Latency-Aware CU Placement/Handover in Dynamic WDM Access-Aggregation Networks

Francesco Musumeci, Omran Ayoub, Monica Magoni, and Massimo Tornatore
Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Milan, Italy
francesco.musumeci@polimi.it; omran.ayoub@polimi.it; monica.magoni@mail.polimi.it;
massimo.tornatore@polimi.it

***Abstract*—Centralized Radio Access Networks (C-RANs) have been recently proposed to cope with the unprecedented requirements of future 5G services, in terms of network capacity, latency, service availability, and network coordination. C-RANs are based on the idea of separating baseband signal processing from the Radio Units (RUs), namely, antenna-sites in the mobile network, in such a way that baseband processing can be eventually concentrated in common locations, the Central Units (CU), that can be shared among several RUs. Although C-RAN brings significant CapEx/OpEx savings, it also requires transport of high-capacity and low-latency fronthaul traffic. Hence, due to the highly-dynamic nature of mobile traffic, a proper placement of CUs in the optical access-aggregation network should adapt to spatio/temporal traffic variation, while maintaining a high degree of RAN centralization and a low service blocking. In this paper, we provide an adaptive latency-aware algorithm for dynamic CU placement in optical access-aggregation networks, which targets the minimization of the number of CUs and also preforms Grooming, Routing and Wavelength Assignment (GRWA) for mobile network traffic demands. When given the possibility to perform *CU handover*, i.e., to move CUs even when they are active, our algorithm, also in high load situations, provides low number of CUs compared with fixed CU placement, and keeps blocking probability within an acceptable range.**

***Index Terms*—C-RAN, eCPRI, fronthaul, low-latency transport, GRWA, CU handover.**

I. Introduction

Telecommunication networks are experiencing a rapid evolution to support emerging bandwidth-intensive and/or low-latency Internet services, such as video streaming, online gaming, augmented reality, Internet of Things, autonomous driving etc., and to sustain the huge growth in the number of devices (e.g., smartphones, tablets, sensors, industrial machineries, etc.) connected to the network. The deployment and management of future-, i.e., fifth-generation (5G) telecommunication networks is challenged by the extremely high performance required by 5G services, in terms of latency, availability, bit-rate, data loss, etc. Such challenges not only impact on the radio interface between eNodeBs and end-users in the mobile Long Term Evolution (LTE) network, but also affects the deployment of the underlying Radio Access Network (RAN), which supports traffic aggregation from eNodeBs and its transport towards the core network infrastructure.

Centralized-Radio Access Network (C-RAN) is a promising architecture to mitigate the aforementioned issues in 5G networks. In C-RANs, the cell-site (CS) equipment is functionally separated into two elements, i.e., a Remote Radio Head (RRH), also known as Remote Unit (RU), which remains located at the antenna premises and is responsible for wireless signal transmission and reception, and a BaseBand Unit (BBU), which performs baseband processing, and which can be located remotely and centralized into common sites.

C-RAN provides significant CapEx/OpEx savings, mainly enabled by simplified antenna architecture, sharing of processing resources and housing facilities among different BBUs, and can effectively support advanced coordination techniques, such as Coordinated Multipoint (CoMP). However, C-RAN requires large amount of *fronthaul* traffic between BBUs and RRHs, which is carried through CPRI interfaces [1]. Morevoer, this traffic must be transported under very low latency constraints, e.g., in the order of few ms. Due to these high-capacity and low-latency requirements, multi-layer optical networks based on OTN over Wavelength Division Multiplexing (WDM) are being deployed for the realization of C-RANs [2].

Considering the expected explosion of 5G traffic and massive deployment of small-cells [3], the aggressive RRH-BBU separation in the original C-RAN architecture is expected to face serious scalability issues due to fronthaul requirements. Therefore, more flexible functional separations are under study [4], which are referred to as *RAN functional splits*. Such flexible solutions are envisioned as an outstanding candidate to help supporting high-bandwidth/low-latency fronthaul traffic and enable effective network reconfiguration and re-adaptation.

Recently, a three-layer functional separation of 5G eNBs (often called gNBs) has been identified and agreed in the context of standardization bodies [5]. As shown in Fig. 1, these three layers are referred to as 1) Remote (or Radio) Unit (RU), indicating the RRH at the antenna site, 2) Distribution Unit (DU) as the element including part of the digital signal processing, possibly providing an amount of functions sharing between several RRHs, and 3) Central Unit (CU), including higher layer (e.g., packet-based) processing, typically located at higher layers in a metro network and associated to several DUs. Correspondingly, besides fronthaul traffic exchanged between RUs and DUs, also the so-called *midhaul* traffic must be supported, which is exchanged between the DU and the CU. Then, as in traditional 4G LTE architectures, mobile traffic is backhauled towards the core network after the CU. Note that, in the following, we consider only a two-layer separation of eNBs, assuming co-location of DU and CU and we refer to this element as the CU (this co-location is commonly

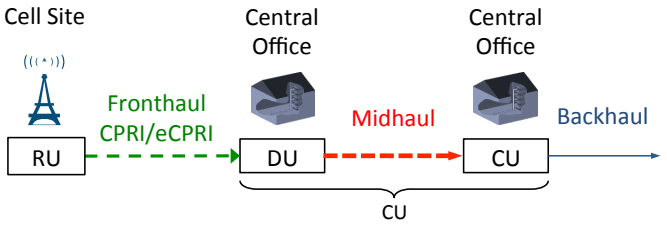


Fig. 1. eNB functional separation in 5G networks. In this paper we assume that the DU is co-located with its corresponding CU.

assumed in various architectures [5]). Therefore only fronthaul and backhaul traffic are considered in this paper. Note that, according to the adopted RAN split, various interfaces has been defined for fronthaul transport, such as the *enhanced CPRI* (eCPRI) [6], described in detail in Sec. II-A. As a future work, we will target the study of DU/CU placement by considering the more flexible three-layer separation of eNBs.

In this paper, we focus on the development of adaptive algorithms for the dynamic placement of CUs to enhance the utilization of processing and transport resources. E.g., following spatio/temporal dynamics of 5G tidal traffic, in low-traffic conditions several virtualized CUs can be centralized at so-called *CU pools* located in higher layers of the metro-access network, so as to promote power savings and enhanced coordination; on the other hand, when traffic increases, CU pools can be located at lower layers, i.e., closer to antenna sites, to avoid excessive fronthaul traffic insertion. Hence, the ability to dynamically reconfigure the CU location allows network operators to achieve the desired balance between baseband-resources consolidation and network capacity utilization.

We consider a multilayer OTN over WDM network as underlying transport technology, so our algorithm must perform grooming, routing and wavelength assignment (GRWA) in an OTN over WDM aggregation network, and explore the interaction of GRWA with CU placement to reach the objective of minimizing the average number of active pools[1], i.e., nodes hosting CUs, while achieving a satisfactory blocking probability. Adopting a multilayer OTN over WDM transport architecture to perform fronthaul traffic grooming has an impact on the latency between CUs and RUs[2], which plays a key role in the CU placement. In turn, the location of the CUs influences the amount of fronthaul traffic inserted in the network. Therefore latency has a direct impact on network resources utilization and CU consolidation. In our previous work [8], we investigated the dynamic CU placement for CU consolidation[3], but the location of a CU could not be modified during operation (e.g., if it is receiving traffic from an RU). In this paper, we consider also the case in which CUs can be moved during their activity, i.e., we allow *CU handover*.

[1]Minimizing the number of active pools is an indirect minimization target to enable reduction of network OpEx, as the energy consumed at CU pools.

[2]Note that operators deploying OTN for fronthaul/midhaul transport are already working on optimizing todays OTN technology to fit with 5G services requirements, e.g., to reduce mapping latency from 10 μs to around 1 μs or less through the so-called Mobile-optimized OTN [7].

