Debunking the "Green" NFV Myth: An Assessment of the Virtualization Sustainability in Radio Access Networks

Raffaele Bolla
S2N National Lab., CNIT, Italy
DITEN, University of Genoa, Italy
raffaele.bolla@unige.it

Roberto Bruschi
DITEN, University of Genoa, Italy
roberto.bruschi@unige.it

Franco Davoli
S2N National Lab., CNIT, Italy
DITEN, University of Genoa, Italy
franco.davoli@unige.it

Chiara Lombardo
S2N National Lab., CNIT, Italy
chiara.lombardo@cnit.it

Jane Frances Pajo
S2N National Lab., CNIT, Italy
DITEN, University of Genoa, Italy
jane.pajo@tnt-lab.unige.it

***Abstract*— Since the adoption of virtualization paradigms is seen as a viable way to fulfil the requirements of next-generation applications in a sustainable way, this paper examines the virtualization of the Radio Access Network (RAN), focusing on the C-RAN architecture, in order to better understand the impact of NFV technologies on power consumption and costs in real networks. The evaluation compares the power consumption obtained by deploying the Base Band Unit (BBU) using commercial devices or pools of Virtual Network Functions (VNFs). Publicly available datasets describing the traffic and the eNodeBs have been used for the evaluation, as well as datasheets for both the commercial devices and the VNF pools. Results show that the usage of the virtualized BBU causes consumptions around 250% higher with respect to the commercial deployment, and operation and capital costs over 66% higher, contradicting the common belief of NFV being a "green" technology. Further estimates conducted in this paper, however, highlight how the deployment of VNFs alongside specialized hardware solutions can represent a successful approach for telecom providers, with energy savings up to 20% and costs in line with the ones of dedicated hardware deployments.**

Keywords—NFV, C-RAN, power consumption, TCO

I. INTRODUCTION

With the upcoming fifth-generation (5G) mobile networks finally drawing close, telecom providers are facing a number of challenges to satisfy growing demands in a sustainable way. In fact, the deployment of small cells, the support of multiple access technologies and new services and business models, required to support the extreme low latency vertical applications and services [1], will increase the amount of data traversing the network and, consequently, the costs to build, operate and upgrade mobile networks.

In order to address such challenges, Network Functions Virtualization (NFV) [2] is seen as a key solution, to the point that the 5G Service-based Architecture (SBA) has been conceived with the design of the core functionalities as highly pluggable Virtualized Network Functions (VNFs).

The allocation of VNFs on demand allows horizontally scaling the individual instances with respect to traffic load, over both time and area/population density, to fulfil the heterogeneous and even more stringent latency requirements of next-generation mobile applications. As a consequence, it would be possible to reduce power consumption and, additionally, the presence of a virtualized environment makes for easier upgrades and maintenance of the infrastructure.

However, the lack of clear figures supporting these claims leaves room for debate on whether the adoption of the NFV paradigm by itself would lead to straight improvements in either the energy efficiency or the performance of the mobile networks, and which would be the overall impact on the OPerating Expenses (OPEX) and CAPital Expenses (CAPEX).

In an attempt to seek clarification on this point, and following the path of [3], this paper examines the virtualization of the Radio Access Network (RAN), focusing on the C-RAN architecture, in order to provide a breakdown of the power consumption and costs that could be expected with the deployment of NFV technologies in real networks. The reported evaluations are based on real, publicly available datasets about the infrastructure and traffic of one of the main Italian mobile operators, and on datasheets for both commercial devices and pools of VNFs, under different deployment scenarios.

Results show that the usage of the NFV paradigm alone, although it can improve proportionality with the incoming traffic load, causes significantly higher power consumption. However, while the sole introduction of NFV does not lead to straight improvements in either energy efficiency or costs, its flexibility level allows for an effective integration with specialized commercial products to leverage on the strength of both technologies in an adaptive way with respect to the time, area and traffic load demands.

