Power-aware optimization of baseband-function placement in cloud radio access networks

Ligia M. Moreira Zorello
Politecnico di Milano
Milan, Italy
ligiamaria.moreira@polimi.it

Sebastian Troia
Politecnico di Milano
Milan, Italy
sebastian.troia@polimi.it

Marco Quagliotti
Telecom Italia
Turin, Italy
marco.quagliotti@telecomitalia.it

Guido Maier
Politecnico di Milano
Milan, Italy
guido.maier@polimi.it

***Abstract*—Because of the advent of highly diverse and heterogeneous services, 5G networks required the design of a new paradigm in terms of Radio Access Networks. The Centralized-RAN appeared as a suitable solution, enabling the reduction of costs by centralizing the baseband functions. Nevertheless, it requires very high bandwidth, making it unfeasible. Functional splits were proposed to divide the protocol stack and place the baseband functions in Central Units (CU) or Distributed Units (DU). It is an efficient way of enabling different types of services while reducing costs for operators. However, each split option introduces very stringent constraints in terms of latency, throughput and processing. The problem that arises is how to optimally place the CU functions in the network nodes. This paper proposes an optimization framework to minimize link and node power, subject to the split requirements. We also present models to compute bandwidth, computational effort and latency for splits 2 (PDCP-RLC) and 6 (MAC-PHY). The solution proposed was compared to the fully centralized and distributed scenarios over a realistic topology of a metro network with 24-hour traffic. Results show that the proposed CU-placement optimization reduces the power consumption in respect to both scenarios in split 2. In addition, even if the consumption is very similar to the centralized scenario in split 6, the ILP solution meets the latency requirements. Finally, when comparing the split options, the overall power consumption in split 2 is 5% lower than split 6.**

Keywords—C-RAN, functional split, optimization, 5G.

I. INTRODUCTION

The fifth generation of mobile networks (5G) is emerging to address very stringent and heterogeneous requirements of myriad services and applications. New technologies and use cases, mapped into the categories enhanced Mobile Broadband (eMBB), ultra-Reliable and Low Latency Communications (uRLLC) and massive Machine Type Communication (mMTC), introduce new Quality of Service (QoS) requirements and the need for performance enhancements in terms of throughput, millisecond-scale end-to-end latency, and reliability [1].

From the perspective of Radio Access Network (RAN) architecture, two main concepts are expected to bring improvements in the system performance. First of all, the centralization of baseband processing functions enables the improvement of radio coordination, system cost, and energy consumption. Furthermore, the adoption of Network Function Virtualization (NFV), in which the processing resources are virtualized and implemented in general purpose hardware [2].

Centralized RAN (C-RAN) takes advantage of these concepts and is based on decoupling the Remote Radio Head (RRH) and BaseBand Unit (BBU) and placing the former in a centralized/cloud radio access architecture. The RRH is responsible for the transmission and reception of radio signals, while the BBU contains all baseband processing functions [3]. By separating the BBU from the antenna sites and centralizing it in a Central Office (CO), it facilitates scaling the system, reduces costs to operators and eases the deployment of cooperative algorithms, such as the Coordinated Multi-Point (CoMP) transmission and reception [1].

With the complete centralization of BBU functions, the fronthaul network would rely on protocols such as the Common Public Radio Interface (CPRI) to realize the connection between RRH and BBU. Because it is traffic-independent, the adoption of such a protocol would generate extremely high and constant capacity demands on the fronthaul network [3], [4]. Redistributing part of the BBU functions is therefore shown to be an effective solution to optimize the trade-off between cost and bandwidth capacity [5]. This partial decentralization is endorsed by the 3rd Generation Partnership Project (3GPP) as functional splits [6]. This approach determines the functions to remain close to the antennas, and those to be centralized according to the application. These functional splits generally divide the BBU into two distinct modules: Centralized Units (CU), and Distributed Units (DU). The DUs are composed by the lower network layer functions and are implemented in the cell sites closer to the antennas. The CUs contain higher layers functions and can be implemented in general purpose datacenters as Virtual Network Functions (VNFs).

The placement of CUs in the metro network nodes can thus be treated as a VNF-placement problem, in which the selection of the appropriate location must be optimized under constraints related to the traffic and to the functional split. Indeed, they are strictly related to the split option, as they present tight requirements in terms of latency and throughput. In addition, the split option also dictates the amount of resources necessary in CU and in DU nodes to process their specified functions.

