Implementation Analysis of Cloud Radio Access Network Architectures in Small Cells

Tiago J. Monteiro
Instituto Superior Técnico / INESC-ID
University of Lisbon
Lisbon, Portugal
tiago.monteiro@tecnico.ulisboa.pt

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

Abstract— The main objective of this work was to analyse the performance parameters of a Cloud Radio Access Network in an already deployed LTE-A network of a mobile operator by taking advantage of the functional Remote Radio Head and Baseband Unit split, with the centralisation of the processing power in standard data centres. The work consists of the analysis of the current macro base station location and a list of the possible data centre to present a possible C-RAN deployment under different strategies and algorithms that a mobile operator may be interested in. The metrics studied are fronthaul latency, pool capacity in traffic per hour, processing power capacity and multiplexing gain. A total cost of ownership model is also presented to estimate cost savings possible with the technology. The model is implemented in a computational tool to provide a generic study of any scenario. The results obtained in the central area of Porto have proven that C-RAN implementation with 19 pools is possible without introducing latency problems. In this kind of implementation, the operator can combine different traffic profiles and achieve a multiplexing gain of up to 1.31 that can be translated in capacity savings. A cost reduction of 63% in capital investment and 31% in operational expenditures was predicted for a centralised deployment when compared to another green field deployment without centralisation. A futuristic approach was also added to the approach to simulate how massive small cell deployment to handle an 8-time increase in traffic affect the results.

Keywords- LTE, LTE-A, Cloud RAN, Radio Access Network, RRH, BBU, Multiplexing Gain, Small cells.

I. INTRODUCTION

Mobile communications systems provide their users the possibility to communicate to each other wherever they are through their mobile equipment. Driven by the evolution of the internet and the increased computational resources of the user’s mobile equipment, operators have been expanding and upgrading their networks to cope with the new traffic demands. The number of mobile subscribers has been globally increasing around 5% per year [1]. The increasing number of smart devices connected to a mobile network means that the average generated traffic per user is also growing. The aggregated traffic by 2020 is expected to be around 9 times greater than the one measured in 2015, with video data representing around 60% of that number [1].

However, the revenue obtained is not growing fast enough to ensure profitability. Not only the traffic and the revenues growth tendencies are decoupled, as revenue growth tends to stabilise or even to decline slowly. This situation creates new challenges to operators. The Total Cost of Ownership (TCO) that includes the Capital Expenditures (CAPEX) and the Operational Expenditures (OPEX) is increasing in order to improve network infrastructures. In this way, network operators can satisfy their users and remain competitive between each other. Nevertheless, the Average Revenue per User (ARPU) is not growing significantly, as the typical mobile user is data-dependent but still expects to pay less for data usage. With that said, operators need new architectures to optimise costs, while providing high-capacity to their subscribers. In addition, more efficient and greener solutions are expected due to energy-saving concerns.

The concept of Cloud-RAN emerges as a solution of the mentioned problems for mobile operators. A centralised approach in which all the signal processing needed for multiple base stations (BS) can be performed in one single physical location with increased computational resources can reduce the global costs of the physical infrastructure, rent and operational expenditures in energy. Figure 1 represents a global view of this solution in which all the network is running over data centres. Savings in OPEX, such as electricity (mostly associated with air conditioning), as well as in CAPEX in civil works and transmission equipment appear as an advantage. Furthermore, upgradability and multi-standard support are facilitated with software-defined equipment in data centres. The concepts of C-RAN combine with recent trends in networks such as Software-Defined Networks (SDN) and Network Function Virtualisation (NFV) to define a scalable and adaptable virtualised network with a control plane separated from the data plane.

Figure 1 - Mobile Cloud Architecture (extracted from [2]).