The remaining of the paper is organized as follows. Section II presents the main features of the C-RAN technology, and the challenges and open points of applying virtualization technologies to RAN. Section III describes the C-RAN deployment that has been considered in the paper, including the datasets characterizing the access network deployment and the traffic in the considered different scenarios, while the related numerical evaluation is reported in Section IV. Finally, conclusions are drawn in Section V.

II. C-RAN PRINCIPLES AND CHALLENGES IN THE UPCOMING 5G PERSPECTIVE

In order to find a response to the increasing costs of building and operating the RAN segment, C-RAN was introduced with the goal of making a more efficient utilization of eNodeB resources. In C-RAN, the radio and baseband processing functionality made available by means of the E-UTRAN Node B (eNodeB) are split, with the Remote Radio Head (RRH), providing the interface to the front-haul and the

signal digitalization, left in the cell site, and the Baseband Unit (BBU) moved to a remote cloud facility offering lower rental and/or maintenance costs. Following along, lower Average Revenue Per User (ARPU) can be achieved by co-locating baseband processing functionality for multiple sites in the same facility: matter-of-factly, the C- in C-RAN can stand for both Cloud and Centralized.

Although the original concept of C-RAN is almost ten years old now [4], this framework has recently gained momentum thanks to the maturity level achieved by enabling paradigms such as cloudification and NFV. In fact, with resources being flexibly allocated according to the current demand, and executed on general-purpose hardware rather than on the proprietary platforms previously used for base stations, costs can be reduced both when planning/deploying a new infrastructure and in the successive updates/upgrades. As a result, a number of proprietary and open source solutions, such as OpenAirInterface [5] have been released to provide various levels of virtualization of the RAN.

The relevance of C-RAN further grows in the perspective of the upcoming 5G networks. In fact, while previous mobile access networks had a monolithic architecture, in which all functionality was provided by a single building block (e.g., the eNodeB in 4G), 5G has been conceived from the start [6] to be service-based, with most access and core functions deployed as VNFs running in virtual machines on standard servers within cloud computing infrastructures. In particular, for the access, gNodeB functionality is logically split between a Central Unit (CU) and one or more Distributed Units (DUs). Different options for deciding which functions are to be centralized or distributed are still under study. The functional split can be further mapped into different deployment scenarios, with a number of Control Plane (CP) and User Plane (UP) network functions potentially deployable in a centralized or distributed fashion.

Considering these architectural characteristics, the adoption of C-RAN could foster the transition from 4G to 5G in a sustainable way. In fact, 5G requires supporting multiple access technologies and, at least in an initial stage, integration with LTE-A RAN, and the new services and business models made possible by 5G will increase the amount of data traversing the network. Furthermore, the deployment of small cells clearly requires a higher number of eNodeBs, resulting in an increase of the costs ascribable to RAN.

A potential C-RAN deployment including both 4G and 5G access is shown in Fig. 1. We can see eNodeBs and gNodeBs functions split and deployed partly in the cell site and partly in a centralized cloud-based infrastructure, connected to the cell sites by fiber. Here, pools of BBUs/CUs are deployed on general purpose IT servers and made available to be dynamically allocated on demand by a hypervisor according to the current traffic load.

The benefits of such a deployment are abounding. The allocation of VNFs on demand promotes scalability with respect to traffic load over both time and area/population density. As a result, it would be possible to reduce power consumption and consequently OPEX. Additionally, the presence of a virtualized environment makes for easier upgrades and maintenance of the infrastructure.

However, while the usage of NFV, by promoting flexibility and scalability thanks to the usage of general-purpose hardware, is widely considered as a valid way of addressing most of the current 4G and upcoming 5G RAN challenges, the lack of clear figures supporting this claim leaves room for debate on whether the introduction of this paradigm by itself would lead to straight improvements in either energy efficiency or performance levels, and which would be the overall impact on the OPEX and CAPEX. In the interest of gaining a better understanding on the impact of softwarization on energy efficiency and cost reductions in the RAN, in the next sections we will compare the power consumption and the OPEX and CAPEX of a C-RAN using commercial devices or pools of VNFs.

