Energy-Efficient BaseBand Processing via vBBU Migration in Virtualized Cloud-Fog RAN

Rodrigo Izidoro Tinini[1], Daniel Macêdo Batista[1], Gustavo Bittencourt Figueiredo[2], Massimo Tornatore[3] [4],
Biswanath Mukherjee[4]

University of São Paulo[1], *Federal University of Bahia*[2], *Politecnico di Milano*[3], *University of California, Davis*[4]

rtinini, batista@ime.usp.br[1], gustavo@dcc.ufba.br[2], massimo.tornatore@polimi.it[3], bmukherjee@ucdavis.edu[4]

***Abstract*—Cloud-Fog Radio Access Networks (CF-RAN) were proposed as an alternative network architecture to alleviate the high fronthaul capacity requested in traditional Cloud RAN (CRAN) by moving some BaseBand Units (BBUs) from the cloud nodes to fog nodes closer to users. However, when BBU processing is moved into fog nodes, OPEX and CAPEX will increase, and the cost and energy savings introduced by CRAN will also reduce. Moreover, mobile traffic fluctuations may lead to an unbalanced resource utilization and energy-inefficient operation in fog nodes. To address this problem, processing functions in fog nodes could be activated and deactivated in function of network traffic and BBUs placed on fog nodes could be migrated to cloud nodes when network traffic is low. In this paper, we propose an Integer Linear Programming (ILP) formulation to address this dynamic resource allocation problem. By means of Network Functions Virtualization (NFV), virtualized BBUs (vBBUs) can be dynamically allocated and deallocated in fog nodes. Furthermore, considering the availability of cloud nodes and the optical fronthaul, vBBUs can be migrated from fog nodes to cloud nodes in order to balance processing loads and save energy. Compared to a baseline incremental algorithm without vBBU migration, our proposal reduces blocking probability in 89% and achieves power savings of 38%, while providing a very small rate of service interruption due to vBBUs migration.**

***Index Terms*—5G networks, CF-RAN, VPON, NFV**

I. Introduction

Cloud Radio Access Networks (CRANs) were proposed to reduce cost and energy footprint of traditional Distributed RAN (DRAN). In CRAN, BaseBand Units (BBU) are moved from cell sites to a BBU pool located in the cloud so that a single infrastructure can be used to implement BBU processing of several cell sites, saving OPEX and CAPEX. At cell sites, High Power Nodes (HPN) are replaced by Remote Radio Heads (RRH) responsible for gathering user equipments (UEs) baseband signals and transmitting them for processing in the cloud. The cloud is composed of dedicated servers where Virtual Digital Units (VDU) are executed, working as containers of virtual BBUs (vBBUs), thus forming a virtualized BBU pool [2], in which virtualized baseband processing is performed. A virtualized BBU (vBBU) pool can offer many advantages for the network operation, e.g., it enables dynamic vBBU deployment to adapt to varying traffic profiles [9]. Furthermore, as BBUs from different cell sites are virtualized on the vBBU pool, vBBU intercommunication can support mechanisms such as enhanced ICIC and Coordinated Multi-Point (CoMP), that require low delay between the vBBUs.

The centralization of BBUs enables significant cost savings, but it also imposes high traffic demands on the optical fronthaul, as baseband signals generated by RRHs must be transported by the bandwidth-intensive Common Public Radio Interface (CPRI) protocol [2]. In CPRI protocol, line rates range from 614.4Mbps to 24.3Gbps depending on Multiple Input-Output (MIMO) configuration of RRHs. Moreover, the CPRI-based fronthaul transport is subject to a delay limit of 3ms between the RRH and its corresponding BBU [10]. So, as the fronthaul becomes more congested it will be harder to operate under the delay threshold due to possible queuing of RRHs transmissions and processing demands in the cloud [7], which may lead to reduced wireless coverage or even denial of services.

To alleviate the bandwidth requirements of the fronthaul, in a previous work [6], we proposed an architecture called Cloud-Fog RAN (CF-RAN) that exploits fog computing to place vBBUs into fog nodes close to users. In CF-RAN, the optical fronthaul is extended with optical links connecting RRHs to fog nodes. As traffic demands grow, fog nodes are activated and vBBUs are dynamically instantiated using the emerging Network Functions Virtualization (NFV) paradigm to support BBU processing. Thus, when a RRH is active, the operator must find a processing node to instantiate a vBBU for baseband processing and, when a RRH is deactivated, its instantiated vBBU may be turned off. However, in [6] we only focused on static network operation and did not investigate the effects of migrating vBBUs to balance load in CF-RAN. In realistic scenarios, mobile traffic load is highly dynamic and follows different patterns according to localization and time of the day. Hence, a different number of RRHs would be activated along the day to support such variable load. Furthermore, this load fluctuation also leads to an unbalanced utilization of fog nodes, since VDUs get less loaded when RRHs are turned off, and their corresponding vBBUs are deactivated. Hence, some vBBUs in lightly-loaded fog nodes could be migrated to the cloud in order to turn the fog nodes off, balance the load and save power.