Midhaul links connecting DUs and CUs are typically carried over a fiber network [7]. In particular, we focus on a metro-core network based on IP-over-optical, according to the infrastructure proposed by the European project Metro-Haul [8]. It comprises metro nodes interconnected by a high capacity dynamic and flexible WDM network, composed by fiber links, containing wavelengths with maximum capacity of 100 Gbit/s [9]. The nodes are considered as mini datacenters hosting both IT and Telecommunication (TLC) equipment, following the 5G Multi-access Edge Computing (MEC) [10] model, so that they can support the instantiation of VNFs.

Horizon 2020 project no. 761727 Metro-Haul

978-3-903176-21-8 ©2020 IFIP

The CU-VNF placement problem can hence be modeled as a power consumption minimization, resulting in a reduction of operational expenditure (OpeEx) costs for the operator. In this work, we propose a power-aware optimization algorithm based on Integer Linear Programming (ILP). The target is to minimize the power consumption of the IT and TLC components subject to all functional-split constraints. In addition, we assess the differences between functional split options by contrasting our solution to the traditional centralized and distributed scenarios.

The remainder of this paper is organized as follows. Section II describes the research works in the literature that address the VNF placement problem in the C-RAN scenario. Section III presents the models used to detail the functional splits characteristics in terms of bandwidth, latency and computational effort, and the ILP proposed. Section IV shows the simulation setup scenario and explains the results obtained. Finally, Section V concludes the paper.

II. Related Work

The problem of placing baseband functions as VNFs in network nodes has attracted significant attention in the past years. In [11], authors propose an ILP to optimize resource allocation and, thus, minimize the required computational capacity subject to the split delay constraints. However, this work does not evaluate the bandwidth constraint. Similarly, Hu *et al.* [12] propose a placement optimization algorithm aiming at minimizing the number of active nodes to handle baseband processing functions. Unlike [11], the latter work considers all the split constraints. Nevertheless, these works do not consider the optimization of network components usage.

Differently from the previously described works, several others consider a broader scenario of optimization, including both network and IT components. An algorithm aiming at minimizing the energy consumption of the servers and optical components in mobile core and radio access networks was presented in [13]. For this, they optimally allocate and route baseband VNFs such that the computational and link capacities are met, but this approach does not consider latency constraints. Ref. [14] provides a VNF placement scheme ensuring low latency and enough capacity to different categories of services. Arouk *et al.* [15] model the network planning by optimizing the deployment of baseband VNFs based on the costs related to the network links and to the nodes, respecting all split constraints. In [16], authors integrate optical and wireless transport technologies and optimize utilization of network and computing resources. In their formulation, the transport technology is selected to minimize the costs, and the baseband functions are placed such that all split constraints are met.

Most of the above-mentioned works consider low-layer functional splits, enabling greater centralization of baseband functions, but none of them compare different functional splits within the same problem formulation. Understanding the impact of other functional splits on the performance of the network is important; thus, our work aims at filling this gap. In addition, we investigate how the baseband processing impacts the one-way and end-to-end latency. It depends on the type of functions executed in the node as well as the total traffic processed: jointly with the propagation delay, it can deeply influence the choice between different functional splits.

III. Functional Split Problem Formulation

This section describes the models that characterize the functional split, and the optimization problem formulation.

A. Functional split model

In order to optimally place CUs, we must assess the requirements associated to the placement of the nodes and to the traffic routing. They are strictly related to the split option selected, as they present tight requirements in terms of latency, throughput and processing. Therefore, the split option dictates the amount of resources necessary in network links, and in the nodes to process their specified functions. 3GPP recommends in [6] to divide the protocol stack of the baseband, as in Fig. 1.

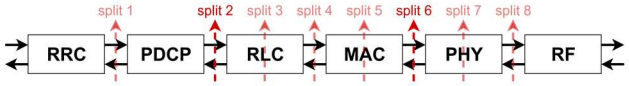

Fig 1. Functional split options and their respective baseband functions: Radio Resource Control (RRC), Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), Medium Access Control (MAC), Physical layer (PHY), and Radio Frequency functions (RF).

Among all proposed functional splits, we will evaluate the following two:

- Split 2: divides the datalink layer functions. PDCP and the network-layer function RRC are placed in the CU. The remaining datalink and all physical layer functions are located in the DU.
- Split 6: separates the physical and datalink layer functions. All physical processing is performed locally, while the remaining RRC, PDCP, RLC and MAC functions are handled by the CU.

As the functional split options have distinct requirements in terms of bandwidth, latency and computational capacity, we present the models used to evaluate these parameters.