III. Reference Scenario

We considered a C-RAN deployment over a reference metropolitan area that includes Milan, Italy, and neighboring cities, covering an area of 552.25 km^2. We consider the Internet traffic activity over the area for the Telecom Italia Mobile (TIM) customers, which comes from the Open Big Data initiative ("Milano Grid" dataset), publicly available at [7].

The entries of the dataset are spatially aggregated in squares according to the GeoJSON format and report the level of interaction of the users with the mobile phone network over time windows of ten minutes. In order to translate these levels into traffic rates in line with the current loads, they have been multiplied by a random traffic rate between 6 and 25 Mbps, in accordance with the Cisco Mobile Visual Networking Index (VNI) mobile speed forecasts [8]. For the results presented in Section IV, we have compared the traffic activity on a working day (December 6th) and a holiday (December 25th). Heatmaps representing the distribution of users over the reference area are shown in Fig. 2 and Fig. 3.

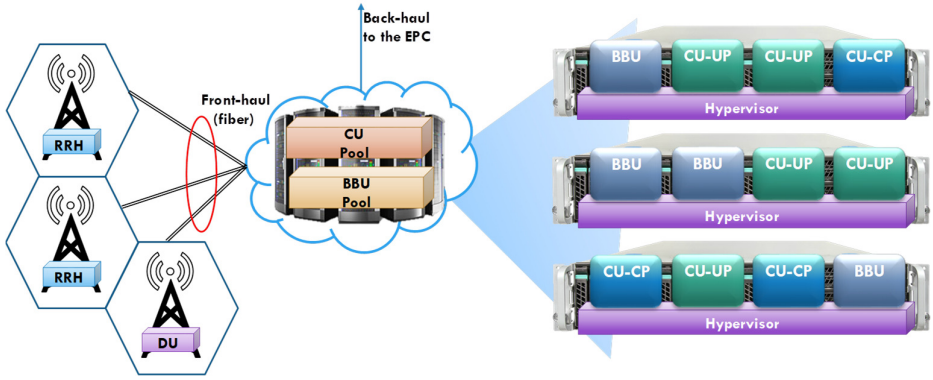

Fig. 1. Potential C-RAN deployment including both 4G and 5G access.

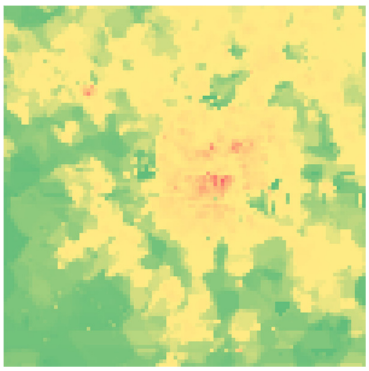

Fig. 2. Heatmap of the traffic distribution over the reference area at 2 PM (rush hour) on December 6th.

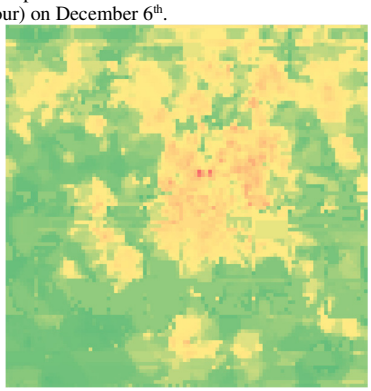

Fig. 3. Heatmap of the traffic distribution over the reference area at 2 PM (rush hour) on December 25th.