When a vBBU is deployed or migrated, the network oper-

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.

ator must also allocate enough bandwidth to transmit CPRI traffic from the RRH to the vBBU. Several works assume that the fronthaul links can be implemented over the channels of a Time-and-Wavelength Division Multiplexed Passive Optical Network (TWDM-PON) due to its high bandwidth, low transmission delays and low costs of operation [6] [8]. An alternative architecture that allows the creation of Virtual PONs (VPON) on top of a TWDM-PON was suggested in [2] [7] [8]. In such architecture, a dedicated PON is created to transmit the CPRI traffic from a group of RRHs to a common processing node, so that several RRHs share a virtualized BBU pool as well as the PON capacity through Time-Division Multiplexing (TDM).

As vBBUs are dynamically deployed and deactivated in cloud or fog nodes, VPONs must be dynamically created to support such variable bandwidth, and they must be properly scheduled to efficiently utilize the limited wavelength capacity of the TWDM-PON. This characterizes a joint optimization problem in which operators must decide how to optimally place or migrate vBBUs and how to create VPONs to support CPRI transmission of a newly deployed/migrated vBBU.

So, in this paper we propose an approach based on an Integer Linear Programming (ILP) model to dynamically decide the vBBU placement, creation of VPONs and migration of vBBUs via reconfiguration of VPONs and reallocation of processing resources, following traffic fluctuations. Simulations show that the dynamic placement and migration of vBBUs lead to significant savings in power consumption and also to low blocking probability rates. The rest of the paper is organized as follows: Section II discusses some relevant related works; Section III introduces the proposed virtualized CF-RAN architecture; Section IV describes the problem of dynamically placing vBBUs and creating VPONs; Section V presents the proposed ILP model; in Section VI results obtained from simulations are shown; Section VII concludes the paper.

II. Related Work

Some recent works have investigated how to efficiently place BBUs and/or create VPONs to support transmissions to the cloud and to fog nodes under static traffic settings. However, these works only focus on static network scenarios and algorithms to handle dynamism and migration of baseband processing in mobile traffic are often neglected. In [12], authors proposed the reconfiguration of VPONs as traffic changes in order to reduce power consumption. However, this study only considered the reconfiguration of the VPONs in a centralized scenario. The novelty of our work resides in minimizing active resources by promoting the reconfiguration of not only VPONs, but also processing resources (VDUs) in a distributed fog scenario where each processing resource and even processing nodes can be dinamically turned off. In [7], authors proposed a TWDM-PON based fronthaul that uses VPONs to reduce processing latency in CRAN. In [8], the dimensioning of wavelengths and VPONs creation to support transmissions both to the cloud and to fog nodes were explored. In this work, the amount of VPONs created on the fronthaul was decided as a function of the free baseband processing capacity on the cloud. Authors also found that power consumption from processing and transmission can be reduced when the number of created VPONs is minimized. However, the impact of traffic fluctuations on the fronthaul and allocated resources was not considered, and only static traffic profiles were considered. In [2], the relation of bandwidth availability and power consumption in static networks was explored when CPRI traffic is split between cloud and fog nodes in a Hybrid CRAN (H-CRAN). Authors showed that, when more bandwidth is available on fronthaul, BBUs can be moved to the cloud to save energy. However, the specific access network technology to connect fog nodes and the allocation of transmission channels for fog nodes was left as an open problem. Overall, fog architectures and service migration introduce some trade offs between power, blocking, service disruption and bandwidth usage that shall be accurately explored in order to maintain a balanced network operation. However, none of these works explored the characteristics and reasons of these trade offs. In next section we present the problems of dynamic vBBU placement and creation of VPONs.

