# Low-Latency and Energy-Efficient BBU Placement and VPON Formation in Virtualized Cloud-Fog RAN

Rodrigo Izidoro Tinini*, Daniel Macêdo Batista*, Gustavo Bittencourt Figueiredo†, Massimo Tornatore‡ and Biswanath Mukherjee§
*University of São Paulo, São Paulo, Brazil, Email: {rtinini, batista}@ime.usp.br †Federal University of Bahia, Bahia - Brazil, Email: gustavo@dcc.ufba.br ‡Politecnico di Milano, Milan - Italy, Email: massimo.tornatore@polimi.it §University of California, Davis - USA, Email: bmukherjee@ucdavis.edu

***Abstract*—Cloud Radio Access Networks (CRAN) allow to reduce power consumption in future 5G networks by decoupling BaseBand Units (BBU) from cell sites and centralizing the baseband processing from Remote Radio-Heads (RRH) in BBUs pools in a cloud. Although this centralization can enable power savings, it imposes much higher traffic on the optical transport network used to connect RRHs to the BBU pool, i.e., the fronthaul. In this paper we propose a hybrid Cloud-Fog RAN (CF-RAN) architecture that resorts to fog computing and to Network Functions Virtualization (NFV) to replicate the processing capacity of CRAN in local fog nodes closer to the RRHs that can be activated on demand to process surplus fronthaul/cloud traffic. We devise an ILP formulation and graph-based heuristics to decide when to activate fog nodes and how to dimension wavelengths on a Time-and-Wavelength Division Multiplexing Passive Optical Network (TWDM-PON) to support the fronthaul. Our results show that our architecture can consume up to 96% less energy than a traditional Distributed RAN (DRAN), providing a maximum transmission latency of about $20\mu$s between RRHs and BBUs even in large traffic scenarios. Moreover, we demonstrate that our graph-based heuristics can achieve same optimal solutions of the ILP formulation but with a reduction of 99.86% in the execution time.**

***Index Terms*—5G networks, Optical Fronthaul, Cloud-Fog RAN, TWDM-PON, VPON.**

## I. Introduction

Optical networks will play an important role in the deployment of 5G mobile networks. Due to their high transmission rate and low latency, they have been considered the best solution to transport fronthaul traffic generated in Cloud Radio Access Networks (CRAN) [1]. CRAN has already been adopted by mobile network operators to increase network efficiency while reducing CAPEX and OPEX [2] [3]. It takes advantage of Cloud Computing to implement centralized baseband processing and reduce power consumption. This is done by moving BaseBand Units (BBU) responsible for processing the received baseband signals from cell sites to a BBU pool located at the cloud, while leaving only low-energy Remote Radio-Heads (RRHs) at cell sites.

Study supported in part by CAPES - Finance Code 001, the INCT of the Future Internet for Smart Cities (CNPq 465446/2014-0, CAPES 88887.136422/2017-00, and FAPESP 14/50937-1 and 15/24485-9) and CNPq 311608/2017-5, 420907/2016-5, and 312324/2015-4.

The fronthaul data between RRHs and the BBU pool is typically encapsulated using the Common Public Radio Interface (CPRI) protocol [4]. CPRI is imposed with a strict round-trip latency of 3ms [5] and jitter of 65ns [6] between the RRHs and BBUs due to clock synchronization and the Hybrid Automatic Retransmit reQuest (HARQ) protocol. Moreover, it establishes line rates that may vary as a function of the Multiple Input Multiple Output (MIMO) configuration of each RRH, resulting in bandwidths demands ranging from 614.4Mbps up to 24.3Gbps [7]. So, to cope with the strict CPRI requirements in latency and bandwidth, optical fronthaul must be carefully planned and dimensioned.

Time-and-Wavelength Division Multiplexing Passive Optical Networks (TWDM-PON) [8] is a potential candidate to implement optical fronthaul due to its low latency, bandwidth efficiency and low cost of operation. In a TWDM-PON fronthaul, Optical Network Units (ONU) are placed on the transmitter side to tune RRHs to a wavelength and an Optical Line Terminal (OLT) is placed at the BBU pool to demultiplex several wavelengths transmitted on a single fiber carrying the data from multiple RRHs/cell sites. Moreover, relying on the concept of Virtual PONs (VPON) [9], virtualized PONs can be created and shared by several RRHs to transmit CPRI traffic to the BBU pool.

However, although CRAN can greatly reduce network costs, it may suffer from scalability issues if all baseband processing is centralized in a single cloud location, because it will become difficult to meet CPRI bandwidth and latency requirements [10]. To alleviate this pressure for bandwidth and to decrease latency, we consider an alternative RAN architecture where some of the BBU processing is moved closer to users [11].

In our previous work [12], we proposed an architecture called Cloud-Fog Radio Access Network (CF-RAN), that, by taking advantage of the emerging Fog Computing [13] paradigm, extends CRAN, by placing fog nodes closer to the edge user to receive baseband processing. The TWDM-PON is used to implement both fronthaul links and optical links connecting RRHs to fog nodes.

CF-RAN also leverages the concept of Network Functions Virtualization (NFV) [14] to migrate BBUs between cloud to fog nodes according to traffic demands. This is achieved by

implementing virtualized BBU (vBBU) processing functions in Virtual Machines (VM) that can be instantiated on demand via NFV. So, depending on the network traffic conditions, virtual processing functions can be dynamically activated on fog nodes in order to receive vBBU processing to alleviate the load on fronthaul and the BBU pool. An optimal placement of baseband processing based on an Integer Linear Programming (ILP) model was proposed to decide when to activate local processing functions and when to move BBUs from the BBU pool to fog nodes to save fronthaul bandwidth and balance baseband processing.

However, in [12], we only studied the placement of baseband processing in static traffic scenarios and its impact on power consumption. Moreover, we do not considered the dimensioning of a wavelength-limited TWDM-PON fronthaul. In realistic scenarios, when vBBUs are moved from the cloud to fog nodes, it is necessary to dimension the amount of VPONs to be used to support transmissions in fronthaul links and optical links connecting RRHs to fog nodes. The study conducted in [15] proposed that the dimensioning of a limited set of wavelengths in a hybrid RAN architecture is essential to avoid wavelength collisions in the fronthaul links.

So, in this work we extend the study in [12] by proposing an extended ILP model that dimension a limited set of wavelengths to support transmission to both cloud and fog nodes through VPONs in static scenarios. Furthermore, we also propose a new graph-based model and wavelength dimensioning heuristics to perform the placement and transmission of baseband processing in dynamic traffic scenarios where VPONs and processing functions at fog nodes needs to be dynamically activated or deactivated. Moreover, in these algorithms we seek to provide low times for the setup of baseband processing. Finally, we also extend the study in [12] by performing a deep analysis on the impact of trade-offs between the minimization of propagation latency, power consumption and network blocking probability in CF-RAN.

The rest of this paper is organized as follows: Section II presents related works, Section III presents the CF-RAN architecture and the TWDM-PON fronthaul, Section IV introduces the problem of placement of vBBUs and dimensioning of wavelengths, Section V presents the ILP formulation, the graph-based model and heuristics, Section VI presents a power model used to model the power consumption in CF-RAN and Section VII presents the results from our simulations. Finally, SectionVIII concludes the paper.

## II. Related Works

The operation of optical networks in support of CRAN and the placement of baseband processing functions in this scenario is currently attracting researchers attention.