[3]Note that in [8] CUs are referred to as Digital Units.

A. Related Work

In recent years, the idea of using optical access-aggregation architectures for C-RAN has attracted lot of attention (see, e.g., overviews in [9], [10] and [11]). Studies of the CU placement in C-RANs can be found in [12], where an ILP-based CU placement model is provided to minimize the number of CU pools, and in [13], where the authors consider resilience/availability and propose a CU placement strategy to guarantee that the fronthaul latency requirement is respected for both primary/backup CUs. Both these works consider a static placement of CUs and do not consider the impact of RAN splits on the CU placement. Dynamic network resources allocation has been studied in [14] in the general context of virtual network function placement for service chaining, and in [15] [16] for the specific C-RAN context. In particular, in [16] the authors consider different types of network slices, including a "Radio tenant" which represent the connectivity requests between RUs and CUs, and target the minimization of service blocking. However, this work does not consider the RAN splits and latency constraints in slice provisioning.

To the best of our knowledge, no existing work has evaluated the interplay between fronthaul latency and traffic grooming on the CU placement in a dynamic OTN-over-WDM access-aggregation network. Besides this, in our work we also consider how the flexibility brought by CU handover impacts on the C-RAN resources utilization.

B. Paper Contribution

The main contributions of this paper are as follows: 1) after providing a schematic overview of different RAN split solutions, we model the impact of fronthaul transport solutions, with particular focus on the impact of traffic grooming, on the tolerated fronthaul latency; 2) we define the Dynamic CU Placement/Handover (DCPH) problem in OTN-over-WDM access-aggregation networks and propose an adaptive algorithm for this problem, namely, the *MaxC-h* algorithm, which minimizes the number of active pools while achieving low network blocking; 4) through a simulative study, we analyze the impact of i) CU handover, ii) traffic grooming and iii) traffic bifurcation on the C-RAN performance, evaluated in terms of CU consolidation, latency and number of lightpaths.

The rest of the paper is organized as follows. Section II overviews RANs and describes the technological/architectural solutions adopted to implement C-RAN. In Sec. III we provide details on the impact of latency on C-RANs and how latency is affected by traffic grooming. In Sec. IV we introduce the DCPH problem in OTN-over-WDM access-aggregation networks, and describe the heuristic algorithm designed to address the problem in Sec. V. Illustrative numerical results are presented in Sec. VI., whereas Sec. VII draws paper conclusion.

II. Background on Radio Access Networks

We focus on optical access-aggregation networks used for the backhauling of mobile traffic. As shown in Fig. 2, RANs include several Cell Sites (CSs), i.e., eNodeBs, and a set of Central Offices (COs) of different hierarchical levels, which

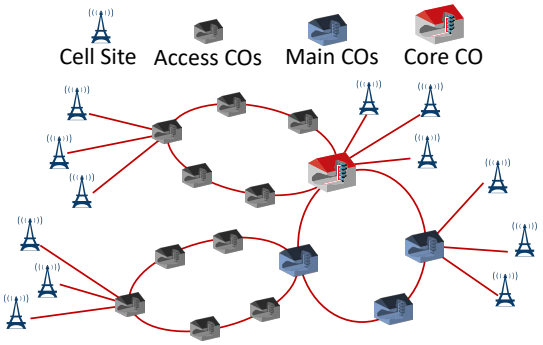


Fig. 2. Hierarchical access-aggregation network architecture.

are organized in "ring-and-spur" topologies and consist of Access COs, Main COs and one Core CO, which represents the RAN Point of Presence (PoP) and the interface towards the core network.

A. Mitigating C-RAN issues: enhanced CPRI

A first evolution of RAN is represented by C-RAN, where digital processing is performed in CU pools which are located in common sites (e.g., Access COs, Main COs or even the Core CO) and shared by several RUs. Although CU centralization in C-RAN enables CapEx/OpEx savings compared to traditional distributed RAN (D-RAN), it introduces new challenges due to the high-capacity (up to tens of Gbit/s per cell site) and low-latency (i.e., below few milliseconds[4]) *fronthaul* traffic, exchanged between a CU and its corresponding RU and transported via CPRI interface[5]. For this reason, despite the success of CPRI, many network operators have started to question their suitability, especially in view of the massive small cells deployment and traffic increase envisioned for 5G [3]. As a matter of fact, 5G small/micro/pico-cells "densification" will induce serious scalability issues in the fronthaul traffic transport, mainly due to the fact that fornthaul traffic is typically transported at a fixed line rate, which is independent on the end-users transported traffic. Thus, alternative solutions for the RAN functional separation are now under analysis in various consortia [17], [18], and standardization bodies, e.g., the IEEE 1914 working group [4], and they are often referred to as *RAN functional splits*.

One example of RAN functional splits specifications is the *enhanced CPRI* (eCPRI) [6], where a number of solutions have been defined, which, compared to CPRI, reduce fronthaul capacity requirements between the CUs and the RUs, while still enabling limited complexity and footprint of traditional base stations and providing sharing of both processing hardware and housing facilities.

[4]Note that, due to the latency needed to perform traffic processing, even lower latency might be required for signal propagation, resulting into propagation delay of few hundreds of microseconds [12], i.e., corresponding to a CU-RU distance in the order of tens of kilometers.

[5]Note that a variety of ultra-reliable low-latency (uRLL) services is envisioned for 5G network, which may lead to significantly different latency constraints. According to the specific service considered, different hybrid automatic repeat request (HARQ) mechanisms can be designed, corresponding to different latency constraints. In our work, we consider maximum fronthaul latency as driven by the HARQ mechanism, which is assumed to be specifically designed for latency-stringent 5G services.

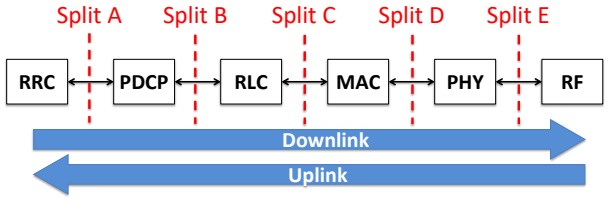


Fig. 3. eNodeB functional chain and split (figure adapted from [6]).

In the eCPRI specification, the base stations are identified by two basic eCPRI nodes, i.e., the eCPRI Radio Equipment Control (eREC), which performs part of the physical layer functions and higher-layers functions of the air interface, and the eCPRI Radio Equipment (eRE), which includes remaining physical layer functions and the analog radio frequency functions. Note that, such two elements correspond, respectively, to the CU and RU defined in the context of C-RAN.

Figure 3 shows the processing functions for in a base station as described in [6]. With reference to the figure, the following functions can be identified in a base station protocol stack, grouped according to the protocol layer, as defined in 3GPP LTE specifications [19], [20], [21], [22], [23]:

- Radio frequency (RF) layer is in charge of performing analog radio frequency functions, such as, e.g., frequency up/down conversion and power amplification;
- Physical (PHY) layer is responsible for preparing the bit stream for transmission by executing baseband functionalities, such as signal filtering, sampling, modulation/demodulation, etc.;
- Medium Access Control (MAC) layer performs radio resources allocation and contentions resolution in the physical medium access;
- Radio Link Control (RLC) layer includes data-link layer functions such as frame error detection and handling of the HARQ mechanism;
- Packet Data Convergence Protocol (PDCP) layer performs ciphering, integrity protection and IP header compression;
- Radio Resource Control (RRC) layer is used to implement coordination of radio channels among several users, handling of users mobility; exploiting information on radio channel quality, retrieved from end users measurement, advanced coordination, such as those provided by Coordinated Multipoint (CoMP), enhanced Inter-Channel Interference Coordination (eICIC), etc., can be accomplished at the RRC layer.