1) Bandwidth: when considering the different split options proposed by 3GPP, the higher is the split layer (split 1 in Fig. 1 is the highest), the less bandwidth is required. As a matter of fact, split 8 (also known as CPRI) presents very high and constant required throughput, since the Fast Fourier Transformation (FFT) is performed only in a higher layer. The model in Eqn. (1) is based on [4], and provides the throughput required for the considered functional splits.

$$B = \begin{cases} n_{pckt}(s_{pckt} + header)TBS, & \text{if } split = 2 \\ n_{pckt}(s_{pckt} + header)TBS + c, & \text{if } split = 6 \end{cases} \quad (1)$$

where n_{pckt} is the number of IP packets per transport block, s_{pckt} is the IP packet size, $header$ is the total header size, which includes only PDCP in split 2, and PDCP, RLC and MAC in split 6, TBS is the number of transport blocks per Transmission Time Interval (TTI), and c is an additional bandwidth cost related to the overhead of the functional application platform interface. Based on this, it is possible to understand by which ratio the traffic increases/decreases when processed by the baseband functions. This expansion factor is therefore calculated as $\epsilon = t_d/B_{split}$, where t_d is the traffic of DU d.

2) Computational capacity: to determine the computational capacity required by each functional split to process the traffic demand, we use a model based on [17], [18]. It considers the capacity to be propotional to the CPRI, as in Eqn. (2).

$$C_{split} = \sigma_{split}\left(n_a^2 + 3 \cdot n_a + \frac{M \cdot C \cdot L}{3}\right)\frac{R}{10} \quad (2)$$

where n_a is the number of antennas per user, M modulation bits, C code rate, L number of MIMO layers, R number of resource blocks per user, σ_{split} scaling factor for each functional split. This parameter is calculated based on the computational complexity of each functional split in respect to the CPRI, and the values are based on [17].

3) Latency: the maximum acceptable transport latency is determined by the functional split option selected. According to [6], this value is 1.5 ms and 0.25 ms in splits 2 and 6, respectively. In this paper, we characterize the latency of a certain demand from a DU by the propagation delay and the baseband processing latency. The propagation delay component takes into account all the links in which the DU demand passes until reaching the CU handling the traffic of this DU. The processing latency, see Eqn. (3), considers the required computational effort to process all the traffic, and the server characteristics [11], [19].

$$\lambda_{split} = \frac{C_{split}}{C_{CPU} \cdot f_{CPU}} \quad (3)$$

where C_{CPU} is the computational capacity of the CPU and f_{CPU} is the CPU operating frequency.

B. CU placement problem description

In order to dynamically deploy baseband functions, we need to assign IT and network resources to the demands such that power consumption is reduced. We assume that a single CU is implemented in each node that the algorithm sets to active. In addition, each CU processes traffic demands from multiple DUs. This optimization is subject to the latency associated to the split and to the service, the network capacity, and the computational capacity. The parameters and decision variables considered in the analysis are described in Table I.

Objective function: the optimization goal is to minimize the total power consumption of the server and network components. The server power is characterized by the idle power whenever the node is active, and by the power related to its utilization. The network power in this paper considers the transponder consumption. The objective function is given in Eqn. 4:

$$\min \sum_{n \in N}\Big((P_{max}^s - P_{idle}^s)\sum_{d \in D} t_d \cdot x_n^d \cdot \pi_n + P_{idle}^s \cdot x_n\Big) + \sum_{d \in D, e \in E, n \in N} P^t \cdot x_e^d \cdot \gamma_e^n \quad (4)$$

Constraints:

1) Routing

$$\sum_{e \in E_n^+} x_e^d - \sum_{e \in E_n^-} x_e^d = \begin{cases} 1, \text{if } n = n_d \\ -1, \text{if } n = n_{gw} \\ 0, \text{otherwise} \end{cases}, \forall d \in D, n \in N \quad (5)$$