In [16], authors studied the energy efficiency of BBU hotelling implemented over a Wavelength Division Multiplexing (WDM) optical network and the energy savings of BBU hotelling was evaluated. In [10] authors considered the case of CRAN over PON to support coordinated multi-point (CoMP) techniques showing that the use of VPON could enable significant savings on the signaling time between cells belonging to the same coordination area. However, even reducing CoMP latency, the centralization of baseband processing showed an increasingly latency related to the queuing of baseband processing as the cloud processor become busy. In [17], using VPON for CRAN, authors were able to demonstrate reductions of the number of handovers and throughput improvement. Authors in [18] proposed the use of a flexi-grid optical network to implement the fronthaul. Thus, the allocation of bandwdith follows finer granularities of bandwidth to meet CPRI line rates. In order to decrease the cost of the fronthaul, authors in [19] proposed a graph-based framework relying on genetic algorithms to split baseband processing in different BBUs. To increase spectral and energy efficiency, authors in [11] propose an architecture called Heterogeneous Cloud Radio Access Network (H-CRAN) that combines the benefits of CRAN and the processing capabilities of conventional base stations. In this architecture, the BBU pool is used for baseband processing whereas the local processing facility is used for user-oriented services. However, in this work, the distribution of processing among cloud and local processing facilities was not explored. The functional splitting of baseband in H-CRAN architectures is another option to deal with a constrained fronthaul. Authors in [5] stated that more baseband processing can be placed at the network but at the cost of increasing power consumption when baseband is split. Furthermore, the influence in the network latency to find a good trade-off between power consumption and the network performance in such scenario was not investigated. Fog-based RANs with focus on content caching are also a recent trend. The latency reduction of different content delivery policies in a constrained fronthaul was studied by authors in [20]. Authors in [21] proposed a graph-based model to coordinate fog nodes in providing cached content. However, while these works mainly focus to propose a framework to enable the caching in a fog-based RAN, they lack in addressing important aspects such as the influence on the power consumption while using a fog scenario. Furthermore, in [11] [20] [21] the technology of the fronthaul and its dimensioning to provide communications both to cloud and fog nodes are not considered. Regarding these works, the novelty of our work relies on proposing and dimensioning a hybrid network architecture with a specific TWDM-PON fronthaul that considers both the impact of power consumption and latency in its operation.

In next section we present our proposed CF-RAN architecture, and, next, we present the problems of placement of vBBUs and dimension of the fronthaul.

## III. System architecture

In this section we present our proposed CF-RAN architecture. First, we will discuss the details of the interconnection network used in the fronthaul to support the traffic connecting each RRH to its vBBU. Then we will present the architecture of the fog node and the operation of fog and cloud nodes.

### A. The Fronthaul Interconnection Network

In our proposed architecture, shown in Figure 1, the interconnection of RRHs and the virtualized BBU pool is done by

a point-to-multipoint TWDM-PON fronthaul. It connects the RRHs to the cloud as well as to local fog nodes equipped with the same processing functions as the cloud. We assume in-band signaling so all control plane communication is also transmitted through the same optical links used to transmit the CPRI traffic. All the processing related to the control plane is entirely done either in the cloud or in the fog nodes, without splitting the baseband processing [5].

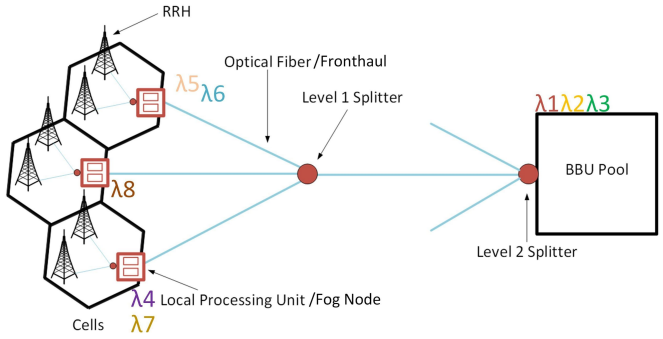



Fig. 1. Proposed CF-RAN architecture

As shown in Fig. 1, each ONU is connected to both a fog node and the cloud and the topology has three levels of multiplexing. In the first level of multiplexing, multiple ONUs are connected to an internal optical splitter inside a fog node. The optical signal transmitted by these ONUs is multiplexed towards the local processing functions of the fog node by an internal optical splitter or towards the BBU pool in the cloud through a distribution fiber.

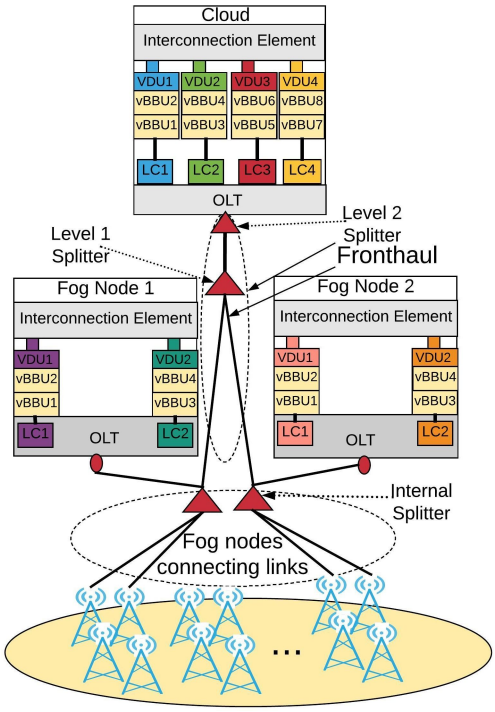
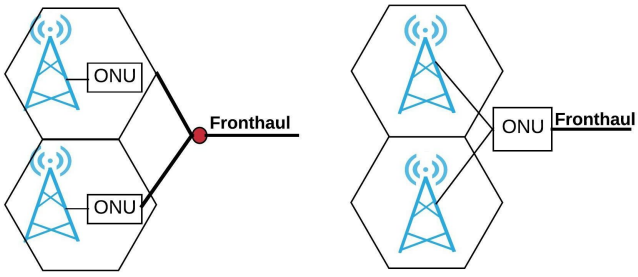



Fig. 2. Details of the processing nodes

In the second level, a Level 1 optical splitter is used to multiplex the traffic transmitted in several distribution fibers into one feeder fiber. In the third level, close to the BBU pool, a Level 2 optical splitter multiplexes several feeder fibers into a single Line Card (LC) that receives the optical signal for a particular wavelength. The receiver node is equipped with an OLT that demultiplexes and switches the traffic received on each LC to its corresponding VDU, that comprises the virtualized environment for the baseband processing. Each ONU is equipped with a tunable laser responsible to assign the wavelengths granted by the OLT to the transmissions, i.e., each ONU can tune its transmission to any available wavelength in its optical link. Finally, many ONUs can also be tuned to a common wavelength in order to form a VPON. Thus, multiple RRHs will share a common virtualized optical channel multiplexed in time.




a) RRHs connected to single port ONUs b) RRHs connected to multi-port ONUs

Fig. 3. Different topologies to RRH/ONU connection

Each RRH is connected to an ONU. Each ONU can provide connections to single or multiple RRHs. It is a choice of the operator to connect each RRH to a single port ONU or multiple RRHs into a multi-port ONU. In this work, we refer to a multi-port ONU connecting to multiple RRHs as an Aggregation Group (AG). Fig. 3(a) shows each RRH connecting to a single port ONU and in Fig. 3(b) a multi-port ONU is used to connect to several RRHs. Using single port ONUs will increase the number of deployed ONUs in comparison with aggregating several RRHs in multi-port ONUs. However, if multi-port ONUs are used, the required CPRI traffic in each ONU may be increased. In this work we consider these two schemes of RRHs-ONUs connection to explore different options in terms of power consumption.