III. System Architecture

Our proposed CF-RAN architecture is composed of both fog nodes and cloud nodes (see Fig. 1 (a)). Each of these nodes can host a set of virtual containers of baseband processing called VDUs. A VDU is responsible for implementing the vBBUs, which are responsible for the virtualized baseband processing of CPRI traffic. In each VDU, the number of vBBUs is limited by its processing capacity. If some VDU has no longer capacity to host a vBBU, a new VDU can be dynamically turned on to deploy new vBBUs. To provide control signaling between RRHs with vBBUs placed in different VDUs, an interconnection element like a backplane ethernet switch is used to redirect traffic between VDUs. In fact, the backplane ethernet switch implements the $X2$ interface used to control signaling in traditional LTE-networks. The transport segment of the CF-RAN is composed of TWDM-PON links connecting RRHs to fog nodes and to the cloud through the fronthaul. A level 1 internal optical splitter is used to multiplex incoming traffic from RRHs to fog nodes and to a feeder fiber connected to a level 2 optical splitter (see Fig. 1 (b)). The level 2 optical splitter multiplexes several distribution fibers to the cloud nodes. Each processing node is equipped with a virtualized Optical Line Terminal (OLT) that is responsible for allocating wavelengths to Optical Nework Units (ONUs) connected to RRHs in order to transmit to that node. Each ONU is equipped with tunable lasers that can be tuned to any available wavelength. A VPON is created for a group of RRHs when the OLT of a processing node allocates the same wavelength to their ONUs. Note that, the proposed architecture is not the ITU-T standard architecture for TWDM-PON, but an upgrade architecture where wavelengths can be received or transmitted in different locations and thus at different optical splitter.

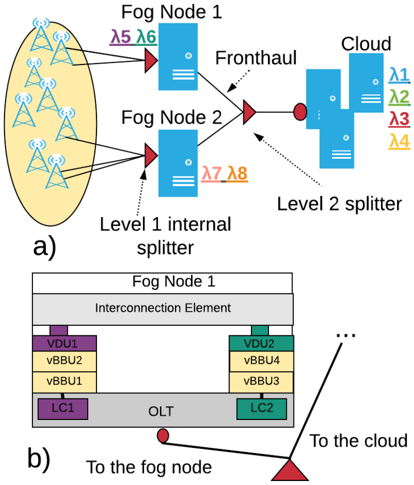

Fig. 1. a) CF-RAN architecture and example of wavelength dimensioning through cloud and fog nodes; b) Internal architecture of fog/processing node

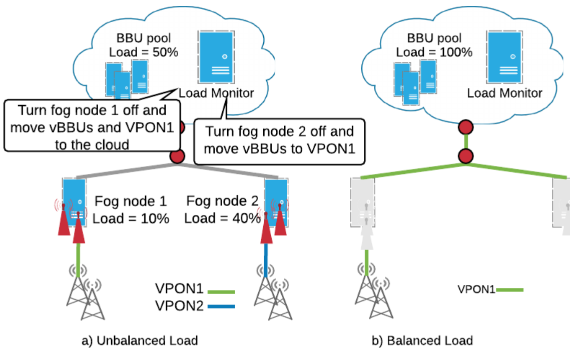

Fig. 2. a) Processing nodes unbalanced before vBBU migration, b) Processing nodes balanced after vBBU migration

Each OLT is equipped with a set of Line Cards (LCs) (Fig. 1 (b)), i.e. transceivers, that are tuned to transmit on a specific wavelength (a VPON) and that is associated with a single VDU. So, the OLT is responsible for switching the incoming traffic from different VPONs to its associated VDU. Moreover, we assume that LCs can be dynamically turned on or off as VPONs are created or deactivated. Hence, if less VPONs are active in a node, less transceivers are used, which results in power savings. Note that, when using the backplane ethernet switch, a single VPON can redirect traffic to VDUs associated to other LCs if the VDU associated to its LC has no longer capacity, but this may increase processing delays and the power consumed by the backplane ethernet switch.

IV. Joint Optimal Baseband Placement and VPON Creation

Placement of baseband processing functions in CF-RAN involves two major steps: first, a processing node must be chosen to support vBBUs; second, a VPON must be created to support the transmission of CPRI traffic to that processing node.

We consider that vBBUs can be deployed according to traffic demands. As soon as a RRH becomes active, a vBBU must be placed in a processing node with enough processing capacity. As power consumption in CF-RAN mostly comes from active processing nodes, to save power, first, only VDUs in the cloud are activated. After vBBUs are deployed in the cloud, the operator must decide how many VPONs will be created. In this study, we assume that the number of VPONs created in the fronthaul is a function of the amount of baseband processing capacity of the cloud. For instance, if 10 RRHs with a 2X2 $20MHz$ MIMO configuration (each of them requiring 2.4Gbps) must be supported in the cloud, three VPONs with 10Gbps capacity are necessary in the fronthaul to support the 24Gbps generated by these 10 RRHs [10].