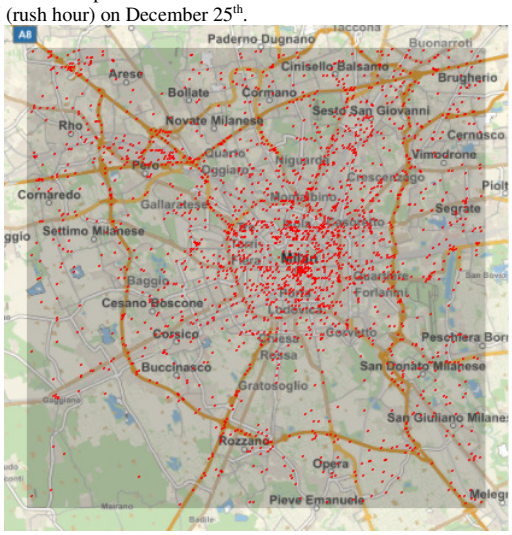

Fig. 4. LTE eNodeBs distribution over the reference area.

The LTE eNodeBs deployed by TIM over the reference area come from the OpenCellID database [9] and can be seen in Fig. 4. In order to characterize the traffic volume for each eNodeB over the selected days, an emulation has been run using Wolfram Mathematica [10], which assigns the users of the "Milano Grid" dataset to the closest eNodeB. Constraints on the maximum allowed traffic volume ensure that, if the closest eNB capacity is nearly saturated, the second closest is selected and so on.

For the deployment of the centralized BBU pool, we compare the power consumption obtained by using commercial [11] and virtual, server-based BBUs (cBBU and vBBU hereinafter, respectively). Information for the characterization of the cBBU are obtained from datasheets; for the sake of this paper, it is worth noting that a board can house up to six BBU modules, and each one can serve one eNB. To characterize the vBBU deployment, we have used the architecture and related performance models from [12], in which EURECOM OpenAirInterface [5] is used to realize the virtualized C-RAN system running on Intel Xeon-based servers, and a CPU utilization model allows determining the required number of servers according to the throughput. Table I reports the complete specifications.

The only consideration made regarding RRHs is that a single BBU pool placed in the center of the reference area allows keeping the maximum distance between RRH and BBU below 20 km, which fulfils the constraint of sub-frame processing delay on a link to be below 1 ms [4]. The presence of a single central office hosting the BBU pool also allows for the same considerations on spectral efficiency, fiber connections and front-haul transmission solutions to hold true for both the cBBU and the vBBU cases.

IV. Numerical Evaluation

In this section, we compare the BBU power consumption of commercial devices and pools of VNFs. We consider the scenario and datasets described in Section III, and evaluate how power consumption changes throughout the day, and between working days and holidays, for the cBBU and vBBU cases.

Since we consider a single, centralized BBU pool for both cases, we only account for the consumption ascribable to the devices: for the cBBU, one module is deployed for each eNodeB in the reference area and is kept constantly powered on, while for the vBBU, we consider a number of active servers, varying according to the load as indicated in [12], and of switches computed by using the k-ary fat-tree topology [13].

Fig. 5 shows the daily trends for the two cases obtained on the 6th and 25th of December. Since the devices deployed for the cBBU case are always active, there is only one line for both days and it shows a constant value throughout the different times of the day. On the other hand, thanks to the power saving mechanisms available in general purpose processors, the vBBU case presents significant differences between the working day and the holiday, with variations up to 70% between the two days and 20% throughout each day. It is worth noting that around 15% of the power consumed is ascribable to the switches.

It is clear how the usage of vBBU results in a higher level of proportionality with the incoming traffic load and provides better results when the traffic is low (for example, between 00:00 and 02:30 AM); however, consumptions are significantly higher with respect to the commercial deployment.

Although the NFV paradigm has been trending for several years now, actual implementations are still in a prototypical

TABLE I. Reference Architectures Characterization

cBBU Specifications		vBBU Specifications	
Board consumption [W]	40	Motherboard consumption [W]	145
BBU modules per board [#]	6	Processors per server [#]	2
BBU module consumption [W]	85	Processor consumption [W]	130
Fan consumption [W]	53.5	NICs per server [#]	8
eNBs per BBU module [#]	1	NIC consumption [W]	6.8

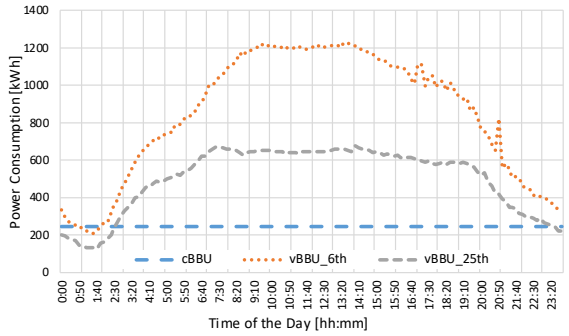

Fig. 5. Power consumption in a working day and a holiday for the cBBU and the vBBU cases.