TABLE I. PARAMETERS AND VARIABLES IN ILP

Parameters	
N	Set of nodes
D	Set of all demands from/to DUs
E	Set of IP links in the network
E_n^-	Set of node n incoming links
E_n^+	Set of node n outgoing links
P_{idle}^s	Server idle power consumption
P^t	Transponder power consumption
t_d	Traffic required by DU d
n_d	Node hosting DU d
n_{gw}	Gateway node
θ_d	Pair of UL-DL demands between DU d and gateway
M	Very large number
C_e	Link capacity
C_n	Node processing capacity
π_n	Workload needed to process 1 Gbit/s
λ_e	Propagation latency over link
λ_n	Processing latency in node to process 1 Gbit/s
λ_{max}^{split}	Maximum one-way allowed latency for split
λ_{max}^{serv}	Maximum end-to-end allowed latency for service
ϵ	Traffic expansion factor after demand is processed in CU
γ_e^n	Binary coefficient taking value 1 if link e starts or ends in node n
$\gamma_e^{n,c}$	Binary coefficient taking value 1 if link e ends in node n, and n is a candidate to host a CU
δ_e^n	Coefficient taking value 1 if link e starts in node n, -1 if e ends in n and 0 otherwise
Decision variables	
x_n	Binary variable taking value 1 if node n is used as a CU
x_n^d	Binary variable taking value 1 if node n processes demand from DU d
x_e^d	Binary variable taking value 1 if link e carries demand from DU d
$x_e^{d,n}$	Binary variable taking value 1 if link e carries demand from DU d between the node that hosts the CU processing the demand from DU d and node $n \in \{n_d, n_{gw}\}$
Λ_n	Total processing latency in node n

$$x_e^{d,n} \cdot \delta_e^n = \begin{cases} 1 - x_n^d, \text{if } n = n_d \\ x_n^d - 1, \text{if } n = n_{gw} \end{cases}, \forall d \in D, e \in E \quad (6)$$

$$x_e^{d,n_d} + x_e^{d,n_{gw}} = x_e^d, \forall d \in D, e \in E \quad (7)$$

$$x_e^d \cdot \gamma_e^{n,c} = x_n^d \quad (8)$$

$$M \cdot x_n \geq \sum_{d \in D} x_n^d, \forall n \in N \quad (9)$$

$$\sum_{n \in N} x_n^d = 1, \forall d \in D \quad (10)$$

2) Capacity

$$t_d \cdot x_e^d \leq C_e, \forall d \in D, e \in E \quad (11)$$

$$t_d \cdot x_n^d \cdot \pi_n \leq C_n, \forall d \in D, n \in N \quad (12)$$

3) Latency

$$\sum_{e \in E} x_e^{d,n_d} \cdot \lambda_e + \Lambda_n \leq \lambda_{max}^{split}, \forall d \in D \quad (13)$$

$$\sum_{d' \in \theta_d}\Big(\sum_{e \in E} x_e^{d,n_d} \cdot \lambda_e + \Lambda_n\Big) \leq \lambda_{max}^{serv}, \forall d \in D \quad (14)$$

$$0 \leq \Lambda_n \leq 1.5 \cdot x_n^d, \forall n \in N \quad (15)$$

$$\sum_{d' \in D}(t_{d'} \cdot x_n^{d'})\lambda_n - 1.5(1 - x_n^d) \leq \Lambda_n \leq \sum_{d' \in D}(t_{d'} \cdot x_n^{d'})\lambda_n, \forall n \in N \quad (16)$$

Eqn. (5) indicates the flow conservation in the network, ensuring that all incoming demands will be forwarded. Eqn. (6) and Eqn. (7) ensure that a request from/to DU d to/from the gateway, passing in the node processing its demand, is carried both in uplink (UL) and downlink (DL). Eqn. (8) determines the node containing CU functions to process the demand from DU d. Eqn. (9) indicates that node n hosts a CU. Eqn. (10) restricts the DU demands to be processed by a single node. Eqn. (11) guarantees that the sum of the bandwidth of all demands routed over link l does not exceed its capacity. Eqn. (12) ensures that the total computational effort necessary to process the demands does not exceed the node capacity. Eqn. (13) constraints the one-way latency of a demand from the DU to the processing node to be under the split maximum accepted latency. Eqn. (14) restricts the end-to-end delay to be under the service maximum accepted latency. Eqn. (13) and Eqn. (14) are composed by two elements: the propagation delay, and the node processing delay, which is computed by Eqn. (15) and Eqn. (16).

IV. RESULTS AND DISCUSSION

This section describes the evaluation of the mathematical programming described in Section III. First, we present the simulation setup, indicating the scenario used for the simulations, and then we explain the obtained results.

A. Simulation environment

To perform the numerical simulations, we relied on Net2Plan open source network planner using the NFV over IP over WDM library [20]. In this platform, we implemented the CU placement optimization proposed in Section III, which will be indicated as ILP scenario, and two other baseline scenarios.

