Modelling of Computational Resources for 5G RAN

Sina Khatibi, Kunjan Shah, Mustafa Roshdi
Nomor Research GmbH
Munich, Germany
Email: {khatibi, shah, Roushdy.stud}@nomor.de

***Abstract*— The future mobile networks have to be flexible and dynamic to address the exponentially increasing demand with the scarce available radio resources. Hence, 5G systems are going to be virtualised and implemented over cloud data-centres. While elastic computation resource management is a well-studied concept in IT domain, it is a relatively new topic in Telco-cloud environment. Studying the computational complexity of mobile networks is the first step toward enabling elastic and efficient computational resource management in telco environment. This paper presents a brief overview of the latency requirements of Radio Access Networks (RANs) and virtualisation techniques in addition to experimental results for a full virtual physical layer in a container-based virtual environment. The novelty of this paper is presenting a complexity study of virtual RAN through experimental results, in addition to presenting a model for estimating the processing time of each functional block. The measured processing times show that the computational complexity of PHY layer increases as the Modulation and Coding Scheme (MCS) index increases. The processes in uplink such as decoding take almost twice the time comparing to the related functions in the downlink. The proposed model for computational complexity is the missing link for joint radio resource and computational resource management. Using the presented complexity model, one can estimate the computational requirement for provisioning a virtual RAN as well as designing the elastic computational resource management.**

I. Introduction

The rapidly growing mobile traffic demands [1] together with a massive increase in the number of expected connected mobile devices [2] force the next generation of mobile network, also known as "5G Networks"; to be flexible, scalable, and cost-efficient. Improving the network capacity with the scarce available resources while confronting the severe geographical and temporal variation of traffic demand is a non-trivial task [3]. The densification of the network and virtualisation are the foundations of many proposed approaches to improve network capacity, which leads to increase the CAPital EXpenditure (CAPEX) and OPerational EXpenditure (OPEX) [4]. Hence, the 5G mobile networks require new techniques to reduce the costs and increase the network flexibility.

The recent studies consider network softwarising and network slicing as two key solutions[5]. Regarding network softwarisation, Centralised Radio Access Network (C-RAN) is a practical solution to cope with cost and flexibility challenge [6]. The revolution in C-RAN is implementing the Network Functions (NFs) of Baseband Units (BBUs) on Commercial Off-The-Shelf (COTS) computers or data-centres forming shared computational resource pool instead of proprietary citewang14. Also, virtualisation of network resources and NFs enables sharing the physical infrastructure while offering isolation, network element abstraction, and ease-of-use [4]. Network slicing is another innovation to enhance 5G systems for different vertical use-cases. Network slicing ensures to fulfil various Quality of Service (QoS) requirements of different slices (which may be even contradictory), operating over the same physical infrastructure. Each network slice is an isolated virtual network, optimised for a specific use-case. For instance, optimisation of the network slice makes sure Ultra Reliable Low Latency Communication (URLLC) slice receives the required low latency, while enhanced Mobile Broad Band (eMBB) slice has high throughputs. Hence, the computational resource management algorithm should also consider the slice-specific computational requirement.

However, the realisation of a practical C-RAN requires addressing the key challenges, including fronthaul delay and capacity requirement [7] in addition to reliability and stability including fulfilling real-time constraints [8], and secure implementation of Virtual Network Functions (VNFs) [9]. The Hybrid Automatic-Repeat-Request (HARQ) feature standardised in LTE imposes a constraint on LTE networks to transmit a subframe following 8 ms to its transmission, during which encoding/decoding at User Equipment (UE), propagation over the air interface, fronthaul propagation, and BBU processing time have to be done. Based on [10], the share of BBU processing is only 3 ms. In the virtualised environment, the Virtual Infrastructure Manager (VIM) should allocate adequate computational resources (i.e., physical CPU core/thread) to the VNFs to enable them to complete the process of subframe less than the 3 ms. The shortage of computational resources may lead to performance degradation of RAN functions while excess unused computational resources increase CAPEX/OPEX.

Hence, 5G RAN requires provisioning of computational resources and elastic slice-aware computational resource management algorithms [11] optimise the resource utilisation and maximise the cost efficiency. The first step in efficient computational resource management is to estimate the processing time of each VNF (i.e., the VNFs complexity). This estimation can improve slice optimisation process in addition to achieving higher computational resource utilisation. This paper, first, provides a brief overview of different virtualisation technologies. Due to limited delay budget in RAN, the focus is on the imposed computational overhead compared to bare-metal implementation (i.e., when there is no virtualisation applied). The primary aim of this paper is studying the computational complexity of functions physical layer in the form of processing time as a function CPU clock frequency and MCS. The

experimental results comparing processing time for VNFs over bare metal vs containers have been presented. The novelty of this paper is proposing statistical models for estimating processing time of the network functions of the physical layer in a container-based environment.

