Analysis of the Performance of Multi-Access Edge Computing Network Slicing in 5G

Sérgio Domingues
Luis M. Correia
Instituto Superior Técnico / INESC-ID
University of Lisbon
Lisbon, Portugal
sergiodomingues96@gmail.com
luis.m.correia@tecnico.ulisboa.pt

Ricardo Dinis
NOS SGPS
Lisbon, Portugal
ricardo.dinis@nos.pt

***Abstract*— The main purpose of this thesis is to provide an overview of the technologies essential to support new 5G network services, with higher data rates and ultra-low latency, emphasising edge networking technology used to offload the computation tasks from the centralised network to the edge of the cloud, near the users, in order to reduce latency and support more computation power near the network terminal nodes. This work studies the characteristics of the different network architectures on the C-RAN in other to optimise the network for multiple services and applications from 5G. The model takes into consideration five essential parameters to support 5G services demands, including centralisation gain, network latency, node throughput, node processing power and network cost. The model is used to study the performance of the network in multiple scenarios, where one concludes that, in other to support the 1 ms latency demands, it requires the introduction of at least 5 MEC nodes and 30 MEC nodes on the Minho and Portugal scenarios, respectively. The results obtained show that it is essential to increase the splitting option used on the 4G network fronthaul, sending more processing power near the users, which will compress the signal and reduce node throughput by, at least, 95%.**

Keywords-5G, Cloud-RAN, RU, DU, CU, MEC.

I. INTRODUCTION

Mobile users and services are in constant changes. Mobile networks continuously take more data and the mobile industries do not only need to fulfil those requirements, but also introduce new capabilities and use cases. According to [1] the number of subscriptions grew at 4 percent per year, reaching 7.9 billion subscriptions in the first quarter of 2018. Not only the number of subscribers is exponentially increasing, but also the average data volume per subscription, due primarily to watching video content in higher resolutions, has also increased.

Mobile operators have been increasing their network capacity in order to satisfy consumer demands. 5G was designed not only to secure those demands, but also to deliver connectivity to virtually every product imaginable. For example, fully-autonomous vehicles need Vehicle-to-everything (V2X) wireless communications technologies. Industrial Internet of Things (IIoT) needs mass connectivity, cloud computing resources, big data analytics, and artificial intelligence. These two examples are services that 5G will support, although their requirements are dramatically different., with different throughput, latency and QoS demands.

The introduction of Multi-Access Edge Computing (MEC) is an important concept used in 5G networks. MEC system brings the services close to the devices which provide computation, storage and network resources, different for each application. MEC is essential for the implementation of the services requirements, i.e., latency, scalability, and throughput. According to [2], MEC also ensures 24% Capital Expenditures (CAPEX) and 25% Operational Expenditures (OPEX) cost reduction for network operators. It is expected that data at the edge of the service will increase exponentially, and that by 2022 70% of the produced data will stay at the edge of the network, with just 30% being transported to the data centres of the core.

In order for the MEC technology to work at its full potential, the network needs to be sliced into several isolated networks. This technology is called network slicing and is used to customise and optimise each slice of the network, to address the diverse Quality of Service (QoS) requirements for each 5G service. A combination of MEC and network slicing enables a new vertical-oriented slicing and slice management framework, which not only reduces the CAPEX and OPEX for the operators but also improves user experience, new business models, reduces the time for service creation and reduces time to market.

The paper is organized as follows. Section II presents the state of the art. Section III presents the model development, starting by presenting the model parameters, following by the specification of the network scenarios, the introduction of the services, the model implementation, and finally the presentation of the model assessment. Section IV contains the results analysis, where the scenario is described, followed by the impact of different parameters on the deployment. In Section V, the most important conclusions of this work are presented.

II. STATE OF THE ART

Several studies addressing the major challenges imposed to C-RAN deployments in 5G systems including the implementation of MEC are describes in the literature. This section states the work developed by several authors in the area of implementation analysis of the Cloud Radio Access Networks architectures.

Flexibility is an important aspect of a 5G network. In [3], the author discusses the implementation of a MEC-enabled 5G architecture that supports the flexibility of the network and Virtual Network Functions (VNFs). The MEC architecture is divided into two tiers of computation capabilities - the core tier

cloud, that has a lot of computing resources and can host application VNFs and network VNFs, and the edge tier cloud, which has limited resources allocated to the MEC entity and should be placed closer to the user for specific services requirements. The VNFs are divided into Real-Time Applications (RTAs), Non-Real-Time Applications (NRTAs), and Hybrid Applications (HAs). At the end of the paper, a real implementation of a MEC with VNFs in an LTE network is used to demonstrate the potential of the architecture. For example, depending on the resources available, the network decides to redirect the needed resources of an NRTA from the edge to the core in order to release resources to possible RTA.

In order to achieve the 5G requirements of latency and data traffic, [4] divide the MEC architecture into three tiers: core, edge, and devices. The authors use a two-phase interactive optimisation method to optimise capacity and traffic allocation in a MEC-based architecture. The paper uses a latency percentage constraint metric that calculates the percentage of the latency that satisfies the latency constraint threshold. The latency percentage constraint is calculated according to three different traffic types, depending on the services and application demands, the DC-Type Traffic, served by the device and the core, the EC-Type Traffic, which is served by an edge and the core, and the EE-Type Traffic, the edge to edge traffic. First, the algorithm adjusts the traffic distribution based on the currently allocated capacity to satisfy the latency percentage constraint. Then, the capacity allocation is adjusted based on current allocated traffic in order to minimise the total capacity.

[5] thesis studies the pros and cons of different solution designs for the C-RAN fronthaul on an LTE network. The study focuses on analysing the connection between the RRHs and BBU Pools based on traffic profile, positioning and delay characteristics. The study uses a model divided into three layers in order to analyse the parameters of the network. The first layer is the physical layer used to compute the maximum distance between RRHs and BBUs, which is directly related to the maximum latency of the fronthaul link. The second layer, or technical layer, aims to identify the best connection between the RRHs and BBU Pools, taking into account the demands of the different networks. The third is the cost layer and deals with OPEX and CAPEX of the network.

[6] thesis addressed the implementation of C-RAN in small cells. The goal is to study the assignment of RRHs to BBU pools using different algorithms in order to study the different performance parameters of the C-RAN. The author uses a proliferation algorithm in order to forecast the growth of RRHs and traffic demands in the future, introducing a scale factor to the architecture. To achieve the best results for each traffic profile, this thesis studies multiple algorithms - the Minimise Delay Algorithm, the Load Balancing Algorithm, the Minimise Number of Pools Algorithm, and the Maximise Multiplexing Gain Algorithm.

III. MODEL DEVELOPMENT

A. Model Overview

The purpose of this thesis is to optimise the BS functions splitting between the RU, DU, CU, CN, and MEC for the different use cases considered in this study, concerning the performance parameters assigned to the network demands.

