R_l = 1, 2$). For $N_d = 1$, the system operates over traditional CPRI, so different values of $F_l - R_l$ don't make any difference. For $N_d = 2$ however, there is a difference between $F_l - R_l = 1$ and $F_l - R_l = 2$ for $P_b = 10^{-5}$.

In our study, the Poisson arrival is linked to the increase of PRBs as more users join the network, which can be described as a Poisson process. However, it is true that additional PRBs could be allocated to the same user, when its requested capacity increases thus making the traffic non-Poisson in nature. The theoretical analysis for these additional cases could not be carried out because for non-Poisson distribution the system cannot be decoupled via partitions. In this paper, we carried out simulation using two non-Poisson traffic namely, Weibull arrival process with shape factor (k)= 0.9 and 1.5 respectively. A summary of this distribution is provided in APPENDIX B

Figure 15 shows comparative result of Weibull arrival process with two shape factors ($k = 0.9$ and $k = 1.5$) against already discussed results using Poisson arrivals under the normalized traffic load $a = 0.3$. We notice that our results with Poisson arrival process lies between Weibull process with $k = 1.5$ and $k = 0.9$. The reason for this is that the intensity of the arrival process changes according to the shape factor. For example, if the average inter-arrival time for our original exponential distribution is $1/\lambda$ (or the arrival rate for our original Poisson distribution is λ), then the inter-arrival time for Weibull arrival process is $\Gamma(1+1/k)/\lambda = 1.0522/\lambda$ for the shape factor (k)=0.9. Therefore, we see that the inter arrival time gets larger which implies a decrease in arrival rate. Thus a Weibull distribution with $k = 0.9$ yields a lower blocking probability compared to our Poisson arrival case. The same argument can be used to explain the reason for $k = 1.5$ having a higher blocking probability when compared to our Poisson arrival case. In addition, as the average rate of arrival increases for higher values of a (which is not shown here), the difference becomes less pronounced as higher capacity will move the RRU states towards the higher bandwidth, thus masking the difference in distribution.

VI. Conclusions

In this paper, we have introduced the concept of variable rate fronthaul for cloud-RANs. After providing the description of the network architecture, we have formulated a mathematical description of our model. We have used Queuing Theory with a two-phase approach to solve the model and obtain an analytical form for blocking probability at the fronthaul aggregator. We have then performed simulation of the model using Matlab's discrete event simulator, and the results were compared with those obtained from our analytical model.

Besides showing a close match between analytical and simulation models, our results prove that by dynamically varying the cell's bandwidth, according to the actual end user demand, we can achieve a more efficient fronthaul transport of a group of C-RAN cells, without increasing the complexity, cost and energy consumption of the RRUs. This is especially relevant for next generation of high-density cell deployment, as multiplexing several cells, can sensibly lower fronthaul costs.

Acknowledgment

Appendix A

Since $\mathbf{k}$ is reversible, the local balance equation

$$P(\mathbf{k}^{(i)}).Q_{\mathbf{k}^{(i)}\mathbf{k}^{(j)}} = P(\mathbf{k}^{(j)}).Q_{\mathbf{k}^{(j)}\mathbf{k}^{(i)}} \tag{22}$$

holds at the steady state. Without loss of generality, let

$$\mathbf{k}^{(i)} = \{k_1, \ldots, k_m, k_{m+1}, \ldots, k_M\}^T,$$

$$\mathbf{k}^{(j)} = \{k_1, \ldots, k_m - 1, k_{m+1} + 1, \ldots, k_M\}^T$$

substituting (17) into (22), we get

$$P(k_1, \ldots, k_m, k_{m+1}, \ldots, k_M)k_m\lambda_m = \\ P(k_1, \ldots, k_m - 1, k_{m+1} + 1, \ldots, k_M)(k_m + 1)\mu_{m+1} \tag{23}$$

Expression in (23) can be re-written as

$$\frac{P(k_1, \ldots, k_m - 1, k_{m+1} + 1, \ldots, k_M)}{P(k_1, \ldots, k_m, k_{m+1}, \ldots, k_M)} = \frac{k_m}{(k_m + 1)}\frac{\lambda_m}{\mu_{m+1}} \tag{24}$$

After little manipulation of (24), we obtain (25).