Figure 3 also shows different solutions proposed as eCPRI splits, although other splits, especially at the PHY layer, are also possible. For each split, functions at the right of the split are performed at the RU, i.e., at the cell site, whereas functions on the left side are centralized in CU pools, typically located at a CO within the aggregation network. With reference to the figure, traditional C-RAN fronthauling (i.e., CPRI) corresponds to split E. The choice of the eCPRI split is determined by a trade-off between functions centralization and

capacity/latency requirements, which become more stringent (i.e., with higher traffic and lower latency) moving from split A to split E [6]. Note that, according to the considered split, fronthaul traffic can be either proportional to the backhaul traffic (i.e., it is scaled via a factor $F > 1$), or be independent from users' activity, e.g., as for CPRI fronthauling, which basically represents the digitized radio-over-fiber signal.

B. Transported traffic types in C-RANs

According to the RAN split chosen and to the placement of CUs within the C-RAN, the following two types of traffic can be distinguished.

- *Backhaul*: it is natively packet-based with some degree of tolerance on delay; in case of distributed RAN, it is exchanged between RU/CU at cell sites and the Core CO; on the other hand, when C-RAN is adopted, regardless , this traffic is exchanged between the CU pools and the Core CO; note that, in principle, CU pools can be located also at the Core CO, in which case no backhaul traffic is present in the access/aggregation network;
- *Fronthaul*: this traffic arises whenever a RAN split is adopted and is exchanged between the RU at the cell site and the corresponding CU, located at one CO in the RAN; in comparison to backhaul, fronthaul traffic has more stringent requirements in terms of both capacity and latency; moreover, according to the selected RAN split, i.e., the eCPRI interface as in Fig. 3, it can be either packet-based or circuit-based, hence it can be proportional to or independent from the actual amount of user traffic (i.e., the backhaul), respectively.

III. Modeling of fronthaul transport latency and impact of traffic grooming

Due to the high capacity required by fronthaul traffic, traffic grooming can be beneficial, i.e., different fronthaul flows originated from various RUs at the cell sites can be aggregated into one (or few) lightpaths and transported towards their CUs. This can be convenient especially in case the aggregated fronthaul flows are destined towards a same CU pool. However, fronthaul traffic grooming is performed at the cost of introducing additional latency due to the switching of multiple traffic flows in the grooming node and inserting them into a single lightpath at the output of the node.

Therefore, a trade-off between capacity utilization and allowed fronthaul latency arises when performing traffic grooming and routing in C-RANs, which, in turn, impacts on the overall network blocking probability and CU consolidation.

In this paper, to evaluate the impact of traffic grooming on fronthaul latency and, in turn, on CU centralization, we consider two different solutions for the fronthaul traffic transport [12], i.e.: 1) *OTN*, where fronthaul flows between any RU and its corresponding CU can be groomed with other traffic into shared lightpaths[6], which can be initiated/terminated also in intermediate nodes along the RU-CU path (namely, we consider *multi-hop grooming* for fronthaul traffic, assuming an OTN-over-WDM network architecture); 2) *Overlay*, where each fronthaul flow is transported over a dedicated lightpath between the RU and the corresponding CU (i.e., we only allow *single-hop grooming* for the fronthaul traffic between an RU-CU pair).

According to the considered case, different latency contributions will impact the maximum allowed fronthaul latency, which are detailed in the following.

- t_{RU} and t_{CU}: these two terms represent the switching and processing latency needed at the end points of the fronthaul transmission, i.e., the RU and the CU, for the accomplishment of L1, L2 and L3 processing functions described in Sec. II-A.
- τ: this term represents the propagation delay and is related to physical distance traversed by fronthaul traffic in optical fiber links, for which we assume 5 $\mu s/km$ propagation speed.
- t_{sw}: such contribution is due whenever an electronic switch is used to perform optical/electronic/optical (OEO) signal conversion, e.g., to perform traffic grooming. As in [12], we assume "low-latency" switches specifically tailored for fronthaul applications, providing $20\mu s$ delay contribution per traversed switch.

To clarify the impact of grooming on fronthaul latency contribution, we show an illustrative example in Fig. 4 for the OTN and Overlay cases, considering the transport of fronthaul flows originated by three different RUs, i.e., "RU A", "RU B" and "RU C". In the example we focus on the latency contributions considered for the fronthaul traffic between "RU A" and the corresponding CU, i.e., "CU A", though similar observations can be drawn for the latency contributions for fronthaul flows originated by RUs B and C.

In the **OTN** case (see Fig. 4(a)) grooming of fronthaul traffic is allowed, but every time a grooming node is traversed a fixed latency contribution equal to t_{sw} must be considered. For the example in Fig. 4(a) we also show the overall set of latency contributions for fronthaul flow A (i.e., between "RU A" and "CU A") in the OTN case, corresponding to:

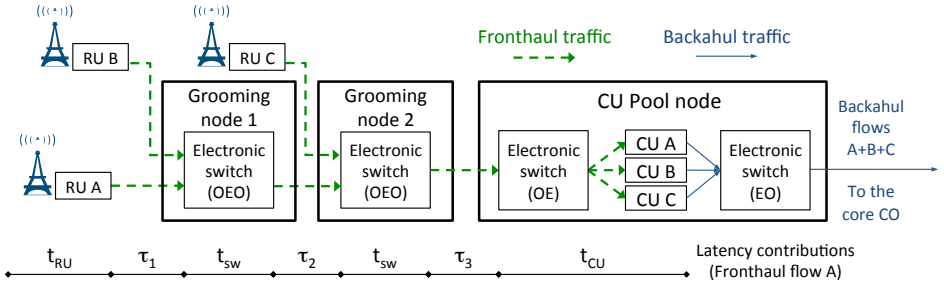
$$t_{A,OTN} = t_{RU} + \tau_1 + t_{sw} + \tau_2 + t_{sw} + \tau_3 + t_{CU}. \quad (1)$$

As specified in [24], a total round-trip latency budget of 3 ms is available between a CU and its corresponding RU, also including latency contributions at the RU and CU (i.e., t_{RU} and t_{CU}, respectively). On the line of [12], in this paper we assume that these two contributions are fixed as they are purely technology-dependent and are not influenced by the CU placement and traffic grooming capability, therefore we concentrate on the propagation (τ) and switching (t_{sw}) contributions. This leads to a maximum fronthaul latency of around 100 μs as in [17], [25].

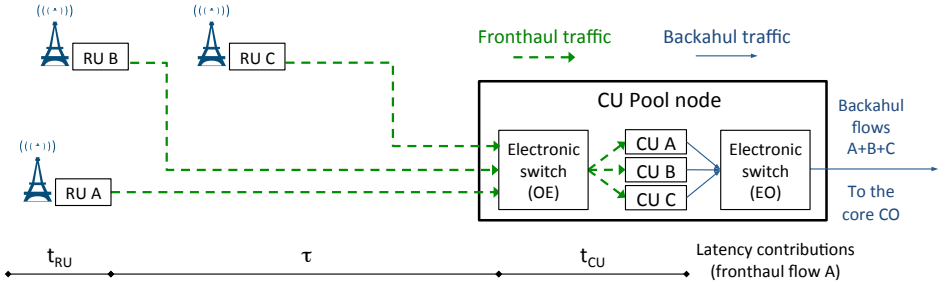
In the example of Fig. 4(a), two grooming nodes are traversed by fronthaul flow A, where fronthaul traffic between "RU A" and "CU A" is groomed with fronthaul flows B and C in grooming nodes 1 and 2, respectively[7]. Moreover, three

[6]Note that, in the OTN case, we assume that fronthaul flows can be also groomed with backhaul traffic.