Figure 1 represents an overview of the model under study considering the relation between the Input and Output parameters. The input of the system is divided into two classes. First, the User Specification, which addresses the specific parameters of the use cases, and the required parameters necessary to establish the number of users on each BS. Second, the Network Specification, that takes into account the input specification of the architecture used by the model, the information on the link distances and the different parameters of the cell.

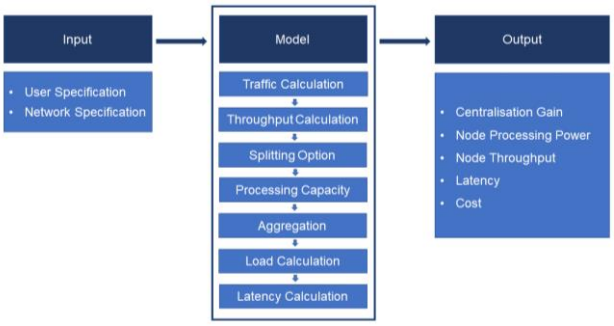

Figure 1. Model overview

The first step of the model consists on the computation of the traffic and the throughput arriving at the RU node, which represents the BS at the cell site. These parameters are calculated based on the number of users assigned to it and use cases specifications. The next step is the splitting options of the BS network functions throughout the RU, DU, CU, and MEC in order to analyse the impact of different architectures of the network on the different use cases considered in this study. The splitting options of the nodes are based on the FH (i.e. Option 8, 7.1, 7.2, 7.3, 6), MH (i.e. Option 2), and BH (i.e. Option 1). To choose a splitting option, the required processing capacity on the nodes is computed depending on the assign computation requirement of each BS function. The aggregation process of the nodes starts with the nodes closest to the users. First, the available RUs are aggregated to the DUs or the CUs, if the DU connection is not possible. Then, the DUs are aggregated to the CUs depending on the distance between the nodes, and the same is done for the connection between CU and the Core or, if necessary, the MEC. Finally, the model computes the latency on the different parts of the network and the total E2E latency depending on the distance between the nodes, the processing delay from the assign BS function that is called GOPS delay, and the queuing delay from the traffic arriving at the node. The two parameters of the processing delay (i.e. GOPS delay and Queuing delay) are computed based on the load of the node that depends on the assign processing power due to data throughput and the BS functions. The required processing capacity is defined so the load of the nodes does not exceed a certain threshold. The purpose of the model is to evaluate the different network architectures performances, and the behaviour of those architectures on each type of service.

B. Architecture scenarios

C-RAN implementations on a 5G network have multiple scenarios, depending on the position of the RU, CU, and DU. Operators may use different deployments scenarios on the

network to address the different applications of 5G, so it is important to address all possible solutions:

- RU-DU-CU (Independent RU, DU, and CU locations) - In this scenario, the distance between the RU and the DU can go up to 10 km, while the distance between DU and the CU can range from 20 km to 40 km.
- RU-DU+CU (Independent RU and Co-located CU and DU) - In this scenario, there is no MH and, in consequence, the CPRI interface between the nodes is heavier.
- RU+DU-CU (Independent CU and Co-located RU and DU) - In this case, the distance between RU and the DU is very small (i.e. in the same building) and, in this case, there is no FH.
- RU+DU+CU (Collocated RU, DU, and CU) - In this scenario, the network only has BH, and the processing on the network is all done in the RU node.

The different implementation approaches correspond to a different mapping of BS function on the C-RAN nodes. Figure 2 illustrates the scenarios for the splitting options of BS functions for the different C-RAN implementations.

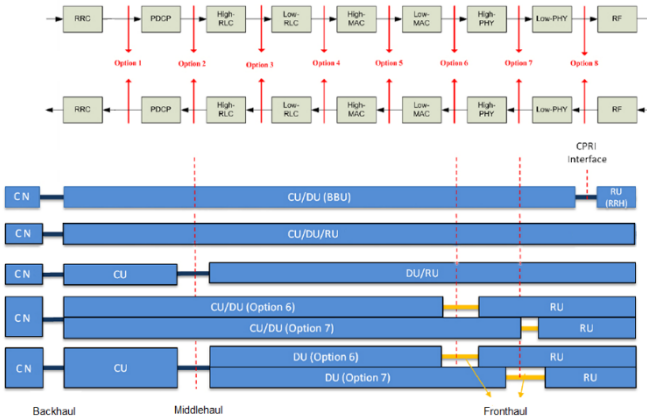

Figure 2. Mapping of CU, DU and RU function according to the split points (adapted from [7]).

C. Services Specification

The network traffic demand on each network node is one of the input parameters necessary for the computation of the link capacity on the links and the load of the different nodes that are directly related to the queuing delay of the nodes.

Table 1. Services characteristics (adapted from [8], [9] and [10]).

Service name	Service Type	Data rate [Mbit/s]	Latency [ms]	Priority
Voice	4G Traffic	0.032	100	3
Video conference		2	150	5
Video streaming		5.12	300	6
Music streaming		0.128	300	7
Web browsing		0.5	300	9
Social networking		2	300	8
File sharing		1	300	10
Email		0.512	300	12
Virtual reality	eMBB	1000	1	4
Realtime gaming	eMBB	1000	1	4
Smart Meters	mMTC	0.1	300	11
Factory automation	URLLC	1	0.25	2
Road safety ITS	URLLC	10	10	2
Remote surgery	URLLC	100	1	1

The traffic assigned to each RU node depends, firstly, on the number of users assigned to each RU node and, secondly, on the different services characteristics. In this study, it is addressed the main network services already supported by the network and the new 5G use cases that will be supported by the 5G network. The reference characteristics of the services are described in Table 1. The 5G use cases were divided according to the three 5G services type: Enhanced Mobile BroadBand (eMBB), massive Machine-Type Communications (mMTC), and Ultra-reliable Low Latency Communications (URLLC).

D. Model Output Parameters

1) Latency

Based on Table 1 the latency critical services in 5G can require an E2E latency from 1 ms to 300 ms. The E2E latency is based on the delay of packet transmission through the network. Two scenarios are considered: one without the implementation of MEC that takes into account the C-RAN, Core backhaul, core network, and external data centre delays, whose delay contribution on the network is presented in (1), and another, with the implementation of MEC. In this second case, there are two possibilities. The first considers that the information does not go to the CN node and takes into account just the C-RAN, MEC backhaul, and the MEC processing delays, whose delay contribution on the network is presented in (2). In this second case the traffic can also be routed to the core if the delay network demands allow extra network latency, or the network is overloaded:

$$\delta_{\mathrm{E2E[ms]}} = \delta_{\mathrm{C-RAN[ms]}} + 2\delta_{\mathrm{BH,C[ms]}} + \delta_{\mathrm{Cor[ms]}} + \delta_{\mathrm{Tran[ms]}} + \delta_{\mathrm{EN[ms]}} \quad (1)$$

$$\delta_{\mathrm{E2E[ms]}} = \delta_{\mathrm{C-RAN[ms]}} + 2\delta_{\mathrm{BH,MEC[ms]}} + \delta_{\mathrm{MEC,UL/DL[ms]}} \quad (2)$$

where:

- δ_{E2E} - End to End Latency.
- $\delta_{\mathrm{C-RAN}}$ - C-RAN associated Latency.
- $\delta_{\mathrm{BH,C}}$ - Backhaul to core transmission Latency.
- $\delta_{\mathrm{BH,MEC}}$ - Backhaul to MEC transmission Latency.
- δ_{Cor} - Core processing delay.
- δ_{Tran} - Transport transmission delay from the core to the Internet data centres.
- δ_{EN} - External Data centre contribution delay.
- $\delta_{\mathrm{MEC,UL/DL}}$ - MEC processing delay.