Only when fronthaul or cloud processing capacity is exhausted, the fog nodes are activated to deploy VDUs. Note that, activating more fog nodes frees up fronthauling capacity, but it also increases OPEX, since cooling and processing functions (VDUs and vBBUs) need to be activated accordingly.

Similarly, after vBBUs are deployed on fog nodes, VPONs must be created by OLTs in fog nodes to transmit CPRI traffic to fog nodes. Since the number of wavelengths is limited, the number of VPONs created by the cloud OLT must be accurately dimensioned so that other wavelengths can be used to create VPONs by the fog nodes OLTs if fog nodes are activated. Instead, if all wavelengths are used to create VPONs in the cloud, it will not be possible to transmit CPRI traffic to vBBUs deployed in fog nodes. An example of an optimal VPONs dimensioning between cloud and fog nodes is depicted in Fig. 1 (a). Considering that vBBUs were placed both in cloud and fog nodes, in order to support transmissions to the cloud, VPONs 1, 2, 3 and 4 were created on the fronthaul by the cloud OLT. To support transmissions to vBBUs in fog nodes, VPONs 5, 6 were created by the OLT of fog node 1 and VPONs 7, 8 by the OLT of fog node 2.

However, as traffic fluctuates, the number of necessary deployed vBBUs and VPONs changes. In the next subsection we propose a scheme for re-deploying vBBUs and turn off fog nodes as traffic fluctuates.

A. *Dynamic Rearrangement of vBBUs and VPONs*

After the initial placement of vBBUs and VPON creation are done, if traffic load fluctuates, it may be necessary to turn off fog nodes and migrate its hosted vBBUs to the cloud. So, as vBBUs are turned off, VDUs in the cloud and in the fog nodes may become lightly-loaded. In this case, vBBUs and VPONs placed in fog nodes could be migrated to the cloud and fog nodes deactivated to save power and bandwidth. In order to achieve this, we propose a vBBU migration mechanism relying on NFV and Live Migration (LM) capabilities.

The vBBU migration mechanism is composed of a load monitor that is implemented on top of a VDU in the cloud. This load monitor measures the current load on each VDU of the network. It considers a load threshold value to be checked every time a vBBU is turned off. When a certain load threshold is reached, it decides if it is convenient to turn off VDUs in a fog node and migrate its vBBUs to the cloud. If the cloud has enough capacity on its VDUs, vBBU migration is performed.

This process is depicted in Fig. 2. As the BBU pool in the cloud has enough capacity to support vBBUs from fog nodes 1 and 2, those vBBUs are migrated to the cloud. Note that before the vBBU migration, two VPONs were used to transmit to fog nodes. As the amount of vBBUs migrated can be supported by VPON1, VPON2 is turned off and its ONUs are reconfigured to transmit to VPON1. However, while the migration is in process, the mobile service gets interrupted. So, the vBBU migration scheme must seek to optimize the network resources performance while minimizing the interruption of service.

V. ILP Formulation

We propose an ILP formulation to decide where to deploy vBBUs and the number of VPONs to transmit CPRI load.

Input Parameters

R: set of RRH traffic demands i, N: set of processing nodes (cloud of fog nodes) n, F_{in}: set of binary values representing fog nodes n connected to RRH i, V_{wn}: set of binary values that represent the availability of each VPON w to be placed on node n, W: set of available wavelengths w and VDUs, B_i: bandwidth demand of RRH i, $B_{|W|}$: capacity of wavelength w, I_w: processing capacity of VDU w, $B_{e_{|N|}}$: bandwidth of the backplane switch e at node n, C_n: power cost of node n, C_{lc}: power cost of a LC, C_e: power cost of the backplane switch, B: a very big positive number.

Decision Variables

x_{iwn}: = 1 if the traffic demand of RRH i is processed at node n being transmitted at the VPON w, u_{iwn}: = 1 if RRH i is processed at the VDU w at node n, y_{in}: = 1 if i was allocated to node n, $x_{|N|}$: = 1 if node n is active, z_{wn}: = 1 if wavelength w transmits to node n, k_{in}: = 1 if traffic from RRH i was redirected to VDU w at node n, r_{wn}: = 1 if VDU w was activated to receive a redirected RRH at node n, s_{wn}: = 1 if VDU w is active at node n, $e_{|N|}$: = 1 if the backplane switch e is active at node n, g_{iwn}: auxiliary variable that equals 1 if traffic of RRH i is redirected to VDU w at node n.