[7]Note that, the switching latency contribution shall be accounted also in case fronthaul traffic is groomed with backhaul traffic only.



(a) OTN architecture: multi-hop grooming is allowed for fronthaul and backahul flows.



(b) Overlay architecture: multi-hop grooming is allowed only for backhaul flows. Fronthaul flows are transported over dedicated lightpaths where single-hop groomed is allowed.

Fig. 4. Impact of grooming on latency contributions for (a) OTN and (b) Overlay architectures.

propagation latency contributions are required and accounted for the propagation over the physical routes connecting RU A and grooming node 1 (τ_1), grooming nodes 1 and 2 (τ_2), grooming node 2 and the CU pool (τ_3).

In the **Overlay** case (see Fig. 4(b)), fronthaul traffic is not groomed and each fronthaul flow between a RU and its corresponding CU is routed over a dedicated lightpath. Therefore no switching latency is required in this case and, with reference to the example in the figure, the overall latency in the Overlay case corresponds to:

$$t_{A,Overlay} = t_{RU} + \tau + t_{CU}. \tag{2}$$

Note that, in general, the propagation delay required in the Overlay case, is different from the sum of propagation delay contributions needed in the OTN case, mainly for two reasons: 1) in the OTN case, aiming at efficiently exploit network capacity may lead fronthaul traffic to be transported over longer end-to-end routes between the RU and the CU, due to the presence of grooming nodes which are not necessarily in the shortest physical path between the RU and the CU; 2) in the Overlay case, using dedicated lightpaths for each fronthaul flow may lead to congestion of some network links, hence direct lightpaths between RUs and CUs might be routed over longer routes compared to the shortest path.

IV. The Dynamic CU Placement and Handover (DCPH) problem

A. Problem Statement

The Dynamic CU Placement/Handover (DCPH) problem in WDM access-aggregation networks can be stated as follows. **Given** 1) a hierarchical multi-stage access-aggregation network topology, represented by a graph $G(N, E)$, where N is the set of nodes (including COs and CSs) and E the set of optical fiber links, 2) random dynamically-generated backhaul traffic demands[8] originated by CSs and directed to the Core CO[9], **decide** the placement/handover of CUs and the Grooming, Routing and Wavelength Assignment (GRWA) of backhaul and fronthaul traffic, **minimizing** the average number of *active pools* in the network, **constrained by** 1) network links capacity (i.e. wavelength capacity and number of wavelengths per fiber) and 2) maximum fronthaul latency.

Note that, although only backhaul traffic demands are randomly generated and taken as input of the DCPH problem, in general, once a CU location is selected for the RU source of the backhaul demand, also one fronthaul traffic demand has to be routed from the RU to the CU together with the backhaul demand between the RU and the Core CO. In this context, for a given backhaul demand originated by a CS c, two special cases may arise according to the location selected for the CU, i.e.: 1) in case the CU is co-located with the RU, only the backhaul demand needs to be routed; 2) if the CU is located at the Core CO, only the fronthaul demand is routed.

We define a node in the network (either a CS or a CO) as an *active pool* if it hosts at least one active CU, which can be associated to a RU in another node or to the co-located RU, in case the active pool is itself a CS[10]. As we assume a CU is always hosted at the Core CO and is associated with a co-located RU, by definition, the Core CO is one active pool.

[8]The term "demand" is used in this paper to identify how we model traffic generation. In other words, in our model two or more demands can be originated by a same RU, but they represent the variation of overall mobile end-users' traffic which is aggregated at the CS.

[9]Note that, in this paper, we only consider uplink traffic, though similar considerations can be drawn also for downlink or bidirectional traffic.

[10]Note that we assume also COs have a co-located CS, i.e., also COs can originate backhaul traffic demands directed to the Core CO.

Moreover, as we will explain in detail in Sec. V, upon the arrival of a new traffic demand, in this paper we re-consider the CU placement to find a better location for that CU also in case one or more ongoing demands exist towards that CU, i.e., we allow *CU handover*, which is a main novelty of this paper. Note that this requires the *live* migration of "stateful" virtual machines. Supported by the recent advances in Network Function Virtualization (NFV), we speculate that such CU handover can be performed in the form of live virtual machines migration, in line with [26][11].

B. CU placement

The main objective of the DCPH problem consists of minimizing the *average* number of active CU pools, weighted by the amount of time when each of them is actually serving a demand. This objective captures the benefits of resource sharing provided with the C-RAN approach, i.e., it gives a measure of required OpEx. For example, assume two CUs are co-located in the same node (i.e., the same CU pool) for a given amount of time. In this case, the average number of active nodes is halved with respect to the case where the two CUs were located in two different locations for the same period of time, as two different nodes (i.e., two different CU pools, each hosting only one CU) would be activated. However, pursuing CU centralization (e.g., concentrating as many CUs as possible at the Core CO) leads to a huge increase in network capacity requirements, as a high amount of fronthaul traffic is inserted in the network, thus possibly causing higher demands blocking.

As it is difficult to characterize a cost function capturing the combined impact of CU centralization and network capacity requirement, to compare different solutions of the DCPH problem for a new incoming traffic demand d, in this paper we define a generic cost function which takes into account the activation of a new pool (i.e., in a node without other active CUs) to host the CU for demand d and the establishment of new lightpaths to provision the demand, i.e.:

$$C_d = c_p \cdot X_{pools,d} + c_l \cdot N_{lightpaths,d} \tag{3}$$

where $X_{pools,d}$ is a binary variable, equal to 1 in case a new pool (i.e., a node hosting only the CU for demand d) is activated, whereas variable $N_{lightpaths,d}$ represents the number of new lightpaths established to accommodate demand d. The parameters c_p and c_l represent the cost, expressed in relative Cost Units (CU), of one CU pool and one lightpath, respectively. As the relative values of these two parameters drive the trade-off between CU centralization and demands blocking, and due to the fact that in this paper our main focus is on the minimization of CU pools, we set $c_p \gg c_l$ (e.g., $c_p = 100 \cdot c_l$) so as to privilege CU centralization.

V. CU Placement and Handover Heuristic Algorithm

The objective of the DCPH problem is to minimize the average number of active pools in the network, while limiting demands blocking probability. To this end, the heuristic algorithm developed in this paper aims at maximum CU centralization and, if it is convenient to provide higher centralization, allows CU handover. For this reason it is called *Maximum Centralization with CU handover (MaxC-h)*. An incoming demand d is characterized by a series of parameters, i.e., 1) its source RU located at CS c_d, 2) the required backhaul traffic b_d, and 3) the demand duration t_d. Upon the arrival of demand $d = \{c_d, b_d, t_d\}$, the *MaxC-h* algorithm also takes in input the current network state, consisting of the set of all the deployed CUs along with their location, the installed lightpaths and their residual capacity, as well as the residual capacity in all the optical fiber links in the network. Then, the following main steps are executed, which are also detailed in Algorithm 1. Variables used in the procedure are summarized in Tab. I.

1) Identify optimal CU location. A list of candidate nodes is created to search for the optimal CU location for demand d; the different solutions, i.e., the candidate nodes in the list, are sorted considering their cost as in eq. 3 (lines 1-8). Note that also trivial solutions, i.e., locating the CU at the cell site or at the Core CO are also included in a list Z.