The rest of this paper is organised as follows: Section II provides an introduction to the C-RAN computational requirements and the related studies in addition to a brief description of different virtualisation technologies and their performances. Sections III presents the details of the test-bed and the experiments followed by the numeric results. Section IV proposes a polynomial approximation as a model for estimating the computational complexity of PHY layer VNF. Finally, the conclusions are drawn in Section V.

II. C-RAN COMPUTATIONAL REQUIREMENTS

As it was discussed in the last section, the delay budget for the BBU processing time is only 3 ms, which RX processes receive 2 ms and TX the remaining 1 ms (refer to [11] for more details). Fig. 1 demonstrates the functional block in both down-link and uplink direction. The PHY layer process is divided to three main segments: FFT, modulation, and encoding segment. Authors in [12] studied the feasibility of a full GPP (General Purpose Process) implementation of RAN and the minimum processing requirements. The studies in this paper focuses on the relation of processing time of each segment to the allocated computational resources. The computational overhead caused

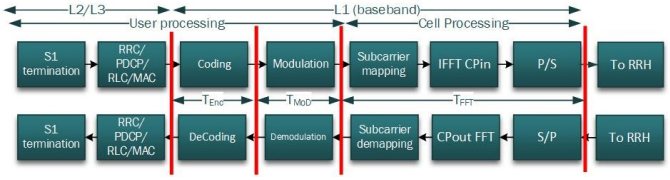

Fig. 1. Functional block in a LTE eNodeB (based on [11]).

by utilisation virtualisation platform imposes an extra delay to the processing time of each NF. Given the tight delay budget, the desired virtualisation solution should provide very small or no delay overhead comparing to bare-metal implantation. Hence, the choice of virtualisation platform is an important design decision. In the following the three main virtualisation technologies with telco cloud perspective are briefly discussed.

IT cloud computing community is considering three virtualisation architectures namely Virtual Machines (VMs), containers, and Unikernels. The comparison of these approaches includes the delay overhead (both deployment and operational delay), level of isolation and the memory footprint (i.e., the amount of the RAM utilisation by the VNF). As described in Fig. 2, telco cloud requirements are much more strict compared to the IT cloud, e.g., Telco cloud solution expects much higher resilience and much lower latency compared to the IT clouds. Hence, we perform a feasibility study of several virtualisation platforms from telco point of view.

In VMs, an extra layer called Hypervisor is added on top of the host Operating System (OS) to provide virtualised hardware resources. Guest OS resides on top of the virtualised hardware resource. Due to multiple layers of OS, VMs offer

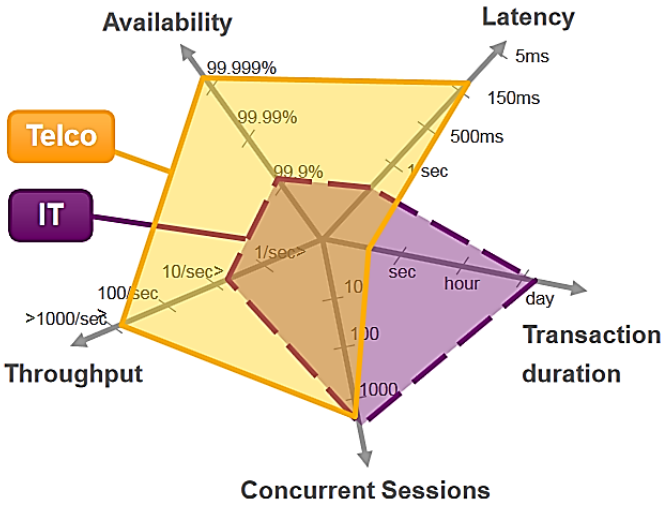

Fig. 2. Telco Cloud vs IT cloud (extracted from [13]).

a superior level (i.e., Hardware-level [14]) of isolation at the cost of higher overhead time, as it is shown in Fig. 3.