Objective Function

The objective function (1) minimizes both the active processing nodes and the VPONs in order to provide a power-efficient operation while optimally using the available bandwidth.

$$(1) Min. \ \sum_{n=1}^{|N|} x_n.C_n + \sum_{w=1}^{|W|} \sum_{n=1}^{|N|} z_{wn}.C_{lc}$$

Constraints

$$(2) \sum_{w=1}^{|W|} \sum_{n=1}^{|N|} x_{iwn}{=}1, \forall i{\in}R, \ (3) \sum_{w=1}^{|W|} \sum_{n=1}^{|N|} u_{iwn}{=}1, \forall i{\in}R$$

$$(4) \sum_{i=1}^{|R|} u_{iwn}{\geq}0, \forall w{,}n{\in}W, N, \ (5) \sum_{n=1}^{|N|} y_{in}{=}1, \forall i{\in}R$$

$$(6) \sum_{n=1}^{|N|} z_{wn}{\leq}1, \forall w{\in}W, \ (7) z_{wn}{\leq}V_{wn}, \forall w{,}n{\in}W{,}N$$

$$(8) y_{in}{\leq}F_{in}, \forall i{,}n{\in}R{,}N, \ (9) \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} x_{iwn}.B_i{\leq}B_{|W|}, \forall w{\in}W$$

$$(10) \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} u_{iwn}{\leq}I_{|W|}, \forall w{\in}W, \ (11) \sum_{i=1}^{|R|} k_{in}.B_i{\leq}B_{e_{|N|}}, \forall n{\in}N$$

$$(12) B.x_{|N|}{\geq}\sum_{i=1}^{|R|} \sum_{w=1}^{|W|} x_{iwn}, \forall n{\in}N, \ (13) x_{|N|}{\leq}\sum_{i=1}^{|R|} \sum_{w=1}^{|W|} x_{iwn}, \forall n{\in}N$$

$$(14) B.z_{wn}{\geq}\sum_{i=1}^{|R|} \sum_{n=1}^{|N|} x_{iwn}, \forall w{\in}W, \ (15) z_{wn}{\leq}\sum_{i=1}^{|R|} \sum_{n=1}^{|N|} x_{iwn}, \forall w{\in}W$$

$$(16) B.y_{in}{\geq}\sum_{w=1}^{|W|} x_{iwn}, \forall i{,}n{\in}R{,}N, \ (17) y_{in}{\leq}\sum_{w=1}^{|W|} x_{iwn}, \forall i{,}n{\in}R{,}N$$

$$(18) B.y_{in}{\geq}\sum_{w=1}^{|W|} u_{iwn}, \forall i{,}n{\in}R{,}N, \ (19) y_{in}{\leq}\sum_{w=1}^{|W|} u_{iwn}, \forall i{,}n{\in}R{,}N$$

$$(20) B.s_{wn}{\geq}\sum_{i=1}^{|R|} u_{iwn}, \forall w{,}n{\in}W{,}N, \ (21) s_{wn}{\leq}\sum_{i=1}^{|R|} u_{iwn}, \forall w{,}n{\in}W{,}N$$

$$(22) B.k_{in}{\geq}\sum_{w=1}^{|W|} g_{iwn}, \forall i{,}n{\in}R{,}N, \ (23) k_{in}{\leq}\sum_{w=1}^{|W|} g_{iwn}, \forall i{,}n{\in}R{,}N$$

$$(24) B.r_{|W|}{\geq}\sum_{i=1}^{|R|} \sum_{n=1}^{|N|} g_{iwn}, \forall w{\in}W, \ (25) r_{|W|}{\leq}\sum_{i=1}^{|R|} \sum_{n=1}^{|N|} g_{iwn}, \forall w{\in}W$$

$$(26) B.e_{|N|}{\geq}\sum_{i=1}^{|R|} k_{in}, \forall n{\in}N, \ (27) e_{|N|}{\leq}\sum_{i=1}^{|R|} k_{in}, \forall n{\in}N$$

