This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2020.3035546, IEEE Transactions on Green Communications and Networking

Energy-Efficient vBBU Migration and Wavelength Reassignment in Cloud-Fog RAN

Rodrigo Izidoro Tinini[1], Daniel Macêdo Batista[1], Gustavo Bittencourt Figueiredo[2], Massimo Tornatore[3,4], Biswanath Mukherjee[4,5]

[1]*University of São Paulo,* [2]*Federal University of Bahia,* [3]*Politecnico di Milano,* [4]*University of California, Davis,* [5]*Soochow University, P.R. China*

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 Network (CF-RAN) is a new architecture that increases network capacity in Cloud RAN (CRAN) by moving some BaseBand Units (BBU) from cloud to fog nodes closer to Remote Radio Heads (RRH). However, fog nodes increases CapEx and OpEx. Moreover, tidal traffic fluctuations may lead to an energy-inefficient operation if resources becomes lightly loaded. To address this problem, BBUs of fog nodes could be dynamically activated, and following traffic fluctuations, migrated to cloud. By leveraging Network Functions Virtualization (NFV), virtualized BBUs (vBBUs) can be dynamically allocated, deallocated, and migrated from fog nodes to cloud. Moreover, considering a Time-and-Wavelength Division Multiplexed Passive Optical Network (TWDM-PON) fronthaul, traffic can be migrated among virtual Passive Optical Network (VPON) channels to optimize bandwidth usage. In this paper, we propose an Integer Linear Programing (ILP) formulation and an algorithm based on its linear relaxation to solve this migration problem. Compared to an algorithm without migration capabilities, our proposals reduces blocking of RRHs demanding processing of 89% and achieves power savings of 38% by reducing activated processing resources and VPONs after migrations, while experiencing small rates of service interruption. Our relaxation-based solution approximates ILP optimality and reduces execution time of ILP up to 50x.

***Index Terms*—5G networks, CF-RAN, NFV, vBBU Migration, TWDM-PON**

I. Introduction

Radio Access Network (RAN) is a very important building block of 5G mobile networks. An important aspect for RAN performance is where the baseband signals generated from user equipment are processed. In traditional Distributed RANs (DRAN), these signals are processed locally at base stations. However, the local processing of baseband signals demands cooling and power facilities in each base station site, which increases the network CapEx and OpEx. This expensive operation limits the scalability of DRANs.

Unlike DRANs, Cloud Radio Access Networks (CRANs) centralize the baseband signal processing so that costs and energy footprint are minimized. The BaseBand Units (BBU), i.e., the computing elements performing baseband processing, are moved from cell sites to a single facility located on the cloud, called BBU pool. Hence, multiple BBUs can be shared among different base stations, and this drastically reduces OpEx and CapEx for cooling and equipment housing. At cell sites, Remote Radio Heads (RRH) replace the eNodeB. RRHs are responsible only for receiving baseband signals from the user equipment and transmitting them for processing in the cloud. The connectivity between RRHs and BBUs is enabled by a high-capacity network called fronthaul [1] [2] [3].

In the BBU pool located in the cloud, several dedicated servers execute Virtual Digital Units (VDU), which are virtualized containers of Virtualized Processing Functions (VPF). The virtualized BBUs (vBBU) are one of these VPFs, where the virtualized baseband processing is performed. The set comprising all vBBUs forms a virtualized BBU pool [4]. Such structure offers several advantages for network operation, e.g., providing the capability of dynamically adapting to variable network load by dynamic vBBU deployment [5], [6].

However, despite its significant cost savings, BBU centralization also imposes high fronthaul traffic due to the high bit rates of the Common Public Radio Interface (CPRI) protocol [4], used to transport the digitized baseband signals from RRHs to the BBU pool. The CPRI line rates vary from 614.4 Mbps to 24.3 Gbps depending on the Multiple Input-Output (MIMO) configuration of RRHs. In this regard, in high fronthaul demands, the cloud processing capacity may not be sufficient to process all of the transmitted baseband signals.

To increase the CRAN capacity, we proposed an architecture called Cloud-Fog RAN (CF-RAN) in [7]. It leverages both Cloud and Fog computing to deploy a more bandwidth efficient RAN. In CF-RAN, we exploit fog computing by placing vBBUs into fog nodes located closer to the users. To do that, we propose the extension of the optical fronthaul with links connecting RRHs to fog nodes.

Hence, whenever traffic load grows, fog nodes can dynamically turn on a set of vBBUs to process the incoming traffic. Thus, when an RRH is active, the operator must identify a processing node to instantiate a vBBU for baseband processing and, when an RRH is deactivated, its corresponding vBBU may be turned off. In this regard, the management of power consumption plays an important role in CF-RAN. Although fog nodes can increase processing capacity in comparison to CRAN, their activation will increase power consumption, so they must be properly activated or deactivated following traffic.

Whether the vBBU is locally instantiated or deployed in the cloud, the network operator must properly provision bandwidth to transport the CPRI frames from the RRH to the vBBU. In this paper, we consider that the fronthaul

A short, summarized version of this work was presented at the IEEE Globecom 2019 conference in Waikoloa, Hawaii, in Dec. 2019.

is implemented over a Time-and-Wavelength Division Multiplexed Passive Optical Network (TWDM-PON). TWDM-PON is considered due to its low transmission delays, high bandwidth, and low cost of operation [8]. The virtualization potential of TWDM-PON has been explored in [9], where it is shown how virtual PONs (VPONs) can be created by sharing a wavelength among several Optical Network Units (ONUs). The use of VPONs increases bandwidth usage and power savings efficiently, and reduces latency in both CRAN and CF-RAN [4] [6] [8]. As the amount of wavelengths is limited, the amount of VPONs that support transmissions to the deployed vBBUs must be efficiently dimensioned [10].

In our previous work [8], the problem of placing vBBUs and creating VPONs in CF-RAN was addressed in static and dynamic traffic scenarios. However, the effects of traffic fluctuations on processing nodes and VPONs were not explored. In more realistic scenarios, mobile traffic load is tidal, which may cause some RRHs to be turned off during the day due to mobility of user equipments. This will lead to the deactivation of the vBBUs of deactivated RRHs, and processing nodes may become lightly loaded as traffic fluctuates and these vBBUs are turned off. So, to save power, the vBBUs of lightly-loaded fog nodes can be migrated to the cloud to deactivate the fog node. Moreover, the traffic from a lightly-loaded VPONs can also be migrated to another VPON in order to optimize the bandwidth usage. As each VPON consumes power from a Line Card (LC) used to terminate its traffic, additional power savings can be achieved if some LCs are deactivated after its VPONs migrate their traffic to another VPON.

In CF-RAN, the vBBU placement and VPON creation (vBP-VC) requires to solve an optimization problem: i) to place or migrate vBBUs, and ii) to create VPONs to sustain CPRI transmission to deployed/migrated vBBUs. In this paper, we present an Integer Linear Programing (ILP) to perform vBBU placement, creation of VPONs and migration of vBBUs via reallocation of processing resources and migration of traffic among VPONs, following traffic fluctuations. In comparison to [8], the proposed ILP and migration mechanism allows to optimally distribute the traffic in the VDUs of processing nodes and to promote intercommunication among different VDUs that exchange CPRI traffic of a single VPON.