The C-RAN delay represents the latency contribution from the network edge. In this case, the delay contributions come from the RU, DU, and CU processing delays and the transmissions delays from the FH and MH.

$$\delta_{C-RAN[ms]} = \delta_{RU,UL[ms]} + \delta_{RU,DL[ms]} + 2\delta_{FH[ms]} + 2\delta_{MH[ms]} + \delta_{DU,UL[ms]} + \delta_{CU,UL[ms]} + \delta_{DU,DL[ms]} + \delta_{CU,DL[ms]} \quad (3)$$

where:

- $\delta_{RU,UL/DL}$ - RU processing delay on UL and DL.
- δ_{FH} - Transmission delay between the RU to the DU.
- $\delta_{DU,UL/DL}$ - DU processing delay on UL and DL.
- δ_{MH} - Transmission delay between the DU to the CU.
- $\delta_{CU,UL/DL}$ - CU processing delay on UL and DL.

The processing delay on the nodes depends on two factors - firstly, the delay from the process of the BS function, which is directly related to the amount of functions that are addressed to the node. Secondly, the queuing delay from the input traffic. (4) illustrates the processing delay on the node:

$$\delta_{Node,UL/DL[\text{ms}]} = \delta_{Node,proc[\text{ms}]} + \delta_{Node,que[\text{ms}]} \quad (4)$$

where:

- δ_{Node} - Processing delay on the node.
- $\delta_{Node,proc}$ - BS function processing delay on the node.
- $\delta_{Node,que}$ - Queuing delay on the node.

To summarise, Figure *3* illustrates the delay contributions from the different parts of the network.

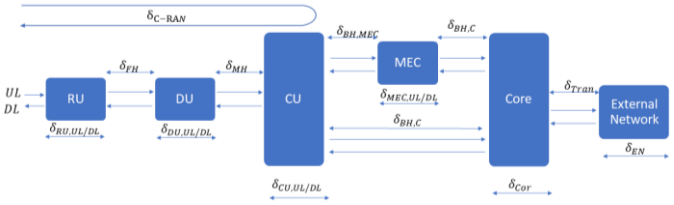

Figure 3. Delay contributions on the network.

2) Node Processing Power

In order to achieve the maximum performance of the network, it is important to balance the processing capacity in RU, DU and CU specific for each use case requirements. The processing power on the node is one of the two parameters that define the processing capacity of the node, and the processing power requirements are directly correlated with the splitting option of the BS function, so it is important to analyse the processing required for each BS function. This parameter is measured in Giga Operations per Second (GOPS), and the model used is based on [11].

The model presented estimate the processing power used in each node instance (i.e. RU, DU, CU, MEC, and CN) for DL and UL, taking into account the multiple physical layer functions processing power, the processing power associated to the data flow management and system control of the MAC and RLC layer, the processing of the PDCL, and the processing power used for the transmission to the core network.

Depending on the chosen splitting option the different power component is assigned to the RU, DU, or CU, assigning the correspondent functions to the different splitting points. The processing capacity in each node can be computed from the following equations:

$$P_{RU\,[\text{GOPS}]} = \sum P_{i[\text{GOPS}]} \quad (5)$$

$$P_{DU\,[\text{GOPS}]} = \sum^{N_{RU}} \sum P_{i[\text{GOPS}]} \quad (6)$$

$$P_{CU\,[\text{GOPS}]} = \sum^{N_{DU}} \sum^{N_{RU}} \sum P_{i[\text{GOPS}]} \quad (7)$$

$$P_{MEC/CN\,[\text{GOPS}]} = \sum^{N_{CU}} \sum^{N_{DU}} \sum^{N_{RU}} \sum P_{i[\text{GOPS}]} \quad (8)$$

3) Node Throughput

The Node throughput is the second parameter that defines the processing capacity on the node, the throughput is an important factor to choose the best splitting option of the network architecture, since higher throughput on the nodes leads to more expensive nodes and interfaces. There are two different splitting options that need to be studied. The lower splitting options, which correspond to the splitting between the RU and the DU, and are directly related to the FH link capacity, and a high splitting which divides the function from the DU to the CU and have a direct impact on the bitrate of the middlehaul. It is considered that the splitting option on the BH is always the splitting option 1. To calculate the throughput on the nodes it is necessary to do an overview of the different link capacities for each functional split option proposed in the model since it is considered that the signal compression factor is proportional to the link capacity for the different splitting options. This study is based on [12] and [13].

It is possible to compute the input and output throughput on the nodes. Since the input throughput on the RU does not suffer any compression, the data rate that arrives at the RU is precisely the data rate generated by the user connected to the RU and can be calculated from (9). Next, the data is compressed in the node, depending on the splitting option and the output throughput of the node given by:

$$R_{Node,out\,[\text{Mbit/s}]} = R_{Node,in\,[\text{Mbit/s}]} \frac{R_{in,split,[\text{Mbit/s}]}}{R_{out,split[\text{Mbit/s}]}} \quad (9)$$

where:

- $R_{Node,in}$ – RU/DU/CU/MEC/CN input throughput.
- $R_{Node,out}$ – RU/DU/CU/MEC/CN output Throughput.
- $R_{i,split}$ – Bit rate of the input splitting option.
- $R_{out,split}$ – Bit rate of the output splitting option.

The input throughput on the other nodes of the network is dependent on the number of connected nodes.

4) Centralisation Gain

First, (10) characterises the aggregation gain, comparing the traffic peaks of each node with the traffic peak in the aggregation node, depending on the number of nodes that are aggregated to it. In an FH link the aggregation node is a DU and the connected nodes are the RU, in an MH link the aggregation node is the CU and the connected nodes correspond to the DU, and in the BH link the connected nodes are the CUs and the aggregation is the MEC or the CN.

$$G_{mux,T} = \frac{\sum_{i=1}^{N_{Node,c}} T_{Node,c,i[\text{GB/h}]}}{\sum_{j=1}^{N_{Node,a}} T_{node,a,j[\text{GB/h}]}} \quad (10)$$

where:

- $G_{mux,T}$ - Aggregation gain.
- $N_{Node,c}$ - Number of nodes connected to the aggregation node.
- $T_{Node,c,i}$ - Peak traffic generated in the connected node.
- $N_{Node,a}$ - Number of aggregation nodes.
- $T_{Node,a,j}$ - Peak traffic generated by the aggregation node.