### *B. Maximum Fronthaul Latency*

The fronthaul of CF-RAN must consider a strict round-trip latency requirement of 3ms for the BBU processing. This latency requirement comes from the Hybrid Automatic Retransmit reQuest (HARQ) protocol used for data retransmission mechanisms between UEs and the RRHs. HARQ imposes that UEs should receive ACK/NACK messages from the BBU pool in three subframes after sending uplink data to the RRH. If no ACK/NACK is received within these three subframes, the UEs will retransmit the data to the RRH.

TABLE I
BBU PROCESSING LATENCY

| | |
|---|---|
| RRH RF UL/DL Processing Time | $\sim 40\mu s$ |
| CPRI Processing Time (RRH + BBU) | $\sim 10\mu s$ |
| BBU Processing Time (UL/DL PHY+MAC) | $\sim 2700\mu s$ |

Considering the HARQ processing, CPRI defines a maximum latency budget of 3ms between the RRHs and the BBUs, that comes from the latency values of the processing functions presented in Table I [22]. Considering that the total BBU processing latency is $40\mu$s+$10\mu$s+$2700\mu$s $= 2750\mu$s, the maximum delay introduced by the fronthaul network must be of at most $250\mu$s [23]. In order to operate under this delay, CF-RAN operates under the TWDM-PON, in which low transmission times can be achieved due to passivity of its equipments.

### C. Operation of the Fog and Cloud Processing Nodes

Figure 2 shows the architecture of the fog and cloud nodes used to provide virtualized processing functions of CPRI traffic or control messages. Both fog and cloud nodes are equipped with a number of LCs to receive optical signals carrying the traffic to be processed. Each LC transmits the traffic to a specific VDU where it is received and processed.

The VDUs implement a set of Virtualized Processing Functions (VPF) that provide control functions and other additional network services. The vBBU is one of these functions. It is responsible for the baseband processing of the CPRI traffic. As processing functions are virtualized, they can be enabled or disabled dynamically according to the network state. In order to efficiently use the virtualized resources in CF-RAN, in the next section we present the problem of placement of baseband processing.

## IV. Placement of BaseBand Processing Functions

For every active RRH in the network, a vBBU must be active to process its baseband signals encapsulated by the CPRI protocol. Before the RRH starts receiving UEs baseband signals, the operator needs to find a processing node with enough free processing resources to accommodate the vBBU. In order to save energy, we assume that vBBUs are first placed only in the VDUs of the cloud. Therefore, as long as the cloud has enough computing capacity to receive vBBUs, all the CPRI data coming from the cell sites is processed in the BBU pool. Furthermore, through the transmission of data from ONUs to the OLT, the operator needs to create VPONs on the fronthaul to support the transmissions between RRHs and its vBBUs.

As the network demand grows, the resources allocated both in the fronthaul (VPONs) and in the cloud (VDUs) may not be enough to support the demands. Hence, local processing functions in the fog nodes are activated on demand by the NFV capability and according to the availability of the wavelengths used for the transmission of data to the fog nodes. As the traffic demand continues to increase, the fog nodes are activated gradually as the capacity of the previous fog node is exhausted.

Furthermore, as local processing functions are activated, the operator needs to find a suitable VPON in fog nodes connecting links to support the transmissions between RRHs and vBBUs that were placed locally. This means that enough wavelengths must be available to connect RRHs to fog nodes. After vBBUs are placed both in cloud or fog nodes, the set of available wavelengths must be dimensioned through the fronthaul and fog nodes links to support transmissions to the newly deployed vBBUs through VPONs.

As vBBUs are first placed at the cloud, firstly wavelengths are given to the fronthaul. The amount of bandwidth to be made available for the fronthaul is calculated in function of the amount of CPRI data processing placed in the BBU pool. As fog nodes get activated, the amount of bandwidth necessary to support the CPRI transmissions to vBBUs placed locally is calculated. If there is enough available wavelengths, VPONs are created by the fog nodes OLTs to support transmissions to fog nodes.

In order to efficient dimension the wavelengths through fronthaul and fog nodes, the less possible amount of wavelengths is used. For instance, if the amount of wavelengths allocated to fronthaul is not properly dimensioned, there may be not enough wavelengths to create VPONs on newly activated fog nodes connecting links. An example of this dimensioning is depicted in Fig. 1 in Section III, where the set of TWDM-PON wavelengths were distributed among the fronthaul and the fog nodes so each processing node has a group of operating wavelengths to receive CPRI traffic.

In the next section we present algorithms developed to optimally place the vBBU processing functions on CF-RAN while energy-efficiently dimensioning and forming VPONs.

## V. ILP Formulation and Graph-based Model and Heuristic

In this section, we propose two approaches to solve the problem of energy-efficient placement of vBBUs and wavelength dimensioning. The first one consists of an ILP model used to perform the vBBUs placement and wavelength dimensioning in network static scenarios where the RRHs demands and the network state are known in advance. The second one is a graph-based model and heuristics suitable for both static and dynamic network traffic scenarios, where traffic demands fluctuate over time.

### A. Problem Formulation

Let $R$ be a set of active RRHs $i$ demanding processing and transmission of CPRI traffic in any processing node of the network using any available wavelength. Given a set $N$ of processing nodes $n$ and a set $W$ of available wavelengths $w$, schedule all the demands of $R$ in the least number of processing nodes $n$ using the least number of wavelengths $w$ to transport the placed demands. This problem is a variation of the 2D Bin-Packing problem, as each RRH CPRI demand must be both packed into a VPON and a processing node, respecting the bandwidth and processing capacities, respectively, and the objective is to minimize the number of active nodes and VPONs to promote energy efficiency.

### B. ILP Formulation for the Static Scenario

The ILP model promotes energy efficiency by activating the processing functions of the nodes as traffic grows in static traffic scenarios. As the demand grows, more processing functions are activated. If the demand is low, some unused processing functions remains deactivated.

**Input Parameters**

$R$: set of RRH traffic demands $i$

$N$: set of all possible processing nodes $n$, including the cloud and fog nodes

$W$: set of available wavelengths $w$

$F_{in}$: set of binary values representing the RRH $i$ that are connected to fog node $n$

$B_i$: bandwidth demand of RRH/ONU $i$

$B_w$: capacity of wavelength $w$

$Proc_i$: processing demand of RRH/ONU $i$

$Proc_n$: processing capacity of node $n$

$C = \{C_1, ...,C_n\}$: Set of power costs of each node $n$

$C_{vdu} = \{C^1_{vdu}, ...,C^n_{vdu}\}$ : Set of power costs of VDUs in each node $n$

$C_{lc}$: power cost of a LC

$B$: a very big positive number

**Decision Variables**

$y^i_{wn}$: = 1 if the traffic demand of RRH $i$ is processed at node $n$ being transmitted at the VPON $w$, 0 otherwise.

$\bar{y}_{in}$: = 1 if RRH $i$ is placed at node $n$, 0 otherwise. This is an auxiliary variable used to ensure that each RRH is placed either in the cloud or in its connected fog node.

$z_{wn}$: = 1 if wavelength $w$ is allocated to node $n$, 0 otherwise.

$x_n$: = 1 if processing functions and infrastructure of node $n$ is activated, 0 otherwise. This variable is used to account the active processing nodes.