2) CU Placement/handover. After computing the amount of required fronthaul traffic f_d, which depends on the backhaul traffic b_d (line 9), the list of candidate CU locations is scanned, starting from the first node in the list (lines 10-44). First, the algorithm checks if a CU is already present in the network for the RU at CS c_d (line 12). If such a CU is present, and it is already located at the optimum location (i.e., the first node in list Z), the available capacity in the lightpaths already used between the RU and the CU (for fronthaul traffic) and between the CU and the Core CO (for backhaul traffic) is decremented by f_d and b_d, respectively (lines 12-15). In such a case, a trivial GRWA is performed for demand d, and the corresponding bandwidth values (f_d and b_d) will be deallocated from the lightpaths after t_d. Note that, if the available capacity in one or more of these lightpaths is not sufficient to provision f_d or b_d, the demand is blocked, and the *MaxC-h* algorithm is considered for a subsequent demand (lines 16-18). On the other hand, if a CU is already present for the RU at CS c_d, but its location does not coincide with the optimum location, CU handover needs to be performed, i.e., a boolean variable *handover* is set as *TRUE* (lines 20-22). In this case, the GRWA for demand d takes place, and it is performed similarly to the case where no CU is already present for the RU at CS c_d.

3) GRWA. In case a new CU is deployed or a CU handover is performed, the GRWA is performed. Note that, in case CU handover takes place, besides the traffic for demand d, also the traffic of all the existing demands originated by the RU at CS c_d must be considered at this step. This process, in general, involves the execution of GRWA for both the fronthaul and the backhaul traffic, and is performed on a shortest-path basis, also considering the possibility of using residual capacity of the existing lightpaths in the network, which are used to transport traffic of other demands. To this end, we build an auxiliary layered-graph [27], where each layer corresponds to a wavelength and replicates the physical topology of the network through a series of *physical-edges*. Edges between two same nodes in different layers, namely *grooming-edges*,

[11]Note that in our numerical analysis, we do not explicitly simulate migration as migration bandwidth for CU handover is negligible with respect to the amount backhaul and fronthaul traffic.

TABLE I
VARIABLES AND PARAMETERS OF THE *MaxC-h* ALGORITHM.

Variable	Description
E	set of bidirectional optical fiber links
N	set of network nodes
P	set of active pools
L	set of active lightpaths
$L_{f,x}$	set of fronthaul lightpaths used from the RU in node x
$L_{b,y}$	set of backhaul lightpaths used from the CU in node y
d	incoming traffic demand
c_d	source CS for demand d
b_d	required backhaul traffic for demand d
f_d	required fronthaul traffic for demand d
F	fronthaul-vs-backhaul scaling factor ($f_d = F \cdot b_d$)
t_d	duration of demand d
$X_{pools,d}$	binary variable, equal to 1 if d originates from a RU with no associated CU
$N_{lightpaths,d}$	number of newly deployed lightpaths used to accommodate demand d
$C_{d,n}$	cost of locating the CU for demand d at node n
Z	current list of candidate CU nodes z, ordered with decreasing value of $C_{d,z}$
$handover$	boolean, it is $TRUE$ if a CU handover is necessary to accommodate current demand d
k	number of shortest paths used in the Yen algorithm
K	current list of candidate GRWA solutions g

are also included to represent the nodes' grooming capability. Moreover, *lightpath-edges* can be also present between two nodes in a given layer to represent an already-established lightpath between the two nodes, and they are associated with the sequence of physical links constituting the lightpath.

The first task of the GRWA step is to perform GRWA for fronthaul traffic, due to the fact that fronthaul has more stringent requirements in terms of latency and required network capacity. The k shortest (i.e., best-cost) paths between c_d and the candidate CU node are calculated using Yen algorithm, and these k GRWA solutions are inserted in a list K (lines 24-25). The main cost metric used in our algorithm is the hop count. However, to favour the utilization of the residual capacity in already provisioned lightpaths, costs are assigned to a given lightpath-edge considering the number of physical links it traverses, divided by two[12]. Moreover, to discourage unnecessary grooming, we assign to *grooming-edges* cost equal to 0.6. The value 0.6 allows to break the tie in case, applying Yen algorithm, equal-cost paths are obtained between a short route where a new lightpath must be established and a longer route re-using existing lightpaths. Furthermore, note that, when fronthaul traffic for a new demand is routed and there are already existing demands from the same CU, the different fronthaul flows can be transported along parallel lightpaths between the RU-CU pair. In general, these lightpaths can be routed along distinct physical paths, therefore, in a first version of the *MaxC-h* algorithm, we assume fronthaul traffic can be physically bifurcated. However, we also consider a variation of the *MaxC-h* algorithm, where fronthaul traffic bifurcation is not allowed. In case one or more additional lightpaths are needed between a RU-CU which already exchange fronthaul for other existing demands, the new lightpaths must be routed along the same physical path of the existing ones, although

[12]As an example, a *lightpath-edge* corrseponding to a lightpath traversing 5 physical links has cost 2.5.

Algorithm 1 *MaxC-h* heuristic algorithm.

INPUT: Network topology and status: $G(N, E)$, P, L. Incoming demand $d = \{c_d, b_d, t_d\}$.
OUTPUT: CU placement/handover; GRWA for d.
Initialization:
1: **for all** $n \in N$ **do**
2: Initialize $X_{pools,d} = N_{lightpaths,d} = 0$;
3: Set $X_{pools,d} = 1$ if n does not host any CU;
4: Calculate shortest-path GRWA between c_d and n using Dijkstra algorithm and set $N_{lightpaths,d}$ equal to the nr. of new lightpaths needed;
5: Calculate cost C_d according to eq. 3;
6: Set $C_{d,n} = C_d$ as the cost of locating the CU at node n;
7: **end for**
8: Sort nodes $n \in N$ in ascending order of $C_{d,n}$ and insert them in a list Z;
CU placement/handover and GRWA:
9: Set fronthaul traffic for demand d: $f_d = F \cdot b_d$;
10: **while** Z is not empty **do**
11: Consider the first node $z \in Z$;
12: **if** c_d already has a CU at node m **then**
13: **if** $m == z$ **then**
14: Add f_d to lightpaths $L_{f,c}$ between c and z;
15: Add b_d to lightpaths $L_{b,z}$ between z and the Core CO;
16: **if** available capacity on $\{L_{f,c}, L_{b,z}\}$ is not enough **then**
17: Block demand d;
18: **end if**
19: EXIT;
20: **else**
21: CU handover: set *handover=TRUE*
22: **end if**
23: **end if**
24: Fronthaul GRWA: compute k shortest-paths GRWA solutions between c_d and z using Yen algorithm;
25: Insert the k solutions in a list K;
26: **while** K is not empty **do**
27: Consider the first element $g \in K$ as a candidate GRWA solution;
28: **if** fronthaul latency budget is respected between c_d and z
29: AND for the other existing fronthaul flows affected by g **then**
30: Backhaul GRWA: compute the shortest-path GRWA between z and the Core CO using Dijkstra algorithm;
31: **if** latency budget is respected for the existing fronthaul flows affected by the Backhaul GRWA solution **then**
32: Provision fronthaul and backhaul flows for d;
33: **if** *handover==TRUE* **then**
34: Deprovision backhaul and fronthaul lightpaths of the existing demands originated from c_d;
35: **end if**
36: EXIT;
37: **else**
38: Remove g from K;
39: **end if**
40: **else**
41: Remove z from Z;
42: **end if**
43: **end while**
44: **end while**
45: **if** Z is empty **then**
46: Block demand d;
47: **end if**

they will use distinct wavelengths.

List K is then scanned starting from the first GRWA solution $g \in K$ (lines 26-43). If fronthaul latency budget is respected for d and for all the existing fronthaul flows possibly affected by g (lines 28-29), GRWA is performed also for the backhaul traffic of d (lines 30-36). Note that, performing traffic grooming for fronthaul and/or backhaul flows of d may affect existing fronthaul flows. Therefore, every time GRWA is performed for demand d, fronthaul latency budget is checked not only for the current fronthaul demand, but also for the other existing fronthaul flows, which may be affected due to the switching latency contribution t_{sw} introduced when performing traffic grooming, as explained in Sec. III.