The centralisation gain can also be calculated based on the distribution of the BS function. Increasing the number of function processes in the aggregation node will increase the centralisation gain, this is called processing gain:

$$G_{proc} = \frac{\sum_{j=1}^{N_{Node,a}} P_{node,a,j[\text{GOPS}]}}{\sum_{i=1}^{N_{Node,c}} P_{Node,c[\text{GOPS}]} + \sum_{j=1}^{N_{Node,a}} P_{node,a,j[\text{GOPS}]}} \quad (11)$$

5) Cost

The study under analysis only takes into account C-RAN cost related, since it is assumed that the operator already as implemented the backhaul and core network, and that it will not

suffer any additional cost from the new C-RAN implementation. The total cost considered is divided into two parts, the CAPEX and OPEX of the network:

$$C_{T[€]} = C_{CAPEX[€]} + N_y C_{OPEX[€]} \quad (12)$$

where:

- C_T - Total cost of the C-RAN.
- C_{CAPEX} - Total CAPEX.
- C_{OPEX} - Total OPEX per year.
- N_y - Number of years considered for OPEX.

Since the model used in this study aims to optimise the splitting options of the BS function, it is proposed a cost function model in order to compare the different options taken into account the different data rate on the link connections, and the different processing power requirements of the nodes. The CAPEX cost of the C-RAN implementation takes into account the hardware, licences and civil work costs, and (13) describes the implementation cost of C-RAN based on [14]:

$$C_{CAPEX[€]} = C_{t,Link[€]} + C_{t,Node[€]} \quad (13)$$

where:

- $C_{t,Link}$ - Total cost of the link.
- $C_{t,Node}$ - Total cost of RU, DU, CU and MEC Nodes.

Regarding the links between the nodes, one assumes that they can be fibre links or microwave links.

The model takes into account constant and variable values from the implementation of the C-RAN, (14) provides a cost model depending on the power capacity of the node in order to emphasize the different processing power depending on the different splitting options proposed in the model.

$$\Delta_{[€]}(i,j) = P_{[GOPS]}(i,j)\beta_{[€/GOPS]} + N_{Inter} C_{Inter_{[€/Interface]}} \quad (14)$$

where:

- β - Cost per unit of resource for the node.
- N_{Inter} - Number of interfaces in each node.
- C_{Inter} - Cost per interface of the node.

Regarding the OPEX cost, the following expression considers three main factors based on [5] and [6]:

$$C_{OPEX[€]} = C_{P[€]} + C_{R[€]} + C_{M[€]} \quad (15)$$

where:

- C_P - Cost related to power consumption per year.
- C_R - Cost related to renting per year.
- C_M - Cost related to maintenance per year.

E. Model Implementation

The main point of the model is to analyse the different architecture implementation scenarios of the network for the different input parameters, depending on the chosen use case and the network specification. In the C-RAN side, the model takes into account the different locations and aggregations scenarios of RUs, DUs and CUs to study the fronthaul, middlehaul connections on the network. It is considered the possibility of a MEC implementation in order to support the ultra-low latency services in 5G. Figure *4* illustrates a detailed implementation perspective of the general model.

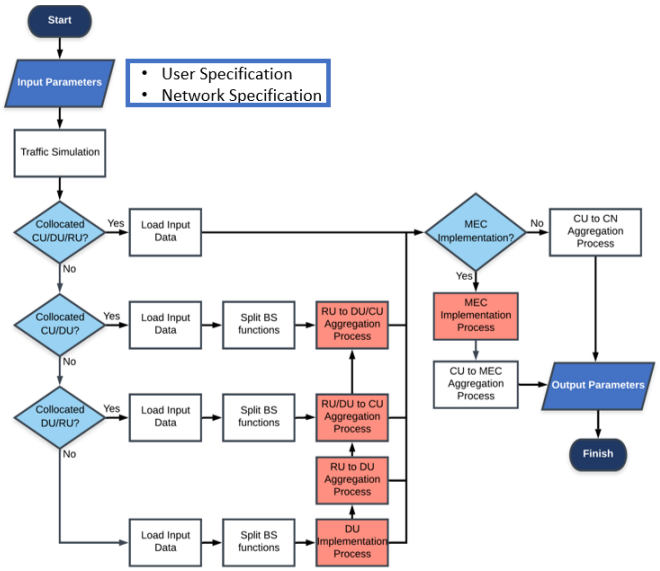

Figure 4. Model Flowchart.

In this study, there are multiple different possible aggregations, and each aggregation considered will have different input parameters, but the algorithm provided will be the same. The model provides different aggregation algorithms developed by [5]. This study considers two different algorithms:

- Minimise Delay Algorithm - This algorithm is used to aggregate the closest nodes with the required processing capacity. The model analyses all the possible connection distances and choose the smallest one. This algorithm aims to reduce link distances, reducing network delay.
- Balance Number of Connections Algorithm - This algorithm aims to balance the number of aggregations for each node. The model checks the maximum capacity of the node with fewer connections until it is capable to aggregate new nodes. The number of aggregations for each node is always updated and available for the next step of the aggregation process evaluation.

Since the model used in this study is based on a 4G C-RAN, it is necessary to convert the RRH and BBU locations to an RU, DU, CU, and MEC implementation.

There are two scenarios where the location problem needs to be addressed. One is on RU, DU, and CU independent location scenarios, and the other is on a MEC implementation scenario. In order to solve the problem, the model aims to efficiently add to the location of the existing node the new node that is being implemented in the network. Considering the first scenario, the model evaluates the density of the RU depending on the FH distance. A threshold level is taken into account in order to consider the possibility to only implement DU nodes in dense traffic areas like urban and dense urban scenarios, if desired. In order to compute the best DU location, it is used the K-means algorithm, a Machin Learning algorithm which first receives the RU locations, dividing the RUs into N_{DU} clusters, and computes the centroid of each cluster that corresponds to the best DU theoretical location.

F. Model Assessment

The model assessment aims to validate the model in the development stage, and uses a set of empirical tests in which the outcome of the results is already expected in order to verify if the model follows the theoretical results. Table *2* described the structure of the empirical tests used to validate the model.

Table 2. List of model assessment tests.

Test ID	Description
1	Validation of the input file read, by verifying if the size and type of inputs values stores in the different variables are the same as in the input files.
2	Scattering the position of the RUs, CUs and CN positions in the Matlab plot over a Google Maps to inspect the node placement on the scenario.
3	In case of implementation of new nodes: • Scattering the position of the new DU or MEC positions, plotting the original nodes and centroids positions in the Matlab plot over a Google Maps to inspect the node placement on the map. • Check if the computational and link capacity values are updated. • Check if the connection is correctly stored.
4	Validation of the maximum distances constraints by checking if there are no connections that do not respect the constraints.
5	Validation of the aggregation process: • Check if the computational and link capacity values are update. • Check if the connection is correctly stored. • Check if the node is marked as served and not assigned again.
6	Validation of the output files, by checking if they are correctly printed and plotting the output results.