Simulation results show that adapting network resource utilization to tidal traffic behaviour leads to significant power savings and to lower blocking probability. In our scenarios, we refer to a RRH being blocked when there is no capacity both on a processing node and on a VPON to support the processing and transmissions of its CPRI traffic. Moreover, to mitigate ILP scalability problems, we also propose an algorithm based on the relaxation of the ILP model. The relaxation-based solution provides good sub-optimal solutions and is able to significantly decrease the ILP execution times.

The rest of the paper is organized as follows: Section II reviews relevant related works; Section III presents the proposed virtualized CF-RAN architecture; Section IV describes the vBP-VC problem with migration following tidal traffic; Section V presents the ILP model; Section VI presents the algorithm based on the ILP relaxation; in Section VII simulation results are presented and discussed; conclusion and future works are presented in Section VIII.

II. Related Work

The vBP-VC problem has been only recently investigated, especially for static traffic scenarios, i.e., when traffic dynamics and workload migrations are not taken into account.

In [11], power consumption is minimized in CRAN by reconfiguring VPONs based on traffic fluctuations. However, only the centralized architecture was considered. Authors in [4] [6] [12] extended the use of VPONs to CRAN, CF-RAN, and Hybrid CRAN (H-CRAN), in which remote processing units are placed between RRHs and the cloud. To do so, a dedicated PON is formed to transport the CPRI traffic from a group of RRHs to a common processing node. Therefore, several RRHs share a virtualized BBU pool and the PON capacity through Time-Division Multiplexing (TDM). Specifically, in CF-RAN, VPONs must be dynamically created following the placement of vBBUs. Ref. [6] showed that the use of VPONs in a TWDM-PON fronthaul in CRAN can help to reduce the baseband-processing latency. An optimal dimensioning of wavelengths for the CF-RAN fronthaul was proposed in [12], where an ILP formulation was designed to decide the amount of bandwidth to be allocated to cloud and fog nodes. It was shown that, in a static traffic scenario, power consumption could be reduced by minimizing the activated VPONs on the fronthaul links. Authors in [4] explored the impact of functional splits used between the cloud and remote processing nodes and, additional to the fronthaul, they proposed the use of a transport interface called the midhaul to connect RRHs to remote nodes. A constraint-programming model was proposed to decide the split point of baseband processing between the cloud and remote nodes for each RRHs. Results showed a strong relation between power efficiency and fronthaul usage. However, the midhaul architecture, its technology and bandwidth allocation on its links were neglected in this work. Authors in [13] showed that latency of mobile applications can be decreased using fog nodes. However, they did not consider the joint operation of fog and cloud nodes.

In general, the migration of workloads between processing resources in CF-RAN may introduce significant tradeoffs between power consumption, blocking probability, resource usage, and service disruption. These tradeoffs are an important aspect in network operation and are neglected in the works discussed above.

In our previous work [8], the vBP-VC problem in CF-RAN is modelled as an ILP formulation, and heuristics are also proposed for computational efficiency. Although dynamic traffic was addressed, no migration of traffic between processing nodes and VPONs was performed when traffic decreases and resources becomes lightly loaded. Processing nodes and VPONs were only turned off when they finish processing all of their requests. Moreover, in [8], our proposal was agnostic regarding how to deal with traffic fluctuations, vBBU/workload migrations and VPON reconfigurations and no intelligent migration could be made when requests leave the network.

In this regard, our study proposes an intelligent migration mechanism that monitors the network state by controlling the departure of each vBBU from a processing node, and seeks to perform three kinds of migration on CF-RAN when processing resources become lightly loaded. The first kind is inter-node migration, where vBBUs from lightly-loaded fog nodes are moved to VDUs on the cloud. The second is an intra-node migration, where the vBBUs of lighlty-loaded VDUs are migrated to other VDUs to turn off some lightly-loaded VDUs to save power. Finally, CPRI traffic from each vBBU can also be migrated among VPONs to optimize power and bandwidth usage if an optical channel also becomes lightly loaded after an inter or intra-node migration. To model the migration of vBBUs among VDUs both in the inter- and intra-nodes cases and monitor its load, in [10] we presented an ILP formulation. In this current paper, we extend our work in [10] by introducing an algorithm based on ILP relaxation to achieve less computational complexity when solving the migration problem. Moreover, we extend our analysis to new performance metrics to assess the efficiency of the ILP relaxation when performing migrations.

Furthermore, recent works proposed migration algorithms for mobile networks. However, they were not conceived to solve problems in a CF-RAN like architecture, taking into account all of the characteristics of CF-RAN, and thus we did not use them as baselines for comparisons. A brief discussion on this algorithms is as follows. In [14], authors proposed a BBU migration algorithm in a CRAN architecture based on the load of each BBU. However, the scheme presented in this work only considers the operation of the CRAN architecture and does not take into account the offloading characteristics of our proposed CF-RAN architecture and the migration of optical channels from one processing node to another. So, this proposal is very different from our proposal and is incompatible for fair comparison with our migration scheme, as it would demand several modifications to the original algorithm. In [15], authors proposed an optical channel reconfiguration scheme between RRHs-and-BBUs to improve the performance of CoMP techniques in a CRAN architecture. Although it explores the migration of traffic of RRHs among lightpaths, the proposed algorithms did not jointly addressed the BBU migration when reconfiguring RRHs in optical channels. In comparison to our proposal, our scheme aims to optimize the network performance when jointly migrating traffic from vBBUs and optical channels (VPONs) in such a way that the network performance is optimized from optimal and sub-optimal solutions that optimizes both vBBUs and VPONs migrations. In [16], authors studied the migration of application and network functions between nodes of a mobile edge computing (MEC) architecture. However, the work presented in this paper is experimental and the migrations are performed following state-of-the-art migration policies, such as proactively migration, retrospectively and on-the-fly migrations, different from our proposal that is based in a optimal/suboptimal ILP approach. Moreover, the optical fronthaul characteristics of our work is not presented in this paper.

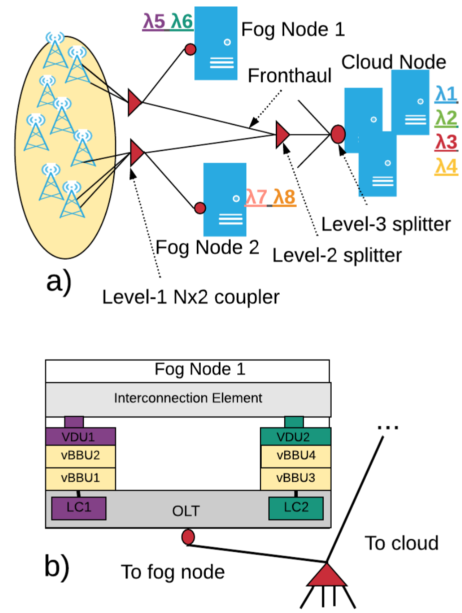


Fig. 1. a) CF-RAN architecture with wavelength dimensioning through fog and cloud nodes; b) Internal architecture of a fog node