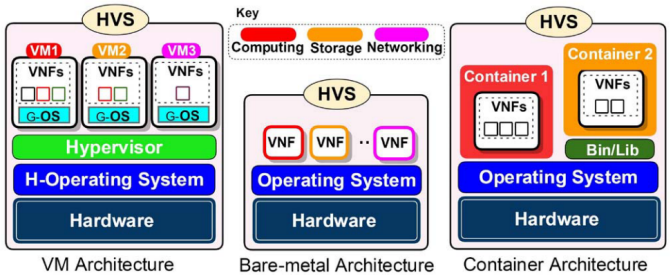

Fig. 3. VNF deployment options (extracted from [15]).

The container-based solutions are lightweight OS virtualisation. It groups and isolates the process and resources from the host OS as well as other containers. The advantage of using containers are the fast deployment time, portability, small footprint, and consolidation. However, the lack of hardware isolation is the main drawback in the container-based approaches. While it is argued that the hypervisor isolation is not infallibly secure, the improvement can be achieved by running each container over its own lightweight VM.

Finally, Unikernels are the trade-off between VM-based and container-based virtualisation of NFs [15]. They are specialised OSs that contain the application code and the minimum OS library. While their security and lightweight make them an interesting choice, they cannot be considered a practical solution yet. This concept is quite new and subject to many development and extensions. The authors believe unikernels can be the primary solution in the near future. After briefly comparing all virtualisation platforms, the authors found containers the most suitable option as virtualisation platform, hence considered as a part of RAN profiling study in the next section.

III. Numeric Results

The selected testbed for software RAN is Open Air Interface (OAI) [16]. OAI is an open-source software-based implementation of the LTE system spanning the full protocol stack of 3GPP standard. OAI provides two simulation tools dlsim and ulsim, which emulate PDSCH and PUSCH respectively. Using these tools enables profiling of the processing time of the LTE PHY layer given different load configurations for the number of Physical Radio Block (PRB) and the MCS. The used physical machine for profiling experimentation has Intel Core i7-4790 CPU @ 3.60GHz and 16 GB RAM. Its operating system is Kubuntu 16.04 LTS with Kernel version 4.4.0-96 generic. The used virtualisation was docker 17.06.1-ce. The processor has four physical cores, eight logical threads, and 8 MB of Cache Memory and supports instruction set extensions SSE4.1/4.2, AVX 2.0. It is worth noting that using processors with instruction set extension support can decrease the processing time to half [12].

Fig. 4 and Fig. 5 show the effect of running the PDSCH and PUSCH profiling on a container in comparison with bare-metal OS (i.e., without any virtualisation). The results show that at high CPU working frequencies the overhead is comparatively negligible and the processing time in the containers is almost equivalent to that of bare-metal. However, in lower CPU working frequencies, in PDSCH in particular, the overhead effect starts to be evident.

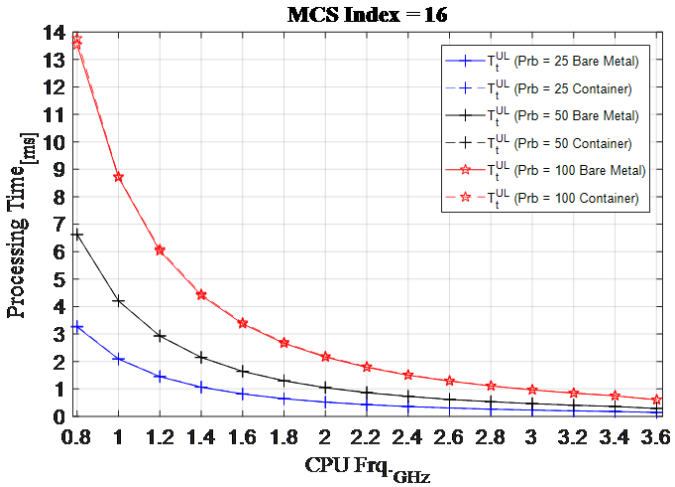

Fig. 4. Comparison of processing time in uplink as a function of CPU working frequency over bare-metal versus containers.

Fig. 6 and fig. 7 present the total processing time for a subframe as the function of MCS index for 5 MHz (25 PRBs) for uplink and downlink while the MCS index is swept from 0 to 27. It is apparent from the figure that by increasing the MCS index, the total processing time increases. Also, the plots present the processing time for the main functional/split as it is indicated in Fig. 1; these partial processing times are the processing time for functional blocks including FFT/IFFT (T_{FFT}), (de)modulation (T_{Mod}) and encoding/decoding (T_{Enc}).