**Objective Function**

$$Minimize \sum_{n=1}^{|N|} x_n * C_n + \sum_{w=1}^{|W|} \sum_{n=1}^{|N|} z_{wn} * (C_{lc} + C^n_{vdu})$$

**Constraints**

$$(1) \sum_{n=1}^{|N|} z_{wn} \leq 1 \mid \forall\ w \in W$$

$$(2) \sum_{w=1}^{|W|} \sum_{n=1}^{|N|} y^i_{wn} = 1 \mid \forall\ i \in R$$

$$(3) \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} y^i_{wn} * B_i \leq B_w \mid \forall\ w \in W$$

$$(4) \sum_{i=1}^{|R|} \sum_{w=1}^{|W|} y^i_{wn} * Proc_i \leq Proc_n \mid \forall\ n \in N$$

$$(5) B.x_n \geq \sum_{i=1}^{|R|} \sum_{w=1}^{|W|} y^i_{wn} \mid \forall n \in N$$

$$(6) x_n \leq \sum_{i=1}^{|R|} \sum_{w=1}^{|W|} y^i_{wn} \mid \forall n \in N$$

$$(7) B.z_{wn} \geq \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} y^i_{wn} \mid \forall w \in W$$

$$(8) z_{wn} \leq \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} y^i_{wn} \mid \forall w \in W$$

$$(9) B.\bar{y}_{in} \geq \sum_{w=1}^{|W|} y^i_{wn} \mid \forall i, n \in R, N$$

$$(10) \bar{y}_{in} \leq \sum_{w=1}^{|W|} y^i_{wn} \mid \forall i, n \in R, N$$

$$(11) \bar{y}_{in} \leq F_{in} \mid \forall i, n \in R, N$$

The objective function aims at packing as many as possible RRHs into a single VPON to reduce the power consumption of LCs and VDUs and to activate the minimum number of processing nodes. Note that, even when a vBBU is placed at the cloud, the cost of the cloud is computed by the objective function through the term $\sum_{n=1}^{|N|} x_n * C_n$, when index $n$ refers to the cloud node. Constraint 1 ensures that each wavelength is assigned to one node at most. Constraint 2 ensures that each RRH is assigned to one processing node and VPON. Constraints 3 and 4 ensure that the bandwidth capacity of each VPON and the processing capacity of a node will be respected. Constraints 5 and 6 assures that node $n$ is activated when RRH $i$ is allocated to it. Constraint 7 and 8 ensures that when a VPON is formed, its wavelength is assigned to the related processing node. Constraint 9, 10 and 11 ensures that, if RRH $i$ is placed on a fog node, it is only placed on the fog node connected to it. To efficiently exploit the TWDM-PON capacity, we consider two policies of bandwidth dimensioning for the ILP formulation, relying on the type of RRH-ONU connection being used, i.e. RRHs connected to single or multi-port ONUs:

- Fully VPON Formation (Fully-VPON): This scheme considers that each RRH is connected to a single port ONU and a common wavelength can be shared by multiple ONUs. So, the OLT at the central office dynamically tune the ONUs of multiple RRHs to a common wavelength. The CPRI traffic of each RRH is used as the bandwidth input variable $i$.
- Aggregation Groups VPON Formation (AG-VF): This scheme considers multiple RRHs connected to a multi-port ONU and allocate one wavelength for each multi-port ONU so that each multi-port ONU will represent a different VPON. The number of RRHs connected to the ONU will define the amount of bandwidth needed to transport the aggregated CPRI traffic at the VPON. At this scheme, the total aggregated traffic at the ONU is used as the input variable $i$ for the ILP.

As ILP formulations are computationally expensive, for a less expensive computing solution, we also propose a graph-based heuristic.

### *C. Graph-based heuristic for Static and Dynamic Traffic Scenarios*

In this section we propose a graph-based model and algorithm heuristics for both static and dynamic cases. The graph model consists in modeling the CF-RAN as a digraph, with vertices representing RRHs and processing nodes and directed arcs representing the fronthaul. The vBBU placement and wavelength dimensioning problems are modeled as a max flow-min cost problem where the objective is to input the maximum flow between source and destination vertices while minimizing the amount of traversed nodes in the digraph. A flow represents the transmitted CPRI traffic of each RRH towards a processing node.

Let $G = (V, E)$ be a digraph, where $V(G)$ represents the set of vertices of $G$ and $E(G)$ the set of arcs. Each arc $e$ has a capacity and cost value. Given $e = (u, v)$, we define a flow from $u$ to $v$ by an arc that is directed from $u$ to $v$ with capacity greater than 0. Let $R$ be the set of RRHs $i$; $F \in V$ be a set of fog nodes $f$ and $F_{bridge} \in V$ a set of intermediate vertices $fb$ called fog bridges responsible for interconnecting

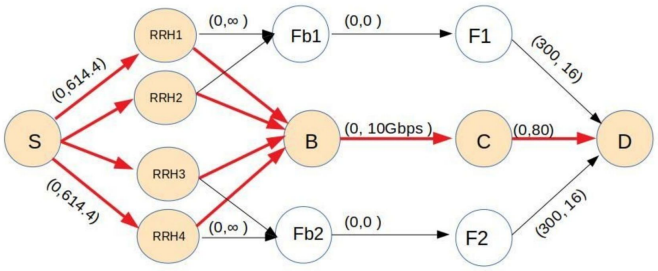



Fig. 4. Digraph representing a flow network for the dimensioning problem

multiple RRHs into a single fog node in our implementation. A vertex $C \in V$ represents the cloud node and vertex $B \in V$ is an intermediate vertex used to implement a link between the RRHs and the cloud. Finally, vertex $S \in V$ is a source node directed to each $i \in R$ responsible to input CPRI flow into the RRHs and vertex $D \in V$ receives all the CPRI that flows through the network. Variables $f_{capacity}$ and $c_{capacity}$ represent the processing capacities of a fog node $f$ and the cloud, respectively.

In our model, the TWDM-PON optical links are represented by the arcs that connect each $i$ to $fb \in F_{bridge}$ and $B$. The wavelengths available in each link are represented by the capacity of each arc. The processing capacity of both fog nodes $f$ and cloud $C$ is represented by its incident arcs $e$ from $fb \in F_{bridge}$ and $e$ from $B$, respectively and their power costs are represented by the costs of these arcs.

The digraph is constructed by the following steps: For each vertex $i \in R$, put a directed arc from $S$ to $i$ with cost 0 and capacity 0. For each RRH $i$ connected to fog node $f$, put a directed arc from $i$ to fog bridge $fb \in F_{bridge}$ with cost 0 and capacity $\infty$. For each $fb \in F_{bridge}$, put a directed arc from $fb$ to $f \in F$ with cost 0 and capacity 0. For each $i \in R$, put a directed arc to $B$ and put a directed arc from $B$ to $C$ with cost 0 and capacity 0. Finally, for each $f \in F$ put a directed arc to $D$ with cost $fog_{cost}$ and capacity $f_{capacity}$ and put a directed arc from $C$ to $D$ with cost 0 and capacity $c_{capacity}$.

The objective of the graph model is to maintain the maximum flow through $S$ to $D$, passing through each $i$ to processing node $f \in F$ or $C$, as depicted in Fig. 4. This operation involves two steps: First, based on the traffic demand, a wavelength dimensioning heuristic is used to calculate the amount of bandwidth (VPONs) to be placed on the arcs representing the fronthaul. When the traffic demand is greater than the overall network available bandwidth (both to transmit to cloud or fog nodes) and there is available wavelengths, new VPONs can be formed in the network links. Second, after the VPONs are dimensioned on the TWDM-PON links, a min-cost-max-flow procedure is executed to maximize the flow between vertices $S$ and $D$ while minimizing the amount of traversed nodes. This model only considers a fully virtualized dimensioning of VPONs, considering that each RRH is connected to a single port ONU because considering multi-port ONUs would increase the number of vertices in the digraph.