III. CF-RAN Architecture

CF-RAN is a hierarchical CRAN architecture with added intermediate fog nodes. As depicted in Fig. 1a, CF-RAN consists of both cloud and fog nodes, divided in two processing layers. Each processing node can support several VDUs. Each VDU is responsible for the deployment of vBBUs. The number of vBBUs instantiated on a VDU is limited by the VDU processing capacity. In case of insufficient VDU capacity, an auxiliary VDU must be turned on to activate more vBBUs. Control signaling (e.g., $X2$ LTE interface) among vBBUs in different VDUs is exchanged through a backplane (e.g., Ethernet) switch. For CF-RAN fronthaul, TWDM-PON links connect the cloud and fog nodes to RRHs by three levels of multiplexing. The traffic from several RRHs is multiplexed in a level-1 optical Nx2 coupler, inside the fog node, and forwarded both to the fog node and to a distribution fiber that is connected to a level-2 optical splitter, as depicted in Fig. 1b. This second level of multiplexing forwards the traffic through a feeder fiber to the cloud. An optional splitter (level 3) can be located in the cloud node to multiplex the traffic from several feeder fibers. CPRI traffic is transmitted through shared optical channels called VPONs. A VPON is a dedicated PON channel created to a group of RRHs that transmits traffic to a common processing node. Virtualized Optical Line Terminals (vOLT) are present in each processing node. The vOLT is responsible for granting wavelength capacity to Optical Network Units (ONUs) connected to the RRHs. So, a VPON is created when a set of ONUs/RRHs shares the same wavelength. Note that this architecture is an upgraded version of the ITU-T standard TWDM-PON architecture [17] where several vOLTs and optical splitters are present in the network, allowing the wavelengths to be transmitted or received in different vOLTs and optical splitters.

As also depicted in Fig. 1b, an vOLT contains a set of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2020.3035546, IEEE Transactions on Green Communications and Networking

transceivers, called Line Cards (LCs), that can be tuned to receive and transmit at a specific wavelength. Moreover, each LC is associated to a VDU. Thus, the traffic of a VPON is switched to its associated VDU by its corresponding LC on the OLT. LCs can be dynamically activated/deactivated as VPONs are created/deactivated. Thus, the minimization of active VPONs leads to less active LCs, which in turn reduces power consumption.

Finally, if the VDU associated to a VPON does not have enough processing capacity to support all the requests being transmitted on this VPON, the backplane switch can be used to redirect the exceeding requests of this VPON to other VDUs, but at the cost of additional switching latency and power consumption from the use of backplane switch ports.

IV. vBBU Placement and VPON Creation (vBP-VC)

We consider a scenario where traffic demands fluctuate daily, and RRHs are activated and shut down according to such variation. Hence, when RRHs become active, the operator must allocate processing capacity in a node to process their CPRI traffic. Note that, to save power, the VDUs of the cloud are activated first as long as the cloud has available processing capacity, since most of the power consumption in CF-RAN is due to the fog nodes' activation.

The vBP-VC optimization problem considered in this paper can be divided into two sub-problems: i) selection and activation of processing nodes to support the vBBUs (vBP); and ii) creation and dimensioning of VPONs to transport CPRI traffic to processing nodes (VC). Formally, this problem can be defined as follows: Let R be a set of ingress CPRI requests i. Given a set N of processing nodes n, a set W of wavelengths, and VDUs w (the cardinality of VDUs in each node is the same of the wavelengths, i.e., $|W|$), find the minimum number of processing nodes, wavelengths, and VDUs to support all requests in R to minimize power consumption. Hence, vBP-VC is a generalization of the well-known 2D Bin-Packing problem, as the traffic demand must be allocated ("packed") on the minimum number of processing nodes and wavelengths while respecting their capacity constraints.

When vBBUs are deployed, one must decide on the number of VPONs that should be created to support transmission to the cloud. During network operation, if fronthaul capacity is exhausted, fog nodes are activated for cloud processing offloading. Likewise, when vBBUs are deployed in fog nodes, additional VPONs must be established to transmit CPRI traffic to them. For that, a vOLT is set in the newly-activated fog node and connected to the ONU located closer to the corresponding RRH.

As an example, in Fig. 1a, we illustrate VPON dimensioning involving cloud and fog nodes (vBBUs are located in both cloud and fog nodes). Each VPON occupies an entire wavelength. VPONs $\lambda 1, \lambda 2, \lambda 3$, and $\lambda 4$ were created under the control of the cloud vOLT to carry traffic to the cloud. VPONs $\lambda 5$ and $\lambda 6$ were created by the vOLT of fog node 1, and VPONs $\lambda 7$ and $\lambda 8$ were created by the OLT of fog node 2 to allow the transmissions to vBBUs in fog nodes.

Note that the creation of VPONs in fog nodes does not increase the network capacity in uplink direction, as each created VPON can only be used to send uplink traffic from its RRHs to its serving fog node. However, it is possible for the fog node to send downlink transmissions to RRHs using any of its activated VPONs. However, a proper Dynamic Bandwidth Assignment (DBA) algorithm to control this operation is out of the scope of this study.

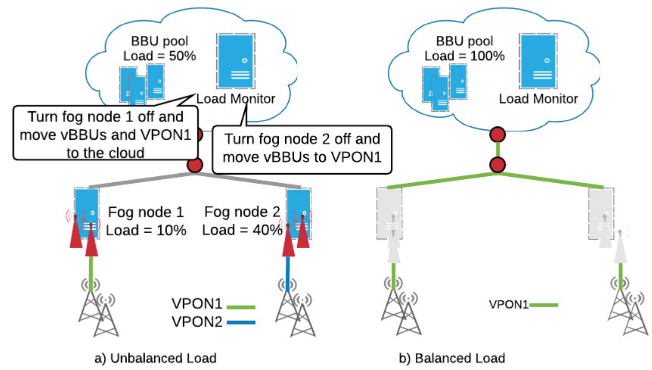


Fig. 2. a) Unbalanced nodes before vBBU migration, b) Balanced nodes after vBBU migration.

A. *Migration of vBBUs and VPONs*

After the preliminary allocation of vBBUs and VPON creation, if traffic load varies and some processing nodes becomes lightly loaded, it may be needed to shut down some fog nodes and migrate their vBBUs to the cloud to save power. It is also possible to migrate vBBUs from a lighlty-loaded VDU to another VDU at the same processing node, thus turning off only the lighlty-loaded VDU. Regarding fronthaul, after migrations are performed among processing nodes, traffic can also be migrated among VPONs, and VPONs can also be migrated to another processing node to optimize bandwidth consumption and deactivate the line cards from lightly-loaded optical channels to save power.

The migration mechanism relies on Live Migration network capabilities and is implemented on top of a VDU in the cloud. It is triggered by a load monitor that keeps track of the actual load on each VDU after a vBBU is turned off due to traffic fluctuation. When a certain load threshold is reached, the load monitor decides on whether to turn off VDUs in a fog node and migrate its vBBUs to another VDU on the fog node or to the cloud. If VDUs on the cloud have available capacity, vBBU migration is executed and vBBUs are moved to the cloud. This process is illustrated in Fig. 2.

Let us consider that, after a traffic fluctuation, several vBBUs, allocated on fog nodes and on the cloud were turned off, and the capacities allocated to them on VDUs and VPONs were released. Given that the BBU pool in the cloud has available capacity to execute the remaining vBBUs of fog nodes 1 and 2, they are migrated to the cloud. Note that, before vBBU migration, two VPONs had been used to transmit their traffic. Considering that the capacity of a VPON is higher than that of a VDU, as traffic demand after migration can still be supported by VPON1, VPON2 is disabled and its ONUs are reconfigured to transmit in VPON1. However, since during migration the mobile service gets interrupted, vBBU

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2020.3035546, IEEE Transactions on Green Communications and Networking

migration might limit the amount of migrations to decrease service interruptions. To do so, in next section, we propose a vBBU migration and VPON/wavelength reassignment subroutine.