Based on the numeric results, the processing time of FFT/IFFT blocks remains constant throughout the experiment

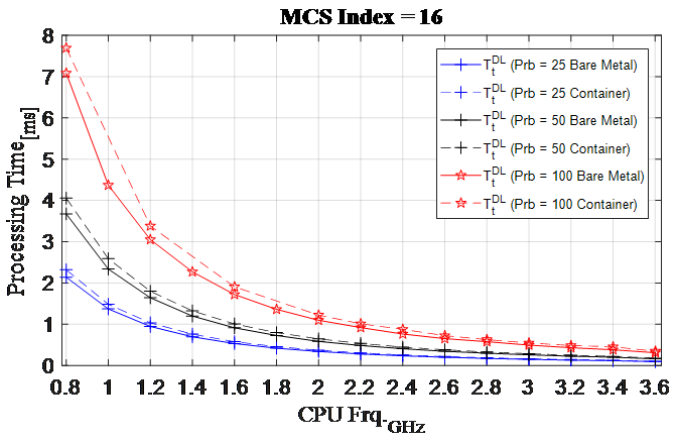

Fig. 5. Comparison of processing time in downlink as a function of CPU working frequency over bare-metal versus containers.

since this block size depends only on the network bandwidth and it does not depend on the choice of the MCS index.

Also, (de)modulation processing time increases in 3 steps by change modulation. The required time for MCS between 0 to 9 is nearly the same since the selected modulation in these cases is QPSK modulation In MCS indices 10 to 16, the modulated 16-QAM, and MCS indices 17 to 27 are modulated using 64-QAM.

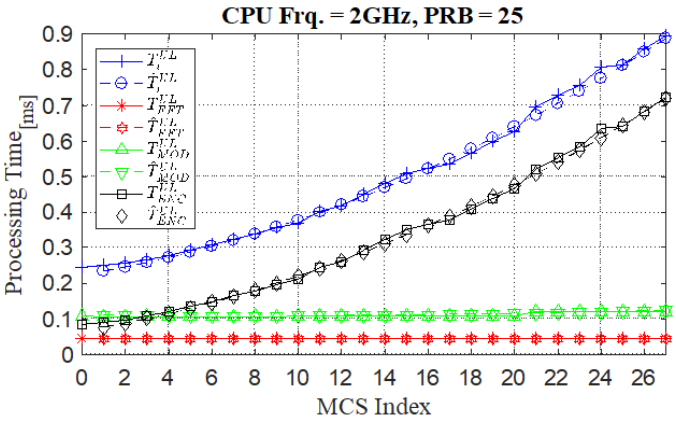

Fig. 6. Processing time in uplink as a function of MCS index.

Comparing the processing time for the uplink and downlink equivalent modules shows the processing time required for decoding is almost twice of the encoding. The time required by FFT is the same as IFFT for the same network bandwidth. The decoding time in eNodeB is almost double than encoding in the same configuration. Hence, decoding has been the most critical process from profiling point of view. The encoding/decoding are the most time-consuming process in PHY layer. The plots of the processing time versus the MCS index shows the exponential increase in processing time as the MCS index increases (i.e., the channel quality improves).

IV. Model for Computational Resources

Having a realistic estimation of required computational resources for RAN network functions is essential for provisioning the computational resource pool as well as dynamically

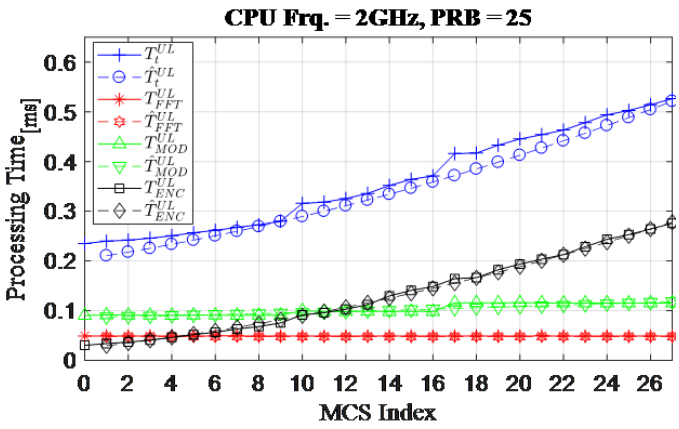

Fig. 7. Processing time in downlink as a function of MCS index.

allocating them to different network slices. The given input variables in the said estimation are network bandwidth, expected traffic demands, and CPU working frequencies. Based on the numeric results, there are three classes of RAN NFs:

- NFs with constant complexity: these NFs (i.e., FFT/IFFT) complexity only depends on the networks configuration and bandwidth. The processing time for these NFs is independent of the networks load or the selected MCS. The required computational resources for these NFs do not change during run-time.
- NFs with complexity depending on the demand and channel quality. The processing time of NFs including (de)modulation and encoding/decoding functional block depends on the number of allocated PRB and the MCS index. The traffic demands plus the effect of the link adaptation algorithms jointly estimate the required computational resources.
- NFs with complexity depending only on the throughputs such as the NFs in layer 2. The required computational resources for these NFs are estimated just based on the demands. These NFs are going to be the next study topic.