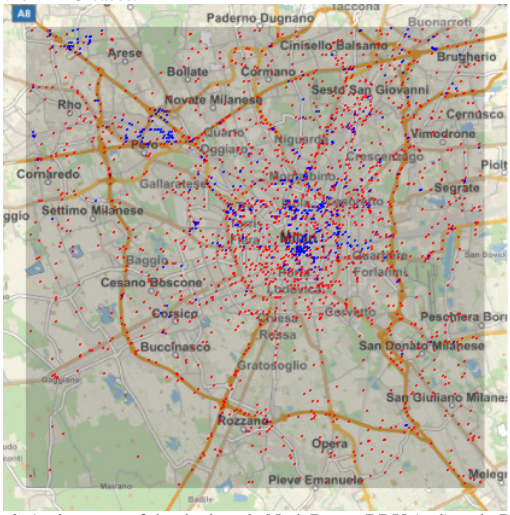

Fig. 6. Assignment of the deployed eNodeBs to cBBU (red) and vBBU (blue).

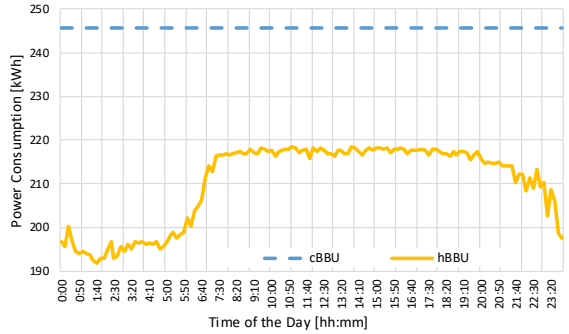

Fig. 7. Power consumption in a working day for the cBBU and the hBBU case.

state and their application in a realistic use case cannot compare to a mature, commercial product. Of course, room for improvement can be envisaged, with one of the most promising enhancements being the decomposition of specific VNFs into sub-tasks to be executed in parallel to fully exploit the performance potentials of multi-core processors [14]. Further solutions can take advantage of the consolidation policies typically applied within datacenters to tune the resources assigned to VNFs. However, all of these solutions result in unfavorable trade-offs towards either power consumption or performance, which leads to believe that NFV technologies will not be able to measure up to commercial ones at least with the current generation of processors.

While the introduction of the NFV paradigm by itself does not lead to straight improvements in either energy efficiency or performance levels, the flexibility brought forth by the usage of general purpose hardware, and the resulting scalability, makes this technology well-suited to be integrated with other solutions and realize heterogeneous infrastructures. For example, the co-location of centralized vBBU pools inside the micro datacenters used for Edge Computing can promote not only a better exploitation of the available physical resources, allocating them to different functions according to the current demands, but even the utilization of the available VNF components for creating customizable network slices. Another feasible design is represented by the deployment of VNFs alongside specialized hardware solutions. In fact, it is expected that a potential transition to full softwarization will be undoubtedly preceded by a period of coexistence with dedicated devices. The cooperation of dedicated hardware and VNFs can generate remarkable benefits by exploiting the scalability of software solutions in the presence of lower traffic loads and the higher performance provided by commercial products where demands are higher.