IV. Results Analysis

A. Scenarios

The study under analysis uses data provided by NOS on the Minho region. The area of Minho is located in the north-west of Portugal, where the main regions into consideration are Porto, Braga, Viana do Castelo, and Vila Real. This scenario has around 3.4 million inhabitants in around 11 600 km² of area. Minho has a population density of 290 inh./ km², and the majority of the population are in the Porto metropolitan area, which has a population density of 843 inh./ km².

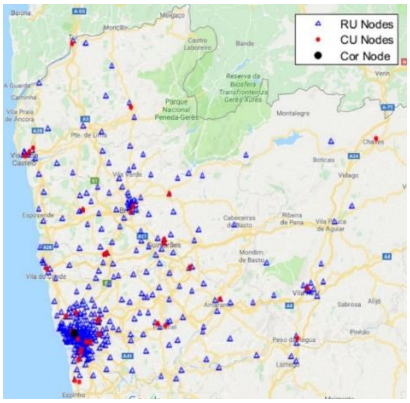

Figure 5. RU, CU and CN locations in Minho.

It is considered the RU nodes as the cell sites, the CU nodes as the aggregation ones, and also the core node. This scenario has 374 RUs, 42 CUs and 1 CN.

The RUs on the map are analysed based on the density of the nodes. Three environments are considered for the classification of RU density types: a dense urban, an urban, and a rural environment. This classification is essential to analyse the processing power of the nodes, as it considers three different bandwidths for the nodes: 100 MHz for dense urban RUs, 50 MHz for urban RUs, and 20 MHz for rural RUs.

B. Centralisation Gain Analysis

Regarding the centralisation gain, one verifies that a higher splitting option achieves higher aggregation gains since the signal is more compressed in the RU nodes and the data rate on the CU nodes are reduced but, on the other hand, the process gain is reduced since fewer functions are assigned to the central node, achieving less benefits from the centralisation of resources. Splitting the physical layer will achieve great traffic and throughput aggregation gains, substantially reducing the requirements of the FH link capacity when compared with the splitting option 8, which leads to belief that to support the high traffic demands of the 5G network it is necessary to assign more processing power on the RU nodes, even though the gains of resource centralisation will be reduced.

Concerning the UL results, the first conclusion is that the traffic on the CU is higher than in the DL, which leads to a reduction in the traffic and throughput aggregation gain for all splitting options. In the FH, one notices a more significant improvement from the physical layer splitting 7.3 (i.e. 10 times higher traffic gain when compared with the option 8). The values for GOPS aggregation gain on the UL are similar to the DL since just a small percentage of the processing power on the node is traffic loaded dependent. The process gain on the UL follows similar results as the DL results, as explained before, and the BS function processing power on the nodes are low traffic load dependent. When giving a general overview on the results of the comparison of DL and UL results, it is worth emphasising that the UL achieves more significant improvements on the network gains on the 7.3 splitting option, since option 7.1 and 7.2 are not beneficial in a UL network environment.

C. Processing Capacity Analysis

1) Throughput Analysis

The output throughput on all the different RU splitting options was analysed, since the chosen link capacity on the link is related to the output throughput on the node. As expected, the most noticeable reduction on the throughput is from the splitting option 8, which corresponds to a 4G architecture, and the splitting option 7.1, where it achieves a 93.7% reduction. When considering the splitting option between the PHY and the MAC (i.e. Option 6), there is a 42.6% throughput reduction from option 7.3 to option 6. From the independent RU splitting option 6 to a collocated RU and DU, that corresponds to an MH splitting option 2, the output node throughput decreases 28.6%, and the throughput from the RU+DU node to the RU+DU+CU node only reduces 0.4%. From these results, one can conclude that the 5G C-RAN architecture should change the FH splitting option 8 to a higher splitting option on the splitting of the physical layer in order to support the high throughputs arriving

on the CU nodes from the new 5G use cases. Figure 6 summarises the results of the output throughput on the RU node

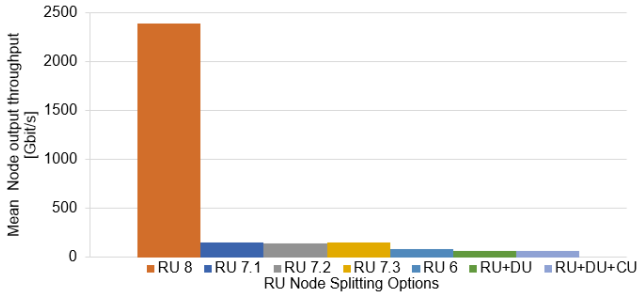

Figure 6. Mean output throughput on the RU nodes in different splitting option on DL.

2) Processing Power Analysis

Regarding the RU-DU+CU architecture it is worth noticing that the options with higher impact on the processing power on the nodes are the transition from an option 8 to an option 7.1 and the transition from an option 7.2 to an option 7.3. In the transition from an option 8 to an option 7.1, there is a 43.3% variation on the processing power on the nodes. The difference between these options is the FFT that modulates the signal, and it is the BS function with higher GOPS assigned to it. The transition from option 7.2 to 7.3 has a 65.2% variation, and this high variation between the options appears because in this transition there is a variation of two BS functions - the MIMO coding/decoding and the baseband modulation/demodulation. Finally, one verifies that confident interval, changes with the splitting option, and this is explained since the functions of MIMO encoder, baseband modulation, and channel coding is dependent on the load of the node, so when these functions are assigned to the RU (i.e. option 7.3 and 6) the processing power changes throughout the day, and when these functions are in the DU+CU node (i.e. option 8, 7.1 and 7.2), the standard deviation of the processing power increases on the DU+CU node.

Regarding RU-DU-CU architecture, the CU node has low processing power requirements since the only BS function assigned to the CU is the PDCP BS function. The splitting option between the DU and CU for all FH splitting options is option 2, so the variation on the CU processing power that is illustrated in Figure 7 happens because, if an RU node does not have the capability to connect to a DU node, the RU is connected to the closest CU node. This architecture achieves a 28.5% reduction of processing power between the DU+CU node to the DU node and 83.9% reduction when compared with the DU+CU and the CU processing power. This architecture can be an alternative to offload CU processing requirements, instead of increasing the number of BS function on the RU nodes.

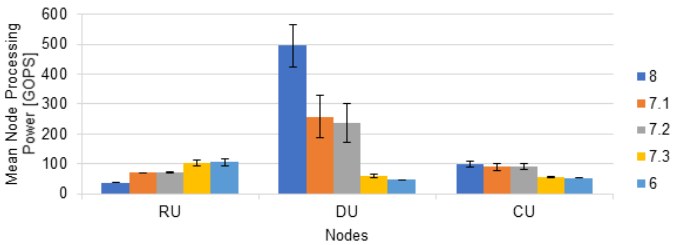

Figure 7. Mean processing power on the network nodes in different splitting options RU-DU-CU architecture on DL.

D. Latency Analysis

In the following section, it is analysed the latency impact on the network, and this section is divided into six subsections in order to analyse the output parameters related to the latency for several different variations of the input parameters.