V. ILP FORMULATION

In this section, we present an ILP formulation to solve the vBP-VC problem presented in Section IV, while minimizing power consumed by processing and network resources.

Input Parameters

R: set of RRH traffic demands i. N: set of cloud or fog nodes (processing nodes) n. F_{in}: set of binary values representing fog nodes n connected to RRH i. V_{wn}: set of binary values representing if VPON w was allocated to node n. W: set of available wavelengths w and VDUs w in each node (as the cardinalities of wavelengths and VDUs on a node are the same, we use the same variable to represent these two elements). 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 backplane switch e at node n. C_n: baseline power consumption in watts of node n. C_{lc}: baseline power consumption in watts of a LC. B: a very big positive number.

Decision Variables

x_{iwn}: = 1 if traffic demand of RRH i is processed at node n being transmitted at VPON w.

u_{iwn}: = 1 if RRH i is processed at 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 vBBU of RRH i is redirected to VDU w at node n.

r_{wn}: = 1 if VDU w receives a redirected RRH at node n.

s_{wn}: = 1 if VDU w is active at node n.

e_n: = 1 if 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 the power consumed by active processing nodes and created VPONs:

$$Minimize \sum_{n=1}^{N} x_n \cdot C_n + \sum_{w=1}^{W} \sum_{n=1}^{N} z_{wn} \cdot C_{lc} \tag{1}$$

Constraints

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

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

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

$$\sum_{n=1}^{|N|} y_{in} = 1, \forall i \in R \tag{5}$$

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

$$z_{wn} \leq V_{wn}, \forall w, n \in W, N \tag{7}$$

$$y_{in} \leq F_{in}, \forall i, n \in R, N \tag{8}$$

$$\sum_{i=1}^{|R|} \sum_{n=1}^{|N|} x_{iwn} \cdot B_i \leq B_w, \forall w \in W \tag{9}$$

$$\sum_{i=1}^{|R|} \sum_{n=1}^{|N|} u_{iwn} \leq I_w, \forall w \in W \tag{10}$$

$$\sum_{i=1}^{|R|} k_{in} \cdot B_i \leq B_{e_n}, \forall n \in N \tag{11}$$

$$B \cdot x_n \geq \sum_{i=1}^{|R|} \sum_{w=1}^{|W|} x_{iwn}, \forall n \in N \tag{12}$$

$$x_n \leq \sum_{i=1}^{|R|} \sum_{w=1}^{|W|} x_{iwn}, \forall n \in N \tag{13}$$

$$B \cdot z_{wn} \geq \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} x_{iwn}, \forall w \in W \tag{14}$$

$$z_{wn} \leq \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} x_{iwn}, \forall w \in W \tag{15}$$

$$B \cdot y_{in} \geq \sum_{w=1}^{|W|} x_{iwn}, \forall i, n \in R, N \tag{16}$$

$$y_{in} \leq \sum_{w=1}^{|W|} x_{iwn}, \forall i, n \in R, N \tag{17}$$

$$B \cdot y_{in} \geq \sum_{w=1}^{|W|} u_{iwn}, \forall i, n \in R, N \tag{18}$$

$$y_{in} \leq \sum_{w=1}^{|W|} u_{iwn}, \forall i, n \in R, N \tag{19}$$

$$B \cdot s_{wn} \geq \sum_{i=1}^{|R|} u_{iwn}, \forall w, n \in W, N \tag{20}$$

$$s_{wn} \leq \sum_{i=1}^{|R|} u_{iwn}, \forall w, n \in W, N \tag{21}$$

$$B \cdot k_{in} \geq \sum_{w=1}^{|W|} g_{iwn}, \forall i, n \in R, N \tag{22}$$

$$k_{in} \leq \sum_{w=1}^{|W|} g_{iwn}, \forall i, n \in R, N \tag{23}$$

$$B \cdot r_w \geq \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} g_{iwn}, \forall w \in W \tag{24}$$

$$r_w \leq \sum_{i=1}^{|R|} \sum_{n=1}^{|N|} g_{iwn}, \forall w \in W \tag{25}$$

$$B \cdot e_n \geq \sum_{i=1}^{|R|} k_{in}, \forall n \in N \tag{26}$$

$$e_n \leq \sum_{i=1}^{|R|} k_{in}, \forall n \in N \tag{27}$$

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

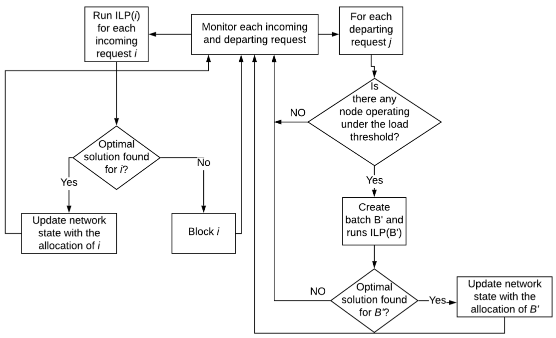


Fig. 3. Flowchart of the vBBU migration scheme.

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

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

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

Constraints 2, 3, 4, and 5 ensure that each RRH request i is processed and transmitted only on a single processing node, VDU and VPON. Constraint 6 and 7 impose that each wavelength can only be allocated to a single processing node, ensuring that two VPONs never share the same wavelength. Constraint 8 guarantees that each RRH will send a CPRI request to be processed either in the cloud or fog node to which it is connected. Constraints 9, 10, and 11 limit the utilization of each wavelength, processing node and internal switches in relation to their capacities. Constraints 12, 13, 14, and 15 activate a processing node n and a VPON w when some demand i uses them for processing and transmission. The remaining constraints ensure the activation of internal switches and auxiliary VDUs in each node in case of traffic switching between VDUs, besides accounting for the number of CPRI requests switched between VDUs.

A. Migration Mechanism

The migration process is based on finding a new vBP-VC solution for all already-allocated requests that yields less power consumption than the current network operation.

The proposed ILP can be used to find the optimal solution for a single request or for a batch of requests. To perform the migration of vBBUs, the ILP performs the re-allocation of a batch of requests being processed on their VDUs. More specifically, the migration sub-routine is implemented using the nfvILP algorithm, that executes the ILP for each single incoming request and triggers the migration mechanism whenever a processing node has its load decreased to a certain threshold after vBBUs are turned off.

The nfvILP algorithm is presented in Fig. 3 and Algorithm 1. For each new request i, the ILP is invoked to solve the vBP-VC for i (line 2). If an optimal solution is found, the network state is updated with the allocation of i (lines 3 and 4). If not, request i is blocked in line 6.

Regarding migration, for each departing request j, it is checked if a processing node is operating under the load threshold value (line 8). If so, the migration is triggered and a batch *B'* containing the requests already allocated is formed in line 9 and the ILP is invoked to search for a new allocation for *B'* in line 10. If a new allocation is found (line 11), the status of the network is updated with the resources allocated for *B'* and migration is performed (line 12). If a new optimal allocation cannot be found for *B'*, migration is not performed (line 14).