In next step, the experimental results are used to fit a polynomial equation (using Lasso technique) and estimate the processing time as the function of CPU frequency and MCS index. Examining the experimental readings suggests that they can be normalised to the number of PRBs. The only exceptions are the FFT/IFFT NFs. While analytical analyses show these blocks are logarithmic time complex, the linear approximation introduces neglectable differences. The input variables described above are used to form the design matrix for regression and the lasso technique is used to fit a curve to the experimental results. Equation (1) presents the polynomial function of MCS index that can estimate the processing time in uplink and downlink.

$$\tau_{p_{[\mu s]}} = \frac{N_{PRB}}{f^2_{[\mathrm{GHz}]}} \sum_{i=0}^{2} \alpha_i i^i_{MCS} \qquad (1)$$

where:

- τ_p:processing time,
- N_{PRB}: Number of PRBs
- f: the working frequency of CPU,
- i_{MCS}: the MCS index,
- α_i: polynomial coefficients,

TABLE I. THE COEFFICIENTS FOR PROCESSING TIME.

	α_0	α_1	α_2
T_t^{DL}	32.583	1.072	0.03
T_{FFT}^{DL}	7.609	0	0
T_{Enc}^{DL}	3.907	0.773	0.027
T_{mod}^{D}	13.851	0.133	0.002
T_t^{UL}	35.545	1.623	0.086
T_{FFT}^{UL}	6.957	0	0
T_{Enc}^{UL}	10.512	1.631	0.083
T_{mod}^{UL}	17.494	-0.08	0.006

Table I present the coefficients value in uplink/downlink for total, FFT/IFFT, modulation/demodulation, and encoding/decoding. The comparison of the estimated equation and the experimental results is shown in the figures above. Equation (1) provides a closed-form approximation for processing time in C-RAN. The infrastructure providers can use this equation to determine the required number of CPUs and their working frequency for implementing each VNF in provisioning phase. The algorithms for elastic computational resource management can use it to estimate the effect computational resource changes and make decisions to cope with networks changes (either the changes in demand or MCS) while meeting the processing time requirements. During the networks run-time, based on the time measurements and observations and using machine learning techniques the approximation can be improved and optimised for the specific running situation (e.g., different implementation of NFs or CPU architecture)

V. CONCLUSIONS

The solutions to realise the requirements for 5G mobile networks and overcome with the consequent problems of IP-tsunami is virtualisation of mobile networks. The virtual networks are composed of different software element referred to as virtual network functions implemented in cloud-data centres. Introduction of network slicing is the next step in network virtualisation, enabling co-existence of multiple virtual networks with different requirements and configuration over the same physical infrastructure. The isolation of the network slices enables to optimise each one them to meet QoS requirements of different services. While computational resource management in IT-cloud is very advanced, it became an interesting research topic. In contrast to IT cloud, the mobile cloud networks have to meet the tight latency and reliability requirement. The total delay budget in RAN is only 3 ms for both uplink and downlink. The tight delay requirements and addition to offering isolation in RAN makes the selection of a virtualisation technique (i.e., VMs or containers) challenging. Comparing the delays imposed by the VMs and containers, in addition to their footprints and deployment time, containers are the better choice for RANs NFs. Although the containers offer lower latency implementation, the isolation level among the containers is comparatively lower than the isolation among