1) Distance Analysis

In the RU-DU+CU scenario, the average total network distance is 88 km. In this case, there is an FH with a maximum distance of 10 km between RU and CU, and a BH to the CN. The RU-DU-CU architecture has an additional MH with a maximum distance of 40 km, but it is worth noticing that the total network distance is reduced 10.8% than in the RU-DU+CU reference Minho scenario, this appears since, with the new DU nodes, the FH distance is substantially reduced and, in the MH link connections, the model follows an algorithm to minimise link distance instead of a balancing algorithm used on the FH connections, so even though there are 3 links in this architecture, the total network distance is lower. It is important to remember that the first link connection of the model uses a balancing algorithm in order to balance the traffic on the network, and the second or third link connection of the network uses a minimise delay algorithm.

The highest network distance is in the RU+DU-CU architecture. In this case, there is a MH with 40 km maximum distance and, since the MH is the first connection of the network, it is using a balancing algorithm. There is an increment of 48.1% total RU to CN distance when compared with the RU-DU+CU scenario. Finally, the all collocated nodes achieve a 5.2% reduction when compared with the RU-DU+CU scenario, and this happens because, in this case, there is only the BH, so the RU node is directly connected to the CN.

2) Impact of the architecture

This subsection analyses the RU-DU+CU scenario latency impact considering a constant processing power on the nodes for all splitting options. The processing power values considered are from splitting option 7.1, while this analysis focused on the processing delay parameters. Those are the GOPS delay, that is associated with the processing of the BS function, and the queuing delay, that is proportional to the input traffic on the nodes. Figure 8 illustrates the results, where one can verify that, for a constant processing capacity, option 8 has 45.4% higher total latency than option 7.1. This variation is mainly due to the queuing delay increase in this option since the throughput on the CU node in option 8 is 94% higher than in option 7.1. For the higher splitting options, the queuing delay is nearly constant and processing delay reduction is achieved due to the BS functions distribution on the nodes. The transition from the splitting option 7.2 to 7.1 has a 58.1% reduction in the mean GOPS delay.

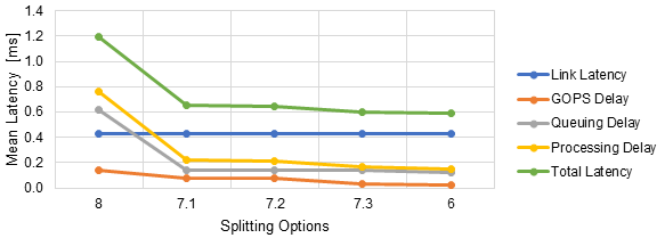

Figure 8. Mean network latency for RU-DU+CU architecture with fix processing power.

3) Impact of maximum latency

Figure *9* shows the mean network latency on the network for different maximum latency demands. In order to support lower maximum latencies, the network is forced to use MEC nodes in order to reduce the network distance. It is important to know that only 5G use cases can be processed in the MEC nodes without going to the CN node.

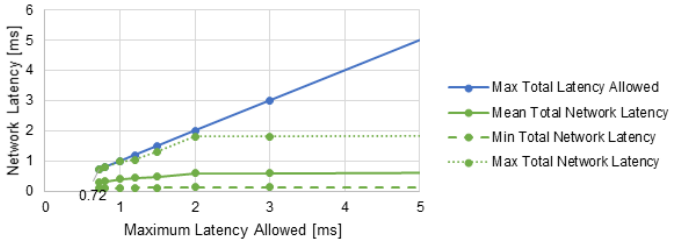

Figure 9. Network Latency variation with maximum network latency.

One can clearly observe that the network reaches a maximum total latency around 2 ms, and this happens when the network no longer needs MEC nodes to support the latency demands. The minimum value for the maximum latency value that the network can support is 0.72 ms, due to the restriction of the minimum processing delays when the network is heavily loaded, like in the evenings, so to reduce that processing latency it is necessary to provide more node capacity.

4) Impact of the node processing capacity

This subsection studies the network latency performance changing the input processing power on the nodes, so the study focuses on the processing delay on the nodes.

When analysing the 5G use cases, one can conclude that the road safety ITS use cases have low processing power requirements, achieving 100% coverage using 10% of the reference scenario processing power. The remote surgery and virtual reality applications coverage stabilise on 90%, which means to increase these use cases cover the network needs to introduce MECs. On the real-time gaming use cases, the processing power on the nodes should be 10 times higher than the reference scenario to achieve the same coverage probability as the other eMBB use cases. This happens since the gaming use case has lower priority level than the VR use case, so if the network nodes have lower processing power the VR overloads the network and the performance of the gaming use case is drastically affected due to its high resource requirements.

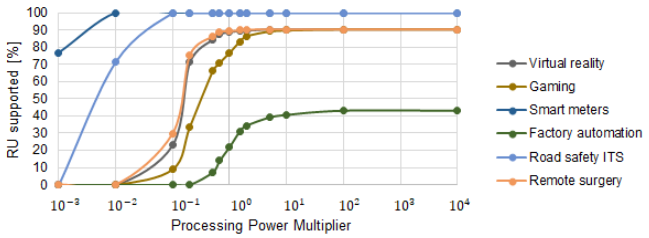

Figure 10. RU use cases coverage for RU-DU+CU architecture with variable processing power.

5) Impact of the users

The number of users on the site is directly proportional to the traffic produced in the network so this subsection focused on the queuing delay parameter that is related to the throughput on the nodes. It is considered the 7.1 splitting option on the RU-DU+CU scenario, so the link latency and the GOPS processing delay will be constant on this subsection analysis.

First, as the usage ratio and penetration ratio variation on the input parameters on the network is analysed, it is considered a reference scenario with a usage ratio of 10% and the penetration ratio 30%. On the reference scenario, the average queuing delay is already 64.5% of the total processing time, which means that normally the majority of the processing time is due to queuing delays. For a scenario with 30% usage and penetration ratio, the average network latency is almost 1 ms and, for an average network time of 1 ms the network cannot support eMBB services. This leads to conclude that increasing the number of users on the cell will have a strong impact on the network latency, reaching an average 25% increase on the network latency, for a 10% variation on the usage and penetration ratio.

Figure *11* illustrates the coverage of the 5G use cases when changing the number of eMBB users. Since the eMBB services uses a lot of network resources, the eMBB is very affected by the increase of users. Doubling the number of eMBB users will reduce by 11% the RUs that support real-time gaming. The VR is more robust to the increment of users since the priority level of this use case is higher than gaming. It is worth noticing that there is no impact on the URLLC services since the priority level of these use cases are higher than the eMBB.

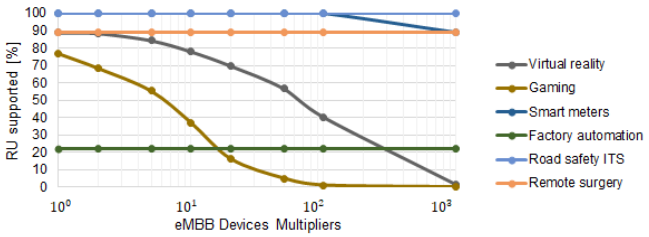

Figure 11. RU use cases coverage with variable eMBB users.