$$(28) g_{iwn}{\leq}x_{iwn} + u_{iwn}, \forall i{,}w{,}n{\in}R{,}W{,}N,$$

$$(29) g_{iwn}{\geq}x_{iwn} - u_{iwn}, \forall i{,}w{,}n{\in}R{,}W, N$$

$$(30) g_{iwn}{\geq}u_{iwn} - x_{iwn}, \forall i{,}w{,}n{\in}R{,}W, N$$

$$(31) g_{iwn}{\leq}2 - x_{iwn} - u_{iwn}, \forall i{,}w{,}n{\in}R{,}W, N$$

Constraints 2 to 5 ensure that each RRH i is allocated to only one processing node, VPON and VDU. Constraint 6 ensures that a VPON is allocated to at most one processing node. Constraint 7 verifies that a RRH i is allocated only to the cloud or to the fog node connected to it. Constraint 8 sets the possible set of nodes that each VPON can be assigned to. Constraints 9, 10 and 11 impose the capacities of VPONs, nodes and the backplane switch capacity. Constraints 11 to 14 activate processing nodes and VPON when demands are allocated to them. The rest of constraints enforce the activation of the backplane switch on each node and the activation of additional VDUs in case of traffic redirection.

The ILP can find the optimal solution for a single request, as well as for batch of requests. As network operation is highly dynamic, we also propose an algorithm that executes the ILP, called nfvILP, to handle dynamic traffic and to implement the vBBU migration mechanism. Algorithm 1 formally describes nfvILP. It first processes each request i incrementally and checks if a vBBU migration is necessary (lines 2 and 10). If so, a batch containing the previously allocated requests is formed (line 3) and the ILP is executed for the batch (function $ILP(.)$ in line 4). If not, it searches the optimal allocation of i (line 10). When an optimal solution is found (both in case of a single request or of a batch), the network state is updated (lines 6 and 12). If no optimal solution is found, an incoming request is blocked (lines 8 and 14). For the vBBU migration, the same algorithm is executed when a request departs. So, when vBBU migration is triggered, a batch of formed allocated vBBUs is created (line(3)) and the ILP is executed to find a new optimum solution (lines 17 and 18) for the current allocated requests in order to minimize the active resources by performing vBBU migration. If a new optimal solution is found, vBBUs are migrated and the network state is updated (line 20). If not, the migration is not performed (line 22).

Algorithm 1 nfvILP

```
Input: RRH request $i$, Departing request $j$, Set of allocated Requests $B$
Output: Optimum Allocation of $i$ — Migration of vBBUs
 1: for all Incoming request $i$ do
 2:     if Load threshold was reached? then
 3:         $B' \leftarrow B + i$
 4:         Run ILP(B')
 5:         if B' has a optimum global then
 6:             nfvUpdate($N_{state}$)
 7:         else
 8:             Blocks $i$
 9:     else
10:         Run ILP(i)
11:         if Optimum global found for $i$ then
12:             nfvUpdate($N_{state}$)
13:         else
14:             Blocks $i$
15: for all Departing request $j$ do
16:     if Load threshold was reached? then
17:         $B' \leftarrow B$
18:         Run ILP(B')
19:         if B' has a optimum global then
20:             nfvUpdate($N_{state}$)
21:         else
22:             Do not migrate vBBUs
```

VI. Illustrative Numerical Examples

To evaluate our proposal, we developed an event-driven simulator and used the DOCPLEX Python API to implement the ILP. Our simulations consider a CF-RAN composed of 1 cloud and 4 fog nodes. Each processing node has 6 VDUs. In the cloud, each VDU can deploy 3 RRHs whereas in the fog each VDU can deploy just 1. The TWDM-PON has 6 wavelengths with capacity of 10Gbps. We use the same power model of [2]. Finally, our scenario has 42 RRHs (maximum load of the processing nodes). We considered a typical 24 hours business traffic profile slotted in periods of 1 hour as shown in Fig. 3 [4]. Regarding the vBBUs migration, we considered a Live Migration scheme where each vBBU takes approximatelly a worst upper bound of 1.5ms to be migrated through a dedicated 10Gbps out-of-band optical channel from the fog node to the cloud [11].

As simulation starts, RRHs are turned off and begin to be activated following a Poisson process with mean equal to $(e/60)$, where e is the erlang for a given hour during the simulation. Moreover, each RRH has a service time uniformly taken from $(0.25$ hour, 1 hour), which means that each RRH can stay turned on from 25 minutes to 1 hour. We compare our proposal to an incremental algorithm called incILP, that incrementally runs the ILP formulation everytime each RRH becomes active, but does not perform vBBU migration when RRHs and vBBUs are turned off due to traffic fluctuations. In our simulations, we studied the effects of vBBU migration for different values of load threshold in VDUs used for initializing the vBBU migration. The load threshold value corresponds to the free processing capacity of a node, e.g., a value of 0.8 means that the VDUs from a node has 80% of free capacity.