$$\frac{P(k_1, \ldots, k_{m-1}, k_m, \ldots, k_M)}{P(k_1, \ldots, k_{m-1} + 1, k_m - 1, \ldots, k_M)} = \frac{k_{m-1} + 1}{k_m}\frac{\lambda_{m-1}}{\mu_m} \tag{25}$$

Clearly, (25) is iterative, therefore we can iterate this equation to obtain the following:

$$P(k_1, \ldots, k_{m-1}, k_m, \ldots, k_M) = \\ P(k_1, \ldots, k_{m-1} + k_m, 0, \ldots, k_M)\left\{\frac{\lambda_{m-1}}{\mu_m}\right\}^{k_m}\left\{\frac{(k_{m-1} + k_m)!}{k_{m-1}!\ k_m!}\right\} \tag{26}$$

Using (26), we can write:

$$P(k_1, \ldots, k_{M-1}, k_M) = \\ P(k_1, \ldots, k_{M-1} + k_M, 0)\left\{\frac{\lambda_{M-1}}{\mu_M}\right\}^{k_M}\left\{\frac{(k_{M-1} + k_M)!}{k_{M-1}!\ k_M!}\right\} \tag{27}$$

Then, starting with (27) and iterating over all the entries except the first position we obtain:

$$P(k_1, k_2, \ldots, k_M) = \\ P(K_s, 0, 0, \ldots, 0)\left[\frac{K_s!}{\prod_{i=1}^{M}k_i!}\prod_{i=2}^{M}\left(\frac{\lambda_{i-1}}{\mu_i}\right)^{\sum_{j=i}^{M}k_j}\right] \tag{28}$$

In (28), $K_s = \sum_{i=1}^{M}k_i$. Now, with the help of (17), we can write the following flow balance equation for $N \leq N_{\text{RRU}}^{\max}$:

$$P(K, 0, \ldots, 0)(N - K)\lambda = P(K + 1, 0, \ldots, 0)(K + 1)\mu_1 \tag{29}$$

Expression in (29) can be simplified using the same process followed in (26) to obtain the following expression:

$$P(K, 0, \ldots, 0) = \binom{N}{K}\left(\frac{\lambda}{\mu_1}\right)^{K}P(0, 0, \ldots, 0) \tag{30}$$

Substituting (30) into (28) and after some manipulation, we get the following final deduction:

$$P(k_1, k_2, \ldots, k_M) = \\ P(0, 0, \ldots, 0)\left[\binom{N}{K_s}\frac{K_s!}{\prod_{i=1}^{M}k_i!}\prod_{i=1}^{M}\left(\frac{\lambda_{i-1}}{\mu_i}\right)^{\sum_{j=i}^{M}k_j}\right] \tag{31}$$

Following the similar procedure, as discussed obtain (29) to (31), we can obtain the expression for $N > N_{\text{RRU}}^{\max}$ as provided in (32)

$$P(k_1, k_2, \ldots, k_M) = \\ P(0, 0, \ldots, 0)\left[\binom{N_{\text{RRU}}^{\max}}{K_s}\frac{K_s!}{\prod_{i=1}^{M}k_i!}\prod_{i=1}^{M}\left(\frac{\lambda_{i-1}}{\mu_i}\right)^{\sum_{j=i}^{M}k_j}\right] \tag{32}$$

Appendix B

The additional simulations are based on the use of a Weibull distribution, which generalizes a large class of distributions (exponential, Rayleigh, chi-squared etc.) depending on the shape factor k. For example, for $k = 1$ we obtain an exponential distribution (Poisson arrival process = exponentially distributed inter-arrival times). Equation-(33) provides the details of an Weibull distribution with mean $= \gamma\Gamma(1 + 1/k)$ and variance $= \gamma^2\left[\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2\right]$. Figure 16 plots the distribution under different shape factor (k).

$$f(x, \gamma, k) = \begin{cases} \frac{k}{\gamma}\left(\frac{x}{\gamma}\right)^{k-1}e^{-(x/\gamma)^k} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{33}$$

References

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016–2021," 2016. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-whitepaper-c11-481360.html