The coverage does not change with the number of mMTC users since the priority level of the service is very low. When increasing the number of devices more than 1 million times, the scenario will start to reduce the coverage of the RU, but as expected with the implementation of smart cities in the future, this number of connected devices will not be achieved.

The final simulation on the network latency analysis is the impact of the URLLC users connected to the network. The URLLC services are characterised by a very low latency requirement (i.e. 1 ms to 10 ms). In this case, the reliability of the service is extremely important to the QoS of the use cases. An application like the road safety ITS require high reliability from the network, so it is important that the service is supported on all the nodes of the network with high reliability.

When considering the coverage of the user cases throughout the network, one concludes that the impact of the URLLC users is uniform for all the use cases. Increasing the number of devices 100 times will lead to a decrease on the performance of the eMBB services and the factory automation service, but it is worth noticing that the Figure *12* shows that the remote surgery use cases support only starts to decrease on 500 times the number of users. This can be explained, firstly, because the reference number of users of remote surgery is the lowest of all

the use cases and, secondly, because the remote surgery priority level is very high so the resources of the network prioritise this use case. The road safety ITS is an important use case of the URLLC services, since it has a network latency requirement of 10 ms and a very high priority level. The network supports one thousand times the reference value of users before coverage starts to drop, but it is noticeable that these values do not account a safety margin, which will reduce the network capacity to support these use cases.

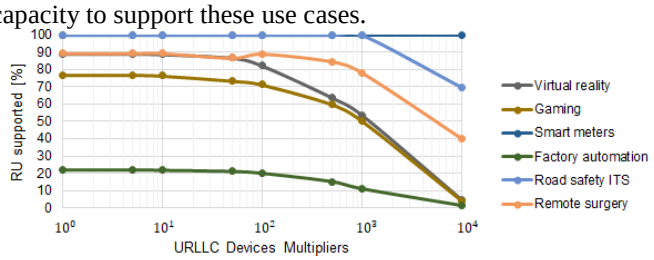

Figure 12. RU use cases coverage with variable URLLC users.

E. *Analysis of Implementation of MEC*

The next analysis measures the impact of MEC nodes on the network distance. One can conclude that, by implementing more than 5 MEC on the Minho scenario (i.e. around 10% of CU nodes), the total network latency impact will not compensate the implementation costs of the MEC. Considering the 5 MEC nodes scenario, the network latency is reduced by 45.3% comparing with an architecture with no MEC nodes.

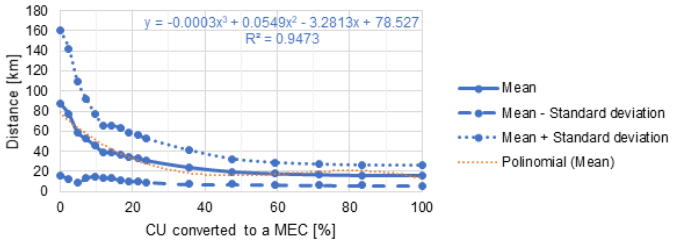

Figure 13. Network distance for RU-DU+CU architecture with a variable number of MEC nodes.

Regarding the use cases coverage on the network, since the 5G use cases have different priority levels, having the same latency requirements like the eMBB services, which does not mean that the use cases (i.e. Virtual Reality and Real-Time Gaming) will need the same network architecture. The results are presented in Figure 14. First, one verifies the previous conclusion that, for 5 MECs on the network, without considering the factory automation cases that have an extremely lower latency demand, all the network use cases considered are supported on more than 98% the RU on the network, which is extremely important since VR and, especially, real-time gaming are services that need to be available for a wide area on the map.

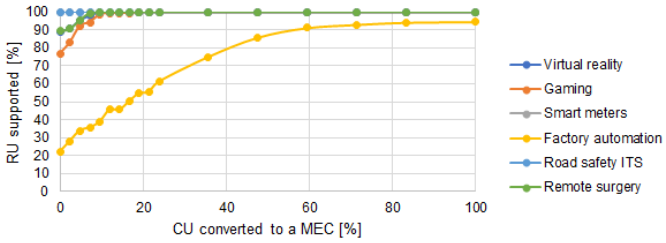

Figure 14. RU use cases coverage for RU-DU+CU architecture with a variable number of MECs.

For factory automation, since it is more region specific, it is not relevant that the network supports this use cases for all RU on the network, but it is worth noticing that, from a zero MEC nodes to a 5 MEC nodes scenario, the coverage of this use case increases 51.5%, covering 45.6% of the RU on the network.

In the previous analysis the network cannot achieve full coverage using an RU-DU+CU architecture for the factory automation use case, so it is considered an independent study using a RU+DU+CU architecture, since this architecture has the smallest network distance due to only having a BH connection and a lower node processing delay since there are only two nodes. In this case, the RU nodes are directly connected to the MEC nodes.

Due to the 0.25 ms latency requirements of the Factory automation, assuming the reference scenario processing power, the MEC node needs to be located at a maximum of 11.5 km from the RU node that is covering the factory, on a fibre link connection, and at a maximum of 19 km on a microwave link.

Table 3. Required MEC nodes on the network to achieve full coverage for different use cases for different architectures.

CU converted to a MEC [%]	RU-DU+CU	RU-DU-CU	RU+DU-CU	RU+DU+CU
Virtual Reality	9.52	9.52	40.48	1.07
Gaming	38.10	Imp.	Imp.	4.28
Factory automation	Imp.	Imp.	Imp.	12.30
ITS	0.00	0.00	0.00	0.00
Remote surgery	9.52	9.52	42.86	1.07

F. *Cost Analysis*

It is considered relative values of CAPEX and OPEX concerning the RU-DU+CU architecture, 7.1 splitting option as reference scenario. Regarding the CAPEX, it is taken into account the cost of the implementation of the links and the cost of implementation of the C-RAN nodes. Concerning the OPEX cost, it is taken into account the expenses of energy, rent, and maintenance of the network per year. This analysis considers the scenario where the processing capacity of the nodes is adjusted to the demands of the different splitting options.

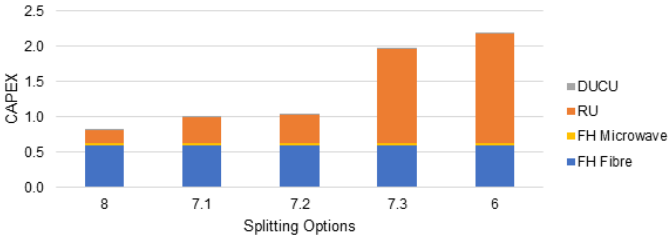

Figure 15. CAPEX for RU-DU+CU architecture with different splitting options.