Besides the multi-standard support and the predicted software reconfigurability, centralised processing offers the possibility of tight coordination between base stations that are
technological or geographically different from each other. Processing all the centralised data in powerful machines is then promising for the latest and future signal processing techniques and distributed algorithms, certainly improving aspects such as the achievable data rates and the energy efficiency. With that being said, is then expected to have such kind of deployment not only being considered for up-to-date networks but also studied for 5G networks, currently aiming at improvements in latency, data rates, energy consumption and connected devices by one or more orders of magnitude. Besides that, the prediction of future denser networks by deploying low-powered base stations can also take some benefits of this new approach.

There are still issues that need to be addressed besides the virtualisation of baseband resources and stringent fronthaul solutions. This work is focused on the problems associated with the implementation of C-RAN architectures in different network configurations. Taking into account the stringent requirements of latency, processing and transmission capacity of LTE-A, an analysis on how to deploy the C-RAN components in an already existent network and the performance achieved with that implementation is needed. Another important metric to quantify is the theoretical gain that can exist by centralising cells with different traffic profiles (residential or office traffic) and the implication of the value might have in the hardware reduction. A particular focus will be given to heterogeneous networks architectures in which the radio frequency components correspond to small cells. The goal of this last study is to propose an improved network with much more RRHs handling the new traffic demand that is expected in the following years. This layer is called proliferation of cells.

The rest of the paper is organised as it follows: Section II presents the C-RAN state of the art. Section III presents the model development and the metrics used to characterise C-RAN. Section IV contains the scenario description and the results analysis. Finally, Section V presents the most important conclusion and possible work to complement this one.

II. STATE OF THE ART

Several studies addressing the major challenges imposed to C-RAN deployments in LTE-A systems are described in the literature. This section states the work developed by several authors in the area of implementation analysis and performance of Cloud Radio Access Networks architectures.

Regarding the statistical multiplexing gain achieved with the centralisation of processing resources into virtual base stations, the authors in [3] introduce a mathematical model to fully analyse it. In this work, a method to calculate the blocking probability of a virtual base station pool and the limit of the pooling gain are derived. The results confirm the statistical multiplexing gain and show that larger pools achieve a negligible gain that might not justify larger economical investments. The model does not cover features such as CoMP and the authors state that verification with realistic data is needed. Working on the statistical multiplexing gain as well, the authors in [4] analyse the traffic profiles of office and residential areas in order to evaluate the pooling strategies that lead to higher pooling gains and higher potential cost savings. The results point to an achievable multiplexing gain of up to 1.6. In this scenario, 20%-30% of the office cells are pooled together with 70%-80% of residential cells. In the same work, the authors also compare the cost of one BBU and one kilometre of fibre to reach the conclusion that the most advantageous deployment situations are the dense urban scenarios (less than 100 km²). In [5] a similar approach is taken and the simulation results in the city of Cologne with traffic prediction for 2017 prove a 75% reduction of BBU resources when compared to traditional RAN architectures. The authors state that the value is higher than expected and that 50% of reduction should be the most optimistic scenario in real situations. In addition, the authors in [6] study the multiplexing gain achieved with processing capacity instead of data rates or traffic, using a power model to characterise the RRHs capacity in Giga Operations per Second (GOPS). The results obtained point to savings of 15% aggregating 57 sectors in the same pool and show a strong dependence on the user distribution throughout the scenario. Using the same power model, [7] builds a model to analyse the overall capacity in GOPS in the city of Lisbon, although does not add traffic variation during the day. The same author builds a proliferation model to analyse the influence of small cells. Still, the authors in [8] build a framework for experimental results on processing savings resulting in about 27% of savings but concluding that only 2% are derived of the centralisation. The main cause of savings is the under-provisioning of resources that could be also done in a non-centralised implementation. Based on the previously presented works, the authors in [9] apply their model to different case studies including heterogeneous network deployments and different traffic profiles. An optimal ratio of 22% of office cells pooled with 78% of residential cells is obtained. The analysis is done based on cost and sensitivity to traffic variations. The authors also point out the benefits of introducing a dynamic BBU-RRH mapping to achieve higher gains and state that their model could be adopted for the re-assignment strategies. NFV and SDN are considered for the re-assignment.