[2] P. Loskot, M. A. M. Hassanien, F. Farjady, M. Ruffini, and D. Payne, "Long-term drivers of broadband traffic in next-generation networks," *annals of telecommunications - annales des télécommunications*, vol. 70, no. 1, pp. 1–10, Feb 2015. [Online]. Available: https://doi.org/10.1007/s12243-014-0424-9
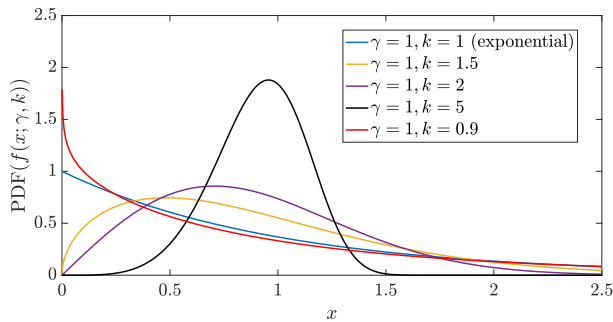
Fig. 16. PDF of Weibull distribution with different shape factors

[3] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks -A Technology Overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[4] NGMN Alliance, "NGMN Overview on 5G RAN Functional Decomposition," Feb. 2018.

[5] NEC Corp. and Nortel Networks SA and Siemens Networks GmbH & Co. KG and Ericsson AB and Huawei Technologies Co Ltd., "V6.1 Common Public Radio Interface (CPRI); Interface Specification," *in Specification CPRI*, July 2014.

[6] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible centralization through Cloud-RAN," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov 2014.

[7] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G Backhaul Challenges and Emerging Research Directions: A Survey," *IEEE Access*, vol. 4, pp. 1743–1766, April 2016.

[8] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 152–159, February 2016.

[9] B. Guo, W. Cao, A. Tao, and D. Samardzija, "LTE/LTE-A signal compression on the CPRI interface," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 117–133, Sept 2013.

[10] 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, 2013.

[11] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "On the statistical multiplexing gain of virtual base station pools," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 2283–2288.

[12] L. Wang and S. Zhou, "On the fronthaul statistical multiplexing gain," *IEEE Communications Letters*, vol. 21, no. 5, pp. 1099–1102, May 2017.

[13] S. Bidkar, J. Galaro, and T. Pfeiffer, "First Demonstration of an Ultra-Low-Latency Fronthaul Transport Over a Commercial TDM-PON Platform," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[14] R. Nadiv and T. Naveh, "Wireless backhaul topologies: Analyzing backhaul topology strategies," *Ceragon White Paper*, pp. 1–15, 2010.

[15] M. Ruffini, "Multidimensional Convergence in Future 5G Networks," *Journal of Lightwave Technology*, vol. 35, no. 3, pp. 535–549, Feb 2017.

[16] P. Alvarez, F. Slyne, C. Bluemm, J. M. Marquez-Barja, L. A. DaSilva, and M. Ruffini, "Experimental Demonstration of SDN-controlled Variable-rate Fronthaul for Converged LTE-over-PON," in *Optical Fiber Communication Conference*. Optical Society of America, March 2018, p. Th2A.49.

[17] M. Ettus, "USRP User's and Developer's Guide," *Ettus Research LLC*, 2005. [Online]. Available: https://www.ettus.com/

[18] L. Golubchik and J. C. S. Lui, "Bounding of performance measures for a threshold-based queueing system with hysteresis," *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, vol. 25, no. 1, pp. 147–157, Jun. 1997. [Online]. Available: http://doi.acm.org/10.1145/258623.258684

[19] P. J. Courtois and P. Semal, "Computable bounds for conditional steady-state probabilities in large markov chains and queueing models," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 926–937, Sep 1986.

[20] P.-J. Courtois, *Decomposability:Queuing and Computer System Applications*. ACM monograph series, Jan 1997.

[21] C. D. Meyer, "Stochastic Complementation, Uncoupling Markov Chains, and the Theory of nearly Reducible Systems," *SIAM Review*, vol. 31, no. 2, pp. 240–272, 1989.

[22] W. Grassmann, "Transient Solutions in Markovian Queueing Systems." *Computers and Operations Research*, vol. 4, pp. 47 – 53, 1977.

[23] J. Kaufman, "Blocking in a Shared Resource Environment," *IEEE Transactions on Communications*, vol. 29, no. 10, pp. 1474–1481, Oct. 1981. [Online]. Available: http://ieeexplore.ieee.org/document/1094894/

[24] H. Akimaru, *Teletraffic Theory and Applications*. Springer London Ltd, 2012, oCLC: 933620017.