In this respect, further results have been computed to evaluate the advantages that can be obtained in the presence of a heterogeneous deployment. In details, we have exploited the emulation outcomes to identify which RRHs on average process a level of traffic low enough to fall in the range in which the vBBU consumes less energy than the cBBU (e.g., to fall in the same consumption as between 00:00 and 02:30 AM in Fig. 5). Then, we have assigned the traffic from such RRHs to be processed by the vBBU pool, while keeping the remaining ones (whose traffic, as can be seen in Fig. 2 and Fig. 3, does not vary significantly) paired with the commercial BBUs. The obtained assignment is depicted in Fig. 6.

Fig. 7 compares the power consumption obtained for the cBBU case and the heterogeneous one (e.g., in which the BBU pool is realized using both commercial products and VNFs, hBBU hereinafter) considering December 6th traffic data. Results for the cBBU are the same as in Fig. 5, with a device module constantly kept on for each eNodeB deployed on the reference area. In the hBBU case, incoming traffic to the eNodeBs that present a low average load, which correspond to the ones located in the outermost part of the reference area in Fig. 2 and Fig. 3, is managed by virtual BBU instances. As a result, the consumption ascribable to them is significantly reduced. Keeping the hosting servers powered off most of the time allows for energy savings that peak up to 20% and allows scaling with the traffic load.

One of the promises of C-RAN is the reduction of the Total Cost of Ownership (TCO), and the mere centralization by itself can actually reduce deployment costs and improve Power Usage Effectiveness (PUE) with respect to a traditional RAN [4]. However, since a significant reduction of such costs can be obtained only by serving an area as big as allowed by the constraints mentioned in Section III, which necessarily increases both the number of required hardware (racks, cabling, internal routers and switches, etc.) and space-related costs (such as, among others, real estate necessary for the data center, for power generation systems and other auxiliary subsystems), care must be taken in planning the infrastructure to host a C-RAN. While a complete analysis of the TCO in a datacenter is out of the scope of this paper, some considerations on how the different cases we have analyzed impact on OPEX and CAPEX can be useful to bring their comparison to a close.

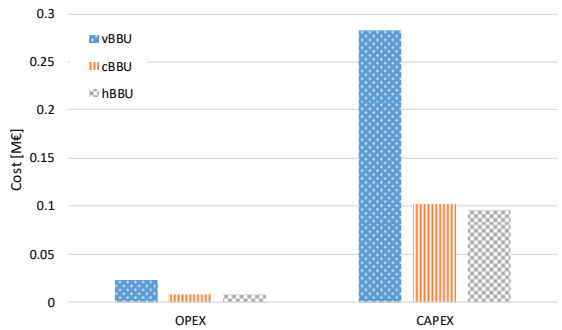

Fig. 8. OPEX and CAPEX over a year for the three cases.

Fig. 8 reports an estimate on OPEX and CAPEX over a year for the three test cases. For the sake of simplicity, but without losing on the comparison, we have considered only the costs ascribable to the hardware performing RAN operations. In more details, OPEX includes the cost of energy consumed by such hardware, so it does not consider other costs such as operation and maintenance, staff or rent. For this reason, the cost is much lower with respect to CAPEX; however, it can be noticed that costs related to the vBBU case are significantly higher than the other two cases. For the computation of CAPEX, we have accounted for the cost of the devices (switches and servers for the vBBU case, dedicated hardware for the cBBU one, and of course all of them for the hBBU) considering a depreciation time of five years. The vBBU case again overcomes the other ones by 66%, while the impact of hBBU over CAPEX is basically the same as that of the cBBU. Although this result may look grim, it is actually extremely promising: in fact, for the same cost, hBBU provides a deployment that is flexible enough to host other applications aside from RAN and that can be extended seamlessly both for scalability reasons and to foster new generations of network/computing resources.

V. Conclusions

This paper has inspected the sustainability and energy requirements obtained by virtualizing the Radio Access Network (RAN) in order to better understand the impact of NFV technologies in real networks. It is well known that the coming of 5G, by fostering new classes of applications with heterogeneous and extremely challenging requirements that will bring along growing demands in terms of high bandwidth, low latency and ultra-reliable communications, will increase the amount of data traversing the network and, as a consequence, the costs to build, operate and upgrade mobile networks.