Due to different characteristics of both static and dynamic traffic, in the next sections we propose specific wavelength dimensioning heuristics to both static and dynamic traffic.

### D. Wavelength Dimensioning Heuristic for Static Traffic

To calculate the amount of necessary VPONs to support CPRI transmission for the static case, we propose the Cloud First-Fog Least (CF-FL) heuristic. This heuristic aims at creating VPONs for the fronthaul as long as the traffic demand can be supported by the cloud to promote energy efficiency. After the cloud processing capacity is exhausted, it calculates how many VPONs would be necessary in each fog node to support traffic transmissions. As long as there are wavelengths available on the network, i.e., there is at least a wavelength that is not being used by the fronthaul or other fog nodes, CF-FL sequentially distribute them among each fog node until the amount of bandwidth available for the fog node can support the transmissions of CPRI traffic of RRHs connected to it. After all CPRI traffic can be supported by the available bandwidth, the wavelength dimensioning is done.

This wavelength dimensioning algorithm is formally presented in Algorithm 1. In line 2, the amount of incoming CPRI traffic is calculated as a function of the activated RRHs requesting for a VPON and a vBBU. Then in line 3 it is checked if the VDUs in the cloud has free capacity to host new vBBUs. After that, it is checked in line 4 if there is enough bandwidth available at the fronthaul to transmit to the cloud. If do not, in lines 5 and 6 VPONs are created in the cloud until the fronthaul has enough VPONs to transmit to the cloud. If the cloud can not host all vBBUs requested by the RRHs (line 7), then, it is checked if some vBBUs can be placed at the cloud while the rest is placed at fog nodes. If so, new VPONs are eventually created on the cloud before activating fog nodes (lines 8, 9 and 10). Then, the amount of residual CPRI traffic, i.e., the amount of CPRI traffic that could not be placed on the cloud, is calculated in line 12. Then, until there are available wavelengths and while the residual traffic is greater than the total bandwidth available in fog nodes (line 13), one VPON is created per fog node until all CPRI traffic can be handled in fog nodes in lines 14, 15, 16 and 17 and the amount of available bandwidth in fog nodes is updated in line 18. Finally, a max flow-min cost algorithm is executed in line 19 to input flow on the digraph.

### E. Wavelength Dimensioning Heuristics for Dynamic Traffic

For the dynamic case, as the amount of activated RRHs varies over time and each fog node can be connected to different amounts of activated/deactivated RRHs, we propose different VPON placement heuristics for both fronthaul and fog nodes. Due to space limitations, we only explain the idea behind these heuristics:

*Dynamic VPON Placement Heuristics:*

- *Least Loaded:* This policy allocates the VPONs first to the cloud and then for the fog nodes starting with least active RRHs first until all incoming traffic can be accommodated by the available VPONs.
- *Most Loaded:* This policy allocates the VPONs first to the cloud and then for the fog nodes starting with most active RRHs first until all incoming traffic can be accommodated by the available VPONs.

**Algorithm 1** CF-FL wavelength dimensioning heuristic

**Input:** Digraph $G$, Set of available wavelengths $w \in W$, Set of active RRHs $R$, fronthaul bandwidth $frtBand$, cloud free capacity $c_{capacity}$, allocated traffic $allocTraffic$, total fog links bandwidth (as a function of the number of VPONs in fog nodes) $fogBand$

**Output:** Placement of baseband processing of RRHs $i \in R$

```
Calculate the total incoming traffic
traffic = |R| * cpriLineRate
if traffic <= c_capacity then
    if traffic > frtBand OR frtBand == 0 then
        while frtBand < traffic AND |W| > 0 do
            frtBand ← w ∈ W
else if traffic > c_capacity then
    Put as many as possible baseband on the cloud
    while frtBand < c_capacity AND |W| > 0 do
        frtBand ← w ∈ W
    Calculate the amount of traffic to be placed on fog nodes
    residual = traffic − allocTraffic
    while residual > fogBand AND |W| > 0 do
        Create VPONs in each fog node until all traffic can be supported in fog nodes
        for all fog node f do
            if residual > fogBand AND |W| > 0 then
                f ← w ∈ W
                fogBand += 10Gbps
Execute a max flow-min cost algorithm
```

- *Fog First:* This policy first allocates VPONs to the fog nodes and then to the cloud.

## VI. Power Consumption Model

We use a power consumption model based on the parameters introduced in [16] and [24] to model the consumption of stand-alone BBUs placed with a RRH on a Distributed RAN (DRAN) node, cloud and fog nodes that consolidates vBBU processing, vBBUs instantiated on processing nodes and LC used to terminate traffic from a VPON. The power consumed to maintain a stand alone BBU at the RRH is equal to 600 W, and this consumption is related to the BBU itself. The consumption of cloud and fog nodes nodes has a fixed value of 600 W for the cloud and 300 W for the fog node and an additional consumption of 20 W per each hosted vBBU per each baseband demand placed on it. For each transmitted wavelength, we assume the power cost of 5 W of using one line card/OLT port. Finally, the OLT has a base consumption of 100 W [25]. Table II summarizes all the power consumption values assumed in our model.

TABLE II
Power consumption parameters

| Element | Cost |
|---|---|
| DRAN node | 600 watts |
| Cloud base consuming | 600 watts |
| Fog node base consuming | 300 watts |
| VDU power cost at cloud and fog nodes | 100, 50 watts |
| vBBU power consuming | 20 watts |
| Line Card | 5 watts |
| OLT | 100 watts |

## VII. Performance Evaluation

To assess the performance of the proposed solution, we performed simulations considering static and dynamic network scenarios. We used the IBM ILOG CPLEX Optimization Studio V12.8.0 to solve the ILP in the static scenario and get the optimal results. For the dynamic scenario, we developed an ad-hoc simulator written in Python using the SimPY library. The computer used to solve our instances is an Intel i7 2.2GHz, 16GB running Ubuntu 18.04.1.

### A. Static Traffic Scenario

We executed the ILP and the CF-FL graph heuristic considering a static scenario where the network traffic demands are known in advance. We simulated the CF-RAN presented in Section III-A. The network is composed of 1 cloud node and 5 fog nodes. Each RRH implements a 10Mhz 1X1 MIMO channel, generating 614.4Mbps [7] [10]. The cloud and each fog node has a processing capacity of 30 and 10 RRHs, respectively. The simulated TWDM-PON fronthaul has 20 available wavelengths of 10Gbps capacity each. The fiber extension from the RRHs to fog nodes is set to 20km and from fog nodes to the clouds is set to 40km. The maximum latency to reach the fog nodes and the cloud is about $98\mu$s and $196\mu$s, respectively. For the ILP executions, we considered the wavelength dimensioning policies in Section V-B (Fully-VPON and AG-VF). For AG-VF, the amount of RRHs connected in a multi-port ONU is set to 4 RRHs, and the number of aggregation groups is 5, 10 or 15. We evaluate the network power consumption, average propagation latency and the execution time of the proposed algorithms.

Fig. 5 shows the power consumption of CF-RAN with the different bandwidth dimensioning strategies in comparison with a traditional DRAN and CRAN. A significantly lower power consumption can be observed in CF-RAN in comparison to DRAN. Note that CRAN shows lower power consumption when the aggregation groups increases, but at the cost of reduced network coverage, as it will be shown in next section. For CF-RAN, the best bandwidth dimensioning in terms of power consumption is achieved for the ILP Fully-VPON and the CF-FL graph heuristic. When the size of the aggregations groups increases, bandwidth consumption of AG-VF tends to increase in comparison to Fully-VPON and the graph-based heuristic. This is because when more ONUs shares the same wavelength, the capacity of each wavelength is best used because as many as possible RRHs will be placed into the same VPON. The power consumption of CF-RAN, when using Fully-VPON and CF-FL graph heuristic is lower than DRAN in the order of 96%. The Fully-VPON and the CF-FL graph heuristic have similar power consumption and in comparison with AG-VF policy, their power consumption is 23.9% more efficient.