- *D-RAN:* fully distributed scenario, *i.e.* all nodes host a CU, except the gateway.
- *C-RAN:* fully centralized scenario, *i.e.* only the gateway hosts a CU.

The optimization proposed was assessed using a dense urban metro-meshed network representing the typical network topology of a big city from the context of Metro-Haul project [8]. Fig. 2 depicts the 35-node topology strongly inspired, even if not exactly coincident, by a metro network of Telecom Italia that was used for simulations. It contains 34 Access Metro Edge Nodes (AMEN) and one Metro Core Edge Node (MCEN). The MCEN is the only node connected to the core network; therefore, it is always the destination in UL, and origin in DL. The remaining nodes, *i.e.* AMENs, are treated as candidate CU, and each of them is assumed to be connected to a set of DUs.

In order to obtain realistic traffic demands to evaluate the optimization problem, we integrate this topology into the context of two open datasets: OpenCellid [21] and TIM Big Data Challenge [22]. The former contains the coordinates of base stations in Milan. The latter reports an anonymized interaction of users with the mobile network every ten minutes, providing the pattern of network usage instead of precise values of Internet

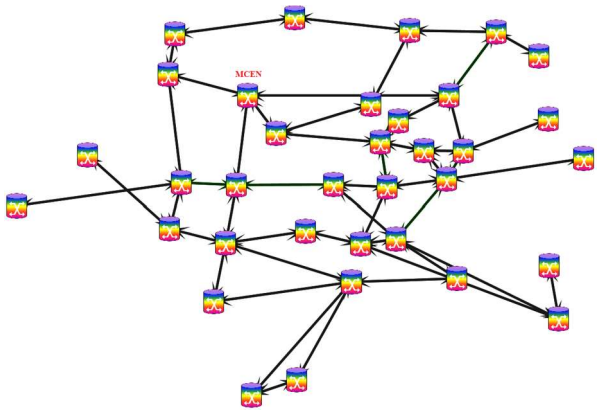

Fig. 2. Topology of the metro network with one MCEN and 34 AMENs.

traffic. The traffic profiles provided by these datasets represent the mobile access traffic and are used as inputs to our optimization model. We consider eight time slots per day in the simulations. Each of them lasts three hours and have a constant traffic equal to the peak of the dynamic traffic of the dataset over the same three hours. The optimization algorithm was therefore executed for each time slot to consider this traffic dynamicity.

Table II summarizes the parameters used in the simulations.

TABLE II. INPUT PARAMETERS USED IN THE SIMULATIONS

Parameters	Value
# MIMO layers	2
QAM modulation	64
Code rate (DL/UL)	0.93 / 1
# resource blocks per user	100
# IP packets per transport block (DL/UL)	6.24 / 4.05 [4]
IP packets size	1500 B [4]
Header size (split 2 / 6)	2 / 9 B [4]
Maximum split latency (split 2 / 6)	1.5 / 0.25 ms [6]
Maximum service latency	10 ms
Server capacity	537.6 GFLOPS [24]
# cores	8
CPU maximum frequency	3.7 GHz
Server power consumption	870 / 130 W
# servers in candidate CU nodes	1
# servers in gateway (split 2 / 6)	2 / 3
Transponder 100 Gbit/s power consumption	110.4 W [23]
# wavelength per fiber	8
Wavelength capacity	100 Gbit/s

The radio configuration considered follows the parameters proposed in [4]. We assume the scenario in which there is a single user per TTI to obtain the maximum throughput transmitting 100 resource blocks of data. The radio is configured with 2x2 MIMO antenna, 64QAM modulation, 200 MHz of channel bandwidth and modulation coding scheme of 28 in DL and 23 in UL. We also consider 6.24 IP packets per transport block in DL, and 4.05 in UL. Each of these packets are 1500 bytes long, without considering the header, which is equal to 2 bytes in split 2, and 9 bytes in split 6.

Regarding the transport network, the nodes are connected via bidirectional fiber links, containing 8 wavelengths with maximum capacity of 100 Gbit/s, following [9]. The nodes are also equipped with a set of 100 Gbit/s transponders which consume 110.4 W [23] whenever active. In terms of the nodes computational capacity required to perform baseband functions

related to each split, we consider the server Intel® Xeon® Gold 6134 with 8 cores, processing capacity of 537.6 GFLOPS, maximum operating frequency of 3.7 GHz. This server idle power consumption is 130 W, and we assume that the idle power corresponds to 15% of the maximum, *i.e.* 870 W. In order to ensure that all above-mentioned scenarios are feasible, we placed one server at each candidate CU node (8 cores), while the gateway was equipped with two servers in split 2 and three in split 6 (16 and 24 cores, respectively).