In [10], an analysis considering only the CAPEX is done to estimate the feasibility of using cheaper microwave links to replace optical fibres. This study considers a cost factor depending on the population density and the average traffic per user. The results show that for high density areas the microwave replacement leads to lower CAPEX. However, the single cell peak data rate is limited due to microwave capacity restriction. The authors also suggest a solution in which eNodeBs and RRH are simultaneously connected to the BBU pool. This last concept simplifies the fronthaul/backhaul connection but it reduces the achievable multiplexing gain.

III. MODEL DEVELOPMENT

A. Model Parameters

The parameters under study in this work are the latency, the traffic capacity, the processing power, the multiplexing gain and the TCO.

Latency or delay is an important metric in telecommunication systems. As the main difference introduced by centralised architectures is the split of the base station into RRH and BBU with the addition of fronthaul connections, the delay limit to be achieved corresponds to the maximum latency allowed for processing in traditional eNBs. In C-RAN, the distributed elements of the base station introduce delays as it is expressed in the following expression:
The eNB latency can be derived from 3GPP specifications and corresponds to 2.0 ms. This restriction of the processing time in an eNodeB limits the delay budget of C-RAN components, especially in fronthaul. This latency is translated into the maximum distance achievable for a link. In (2), one shows the dependence of the maximum distance with the propagation speed and the tolerated delay.

\[
d_{FH} [\text{km}] = \frac{\delta_{RRH} [\text{ms}]}{2} \cdot v [\text{km/ms}]
\]

where:
- \( v \) – Transmission speed in the link.
- \( d_{FH} \) – Fronthaul allowed distance.

The achievable distances are medium dependent as the transmission speed differs if the transport solutions chosen for the fronthaul is optical fibre or microwave links.

A stringent value of 200 \( \mu \text{s} \) is often considered in the literature to account for delay sensitive functions such as CoMP.

The capacity values used to characterise pools are measure in traffic per time (GBph) as these values are the realistic ones obtained in real cells. The values fully characterise the 24 hours of the day and the different traffic curves of each one of them.

The need to model the power required by each base station comes from the virtualisation of base stations in the BBU pool. In order to fully analyse the implementation of C-RAN, the overall processing capacity required for a given network configuration should be considered. This parameter is measured in Giga Operations per Second (GOPS). In [11], a power model is used to characterise the processing capacity required for different types of LTE base stations. A simple adaptation of the proposed model to a centralised architecture such as C-RAN is described as:

\[
P_{bbu,pool} [\text{GOPS}] = \sum_{i=1}^{N_{bbu}} P_{bbu,n} [\text{GOPS}] + P_{\text{fixed}} [\text{GOPS}]
\]

where:
- \( P_{pool} [\text{GOPS}] \) – BBU pool processing power.
- \( P_{bbu,n} [\text{GOPS}] \) – Single BBU processing power.
- \( P_{\text{fixed}} [\text{GOPS}] \) – Fixed processing power for scheduling and signalling, independent of the number of BBU.
- \( N_{bbu} \) – Number of BBU in the pool in pool p.

The model proposed offers the tools needed to estimate the baseband processing power that can be associated to each BBU instance in a BBU pool for downlink and for uplink. The model takes into account the physical layer processing and the communication protocols in the second layer of LTE. Note that one is assuming a splitting point in C-RAN where all the digital functions are centralised. The total power consumption is then given by the processing power required in the physical layer, the processing power required for data flows management and system control, the processing power for high-level protocols of layer 2 and processing power for backhauling to the core network.

\[
P_{bbu,n}[\text{GOPS}] = P_{\text{phy}}[\text{GOPS}] + P_{\text{control}}[\text{GOPS}] + P_{\text{hilp}}[\text{GOPS}] + P_{\text{bh}}[\text{GOPS}]
\]

where:
- \( P_{\text{phy}} \) – Processing power required for the physical layer functions.
- \( P_{\text{control}} \) – Processing power required for data flows management, scheduler and system management.
- \( P_{\text{hilp}} \) – Processing power required for high-level protocols in LTE processing stack.
- \( P_{\text{bh}} \) – Processing power required for the S1 interface depending on the S1 data rate.