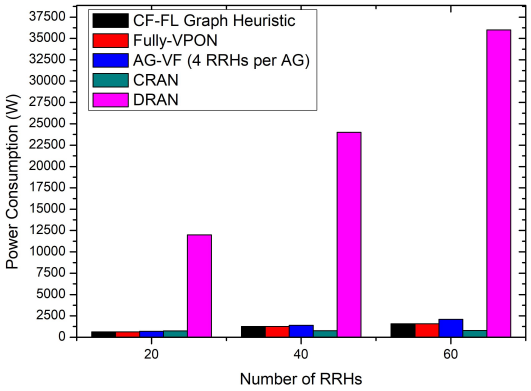



Fig. 5. Power consumption for different amounts of RRHs and considering aggregation groups of 4 RRHs for AG-VF policy

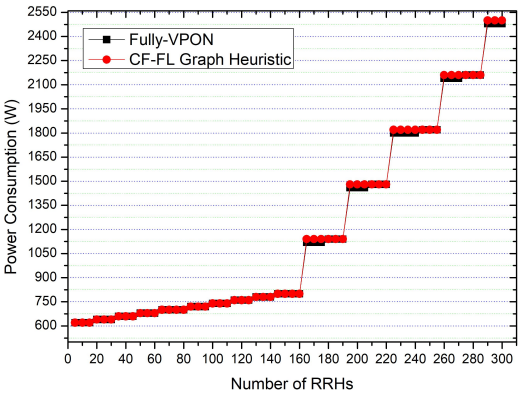



Fig. 6. Comparison of power consumption between ILP and graph heuristic for increased CF-RAN sizes

To compare CF-FL graph heuristic to the ILP Fully-VPON policy in large network scenarios, we increased the CF-RAN size, ranging the amount of RRHs from 5 to 300 in Fig. 6. In this scenario, the processing capacity of the cloud and fog nodes were increased to 160 and 32 RRHs, respectively. As it can be observed in Fig. 6, even for increased amount of RRHs, the CF-FL performance is comparable to the ILP Fully-VPON, being able to achieve the almost optimal solution for every number of RRHs. CF-FL sub-optimal solutions approximated the optimal solution in the order of 92% to 99%.

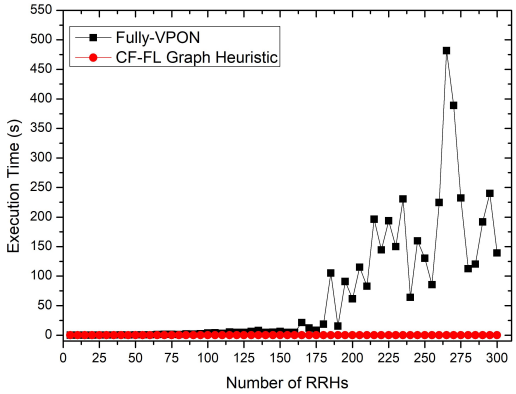



Fig. 7. Comparison of execution times between ILP and graph heuristic for increased CF-RAN sizes

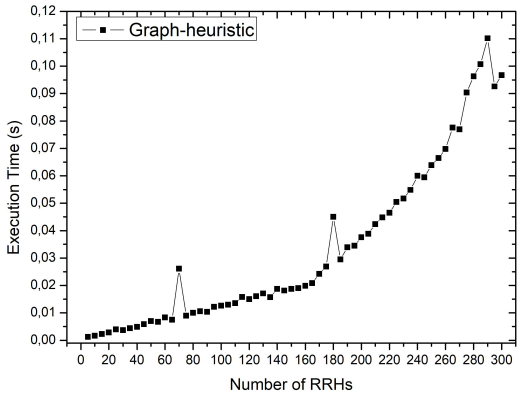



Fig. 8. Graph heuristic execution time for increased CF-RAN sizes

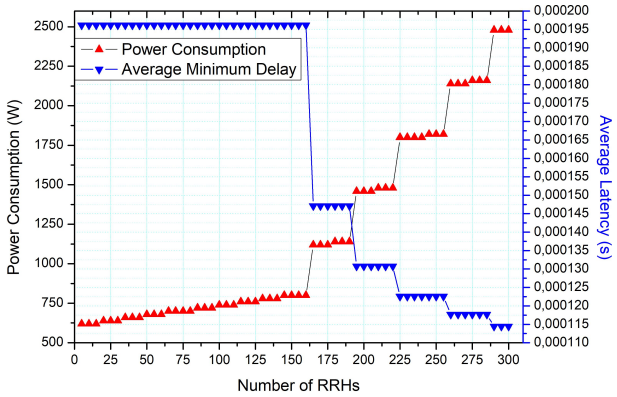



Fig. 9. Trade-off between power consumption and average minimum delay for the transmission of CPRI traffic

In Fig. 7, we show the execution time for both the ILP and the CF-FL. Note that the execution time for the ILP tends to increase as the network size increases. However, the CF-FL graph heuristic provides much less computing effort than the ILP, with execution times in the order of few milliseconds to hundred milliseconds. Fig. 8 shows a zoomed view of CF-FL execution time. Note that execution time slightly increases with the network size, but it stays very low.

Finally, in Fig. 9 we show the average propagation latency between RRHs and processing nodes. Note that there is a clear trade-off between power consumption and the average propagation latency. When power consumption in minimized, all CPRI traffic is sent to the cloud and the propagation latency is higher, but, as fog nodes are activated, RRHs begins to be locally processed and the average propagation latency is decreased at the cost of higher power consumption.

### *B. Dynamic Traffic Scenario*

For the dynamic scenario, we considered a CF-RAN composed of 1 cloud, 5 fog nodes, 160 RRHs and 20 wavelengths of 10Gbps capacity. Each fog node is connected to 32 RRHs, each of them generating basic CPRI traffic (614.4Mbps). The processing capacity of the cloud and each fog node is 80 and 16 RRHs, respectively. We considered the traffic fluctuation from a typical 24 hours operation in a business geographical region following patterns taken from [26] and [9] and shown in Fig. 10. At the beginning of the simulation, all RRHs

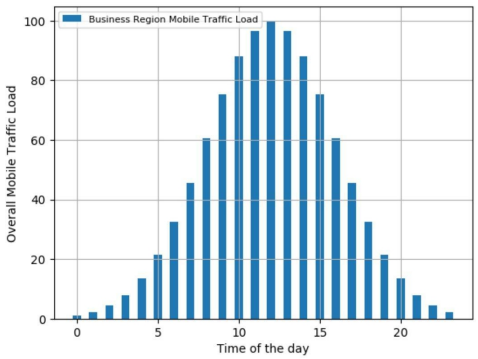



Fig. 10. Typical Traffic Load of a Business Area

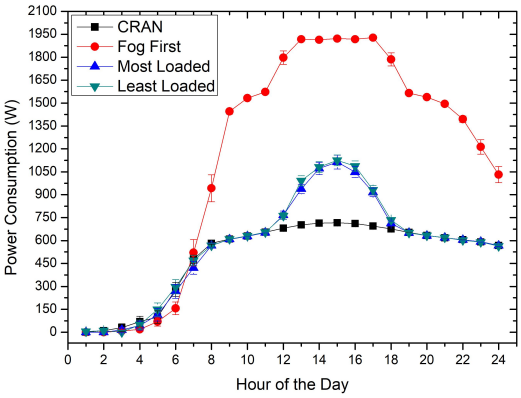



Fig. 11. Power consumption of the CF-RAN proposed heuristics and CRAN

are turned off. Then, they are activated following a Poisson process whose mean is equal to $(e/60)$, where $e$ is the maximum traffic load (erlang) at a given hour. Each RRH stays active during a service time uniformly taken from (0.25 hour, 1 hour). Results show average values obtained from 60 executions of each scenario with a confidence level of 95%.