One can verify that increasing the number of BS function in the RU nodes increases the CAPEX cost of the network, since the network does not benefit from centralisation. From splitting option 8 to spitting option 7.1, the CAPEX cost increases 21%, and it is worth remembering that, from splitting option 8 to 7.1, the network throughput on the CU nodes is drastically reduced. Changing the splitting option from 7.2 to 7.3 will increase the CAPEX cost on 47% due to the high increment of the required processing capacity on the RU nodes. Regarding the OPEX cost, the results follow the same variation with the different splitting options, mainly due to the increase of maintenance cost of the RU nodes. It is noticeable that the rent does not change

with the splitting options since for all splitting options the nodes have the same area.

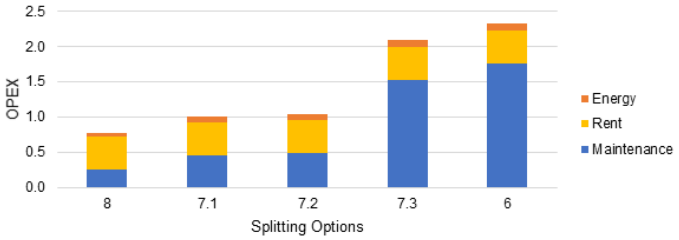

Figure 16. OPEX for RU-DU+CU architecture with different splitting options.

V. Conclusions

Regarding the centralisation gain, one concludes that a higher splitting option achieves higher aggregation gains since the signal is more compressed in the RU nodes and the data rate on the CU nodes are reduced but, on the other hand, the process gain is reduced since fewer functions are assigned to the central node, achieving less benefits from the centralisation of resources. Splitting the physical layer will achieve great traffic and throughput aggregation gains, substantially reducing the requirements of the FH link capacity when compared with the splitting option 8, which leads to belief that to support the high traffic demands of the 5G network it is necessary to assign more processing power on the RU nodes, even though the gains of resource centralisation will be reduced.

Concerning the processing capacity of the nodes, one concludes that changing the splitting option from option 8 to a higher option will greatly reduce the input throughput on the CU, since in the option 8 the signal is not compressed so the throughput that arrives at the CU node is extremely high, and increasing the splitting option from option 8 to option 7.1 shows an 94% reduction on the throughput of the CU node. Another option to offload the central node data rate is to use DU nodes. In this case, the throughput on the DU is 25% lower and the throughput on the CU is 50% lower than in the reference scenario. In the UL scenario, the splitting option 7.1 cannot achieve the same data compression has the DL, and in this case, the best splitting option would be the 7.3, which achieves a 75% reduction when compared with option 8. Since using a splitting option 7.1 is recommended to reduce throughput on the CU node, one verifies that this option requires a 48% higher processing power on the RU. The processing power measure in GOPS on the RU also substantially increases from option 7.2 to 7.3 since the MINO encoding and baseband modulation is performed on the RU (29%).

Considering the network distance, one verifies that assuming a RU+DU-CU architecture will lead to a 48% increase on the network distance since the maximum MH distance is 40 km instead of 10 km on the maximum FH distance. Looking at the other network architectures, the network distances are approximately similar even though the RU+DU+CU architecture only has a BH link that can bring latency benefits on the network, since the RU is directly connected to the CN. Regarding the capacity of the network to support the different 5G use cases, it is important to notice that, when using the 4G C-RAN architecture, it is not possible to achieve full coverage on the ultra-low 2E2 latency use cases, so it is necessary to introduce MEC nodes in the network.

In order to achieve full network coverage on the 5G use cases, the implementation of MEC nodes in the network will be required, in this case, the information of the 5G use cases does not go to the CN nodes, being processed in the MEC node, thus reducing network latency. Considering the Minho scenario, one concludes that it is required at least 5 MEC nodes in order to achieve a maximum network latency below 1 ms.

Concerning the network cost, it was considered relative values on the analysis due to the lack of detailed information on the hardware cost of the nodes since the 5G network is still in its beginning stages. As expected, increasing the BS functions on the RU nodes will increase the overall cost of the network, since the majority of the cost of the nodes on the network comes from the RU nodes. From splitting option 8 to option 7.1, the initial investment on the network is 21% higher just due to the increased cost of the RU nodes.

References

[1] Ericsson, *Ericsson Mobility Report*, Public Consultation, Jun. 2018 [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-june-2018.pdf.

[2] J. Hodges, *Transforming the Edge: The Rise of MEC*, Intel, Heavy Reading, New York, USA, Mar. 2018. Available: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/the-rise-of-multi-access-edge-computing-paper.pdf.

[3] I. Sarrigiannis, E. Kartsakli, K. Ramantas, A. Antonopoulos and C. Verikoukis, "Application and Network VNF migration in a MEC-enable 5G Architecture", in *Proc. of CAMAD18 - 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks*, Barcelona, Spain, Sep. 2018. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8514943.

[4] Y. Lin, Y. Lai, J. Huang and H. Chien, "Three-Tier Capacity and Traffic Allocation for Core, Edges, and Devices for Mobile Edge Computing", *IEEE Transactions on Network and Service Management*, Vol. 15, No. 3, Sep. 2018, pp. 923-933, Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8402110.

[5] H. Silva, *Design of C-RAN Fronthaul for Existing LTE Networks*, M.Sc. Thesis, IST, University of Lisbon, Lisbon, Portugal, 2016.

[6] T.Monteiro, *Implementation Analysis of Cloud Radio Access Network Architectures in Small Cells*, M.SC. Thesis, IST, Technical University of Lisbon, Lisbon, Portugal, Nov. 2016.

[7] ITU-T, *Transport network support of IMT-2020/5G*, Technical Report, Stephen Shew, Ciena, Canada, Fev. 2018.

[8] B. Rouzbehani, *On-demand RAN Slicing Techniques for SLA Assurance in Virtual Wireless Networks*, Ph.D. Thesis, IST, University of Lisbon, Lisbon, Portugal, 2019.

[9] H. Martins, *Analysis of CoMP for the Management of Interference in LTE*, M.Sc. Thesis, IST, University of Lisbon, Lisbon, Portugal, Nov. 2017.

[10] S. Khatibi, *Radio Resource Management Strategies in Virtual Networks*, Ph.D. Thesis, IST, University of Lisbon, Lisbon, Portugal, 2016.

[11] B. Debaillie, C. Desset, and F. Louagie, "Flexible and Future-Proof Power Model for Cellular Base Stations", in *Proc. of VTC Spring 2015 – 81st IEEE Vehicular Technology Conference, Glasgow*, United Kingdom, May 2015.

[12] L. Larsen, A. Checko, H. Christiansen, A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Network, *IEEE Communications Surveys & Tutorials*, Vol. 21, No. 1, Oct. 2018, pp. 146-172. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8479363.

[13] 3GPP, *CU-DU split: Refinement for Annex A (Transport network and RAN internal functional split)*, R3-162102, Sophia Antipolis, France, Oct. 2016.

[14] O. Arouk, T. Turletti, N. Nikaein, K. Obraczka, "Cost Optimization of Cloud-RAN Planing and Provisioning for 5G Networks", in *Proc. Of ICC18-2018 IEEE International Conference on Communications*, Kansas City, MO, USA, May 2018. Available: https://ieeexplore.ieee.org/document/8422744.