VNF Performance Prediction for Optimal O-RAN Functional Splits in 5G Networks

Student: TBD
Mentor: Jesutofunmi Ajayi
Advisor: Prof. Dr. Torsten Braun.

May 11, 2022

1 Project Proposal

The O-RAN architecture is enabled by dis-aggregating a typical Base Station (i.e. evolved Node B (eNB) or next-generation Node B (gNB)) and remodelling it as a set of (chained) Virtual Network Functions (VNFs). RAN VNFs can be deployed across various nodes at the network edge (namely, the Central Unit (CU), the Distributed Unit (DU) and the Radio Unit (RU)) in order to create different Functional Splits, as proposed by the 3rd Generation Partnership Project (3GPP). An overview of the O-RAN architecture, along with the main interfaces that enable communication between RAN VNFs is shown below in Figure 1.

Unlike other virtualized NFs, the signal processing algorithms (e.g. for Modulation/Demodulation, Encoding/Decoding etc.) utilized by various virtualized RAN functions are generally very compute-intensive [1] and therefore require highly available compute and communication resources before they can be placed on an edge node to perform a functional split. In this context, the

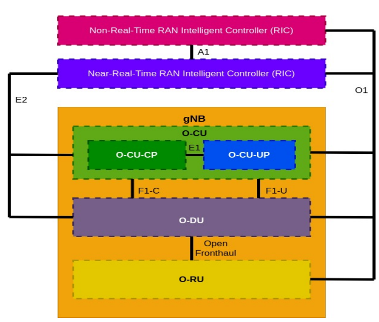

Figure 1: O-RAN Architecture

profiling of the gNB's RAN functions must be conducted in advance before functional splits can be effectively scheduled in O-RAN.

Taking inspiration from [2] and [3], in this project we aim to profile the gNB in order to determine the optimal functional split under various resource configurations and co-location scenarios. This is an important consideration for the scenario in which RAN functions are co-deployed with other edge workloads, which is the likely deployment option in 5G networks, based on a recent proposal[1] by the European Telecommunication Standards Institute (ETSI).

Through the profiling of edge VNFs and applications, we aim to learn the correlation between RAN performance and the availability of communication and compute resources. Specifically, we will consider the performance of the gNB (which provides the RAN in a 5G network) as a typical real-time application that can be deployed in a cloudified edge setup. Previous work [4] showed that such an application can run within fully virtualized environments like a Kernel Virtual Machine or utilizing other virtualization-like techniques such as Linux Containers (LXC)[2] or Docker[3]. The containerization of the gNB through platforms such as Docker opens up new possibilities for migration, resource monitoring and performance assessment of the containerized application under various resource configurations, which are important to profile the applications considered in this work. As we consider the Edge/RAN segment of the 5G network, the VNFs and applications considered will be those that are considered native to such environments.

1. First, the student will need to deploy the NR RAN functions and other selected Edge software applications in docker containers, enabling the provisioning of the RAN as a Service (RANaaS) or Software as a service (SaaS) and following the O-RAN architecture proposed by 3GPP (but focusing primarily on the CU and DU and their respective functionalities). The NR RAN functions/gNB can be provided by the OpenAirInterface (OAI) software. A Docker-enabled version of the OpenAirInterface (OAI) LTE software has been provisioned by us[4] but will need to be updated for the 5G context with the NR RAN/gNB software.

2. In the second task, the student has to familiarize themselves with the Docker API (or collectd tool) in order to monitor the resource utilization of containerized RAN and Edge applications. Optionally, they could develop a visualization tool to view this information in real-time using platforms such as Grafana.

3. Third, the student will profile the RAN VNFs and Edge applications by developing a program to dynamically configure the containerized services/applications with compute and communication resources using the

[1] https://www.etsi.org/images/files/etsiwhitepapers/etsi_wp23_mec_and_cran_ed1_final.pdf
[2] https://linuxcontainers.org/
[3] https://www.docker.com/
[4] https://github.com/tA-bot-git/docker-OpenAirInterface

Docker API, in order to evaluate their performance under various resource configurations.

4. To determine the optimal functional split under various resource configurations, the student will be required to study and assess the resource requirements of the different RAN stacks (i.e. PDCP, RLC, MAC, PHY) to profile them individually. Here, a literature search as well as the experimental setup could be useful to determine the way in which the RAN communication stack is composed and operates to profile individual stack layers and derive their individual resource requirements for performing functional splits in O-RAN.

5. For the fifth task, the student will utilize the profiling data gathered in the previous tasks to learn the correlation between resource configurations and RAN/Edge service performance by developing an AI/ML framework to train models for prediction of performance.

6. In the final task, the student will be required to evaluate the whole system by assessing the performance of their AI/ML framework to determine optimal splits during the deployment of the RAN under different resource configurations and co-location scenarios.

2 Project Milestones

The official start date of this thesis is to be defined.

The project milestones include, but are not limited to, the following key activities:

MS1 Definition of primary Use-Cases (M1)

MS2 Definition of KPIs, Metrics (M1)

MS3 Definition of the Edge/O-RAN Architecture and Infrastructure (M1)

MS4 Design and implementation of a cloud and edge environment to instantiate the 5GC and NR RAN (gNB), respectively, for the 5G network. (M1)

MS5 Deployment of a containerized edge network infrastructure composed of RAN VNFs for the 5G network delivered and co-located Edge applications. (M2)

MS6 Evaluation of State-of-the-Art/Related Work (M2)

MS7 Profiling of NR RAN functions and edge applications with respect to the available edge resources (considering different co-deployment scenarios) (M3-M4)

MS8 Development of AI/ML framework to learn correlation between Edge application/RAN service performance and resource availability, based on profiling data collected in MS7 (M5)

MS9 Evaluation of AI/ML framework developed in MS8 for determining optimal Functional Splits in certain scenarios (M5)

MS10 Initial draft of the thesis (M6)

MS11 *Optional*: Scientific Paper delivered (M6)

MS12 Final version of the thesis (M6-M7)

References

[1] X. Foukas and B. Radunovic, "Concordia: Teaching the 5G vRAN to Share Compute," in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, ser. SIGCOMM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 580–596. [Online]. Available: https://doi.org/10.1145/3452296.3472894

[2] J. Wang and Y. Hu, "Performance Analysis of 5G NR vRAN Platform and its Implications on Edge Computing," in *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2020, pp. 215–217.

[3] S. Mollahasani, M. Erol-Kantarci, and R. Wilson, "Dynamic CU-DU Selection for Resource Allocation in O-RAN Using Actor-Critic Learning," 2021. [Online]. Available: https://arxiv.org/abs/2110.00492

[4] E. Schiller, J. Ajayi, S. Weber, T. Braun, and B. Stiller, "Toward a Live BBU Container Migration in Wireless Networks," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 301–321, 2022.