B. Numerical results

Fig. 3(a) depicts the overall power consumption in split 2. This graph shows both IT and the TLC components power consumption on the D-RAN, C-RAN, and ILP scenarios. The IT share of this value represents the power consumption of servers in the entire network, *i.e.* in DU and CU nodes, to assess the total baseband processing consumption. The TLC comprises the used transponders in each node.

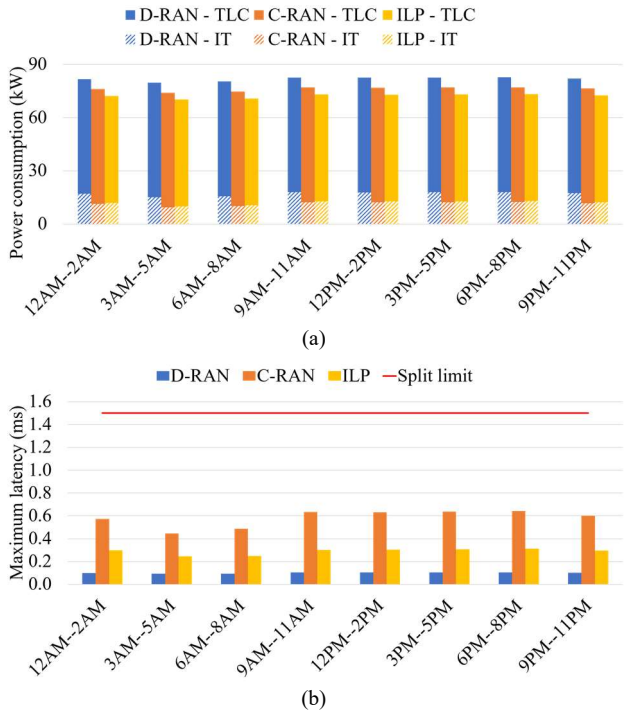

Fig. 3. Analysis of split 2. (a) Power consumption of IT (pattern) and TLC (solid). (b) Maximum one-way latency over all demands.

The IT power consumption illustrated in Fig. 3(a) shows that the D-RAN scenario consumes the highest amount in all time slots, followed by the ILP and C-RAN. This characteristic comes from the fact that, although underutilized, all nodes are active, hence consuming at least the idle power. Considering the C-RAN scenario, the node power consumption represents almost 70% of the D-RAN, because, even if a single node is used, it has the greatest capacity and, consequently, power. Lastly, the ILP is an in-between scenario, because it selects fewer nodes to host CUs. Considering the total power consumption, the D-RAN scenario is still the one with highest power demand. Nevertheless, because of the network component in the ILP optimization, it is the less power consuming. It enables reducing in average 5% and 12% in comparison to the C-RAN and D-RAN scenarios, respectively.

Fig. 3(b) shows the maximum latency for all DU demands in split 2 obtained in this simulation compared to the limit latency allowed by split 2. As expected, the D-RAN reaches the lowest latency as there is no propagation delay in the metro network because the DU is directly connected to the CU. Furthermore, the C-RAN scenario is the one with the highest reached latency, with values over 2 and 5 times the distributed and ILP solutions, respectively. Nonetheless, all scenarios present latency values under the maximum allowed by split 2, *i.e.* 1.5 ms.

Fig. 4(a) and (b) present the power and maximum latency, respectively, of the simulations using split 6 in the evaluated scenarios.