Fig. 11 shows the power consumption for the three bandwidth dimensioning heuristics (Least Loaded, Most Loaded and Fog First) in comparison to CRAN. Note that CRAN has the lowest power consumption in comparison with CF-RAN. However, this low power consumption will lead to high blocking probabilities as no auxiliary processing nodes will be activated when the cloud became stressed. Regarding the proposed heuristics for CF-RAN, it is observed that more power efficiency is achieved when vBBUs are first placed in the cloud and both Least Loaded and Most Loaded policies have similar power efficiency. Note that, even in hours with small loads (6a.m. and 12p.m.), a relatively high power consumption is expected because even when a small number of RRHs is active, they are placed on the cloud, thus, accounting the cloud power consumption even in lowest loaded hours. The same behaviour can be observed for the Fog First heuristic because several fog nodes can be active to support even a small number of RRHs.

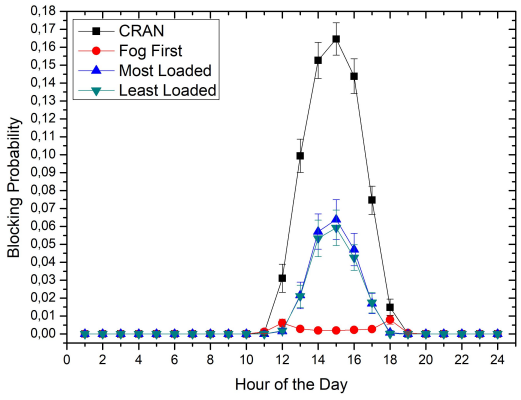



Fig. 12. Blocking Probability for CF-RAN heuristics and CRAN

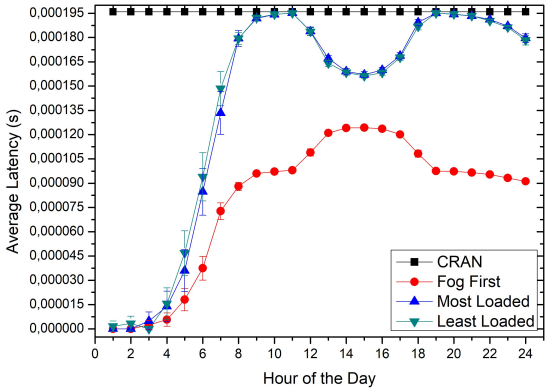



Fig. 13. Average propagation latency

Fig. 12 shows the blocking probability for CF-RAN and CRAN. Note that CRAN has the highest blocking probability. This is because no fog node is used when the capacity of the cloud or the fronthaul is exhausted. In CF-RAN, it is possible to observe that the Fog First heuristic achieves the lowest blocking probability, but at the cost of higher power consumption from the cost of first placing vBBUs at fog nodes, as shown in Fig. 11. Moreover, it is possible to see that first placing vBBUs in the cloud also leads to small blocking probabilities. It is also possible to observe that the Least Loaded heuristic tends to achieve lower blocking probability compared to Most Loaded heuristic in peak hours, with a reduction in the order of at most 7.2%.

The average propagation latency for CF-RAN is shown in Fig. 13. Note that, CRAN will always achieve the higher average propagation latency due to the full centralization of baseband processing. It can be observed that, the average latency in our proposed heuristics grows in function of the network load. Note that the Fog First heuristic will always achieve the lowest average latency, with a reduction of at most 57.3% in lowest loaded hours, because more baseband processing will be placed on the fog than in the cloud. However, although Most and Least Loaded heuristics have a higher average latency than Fog First, their difference tends to be reduced to at most 20% in peak hours. An interesting trade-off between the traffic load and the average latency can be observed. For a fog first scenario, the average latency tends to grow in function of the traffic load, however, when the use of the cloud is prioritized, higher traffic loads results in the reduction of the average propagation latency.

Fig. 14 shows the trade-off between the average propagation

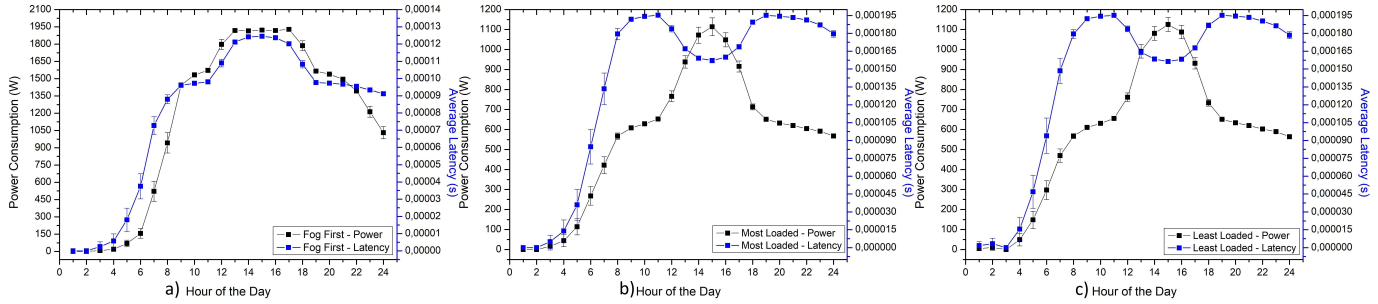



Fig. 14. Trade-offs between latency and power consumption in CF-RAN: a) Fog First heuristic, b) Most Loaded heuristic, c) Least Loaded heuristic

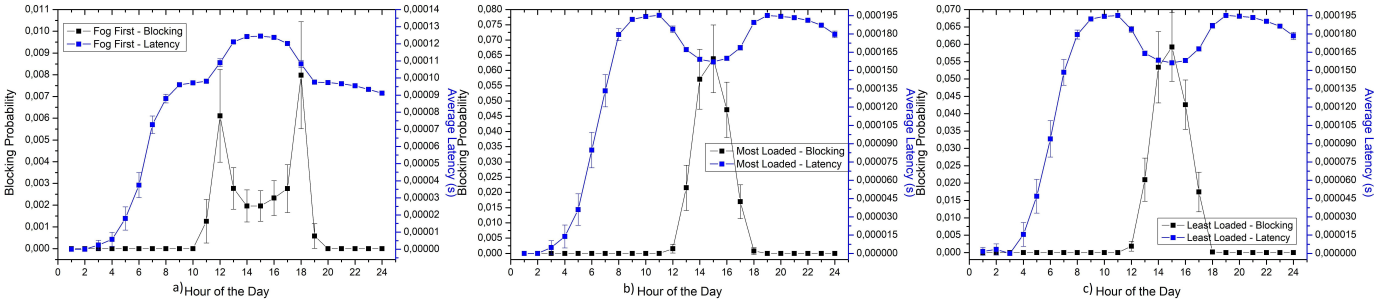



Fig. 15. Trade-offs between latency and blocking probability in CF-RAN: a) Fog First heuristic, b) Most Loaded heuristic, c) Least Loaded heuristic