Algorithm 1 nfvILP

Input: RRH request i, Departing request j, Set of allocated Requests B
Output: Optimum Allocation of i / Migration of vBBUs and Reconfiguration of VPONs

```
1:  for all Incoming request i do
2:      Run ILP(i)
3:      if Optimum global found for i then
4:          Update network state with allocation of i
5:      else
6:          Block i
7:  for all Departing request j do
8:      if Workload threshold was reached in a processing node then
9:          B'←B
10:         Run ILP(B')
11:         if B' has a optimum global then
12:             Update network state with allocation of B'
13:         else
14:             Do not migrate vBBUs neither traffic among VPONs
```

Execution time of the ILP may become prohibitive in large-scale scenarios. So, in the next section, we propose an approximate solution based on linear relaxation of the ILP, such that a solution can be found in most scenarios.

VI. A Scalable Algorithm Based on Linear Relaxation

The linear relaxation of the ILP consists in removing the integrality constraints of decisions variables, changing the complexity of the problem to polynomial. Hence, the possible values of the decision variables now fall within the interval $[0,1] \in \mathbb{R}$. This means, for instance, that the allocation of CPRI request 1 following variable x_{iwn} may present fractional values; e.g., $x_{111} = 0.5$, $x_{112} = 0.3$, and $x_{113} = 0.2$. This means, in a scheduling context, that CPRI request 1 is allocated in processing nodes 1, 2, and 3. This clearly has no practical sense, as each CPRI request can only be allocated in one processing node.

So, our solution considers the results of relaxed ILP for each decision variable as "probabilities of allocation". Hence, the value for variable x_{iwn} indicates the probability of the *i-th* CPRI request to be transmitted by the *w-th* wavelength/VPON and to be processed on the *n-th* processing node.

Note that the relaxed solution produces degraded solutions in comparison to those provided by the ILP formulation, as in the given example, three processing nodes would be activated to receive CPRI request 1. On the other hand, the execution time is drastically reduced with respect to ILP.

To obtain feasible sub-optimal solutions from linear relaxation, it is necessary to develop effective post-processing algorithms. In this work, we use the randomized rounding approach [18] where a post-processing algorithm chooses the decision variable with the higher "probability value" and rounds it to 1 while the other decision variables are rounded to 0. Then, capacity constraints are checked. The attribution of continuous values to the decision variables might create a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2020.3035546, IEEE Transactions on Green Communications and Networking

situation in which the capacity constraints (Eqs. 9, 10, and 11) are not guaranteed, because those constraints were based on the integrality of the variables to control the use of resources (a node can receive more requests than it can support). Also, Constraint 8 can be removed from the relaxed ILP, as it also relies on a integrality constraint to assure the association of each RRH to a single fog node. Eqs. 8-11 can be removed from the relaxed ILP model, and note that this means that the amount of constraints will be reduced, further contributing to problem complexity reduction. However, all the removed constraints must be treated by the post-processing algorithms so that the relaxed solutions can be feasible. The randomized rounding-based post-processing algorithm will be presented in the next sub-section.

A. Post-Processing Algorithms

The post-processing algorithm to transform the relaxed solution into an integer feasible solution is presented here. Note that only the key decision variables x_{iwn} and u_{iwn} need to be rounded to obtain a sub-optimal solution, as the other decision variables are auxiliary variables used to maintain the constraints.

Algorithm 2 relaxedILP

Input: Set of decision variables x_{iwn} and u_{iwn} provided by the relaxation solution, list of active RRHs i on the network R_{act}, list of previous allocations A, list of fog nodes connected to RRH i R_{fog}
Output: Allocation a of i at node n, wavelength w and VDU w

for all Repetitions **do**
  for all Decision Variable $x \in x_{iwn}$ **do**
    if checkNodeCapacity($x.node$) ¿ 0 **then**
      $R_{act}[x.rrh].node \leftarrow x.node$
    else
      $R_{act}[x.rrh].node \leftarrow R_{fog}.rrh$
    if $R_{act}[x.rrh].node$ is $\emptyset$ **then**
      Blocks request i
    else
      if !checkLambdaInNode($R_{act}[x.rrh].node$) **then**
        if checkLambdaCapacity($x.wavelength,\ R_{act}[x.rrh].node$) **then**
          $R_{act}[x.rrh].wavelength \leftarrow x.wavelength$
        else
          $lambda \leftarrow$ getFreeLambda()
          if $lambda \neq \emptyset$ **then**
            $R_{act}[x.rrh].wavelength \leftarrow lambda$
      else
        $lambda \leftarrow$ getFirstFitLambda()
        if $lambda \neq emptyset$ **then**
          $R_{act}[x.rrh].wavelength \leftarrow lambda$
        else
          $lambda \leftarrow$ getFreeLambda()
          if $lambda \neq \emptyset$ **then**
            $R_{act}[x.rrh].wavelength \leftarrow lambda$
      if $R_{act}[x.rrh].wavelength \neq \emptyset$ **then**
        Blocks request i
      else
        $vdu \leftarrow$ getVDU(x)
        if checkVDUCapacity(vdu, $R_{act}[x.rrh].node$) **then**
          if checkSwitchUse(vdu, $R_{act}[x.rrh].wavelength$) **then**
            $R_{act}[x.rrh].VDU \leftarrow vdu$
            updateSwitch($R_{act}[x.rrh].node$)
          else
            $R_{act}[x.rrh].VDU \leftarrow vdu$
        else
          $Random_{vdu} \leftarrow$ getRandomVDU($R_{act}[x.rrh].node$)
          if checkSwitchUse($Random_{vdu}$, $R_{act}[x.rrh].wavelength$) **then**
            $R_{act}[x.rrh].VDU \leftarrow Random_{vdu}$
            updateSwitch($R_{act}[x.rrh].node$)
          else
            $R_{act}[x.rrh].VDU \leftarrow Random_{vdu}$
    if $R_{act}[x[i]].VDU \neq \emptyset$ **then**
      update($R_{act}[x.rrh].node$, $R_{act}[x.rrh].wavelength$, $R_{act}[x.rrh].VDU$)
    else
      Blocks request i
  $A \leftarrow a$
getBestSolution(A)

This algorithm, called *relaxedILP*, is formally described in Alg. 2. Its first step is to choose the processing node with the highest probability value to receive the CPRI request and then allocate a VPON to support the transmission of this request. Regarding VPON allocation, a previously-allocated VPON on the selected processing node is searched. If more than one VPON is activated on the processing node, the CPRI request is allocated to the one with more free bandwidth. If no VPON can be found on this node, a new VPON will be created from the wavelength with the highest probability value returned from the relaxed algorithm. If this wavelength is already being used, a wavelength that is not yet allocated to any processing node will be selected from a list of non-allocated wavelengths.

If no VPON/wavelength can be found, the request is blocked. Otherwise, the algorithm proceeds to the allocation of a VDU. For the VDU, the algorithm also tries to allocate the VDU with the highest probability value of decision variable u_{iwn}. If it has capacity, it is verified if this VDU demands the use of the internal switch (the chosen VDU is not associated to the LC of the allocated VPON). If the switch is needed, its capacity is checked. If the switch has enough capacity or the VDU does not need it, this VDU is set to accomodate the CPRI request. Otherwise, a random VDU is allocated. However, if no VDU can be found, the request is blocked. Finally, after all repetitions are executed, the function getBest-Solution(.) chooses the best resource allocation in terms of power consumption, receiving the list A as argument, where A contains all the allocations performed by the algorithm.