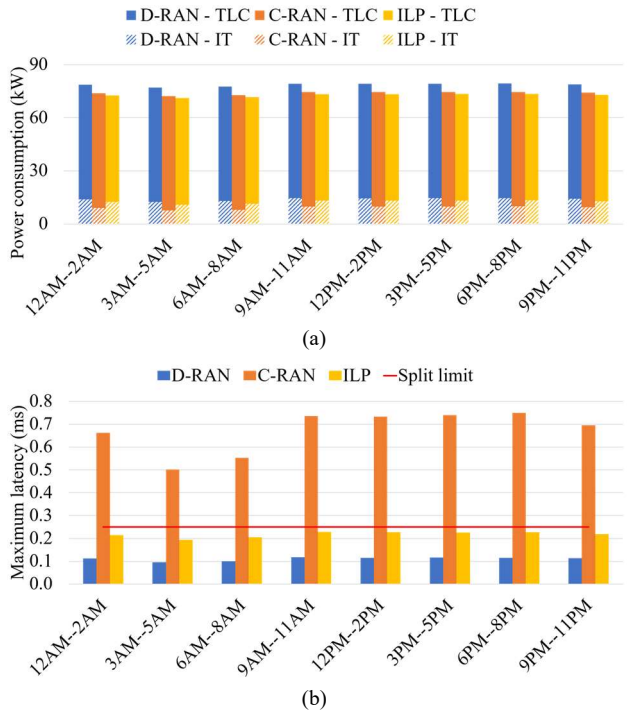

Fig. 4. Analysis of split 6. (a) Power consumption of IT (pattern) and TLC (solid). (b) Maximum one-way latency over all demands.

As in split 2, D-RAN presents the highest power consumption both in terms of servers and transponders. Indeed, the ILP optimization proposed in this paper enables reducing more than 7% the consumption. The difference to the previous simulation arises when analyzing the C-RAN and ILP scenarios. In split 6, the difference in terms of power consumption decreases, and the ILP reduces at most 2% with respect to the C-RAN. This result is explained by the tight constraints in terms of latency in split 6 (0.25 ms). Analyzing Fig. 4(b), it is possible to observe this fact. C-RAN presents very high latency, reaching 0.7 ms in average and, thus, not respecting the constraint. The ILP solution, on the other hand, can satisfy the maximum latency for all demands and at any time of the day.

Evaluating Fig. 3(a) and Fig. 4(a) in terms of functional split option, the overall power consumption of split 2 is 5% lower than split 6, which is driven mainly by the IT part. Fig. 5 compares in more detail the IT consumption using the ILP in both splits. The solid fill represents the CU component of IT power consumption, while the pattern shows the DU part.

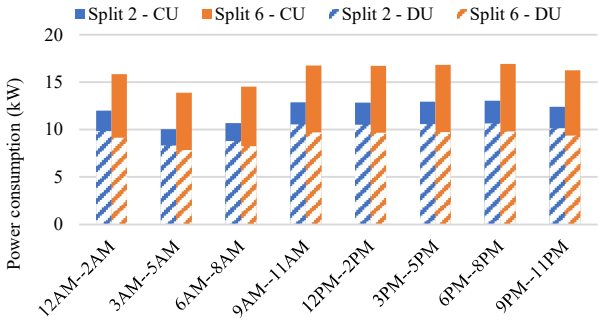

Fig. 5. Power consumption of DUs (pattern) and CUs (solid) per split.

The splits present very similar power consumption in terms of the DU. The light reduction of split 6 with respect to split 2 (1%) is due to the lower number of functions that need to be handled in these nodes. However, analyzing the CU consumption, the use of split 6 increases in more than 30% compared to the split 2. Because of the tight requirements in terms of latency of split 6, the algorithm needs to activate more nodes so that the one-way latency remains under 0.25 ms.

V. Final Remarks

This paper presented an evaluation of the baseband-VNF placement in an IP-over-optical metro network. First, we described the modeling of the functional split constraints, *i.e.* bandwidth, processing and latency. Then, we proposed an ILP whose objective is to minimize the power consumption of IT and TLC components subject to the functional split constraints. We further evaluated the proposed solution over a telco operator's inspired metro topology using realistic traffic and compared it to two baselines: D-RAN (fully distributed) and C-RAN (fully centralized). Results showed that the ILP enables decreasing on average 5% and 12% the power consumption when compared to the C-RAN and D-RAN, respectively, in split 2. In split 6, the ILP solution has a lower cost than D-RAN but slightly higher than C-RAN. However, the latter is not able to satisfy the latency requirement (0.25 ms one-way) and therefore the ILP is in fact the lowest-cost feasible solution. Furthermore, when comparing the splits, split 2 enables reducing the overall power consumption by 5% and by more than 30% if only IT power is considered.

Future work will focus on a sensitivity analysis to understand the impact of each power component (servers and transponders) on the final result. Moreover, we plan to consider dynamic placement of baseband VNFs along the day.

ACKNOWLEDGMENT

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

References

[1] M. Agiwal, A. Roy and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1617--1655, 2016.

[2] C. I, H. Li, J. Korhonen, J. Huang and L. Han, "RAN Revolution With NGFI (xhaul) for 5G," in Journal of Lightwave Technology, vol. 36, no. 2, pp. 541--550, 2018.