Since the adoption of virtualization paradigms is seen as a viable way to fulfil the requirements of next-generation applications in a sustainable way, this paper has made an effort to quantitatively assess the accuracy of this claim by comparing the BBU power consumption obtained using commercial devices or pools of VNFs.

The evaluation has been conducted by using publicly available datasets describing the traffic and the LTE eNodeBs deployed by one of the main Italian mobile operators over the Milan and neighboring cities metropolitan area, and on datasheets for both the commercial devices and the VNF pools.

Despite the widespread impression of NFV being a "green" technology, the obtained results have unveiled that, while the usage of the virtualized BBU pool provides proportionality with the incoming load and better results when the traffic is low, consumptions are on average around 250% higher with respect to the commercial deployment, and further estimates on OPEX and CAPEX have highlighted costs above 66% higher. On the other hand, further results evaluating the deployment of VNFs alongside specialized hardware solutions exhibited energy savings up to 20% and costs in line with the ones of dedicated hardware deployments, showing that remarkable benefits can be obtained by exploiting the scalability of software solutions in the presence of lower traffic loads and the higher performance provided by commercial products where demands are higher.

Acknowledgment

This work has been supported by the Horizon 2020 5G-PPP Innovation Action MATILDA (Grant Agreement no. 761898) and by the Horizon 2020 Innovation Action SPIDER (Grant Agreement no. 833685).

References

[1] "5G Vision - The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services", URL: https://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf. (Accessed on December 24th, 2019).

[2] M. Chiosi et al., "Network Functions Virtualization: An Introduction, Benefits, Enablers, Challenges and Call For Action," In Proceedings of the SDN and OpenFlow World Congress, Darmstadt, Germany. ETSI White Paper. URL: https://portal.etsi.org/nfv/nfv_white_paper.pdf (Accessed on December 24th, 2019).

[3] R. Bolla, R. Bruschi, F. Davoli, C. Lombardo, J.F. Pajo, O.R. Sanchez, "The Dark Side of Network Functions Virtualization: A Perspective on the Technological Sustainability", Proc. IEEE Int. Conf. Commun. (ICC 2017), May 2017.

[4] "C-RAN the road towards green ran," China Mobile Research Institute, Beijing, China, Oct. 2011, Tech. Rep.

[5] EURECOM, "Open air interface." URL: http://www.openairinterface.org/, Oct. 2014 (Accessed on December 24th, 2019).

[6] Next Generation Mobile Networks (NGMN) Alliance, "NGMN 5G White paper", February 2015, URL: https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf (Accessed on December 24th, 2019)

[7] https://dandelion.eu/datagems/SpazioDati/telecom-sms-call-internet-mi/resource/. (Accessed on December 24th, 2019)

[8] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper, February 2019, URL: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html (Accessed on December 24th, 2019).

[9] https://opencellid.org/downloads.php (Accessed on December 24th, 2019).

[10] "Wolfram Mathematica - The world's definitive system for modern technical computing", URL: http://www.wolfram.com/mathematica/ (Accessed on December 24th, 2019).

[11] https://e.huawei.com/en/material/onLineView?materialid=ebfb3e1e7ed14d48b72c3eb4f37ddb2e. (Accessed on Dcember 24th, 2019).

[12] A. Younis et al., "Bandwidth and Energy-Aware Resource Allocation for Cloud Radio Access Networks," IEEE Transactions on Wireless Communications, vol. 17, no. 10, pp. 6487 – 6500, Oct. 2018.

[13] M. Al-Fares, A. Loukissas, A.Vahdat, "A Scalable, Commodity Data Center Network Architecture," Proc. ACM SIGCOMM 2008 Conf. on Data Communication, Seattle, WA, USA, Aug. 2008, pp. 63-74.

[14] V. Q. Rodriguez, F. Guillemin, "Cloud-RAN modeling based on parallel processing", IEEE J. Sel. Areas Commun., vol. 36, no. 3, pp. 457-468, Mar. 2018