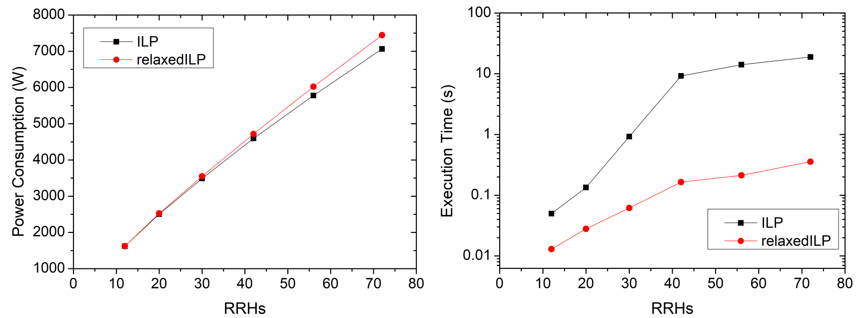
VII. Illustrative Numerical Examples

In this section, we provide numerical evaluations of the proposed ILP and *relaxedILP* in static and dynamic traffic scenarios. All executions of the ILP and *relaxedILP* and simulations on the dynamic traffic scenario were performed using 5GPy simulator [19] with IBM CPLEX on an I7 with 16GB RAM machine running Windows 10.

We performed evaluations on power consumption (defined as Eq. 1), blocking probability (defined as the ratio $R_{Blocked}/R_{Generated}$, where $R_{Blocked}$ is amount of blocked CPRI requests and $R_{Generated}$ amount of generated CPRI requests), bandwidth usage (defined as the ratio T_{cpri}/T_{vpons}, where T_{cpri} is total amount of CPRI flows and T_{vpons} is total amount of available bandwidth in allocated VPONs), and probability of service interruption (defined as the ratio $T_{DownTime}/T_{ProcTime}$, where where $T_{DownTime}$ is average interruption time of the vBBUs when migrated and T_{OpTime} is average operation time of the RRHs). Power consumption parameters are expressed in watts and taken from [8] and [4].

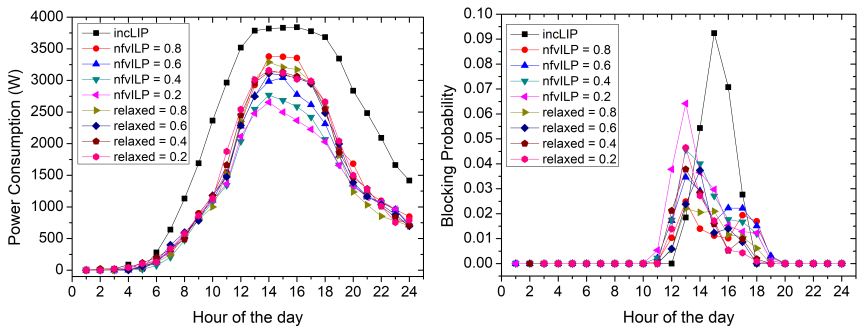
A. Static Traffic Scenario

First, we provide an evaluation of the proposed algorithms for static traffic. For each ILP execution, the network size follows the parameters from Table 1. RRHs are uniformly connected to each fog node. In Fig. 4a, power consumption results



(a) Power Consumption (b) Execution Time in logarithmic scale

Fig. 4. Comparisons between ILP and its relaxed version for static traffic scenario (Power Consumption and execution time).



(a) Power Consumption (b) Blocking Probability

Fig. 5. Comparisons between the ILP and its relaxed version for dynamic traffic (Power consumption a) and blocking probability b)).

TABLE I
NUMBER OF RRHS, PROCESSING NODES, AND WAVELENGTHS FOR EACH ILP EXECUTION.

RRHs	Processing nodes	Wavelengths
12	2 (1 cloud, 1 fog)	3
20	3 (1 cloud, 2 fog)	4
30	4 (1 cloud, 3 fog)	5
42	5 (1 cloud, 4 fog)	6
56	6 (1 cloud, 5 fog)	7
72	7 (1 cloud, 6 fog)	8

for ILP and *relaxedILP* are provided. It can be observed that the *relaxedILP* algorithm provides very good approximations of the ILP, showing a small additional power consumption only for the larger problem instances. Execution times are shown in Fig. 4b. In comparison to the ILP, *relaxedILP* reduces the execution times of up to 50x. The integrality gap for the relaxed ILP and the ILP is equal to 1, meaning that relaxedILP is capable to provide integer optimal solutions.

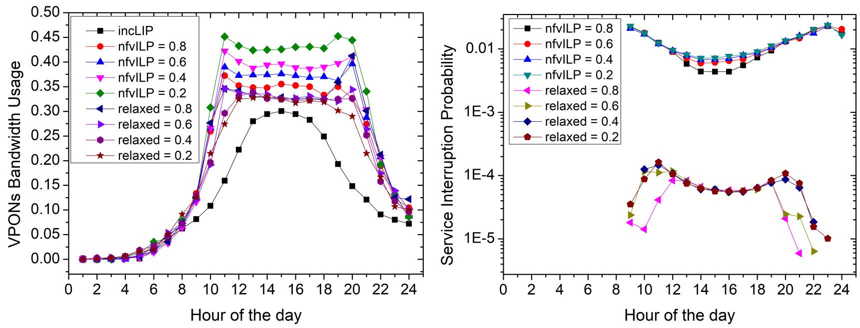
B. Dynamic Traffic Scenario

We also performed simulations for dynamic traffic to assess the efficiency of the proposed algorithm when arrival and departure times of each CPRI request are not known in advance, and resource allocation must be performed at run time for each request. The network is formed by 42 RRHs, 1 cloud node, and 4 fog nodes. The RRHs are uniformly connected to fog nodes. Each processing node implements 6 VDUs, each VDU in the cloud hosts up to 3 vBBUs, and each VDU in a fog node hosts 1 vBBU. The fronthaul is composed of 6 wavelengths of 10 Gbps of capacity. The traffic follows the same daily pattern from [10], and the CPRI requests follow a Poisson process with mean equal to $(erl/60)$, where erl is the traffic load in erlang (each hour of the day has a different erlang value) and each request has a service time uniformly taken from $(0.25$ hour, 1 hour).

We compared the *nfvILP* (proposed ILP with migration sub-routine) and *relaxedILP* with a non-migration baseline algorithm called *incILP*. *incILP* finds the allocation of each incoming CPRI request and is able to move vBBUs from the cloud to the fog when the cloud has no available processing capacity, but does not perform any kind of intra or inter node migration in function of traffic fluctuations. Using *incILP*, we are able to evaluate how the periodical vBBUs migration, among VDUs or among different nodes, affects the daily operation of the network. The fractional values next to the algorithm represents the free load threshold value for a node to trigger the migration process. For instance, a value of 0.2 means that the migration must initiate when any node has 20% of free processing capacity. Simulation results were obtained from the average of 30 runs, obtaining confidence intervals smaller than 5% in all cases, considering a confidence level of 95%.