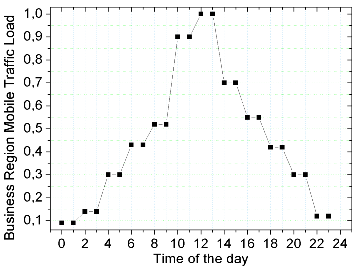

Fig. 3. Traffic load pattern

Fig. 4 (a) show the power consumption achieved by our algorithm. We observed that our proposal achieves higher power savings. Compared to incILP, nfvILP achieves up to 38% power savings in peak hours. It is also possible to see that power consumption is small when vBBU migration is triggered sooner. This is so because the sooner the vBBU migration occurs, the sooner the traffic is moved to the cloud and unbalanced resources are turned off. Regarding blocking probability, we can observe that the vBBU migration plays an important role in reducing blocking, being able to reduce it up to 89% in comparison to incILP for the load threshold of 0.8. We can also observe that, when more time is taken to start vBBU migration, blocking tends to be more reduced. The bandwidth wastage (defined as $1-(T_{cpri}/T_{vpons})$, where T_{cpri} is the total CPRI flow and T_{vpons} the amount of available bandwidth) is also reduced when vBBU migration is performed, as shown in Fig. 4 (c). This happens because, when vBBU migration is performed, nfvILP re-arranges the active resources in an optimal scheduling that leads to the minimization of necessary VPONs to support the network demands. In comparison to incILP, the rate of bandwidth wastage can be reduced to at most 32% by nfvILP.

In Fig. 4 (d), the average amount of vBBUs migrations are presented. Note that the highest amount of migrations occurs at the load threshold of 0.2, when fog nodes workload are close to 100% and a higher number of vBBUs can be migrated. This shows a clearly interplay between the power consumption and the triggering of vBBUs migration, as the lowest power consumption in achieved with the load threshold of 0.2. On the other hand, although a higher number of vBBUs expecting to be migrated leads to power-efficiency, an early re-arrangement of the network resources leads to an inefficient performance and use of the network resources, leading to the highest blocking probability experienced by nfvILP.

Fig. 5 (a) shows the average time that the baseband processing is interrupted in the network due to vBBU migration. Note that the vBBU migration only interrupts the baseband processing for a few seconds, which will leave to significantly gains in power consumption and blocking probability, as we observed in Figs. 4 (a) and (b). The percentage of time that the baseband processing is interrupted is shown in Fig. 5 (b). Note

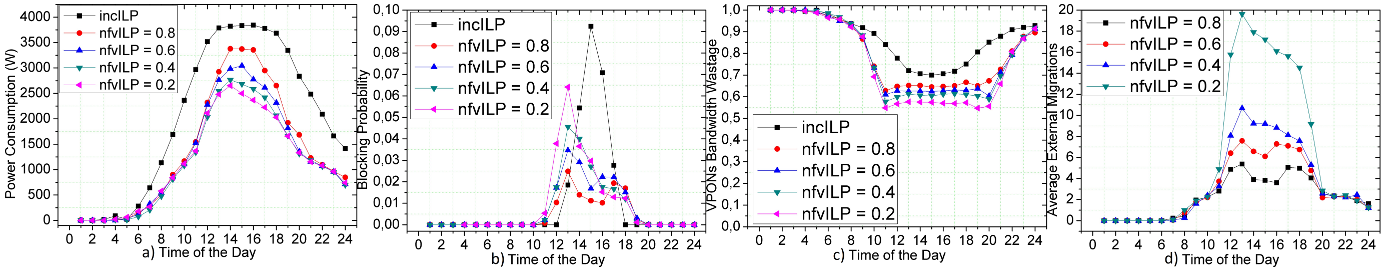

Fig. 4. a) Power Consumption, b) Blocking Probability, c) Bandwidth Wastage, d) Average of External vBBUs Migrations