Moreover, in case the GRWA solution $g \in K$ cannot be used due to the violation of a latency constraint, the first solution is removed from lisk K, and the subsequent solution is analyzed (line 37-38). In case no solution is found from list K, the current candidate CU location z is removed from list Z (lines 40-42) and the subsequent candidate CU location is analyzed, i.e., the process is repeated from line 11. If no solution is found for any of the candidate location in Z, e.g., due to the lack of network capacity and/or the violation of the fronthaul latency constraint, the demand d is blocked (lines 45-47). Conversely, if a solution is found for d, corresponding backhaul and fronthaul traffic are deprovisioned after t_d and, in case the used lightpaths are not used for any other demand, such lightpaths are torn down.

A. Alternative versions of MaxC-h Algorithm

The MaxC-h algorithm described in Sec. V is a complex procedure which encompasses several optimization aspects, i.e., 1) fronthaul transport architecture, which has an impact on traffic grooming, 2) CU handover, and 3) traffic bifurcation. Therefore, to capture the impact of the various aspects, we developed different flavours of the MaxC-h algorithm, as summarized in Tab. II. In particular, compared to the *MaxC-h* algorithm, we also consider: 1) the *Overlay* MaxC-h, where traffic grooming is not allowed as fronthaul traffic is transported over dedicated wavelengths between CU-RU pairs; 2) the *T-constrained* MaxC-h, where handover can be performed for a given CU only after T seconds from the previous handover performed for that CU[13]; and 3) *Non-bifurcated* MaxC-h, where all the lightpaths between an RU-CU pair are provisioned along the same physical path by using different wavelengths.

VI. Numerical Results

A. Case study

To perform our numerical evaluation, we developed a C++ event-driven simulator, where we randomly generate the arrival of 55000 demands originated by the RUs. Arrivals are generated according to a truncated-Poisson distribution, used to capture the fact that CSs support a limited backhaul traffic,

[13]This constraint is adopted to limit the number of CU handover operations, which require signalling between source and target CUs and potential additional blocking, which are not considered in this paper.

TABLE II
Different flavours of *MaxC-h* algorithm and their features.

DCPH Algorithm	Fronthaul transport architecture	CU handover	Traffic bifurcation
MaxC-h	OTN w/ grooming	Unconstrained	Allowed
Overlay	Overlay w/o grooming	Unconstrained	Allowed
T-constrained	OTN w/ grooming	Allowed after T seconds from previous handover	Allowed
Non-bifurcated	OTN w/ grooming	Unconstrained	Not allowed

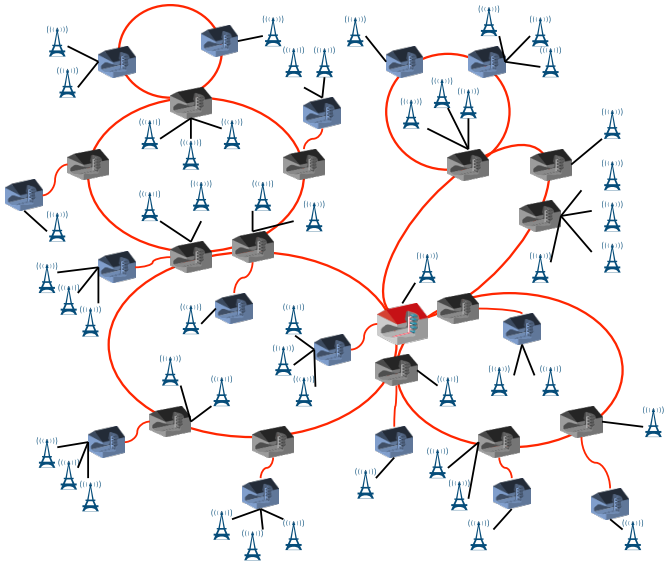


Fig. 5. 5G HetNet topology used for the numerical evaluation.

and are uniformly distributed among RUs in the network[14]. Demands duration t_d is assumed as exponentially-distributed with mean $\mu = 1$ s.

We consider a 5G HetNet scenario, where 80 nodes, consisting of 50 Macro CSs (MC) and 30 COs also inserting mobile traffic, cover a square region of 200 km^2 and are interconnected via a 4-stages topology as shown in Fig. 5. Each MC is interconnected via a lower-layer tree to 10 Small Cells (SCs), not shown in Fig. 5 for the sake of figure clarity, via optical fiber links at a maximum distance of 2 km. Each fiber supports 8 wavelengths at 100 Gbit/s each, in line with [29]. This scenario follows the guidelines of a 5G urban mobile aggregation network, as identified in [30]. MCs are assumed as 3-sectored sites with maximum backhaul traffic of 15 Gbit/s each, corresponding to an antenna configuration with 125 MHz spectrum, 256 QAM and 8×8 MIMO. We consider the same

[14]Other choices for arrivals distribution are possible, such as, e.g., "simple" Poisson, or Bernoulli distributions [28]. However, note that, in Poisson there is no theoretical limit in the amount of traffic that can be generated by a single CS, which is not realistic, whereas in the Bernoulli (that we already considered in [8]), the probability of a new demand from a given CS is inversely proportional to the current traffic generated by that CS, which is again not in line with a realistic mobile user behaviour.

configuration for SCs, though we assume single-sectored sites, thus requiring a maximum of 5 Gbit/s traffic.

Each demand d requires a fixed b_d =300 Mbit/s backhaul and, as we assume RAN split option II_d as in [6] (i.e., an intermediate split between splits D and E in Fig. 3), the corresponding fronthaul traffic is $f_d = 1.2$ Gbit/s (i.e., the bandwidth scaling factor is $F \simeq 4$), leading to a maximum fronthaul of 60 Gbit/s and 20 Gbit/s per MCs and SCs, respectively. The maximum tolerated latency for the considered RAN split is set to 100 μs. The choice of the RAN split is motivated by the fact that, among the eCPRI splits with fronthaul traffic proportional to backhaul, eCPRI split II_d enables the highest degree of functions centralization. Note that, considering a RAN split with backhaul-proportional fronthaul traffic allows to evaluate the importance of traffic grooming when solving the DCPH problem.

Moreover, for the various flavour of the *MaxC-h* algorithm, we consider $k = 10$ as the number of shortest paths GRWA solutions to be evaluated in Alg. 1, as higher values of k do not provide relevant gains, while negatively impacting complexity. For the *T-constrained* MaxC-h algorithm we set $T = 0.5$ s to impose that, on average, each demand undergoes at most one CU handover (note that the mean demands duration is 1 second).

The parameters used in the numerical evaluation are summarized in Tab. III.

We evaluate the performance of the developed algorithms considering the following metrics: 1) average number of active pools, P_{av}, 2) average number of lightpaths Λ_{av}, 3) average fronthaul latency, L_{av}. Concerning metrics P_{av} and Λ_{av}, the contribution provided by each demand d is weighted by the fraction t_d/D, where t_d is the demand duration and D is the total simulated time.

B. Discussion

To validate the effectiveness of the *MaxC-h* algorithm we first compare its performance with that of the *Adaptive* algorithm in [8], which, among the algorithms in [8], is the one providing the lowest P_{av} while maintaining low blocking probability.

TABLE III
PARAMETERS USED IN THE NUMERICAL EVALUATION.