The daily power consumption is shown in Fig. 5a. The best power consumption for the ILP is achieved at the threshold value of 0.2, especially at peak rate times, between noon and 18p.m.. At these hours, processing nodes will experience higher workloads faster than in low rate hours, meaning that it is better to migrate workload as soon as a processing node is experiencing high workload rates, which will happen more often at peak rates. This can be explained by the fact that, in peak times of operations, more vBBUs will leave the network, leaving gaps in the resources of a processing node, hence leading to more opportunities for the ILP to optimize the operation, as, at this moment several vBBUs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2020.3035546, IEEE Transactions on Green Communications and Networking



(a) Bandwidth Usage

(b) Probability of Service Interruption in logarithmic scale

Fig. 6. Comparisons between the ILP and its relaxed version for dynamic traffic (Bandwidth usage a) and service interruption rate b)).

has left the network and the overall network load may not be optimally distributed among processing resources. Note that *relaxedILP* shows a good approximation of ILP results, with relative increase of power consumption of only 18%. Regarding blocking probability in Fig. 5b, note that the best result for peak hours is obtained for the threshold value of 0.8. This means that, to reduce blocking, migration must be triggered as soon as possible at these hours. This is more clear for peak rates times, and it happens because at higher loads, more resources may be occupied and, due to the high rate of vBBUs leaving the network, the distribution of traffic may be no optimal, so it may be hard to find suitable resources for a incoming request. Hence, there is a clear tradeoff between power consumption and blocking probability when deciding when to migrate vBBUs and VPONs. It is also possible to see that *relaxedILP* provides low blocking probabilities and, compared to ILP, it presents a relative increase of about 88% on blocking rate.

The bandwidth usage is shown in Fig. 6a. The ILP is able to provide a relative increase on bandwidth usage of about 20% in comparison to *relaxedILP*. The best bandwidth usage is obtained for the ILP at the threshold of 0.2. This means that the time taken to start the migration has a significant impact on the bandwidth usage. At peak hours, there is a little decrease on the bandwidth usage, following a little decrease on the number of migrations as it will be shown in Figs. 7.

The probability of service interruption due to migrations is shown in Fig. 6b in log scale. Note that our algorithms are efficient in providing migrations with very small probability of service interruption, especially for the threshold value of 0.8, meaning that, when migration is triggered soon, the rate of interruptions will subsequently decrease in peak hours (i.e., less migration will need to be performed later) due to an early optimal allocation of resources. Note that both the ILP and *relaxedILP* provide very small interruption probabilities, and *relaxedILP* reduces the interruption of service by about 2 orders compared to ILP. Note that in peak hours the service interruptions tends to decrease. This can be explained by the fact that, as each performed migration in peak hours increases cloud workload, cloud capacity will limit the possibility of future inter-node migrations. The little decrease from 8 to 10pm happens because the network demand is still low at these times and only a few vBBUs were migrated to cloud.

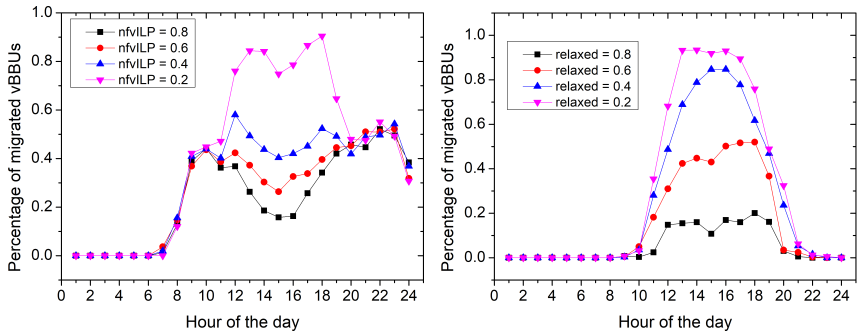
The percentage of migrated vBBUs is shown in Figs. 7a and 7b for the ILP and *relaxedILP*, respectively. Note that threshold value of 0.2 accounts for more migrations in peak hours, which contributes to more power savings as seen in Fig. 5a. Note that, when nodes become close to lightly-loaded state (threshold of 0.8), less migrations are performed, which leads to more blocking as shown in Fig. 5b. As stated before, the little decrease on the number of migrations comes from the fact that each peformed migration will limit the cloud to receive new migrated vBBUs.

Finally, Fig. 8 shows the power consumption as the fog capacity increases in the peak hours for values of 0.2 and 0.8. Note that, if more capacity is available, power is considerably reduced if migration is triggered sooner (0.8). This happens because, when more capacity is available at fog nodes, a single fog node can host more vBBUs, which will lead to less active fog nodes. However, if less migrations are performed (0.2), it will take more time to optimize the distribution of traffic among fog nodes with more free capacity. Note that, when migration is performed later, the network tends to behave like incILP, where no migration is performed.

VIII. Conclusion

In this work, we studied the impact of vBBU migration between cloud and fog nodes and of the reconfiguration of VPONs in a CF-RAN architecture. We modelled the problem as an ILP formulation. To reduce computational complexity, we also proposed an algorithm based on the linear relaxation of the ILP, called *relaxedILP*. The relaxation-based algorithm was able to reduce execution times up to 50x and the quality of solutions were good approximations to the optimal solutions provided by the ILP. We observed that the migration of vBBUs and VPONs among processing nodes allows to optimize power consumption, especially in peak hours. Regarding the moment to trigger the migration process, we observed that, to save power in peak hours, it is better to migrate vBBUs when the node workload is close to its maximum capacity. On the other hand, we observed that blocking and service interruption probabilities are better optimized when migration occurs when the workload on nodes are increasing from a lightly-loaded state to a heavy-loaded one, imposing a clearly tradeoff to power consumption. So, future works should investigate how Machine Learning-based algorithms can perform migration

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2020.3035546, IEEE Transactions on Green Communications and Networking



(a) Percentage of migrated vBBUs for the ILP

(b) Percentage of migrated vBBUs for *relaxedILP*

Fig. 7. Comparisons of the percentage of migrated vBBUs between ILP a) and its relaxed version b) for dynamic traffic.

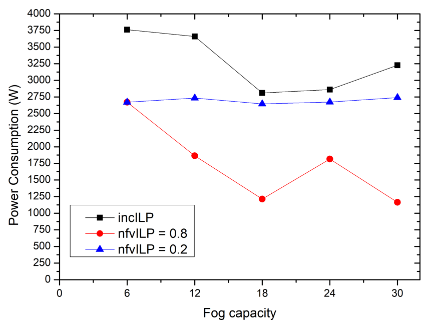


Fig. 8. Power consumption for different fog capacities

while better balancing the tradeoffs between power consumption, blocking, and service interruption probability.

Acknowledgment

This work was 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), CNPq 311608/2017-5, 420907/2016-5, and 312324/2015-4, and FAPESP 18/22979-2. M. Tornatore and B. Mukherjee acknowledge support by the National Science Foundation (NSF) grant number: 1716945.

References

[1] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, and S. Gosselin, "Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks," *Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1963–1970, 2015.

[2] J. Zhang, Y. Ji, S. Jia, H. Li, X. Yu, and X. Wang, "Reconfigurable optical mobile fronthaul networks for coordinated multipoint transmission and reception in 5G," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 6, pp. 489–497, 2017.

[3] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an efficient C-RAN optical fronthaul for the future networks: a tutorial on technologies, requirements, challenges, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 708–769, 2017.

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

[5] 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, 2014.

[6] 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)*. IEEE, 2016, pp. 1–6.

[7] R. I. Tinini, L. C. 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*. IEEE, 2017, pp. 1–6.