latency and the power consumption for the proposed heuristics in CF-RAN. Note that, for Fog First, the growth of latency closely follows the power consumption for all traffic loads (Fig. 14 a)). Although Fog First produces the lowest average latency, we observed that the growth of latency and power consumption is bigger for Fog First in comparison to Most and Least Loaded. For Fog First, from 7a.m. to 15p.m., a growth of about 40% and 74% in latency and power consumption is observed, respectively. On the other hand, there is a greater gap between the growth of latency and power consumption for Most and Least Loaded heuristics. From 7a.m. to 15p.m., the latency and power consumption growth is about 7% and 63% for Most Loaded, respectively and 7% and 59% for Least Loaded, respectively. In lowest loaded hours, an increased latency (i.e., more baseband centralized in the cloud) incurs low power consumption. However, in peak hours, around 13 and 17p.m., there is a strong trade-off between latency and power, as with many more fog nodes activated, latency will be decreased at the cost of a increased power consumption (Fig. 14 b) and c)). Note that, as Fog First achieves the lowest latency imposing the higher power consumption, the prioritized use of the cloud by Most and Least Loaded brings a better balancing between latency and power consumption in comparison to Fog First.

Finally, Fig. 15 shows the influence of latency in blocking probability. As it can be seen in Fog First heuristic (Fig. 15 a)), in peak hours, as latency grows in about $15\mu$s, blocking probability is minimized in about 3 times. This happens because at these peak hours, the cloud begins to be used by the Fog First heuristic and more vBBUs are supported. However, for Most and Least Loaded heuristics, as latency is decreased by the activation of fog nodes, blocking probability tends to increase, as it becomes more complexity to dimension the available wavelengths to all activated fog nodes. For Most and Least Loaded heuristic, when latency decreases in about $40\mu$s, blocking probability increases in about 40 and 32 times, respectively.

## VIII. Conclusion

In this paper we proposed a power efficient and latency-guaranteed network architecture called CF-RAN. By means of fog computing and NFV, CF-RAN greatly reduces the footprint of DRAN architectures and is capable of expanding the capacities of CRAN fronthaul by the dynamic activation of baseband processing functions in fog nodes closer to RRHs. To power-efficiently plan the activation of fog nodes, we proposed an ILP formulation and a graph-based model and heuristics. Our results have shown that the graph-based heuristic is capable of achieving optimal solutions as the ILP but with a huge decrease in execution time. While CF-RAN is able to operate under the latency requirements of CPRI protocol, we also observed that there is an interesting trade-off between propagation latency, power consumption and blocking probability. If the minimization of latency is valued through a prioritized placement of vBBUs in fog nodes over the cloud, CF-RAN can offer the best network coverage as the blocking probabilities are the lowest. However, it imposes a huge increasing in OPEX in comparison to a prioritized placement of vBBUs in the cloud over the fog nodes. Although prioritizing the cloud will increase the average propagation latency, the power consumption will be greatly decreased and CF-RAN will still operate with very low blocking probabilities. In future works, we will propose new heuristics in order to reduce the trade-off between latency and blocking probability to further improve the prioritized power-efficient use of the cloud over fog nodes to perform baseband processing.

## REFERENCES


[1] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, 2015.

[2] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networksa technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[3] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.

[4] A. De la Oliva, J. A. Hernández, 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, 2016.

[5] X. Wang, A. Alabbasi, and C. Cavdar, "Interplay of energy and bandwidth consumption in CRAN with optimal function split," in *Communications (ICC), 2017 IEEE International Conference on, 21-25 May 2017, Paris, France*. IEEE conference proceedings, 2017.

[6] D. Chitimalla, K. Kondepu, L. Valcarenghi, M. Tornatore, and B. Mukherjee, "5g fronthaul-latency and jitter studies of CPRI over ethernet," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 2, pp. 172–182, Feb 2017.

[7] CPRI - Common Public Radio Interface, http://www.cpri.info/, last Access: 02/02/2017.

[8] D. Nesset, "Ng-pon2 technology and standards," *Journal of Lightwave Technology*, vol. 33, no. 5, pp. 1136–1143, 2015.

[9] "Energy-efficient virtual base station formation in optical-access-enabled Cloud-RAN, author=Wang, Xinbo and Thota, Saigopal and Tornatore, Massimo and Chung, Hwan Seok and Lee, Han Hyub and Park, Soomyung and Mukherjee, Biswanath, journal=IEEE Journal on Selected Areas in Communications, volume=34, number=5, pages=1130–1139, year=2016, publisher=IEEE."

[10] G. B. Figueiredo, X. Wang, C. C. Meixner, M. Tornatore, and B. Mukherjee, "Load balancing and latency reduction in multi-user CoMP over TWDM-VPONs," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[11] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 126–135, 2014.

[12] R. I. Tinini, L. C. M. Reis, D. M. Batista, G. B. Figueiredo, M. Tornatore, and B. Mukherjee, "Optimal placement of virtualized BBU processing in hybrid Cloud-Fog RAN over TWDM-PON," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[13] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[14] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vepc)," *IEEE Network*, vol. 28, no. 6, pp. 18–26, 2014.

[15] R. I. Tinini, D. M. Batista, and G. B. Figueiredo, "Energy-efficient VPON formation and wavelength dimensioning in Cloud-Fog RAN over TWDM-PON," in *2018 IEEE Symposium on Computers and Communications (ISCC)*, June 2018, pp. 521–526.

[16] N. Carapellese, M. Tornatore, and A. Pattavina, "Energy-efficient baseband unit placement in a fixed/mobile converged WDM aggregation network," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 8, pp. 1542–1551, 2014.

[17] X. Wang, L. Wang, C. Cavdar, M. Tornatore, G. B. Figueiredo, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Handover reduction in virtualized cloud radio access networks using TWDM-PON Fronthaul," *Journal of Optical Communications and Networking*, vol. 8, no. 12, pp. B124–B134, 2016.

[18] L. Velasco, A. Castro, A. Asensio, M. Ruiz, G. Liu, C. Qin, R. Proietti, and S. J. B. Yoo, "Meeting the requirements to deploy cloud ran over optical networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 3, pp. B22–B32, March 2017.

[19] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "Graph-based framework for flexible baseband function splitting and placement in C-RAN," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 1958–1963.

[20] S. Park, O. Simeone, and S. Shamai, "Joint cloud and edge processing for latency minimization in fog radio access networks," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.

[21] X. Cui, Y. Jiang, X. Chen, F. Zhengy, and X. You, "Graph-based cooperative caching in fog-ran," in *2018 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2018, pp. 166–171.

[22] H. J. Son and S. Shin, "Fronthaul size: Calculation of maximum distance between RRH and BBU," 2014. [Online]. Available: https://www.netmanias.com/en/post/blog/6276/c-ran-fronthaul-lte/fronthaul-size-calculation-of-maximum-distance-between-rrh-and-bbu

[23] N. N. G. M. Networks, "Project ran evolution: Backhaul and fronthaul evolution," 2015. [Online]. Available: https://www.ngmn.org/publications/all-downloads.html

[24] I. Waßmann, D. Versick, and D. Tavangarian, "Energy consumption estimation of virtual machines," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. ACM, 2013, pp. 1151–1156.

[25] J. Baliga, R. Ayre, W. V. Sorin, K. Hinton, and R. S. Tucker, "Energy consumption in access networks," in *OFC/NFOEC 2008 - 2008 Conference on Optical Fiber Communication/National Fiber Optic Engineers Conference*, Feb 2008, pp. 1–3.

[26] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3g cellular networks," in *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 2011, pp. 121–132.