[3] L. M. Larsen, A. Checko and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," in IEEE Communications Surveys \& Tutorials, vol. 21, no. 1, pp. 146--172, 2018.

[4] Small Cell Forum, "Small cell virtualization functional splits and use cases," Jan. 2016.

[5] A. Checko, A. P. Avramova, M. S. Berger and H. L. Christiansen, "Evaluating C-RAN Fronthaul Functional Splits in Terms of Network Level Energy and Cost Savings," in Journal of Communications and Networks, vol. 18, no. 2, pp. 162--172, 2016.

[6] 3GPP TR 38.801 V14.0.0 (2017-03), Radio access architecture and interfaces (Release 14)

[7] J. S. Wey and J. Zhang, "Passive Optical Networks for 5G Transport: Technology and Standards," in Journal of Lightwave Technology, vol. 37, no. 12, pp. 2830--2837, 2019.

[8] Metro-Haul, [Online]. Available: https://metro-haul.eu/

[9] F. Musumeci, O. Ayoub, M. Magoni and M. Tornatore, "Latency-Aware CU Placement/Handover in Dynamic WDM Access-Aggregation Networks," in Journal of Optical Communications and Networking, vol. 11, no. 4, pp. B71--B82, 2019.

[10] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher and V. Young, "Mobile Edge Computing: A key technology towards 5G," ETSI white paper, 2015.

[11] A. De Domenico, Y. Liu and W. Yu, "Optimal Computational Resource Allocation and Network Slicing Deployment in 5G Hybrid C-RAN," IEEE International Conference on Communications, China, 2019.

[12] H. Yu, F. Musumeci, J. Zhang, Y. Xiao, M. Tornatore and Y. Ji, "DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks," Conference on Optical Network Design and Modelling, Greece, 2019.

[13] A. N. Al-Quzweeni, A. Q. Lawey, T. E. H. Elgorashi and J. M. H. Elmirghani, "Optimized Energy Aware 5G Network Function Virtualization," in IEEE Access, vol. 7, pp. 44939--44958, 2019.

[14] J. Yusupov, A. Ksentini, G. Marchetto and R. Sisto, "Multi-objective Function Splitting and Placement of Network Slices in 5G Mobile Networks," IEEE Conference on Standards for Communications and Networking, France, 2018.

[15] O. Arouk, T. Turletti, N. Nikaein and K. Obraczka, "Cost Optimization of Cloud-RAN Planning and Provisioning for 5G Networks," IEEE International Conference on Communications, USA, 2018.

[16] A. Tzanakaki, M. P. Anastasopoulos and D. Simeonidou, "Converged Optical, Wireless, and Data Center Network Infrastructures for 5G Services," in Journal of Optical Communication Networks, vol. 11, no. 2, pp. A111--A122, 2019.

[17] B. Debaillie, C. Desset and F. Louagie, "A Flexible and Future-Proof Power Model for Cellular Base Stations," IEEE Vehicular Technology Conference, Scotland, 2015.

[18] M. Shehata, A. Elbanna, F. Musumeci and M. Tornatore, "Multiplexing Gain and Processing Savings of 5G Radio-Access-Network Functional Splits," in IEEE Transactions on Green Communications and Networking, vol. 2, no. 4, pp. 982--991, 2018.

[19] X. Wang, L. Wang, S. E. Elayoubi, A. Conte, B. Mukherjee and C. Cavdar, "Centralize or distribute? A techno-economic study to design a low-cost cloud radio access network," IEEE International Conference on Communications, France, 2017.

[20] J. L. Romero-Gázquez, M. Garrich, F. M. Muro, M. Bueno Delgado and P. P. Mariño, "NIW: A Net2Plan-Based Library for NFV over IP over WDM Networks," International Conference on Transparent Optical Networks, France, 2019.

[21] Unwire Labs, OpenCellid, [Online]. Available: http://opencellid.org/.

[22] TIM, Big Data Challenge, 2014 [Online]. Available: https://dandelion.eu/datamine/open-big-data/.

[23] J. M. H. Elmirghani, T. Klein, K. Hinton, L. Nonde, A. Q. Lawey, T. E. H. El-Gorashi, M. O. I. Musa and X. Dong, "GreenTouch GreenMeter Core Network Energy-Efficiency Improvement Measures and Optimization," in Journal of Optical Communications and Networking, vol. 10, no. 2, pp. A250--A269, 2018.

[24] Intel, "CTP Metrics for Intel Mocroprocessors - Intel Xeon Scalable Processors," 2019