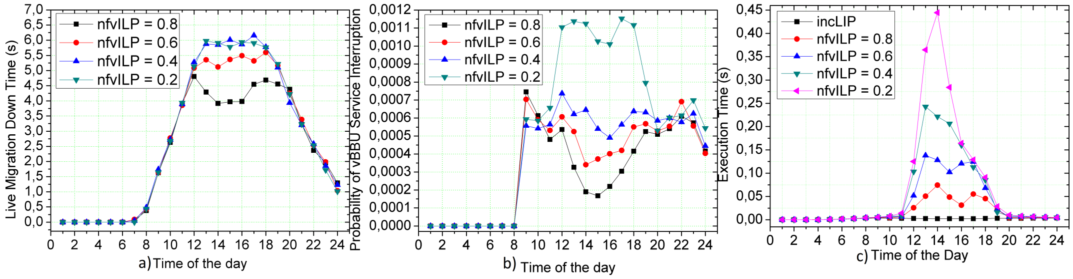

Fig. 5. a) Average down time of vBBUs during migration, b) Percentage of time that vBBUs service are interrupted due to migrations, c) Execution time of the algorithms

that the precentage of time in which the service is interrupted is extremely low for all values of load threshold, being under 0.01%. It is important because it shows that our proposal does not interrupts the UEs connectivity constantly, which is important in order to keep a good Quality of Service (QoS).

Finally, Fig. 5 (c) shows the execution time of nfvILP. In comparison to incILP, it shows an increased execution time, due to the fact that when vBBU migration is triggered, the size of the input to the algorithm is increased in function of the size of the batch of RRHs. The execution time of nfvILP tends to increase as the threshold value decreases, because with low values of load threshold, more vBBUs in fog nodes expects to be migrated as the migration process occurs early.

VII. Conclusion

In this paper we proposed an ILP formulation to dynamically allocate processing and network resources and perform vBBU migrations in a 5G CF-RAN. This ILP formulation allows to dynamically solve the problem of placing vBBUs and creating VPONs for groups of RRHs in CF-RAN as the traffic demands fluctuates over a day. Our proposal is able to handle the network dynamism and achieves significant power savings and reductions in blocking probability in comparison to an incremental algorithm without vBBU migration. Finally, our proposed vBBU migration scheme provides very small times and rates of service interruption. In our future works we will propose a relaxed version of the ILP to solve this problem to even larger instances.

References

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

[2] X. Wang, A. Alabbasi and C. Cavdar, ”Interplay of energy and bandwidth consumption in CRAN with optimal function split,” 2017 IEEE International Conference on Communications (ICC), Paris, 2017, pp. 1-6.

[3] Y. Y. Shih, W. H. Chung, A. C. Pang, T. C. Chiu and H. Y. Wei, ”Enabling Low-Latency Applications in Fog-Radio Access Networks,” in IEEE Network, vol. 31, no. 1, pp. 52-58, January/February 2017.

[4] C. Peng, S. B. Lee, S. Lu, H. Luo, and H. Li. 2011. Traffic-driven power saving in operational 3G cellular networks. In Proceedings of the 17th annual international conference on Mobile computing and networking (MobiCom ’11). ACM, New York, NY, USA, 121-132.

[5] M. Peng, S. Yan, K. Zhang and C. Wang, ”Fog-computing-based radio access networks: issues and challenges,” in IEEE Network, vol. 30, no. 4, pp. 46-53, July-August 2016.

[6] 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,” GLOBECOM 2017 - 2017 IEEE Global Communications Conference, Singapore, 2017, pp. 1-6.

[7] 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,” 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, 2016, pp. 1-6.

[8] R. I. Tinini, D.M. Batista and G. B. Figueiredo, ”Energy-efficient VPON formation and wavelength dimensioning in Cloud-Fog RAN over TWDM-PON,” 2018 IEEE International Symposium on Computers and Communications, Natal-Brazil, 2018, pp. 1-6.

[9] A. Checko et al. ”Cloud RAN for mobile networksA technology overview.” IEEE Communications surveys and tutorials 17.1 2015: 405-426.

[10] A. de la Oliva, J. A. Hernández, D. Larrabeiti, and A. Azcorra. (2016). An overview of the CPRI specification and its application to C-RAN-based LTE scenarios. IEEE Communications Magazine, 54(2), 152-159.

[11] S. Akoush, R. Sohan, A. Rice, A. W. Moore and A. Hopper, ”Predicting the Performance of Virtual Machine Migration,” 2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Miami Beach, FL, 2010, pp. 37-46.

[12] R. Wang, H. H. Lee, S. S. Lee and B. Mukherjee, ”Energy Saving via Dynamic Wavelength Sharing in TWDM-PON,” in IEEE Journal on Selected Areas in Communications, vol. 32, no. 8, pp. 1566-1574, Aug. 2014.