[8] R. I. Tinini, D. M. Batista, G. B. Figueiredo, M. Tornatore, and B. Mukherjee, "Low-latency and energy-efficient BBU placement and VPON formation in virtualized cloud-fog RAN," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 11, no. 4, pp. B37–B48, 2019.

[9] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Energy-efficient virtual base station formation in optical-access-enabled cloud-RAN," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1130–1139, 2016.

[10] R. I. Tinini, D. M. Batista, G. B. Figueiredo, M. Tornatore, and B. Mukherjee, "Energy-efficient baseband processing via vBBU migration in virtualized cloud-fog RAN," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[11] R. Wang, H. H. Lee, S. S. Lee, and B. Mukherjee, "Energy saving via dynamic wavelength sharing in TWDM-PON," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 8, pp. 1566–1574, 2014.

[12] 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)*. IEEE, 2018, pp. 521–526.

[13] 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," *IEEE network*, vol. 31, no. 1, pp. 52–58, 2016.

[14] B. Han, L. Liu, J. Zhang, C. Tao, C. Qiu, T. Zhou, Z. Li, and Z. Piao, "Research on resource migration based on novel rrh-bbu mapping in cloud radio access network for hsr scenarios," *IEEE Access*, vol. 7, pp. 108 542–108 550, 2019.

[15] J. Zhang, Y. Ji, S. Jia, H. Li, X. Yu, and X. Wang, "Reconfigurable optical mobile fronthaul networks for coordinated multipoint transmission and reception in 5g," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 6, pp. 489–497, 2017.

[16] I. Sarrigiannis, E. Kartsakli, K. Ramantas, A. Antonopoulos, and C. Verikoukis, "Application and network vnf migration in a mec-enabled 5g architecture," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2018, pp. 1–6.

[17] G. ITU-T, "989.2 40-gigabit-capable passive optical networks 2 (NG-PON2): Physical media dependent (PMD) layer specification," *International Telecommunication Union*, 2014.

[18] P. Raghavan and C. D. Tompson, "Randomized rounding: a technique for provably good algorithms and algorithmic proofs," *Combinatorica*, vol. 7, no. 4, pp. 365–374, 1987.

[19] R. I. Tinini, M. R. P. dos Santos, G. B. Figueiredo, and D. M. Batista, "5GPy: A simpy-based simulator for performance evaluations in 5G hybrid cloud-fog RAN architectures," *Simulation Modelling Practice and Theory*, p. 102030, 2019.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2020.3035546, IEEE Transactions on Green Communications and Networking

Rodrigo Izidoro Tinini received his Computer Science (2011) degree from Municipal University of São Caetano do Sul, M.Sc. in Computer Science (2014) from Federal University of ABC and Ph.D. in Computer Science (2019) from University of São Paulo. He currently works as a postdoctoral researcher at Federal University of ABC. His current research interests are: Optical Networks, mobile fronthaul architectures for 5G Networks, Internet of Things and Artificial Intelligence applied to optical and mobile networks.

Daniel Macêdo Batista is a Professor at the University of São Paulo (USP), the major institution of higher learning and research in Brazil, since 2011. He received his Computer Science (2003) degree from Federal University of Bahia, Brazil, MSc in Computer Science (2006), and Ph.D. in Computer Science (2010) degrees from State University of Campinas (Unicamp), Brazil. Prof. Daniel received the title of Associate Professor in Computer Networks from the University of São Paulo in 2017. Prof. Daniel has published 90+ refereed papers and advised 13 undergrad students, 9 MSc students, and 5 Ph.D. students. His students received several awards, including the IEEE ISCC 2018 Best Student Paper Award and second place at the ACM SAC 2014 Student Research Competition. His main research interests are: IoT Security, B5G, and Data Analytics applied to Computer Networks.

Gustavo Bittencourt Figueiredo received the B.Sc. (2000) in computer science from Salvador University, M.Sc.(2003) and Ph.D. (2009) degrees in computer science, both from University of Campinas (UNICAMP). He is currently an associate professor in the Department of Computer Science of the Federal University of Bahia (UFBA). His main research interests include problems involving Algorithms and optimization, Network Performance Evaluation, Planning, Dimensioning and Optimization of Optical networks.

Massimo Tornatore (S'03–M'06–SM'13) is currently an Associate Professor with the Department of Electronics, Information, and Bioengineering, Politecnico di Milano. He also holds an appointment as Adjunct Professor at University of California, Davis, USA and as visiting professor at University of Waterloo, Canada. His research interests include performance evaluation, optimization and design of communication networks (with an emphasis on the application of optical networking technologies), cloud computing, and machine learning application for network management. In these areas, he co-authored more than 350 peer-reviewed conference and journal papers (with 18 best paper awards), 2 books and 1 patent. He is a member of the Editorial Board of IEEE Communication Surveys and Tutorials, IEEE Communication Letters, Springer Photonic Network Communications, and Elsevier Optical Switching and Networking. He is active member of the technical program committee of various networking conferences such as INFOCOM, OFC, ICC, and GLOBECOM. He acted as technical program chair of ONDM 2016 and DRCN 2017 and DRCn 2019 conferences. He has participated in several EU R&D projects (among others FP7 COMBO, H2020 MEtroHaul, and Cost Action RECODIS) as well as in several projects in USA, Canada and Italy.

Biswanath Mukherjee (S'82–M'84–SM'05–F'07) received the B.Tech. degree from the Indian Institute of Technology, Kharagpur, India, in 1980, and the Ph.D. degree from the University of Washington, Seattle, WA, USA, in 1987. He is a Distinguished Professor and Founding Director of the Institute for Broadband Research and Innovation (IBRI) at Soochow University, P. R. China; and a Distinguished Professor Emeritus at the University of California, Davis, CA, USA, where he was the Chairman of Computer Science during 1997–2000. He was the General Co-Chair of the IEEE/OSA Optical Fiber Communications (OFC) Conference 2011, Technical Program Co-Chair of OFC'2009, and Technical Program Chair of the IEEE INFOCOM'1996 conference. He is Co-Editor of Springer's Optical Networks Book Series. He has served on eight journal editorial boards, most notably IEEE/ACM TRANSACTIONS ON NETWORKING and IEEE NETWORK. In addition, he has the Guest Edited Special Issues of PROCEEDINGS OF THE IEEE, IEEE/OSA JOURNAL OF LIGHTWAVE TECHNOLOGY, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and IEEE COMMUNICATIONS. He has supervised 79 Ph.D.s to completion. He is Winner of the 2004 Distinguished Graduate Mentoring Award, 2009 College of Engineering Outstanding Senior Faculty Award, 2016 International Community Building Award, and the 2019 Faculty Distinguished Research Award at UC Davis. He is the Co-winner of 15 Best Paper Awards, including the 2018 Charles Kao Best Paper Award for IEEE/OSA Journal on Optical Communications and Networks; five from IEEE Globecom Symposia, four from IEEE ANTS, and two from National Computer Security Conference. He is the author of the graduate-level textbook Optical WDM Networks (New York, USA: Springer, Jan. 2006). He served a five-year term on the Board of Directors of IPLocks, a Silicon Valley startup company (acquired by Fortinet). He has served on the Technical Advisory Board of several startup companies, including Teknovus (acquired by Broadcom) and Optella (acquired by Cosemi). He is the Winner of the IEEE Communications Society's inaugural (2015) ONTC Outstanding Technical Achievement Award "for pioneering work on shaping the optical networking area."