Parameter	Value
Number of Macro CS	50
Number of COs	30
Number of small cells per Macro CS	10
Macro CS density	0.4 km^{-2}
Number of wavelengths per link	8
Wavelength capacity	100 Gbit/s
Required backhaul traffic per demand	$b_d = 300$ Mbit/s
RAN Split option	II_d [6] ($F \simeq 4$)
Required fronthaul traffic per demand	$f_d = 1.2$ Gbit/s
Maximum backhaul traffic per Macro CS	60 Gbit/s
Maximum backhaul traffic per SC	20 Gbit/s
Maximum fronthaul latency	100 μs
Number of shortest paths used in the Yen algorithm	$k = 10$
Time-constraint in the *T-constrained* MaxC-h	$T = 0.5$ s

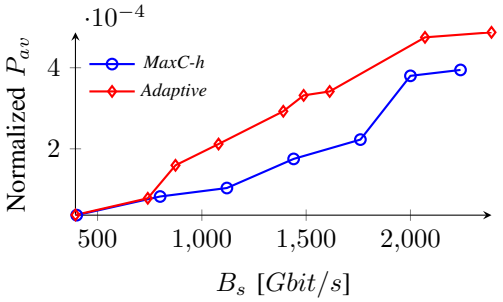


Fig. 6. Comparison of *MaxC-h* and *Adaptive* algorithms for increasing backhaul served traffic B_s.

Figure 6 shows the comparison between the two algorithms in terms of average number of active pools. To better capture the performance difference of the two algorithms, we show the *normalized* P_{av}, i.e., we normalize the average number of active pools with respect to the number of provisioned demands by the two algorithms. For the same reason, in our analysis we show the results as a function of the backhaul served traffic B_s, i.e., excluding the blocked backhaul demands.

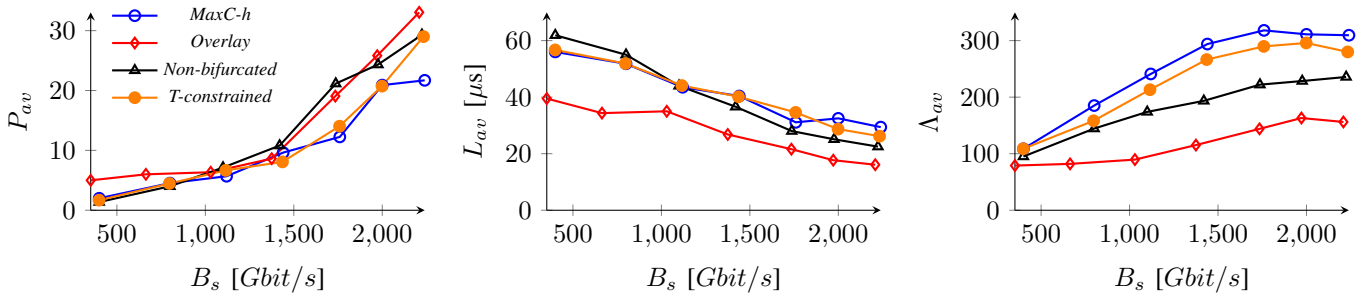
As shown in Fig. 6, *MaxC-h* always provides lower number of active pools per demand, mainly due to the possibility of performing CU handover in case it is convenient to improve CU consolidation. *MaxC-h* and *Adaptive* have comparable performance in terms of normalized P_{av}, only for lower served traffic, confirming that *MaxC-h* is able to better adapt to the dynamic changes of network traffic behaviour. In other words, this demonstrates that the *MaxC-h* algorithm is able not only to reduce the number of active pools, but also supports more users' traffic thanks to the opportunity of moving CUs and consequently reduce the amount of fronthaul traffic which might lead to network congestion. As a matter of fact, for the considered arrival rates no demands are blocked in the *MaxC-h* case. Conversely, the *Adaptive* algorithm provides higher blocking, i.e., in the order of 20%, even for medium traffic (e.g., 20 Gbit/s per RU)[15].

Now we provide in Fig. 7 the results for the different flavours of the *MaxC-h* algorithm as described in Sec. V-A. This comparison allows to quantify the impact of the various features of the *MaxC-h* algorithm on network performance.

Average number of active pools.
Figure 7(a) shows, for the four cases, the average number of active pools (P_{av}) as a function of the served backhaul traffic (B_s). Note that, two benchmark CU placement solutions, i.e., fully-distributed and fully-centralized (not shown in the figures), corresponding to the case of CUs co-located with their RUs at all MCs, and to the case with only one CU pool at the Core CO, would produce a normalized P_{av} of 80 and 1, respectively.

For lower values of B_s the average number of active pools approximates the lower bound of 1 pool for all the algorithms, i.e., only the pool at the Core CO is sufficient for the whole set of RUs. The *Overlay* algorithm is an exception to this,

[15]As a further confirmation, with reference to Fig. 6, the arrival rate of 28 Gbit/s per RU corresponds to $B_s = 2240$ Gbit/s for the *MaxC-h* case, and to $B_s = 1612.8$ Gbit/s for the *Adaptive* algorithm.



(a) Average nr. of active pools (b) Average fronthaul latency (c) Average number of lightpaths

Fig. 7. Comparison of the different flavours of *MaxC-h* algorithm for increasing backhaul served traffic B_s.

i.e., few more pools are activated in this case, due to the fact that using dedicated wavelengths for fronthaul transport corresponds to a higher network capacity requirement, and consequently to lower opportunity for CU consolidation. On the other hand, for increasing B_s, the values of P_{av} increase in all cases. As expected, the lowest P_{av} is obtained, in general, for the *MaxC-h* case, which allows the highest flexibility in performing CU handover and GRWA of the backhaul and fronthaul traffic. On the other hand, when adopting *the Overlay* fronthaul transport, the highest average number of active pools is obtained, due to the fact that using dedicated wavelengths for fronthaul transport leads to underutilization of network capacity. Indeed, to pursue CU consolidation, in the *Overlay* case direct lightpaths are typically deployed on longer physical routes between the RUs and higher stages of the access-aggregation network. Consequently, this quickly leads to network congestion, especially in higher hierarchical levels of the network (i.e., in links interconnecting Main COs and the Core CO), and thus forces new CUs to be placed closer to the corresponding RUs so as not to introduce further fronthaul traffic in the network. The difference between the two fronthaul transport solutions is more evident for increasing load, when the importance of traffic grooming is more relevant. As expected, also in the case of *Non-bifurcated* algorithm P_{av} is higher compared to *MaxC-h*, due to the fact that multiple demands originated by a given RU must be routed along the same physical route. This is not always possible, especially for increasing load, therefore, in order to be able to accommodate new demands, CUs are often placed at lower stages of the network or even co-located with the RUs. Moreover, considering the *T-constrained* algorithm, the number of active pools is comparable with the one in the *MaxC-h* case, except for very high traffic, when the limit of the number of CU handovers per RU plays a role.

It is worth noting that backhaul blocking probability (not shown as a figure) is kept below a satisfactory value of 1% for all values of B_s, especially for the *T-constrained* and *MaxC-h* cases. However, due to the inefficient utilization of lightpaths' capacity, in the *Overlay* case blocking probability is below the 1% threshold only for higher loads. Though counter-intuitive, this behaviour is motivated by the fact that the primary objective of the algorithms is to consolidate CUs, which is easier when for lower loads. Instead, for higher loads CUs are typically placed at lower network stages, leading to lower capacity requirements for fronthaul traffic transport.

Average fronthaul latency.
The difference between the four algorithms in performing CU consolidation can be observed from another point of view in Fig. 7(b), which shows the average latency between an RU and its corresponding CU pool, i.e., L_{av}. In all cases, L_{av} tends to decrease with increasing loads, due to the larger amount of fronthaul traffic inserted, which limits the opportunity for CU consolidation at the Core CO or, in general, at nodes in higher layers of the network. As it is evident from the figure, the lowest values of L_{av} are obtained, independently from B_s, with the *Overlay* algorithm, when distributed placement of CUs (i.e., closer to RUs) is necessary to face network congestion at higher stages of the network. Moreover, at lower loads, the other algorithms provide comparable values of L_{av}, although in the *Non-bifurcated* case latency is slightly higher, mainly due to the fact that lightpaths are typically routed over longer paths to maintain non-bifurcated traffic. Interestingly, at a certain value of B_s (i.e., around 1000 Gbit/s), L_{av} becomes lower for the *Non-bifurcated* case, in comparison to *T-constrained* and *MaxC-h* algorithms, showing that the impact of traffic bifurcation on RU-CU latency is more relevant than the limit in the number of CU handovers.