VMs. The numeric results confirm that the imposed processing delay for high CPU working frequency comparing the bare metal is neglectable. Based the Open Air Interface, an open-source software mobile network, the computational complexity of PHY layer through series of experiments has been studied. According to the numeric results, there are three categories of NFs in RAN: a) NFs with constant computational complexity (FFT/IFFT). The processing time of these functions only depend on the networks bandwidth b) NFs with complexity depending on the demands and MCS are the second group. The processing time of these functions changes based on both allocated number of PRB and the selected MCS index. C) NFs depending only on the demands such as most of layer two functions. Among the tested functional blocks, the encoding/decoding functions are the most time consuming functional blocks and the required time increases by increasing the MCS index. Likewise, the complexity of (de)modulation blocks increases by as the MCS index increases in three steps the relative to the three modulation options. By examining the numeric results concerning CPU working frequency, it can be concluded that an AXV2-enabled CPU with working frequency higher than 2 GHz is the minimum requirement for having full virtual RAN over GPP. Finally, the provisioning of virtual RAN in addition to elastic computational resource management depends on estimating the Probability Density Function (PDF) of processing delay of RAN functions in different situations. The novelty of this paper is studying the complexity of functional blocks to estimate the processing time with different configurations. The main result presented is a closed-form equation to approximate the processing time as the function of CPUs working frequency, number of PRBs, and MCS index) is proposed. Using the equation and provided coefficient the processing time of the total L1 in addition to FFT/IFFT, decoding/encoding, and (de)modulation can be estimated. It can also be used to obtain an analytical estimation of the PDF of required computational resources. It is worth re-emphasising that introduced inaccuracy as the result of simplifications and approximation can be improved in run-time by real-time measurements and applying machine learning techniques, which is the topic of our next publication.

ACKNOWLEDGEMENT

Part of this work has been performed within the 5GMoNArch project, part of the Phase II of the 5th Generation Public Private Partnership (5G-PPP) program partially funded by the European Commission within the Horizon 2020 Framework Program.

REFERENCES

[1] Cisco Systems, "Global Mobile Data Traffic Forecast Update, 2016 - 2021," Cisco Systems, California, USA, Tech. Rep., 2017. [Online]. Available: http://www.cisco.com

[2] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wbben, "Cloud technologies for flexible 5g radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, May 2014.

[3] S. Khatibi and L. M. Correia, "A Model for Virtual Radio Resource Management in Virtual RANs," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 68, 2015. [Online]. Available: http://jwcn.eurasipjournals.com/content/2015/1/68

[4] ——, "Modelling of virtual radio resource management for cellular heterogeneous access networks," in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, Washington, DC, USA, Sept 2014, pp. 1152–1156.

[5] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing & softwarization: A survey on principles, enabling technologies & solutions," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2018.

[6] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[7] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an efficient c-ran optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 1, pp. 708–769, Firstquarter 2018.

[8] C. L. I, J. Huang, R. Duan, C. Cui, J. . Jiang, and L. Li, "Recent progress on c-ran centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[9] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, Feb 2015.

[10] I. Alyafawi, E. Schiller, T. Braun, D. Dimitrova, A. Gomes, and N. Nikaein, "Critical issues of centralized and cloudified lte-fdd radio access networks," in *2015 IEEE International Conference on Communications (ICC)*, London, UK, June 2015, pp. 5523–5528.

[11] D. Gutierrez-Estevez, M. Gramaglia, N. P. A. Domenico, S. Khatibi, K. Shah, D. Tsolkas, P. Arnold, and P. Serrano, "The path towards resource elasticity for 5g network architecture," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW): Workshop on Flexible and Agile Networks (FlexNets)*, Barcelona, Spain, Apr. 2018.

[12] N. Nikaein, "Processing radio access network functions in the Cloud: Critical issues and modeling," in *MCS 2015, 6th International Workshop on Mobile Cloud Computing and Services, in conjunction with MOBICOM 2015, 11 september 2015, Paris, France*, Paris, FRANCE, 09 2015. [Online]. Available: http://www.eurecom.fr/publication/4640

[13] L. Nmeth and J. Br, "Global Mobile Data Traffic Forecast Update, 2016 - 2021," Miyazaki, Japan, Tech. Rep., 2012.

[14] P. Sharma, L. Chaufournier, P. Shenoy, and Y. C. Tay, "Containers and virtual machines at scale: A comparative study," in *Proceedings of the 17th International Middleware Conference*, ser. Middleware '16. New York, NY, USA: ACM, 2016, pp. 1:1–1:13. [Online]. Available: http://doi.acm.org/10.1145/2988336.2988337

[15] J. G. Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, Sept 2016.

[16] S. Khatibi, L. Caeiro, L. S. Ferreira, L. M. Correia, and N. Nikaein, "Modelling and implementation of virtual radio resources management for 5g cloud ran," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, p. 128, Jul 2017. [Online]. Available: https://doi.org/10.1186/s13638-017-0908-1