Average number of lightpaths.
Finally, Fig. 7(c) shows the average number of active lightpaths in the four cases. As expected, for increasing B_s, Λ_{av} increases for all the four algorithms, and saturates to a maximum value. However, the motivation for this increase is different in the various cases. Specifically, in the *Overlay* case grooming can be performed only for backhaul traffic, as dedicated lightpaths are provisioned for fronthaul transport between RUs and their CUs. Therefore, when less CU pools are activated (e.g., around 5 active pools for lower loads, as shown in Fig. 7(a)), typically in medium-higher network stages (i.e., Main COs or the Core CO), grooming backhaul demands is less frequent. Then, as B_s increases, there is more opportunity for backhaul traffic grooming as CUs are placed in lower network stages. On the other hand, when fronthaul traffic grooming is allowed (i.e., in *MaxC-h*, *Non-bifurcated* and *T-constrained* cases), a higher number of shorter lightpaths are typically needed to efficiently exploit network capacity and obtain CU consolidation at higher network stages at the same time. This behaviour is more evident for the *T-constrained* and especially for the *MaxC-h* cases, as the opportunity for

traffic bifurcation provides higher flexibility in performing traffic grooming.

VII. Conclusion

In this paper we focused on the dynamic placement of CUs in optical access-aggregation networks, with the objective of minimizing the number of active CU pools. To this end, we defined the Dynamic CU Placement/Handover (DCPH) problem in WDM access-aggregation networks and provided a latency-aware heuristic algorithm, namely *MaxC-h*, for the CU placement/handover and GRWA of mobile traffic demands. We also evaluated how C-RAN performance are influenced by *MaxC-h* algorithm features, i.e., i) CU handover, ii) traffic grooming and iii) traffic bifurcation. We found that, especially for higher loads, fronthaul latency plays a critical role in reducing the number of active CU pools. Advanced sharing of baseband processing resources can be obtained if the C-RAN is capable of performing CU handover and, especially, if multi-hop grooming capabilities are enabled for fronthaul transport, e.g., by adopting an OTN-over-WDM network architecture. As a future work, we plan to extend our study also considering the three-layer separation of eNBs into RU, DU and CU.

Acknowledgment

The work leading to these results has been supported by the European Community under grant agreement no. 761727 Metro-Haul project.

References

[1] "CPRI (Common Public Radio Interface) Specification V7.0," available online at http://www.cpri.info, Oct. 2015.

[2] C. Zhang, "Optical Networking in the Cloud and 5G Era," Keynote talk at Optical Fiber Communications Conference and Exhibition (OFC) 2018, available online at https://www.youtube.com/watch?v=kN9rSctv-tg.

[3] Next generation Mobile Network (NGMN) Alliance, "5G White Paper - Executive Version," Dec. 2014.

[4] J. Yang, "Latency requirements and analysis," IEEE 1914.1 Working Group Meeting, Aug. 2016.

[5] "Transport network support of IMT-2020/5G," ITU-T GSTR-TN5G Technical Report, Feb. 2018.

[6] "Common Public Radio Interface: eCPRI Interface Specification V1.2," available online at http://www.cpri.info, June 2018.

[7] R. Jing, "China Telecoms Requirements on 5G Transport," in *ITU-T Workshop on the evolution of transport networks to support IMT-2020/5G*, Oct. 2017.

[8] F. Musumeci, G. Belgiovine, and M. Tornatore, "Dynamic Placement of BaseBand Processing in 5G WDM-based Aggregation Networks," in *Optical Fiber Communications Conference and Exhibition (OFC) 2017*, Mar. 2017.

[9] 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, First quarter 2015.

[10] T. Pfeiffer, "Next generation mobile fronthaul and midhaul architectures," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 11, pp. B38–B45, Nov. 2015.

[11] X. Liu and F. Effenberger, "Emerging optical access network technologies for 5G wireless," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 8, no. 12, pp. B70–B79, Dec. 2016.

[12] 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," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1963–1970, Apr. 2016.

[13] B. M. Khorsandi and C. Raffaelli, "BBU location algorithms for survivable 5G C-RAN over WDM," *Computer Networks*, vol. 144, pp. 53–63, 2018.

[14] J. Liu, W. Lu, F. Zhou, P. Lu, and Z. Zhu, "On Dynamic Service Function Chain Deployment and Readjustment," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 543–553, Sep. 2017.

[15] J. Zhang, Y. Ji, J. Zhang, R. Gu, Y. Zhao, S. Liu, K. Xu, M. Song, H. Li, and X. Wang, "Baseband unit cloud interconnection enabled by flexible grid optical networks with software defined elasticity," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 90–98, Sep. 2015.

[16] M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska, and P. Monti, "Dynamic slicing approach for multi-tenant 5G transport networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 1, pp. A77–A90, Jan. 2018.

[17] "Small Cell Virtualization: Functional Splits and Use Cases," Small Cell Forum whitepaper, rel. 7, July 2016.

[18] Next generation Mobile Network (NGMN) Alliance, "Project RAN Evolution: Further Study on Critical C-RAN Technologies," Mar. 2015.

[19] "Evolved Universal Terrestrial Radio Access (E-UTRA); Long Term Evolution (LTE) physical layer; General description," 3rd Generation Partnership Project (3GPP), Technical Specification 136.201, v.8.1.0, available online at https://www.3gpp.org, Nov. 2008.

[20] "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification 36.321, v.15.2.0, available online at https://www.3gpp.org, July 2018.

[21] "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification 36.322, v.15.1.0, available online at https://www.3gpp.org, July 2018.

[22] "Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification," 3rd Generation Partnership Project (3GPP), Technical Specification 36.323, v.15.0.0, available online at https://www.3gpp.org, July 2018.

[23] "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification 36.331, v.15.2.2, available online at https://www.3gpp.org, June 2018.

[24] *3GPP TS-36.213 (Physical layer procedures); http://www.3gpp.org*.

[25] A. Asensio, P. Saengudomlert, M. Ruiz, and L. Velasco, "Study of the centralization level of optical network-supported Cloud RAN," in *International Conference on Optical Network Design and Modeling (ONDM), 2016*, May 2016, pp. 1–6.

[26] T. Wood, K. K. Ramakrishnan, P. Shenoy, J. V. der Merwe, J. Hwang, G. Liu, and L. Chaufournier, "CloudNet: Dynamic Pooling of Cloud Resources by Live WAN Migration of Virtual Machines," *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1568–1583, Oct. 2015.

[27] H. Zhu, H. Zang, K. Zhu, and B. Mukherjee, "Dynamic traffic grooming in WDM mesh networks using a novel graph model," in *IEEE Global Telecommunications Conference (GLOBECOM) 2002*, vol. 3, Nov. 2002, pp. 2681–2685.

[28] J. C. Bellamy, *Digital Telephony*, 3rd ed. New York, USA: John Wiley and Sons, Inc., 2000.

[29] "Cloud Scale Metro Networks," available online at https://www.infinera.com/gometro/, accessed Jan. 2019.

[30] "Assessment of candidate transport network architectures for structural convergence," European Project "COMBO", Deliverable D3.4, available at http://www.ict-combo.eu, Apr. 2016.