The physical processing power depends on multiple digital processing functions such as filtering, pre-distortion, channel estimation, synchronisation, etc. The total power consumed is given by:

\[
P_{\text{phy}} = \sum_{t \in \mathbb{T}} p^t \cdot O_t
\]

The scaling factors are computed through the reference scenario as it follows:

\[
p^t = \prod_{x \in \mathbb{X}} \left( \frac{x_{\text{act}}}{x_{\text{ref}}} \right)^{e_{tx}}
\]

where:
- \( x_{\text{act}} \) – Real value of the parameter used in the BS.
- \( x_{\text{ref}} \) – Reference value of the parameter used in the BS.
- \( e_{tx} \) – Scaling factor of physical function t [11].

The scaling parameters are the following: Bandwidth, number of antennas, number of spatial streams, spectral efficiency, frequency domain load and number of quantisation bits. The frequency-domain duty cycling is the fractional load of the system RBs. This parameter is used to quantify the effect of the load in the overall processing power. It is also possible to adjust the system load with a time domain duty cycle factor to represent the fraction of time in which the BS is sleeping for power savings. This last factor is actually important in modern base stations with power saving features.

The control and network processing power are also described in the model in [12]. Both values are computed through the formula (6). The backhaul power model was not derived in this work.

Multiplexing gain (MG) is a metric of the gains obtained with the centralisation of different traffic profiles. It is the ratio between the sum of isolated RRHs peak traffic and combined peak traffic in pools. Based on [5], one uses the following expression to quantify this parameter:
\[ G_{\text{mux}} = \frac{\sum_{i=1}^{N_{\text{RRH}}} T_{\text{peak},i}[\text{Gbps}]}{\sum_{p=1}^{N_{\text{pools}}} T_{\text{peak},p}[\text{Gbps}]} \] (7)

where:
- \( G_{\text{mux}} \) – Multiplexing Gain of the network.
- \( N_{\text{RRH},p} \) – Number of RRHs connected to the \( p \)th pool.
- \( T_{\text{peak},i} \) – Peak traffic generated in the \( i \)th RRH.
- \( T_{\text{peak},p} \) – Peak traffic handled by the \( p \)th pool.

The same formula is used to compute the processing gain but using the peak traffic as the peak processing power (in GOPS).

Another gain can be defined that expresses not only the computational gain but also eventual capacity savings in terms of under-provisioning of cells. What the parameters measures is the centralisation (as processing gain) and also the computational resources that are being wasted in decentralised implementations and that can be discarded in C-RAN with the scalability offered by data centres. The name of this parameter is the total gain and can be computed by the following expression:

\[ G_{\text{total}} = \frac{N_{\text{RRH}} T_{\text{peak},i}[\text{GOPS}]}{\sum_{p=1}^{N_{\text{pools}}} T_{\text{peak},p}[\text{GOPS}]} \] (8)

where:
- \( T_{\text{peak},i} \) – Peak processing power required for an RRH under the most demanding radio conditions.

One of the benefits often associated with C-RAN is its cost saving potential. This savings can be either in CAPER or in OPEX.

As for the investment expenses, one needs to account for the hardware cost, the licences paid and the cost of civil work required. One can also include the fronthaul costs as an initial investment. In [12], a hardware analysis for centralised RAN is evaluated. The authors of the work propose 3 types of technological implementations to be compared with a traditional approach in which the BBU is placed in a site and usually serving three cells (three RRHs). The local approach and the complete centralised one with share of resources are adapted in this work. Although this hardware model offers a complete dimensioning of the equipment, most of the factors needed are hard to estimate for future technology. Besides that, the equipment currently available is usually sold as a whole, so the price of individual components is not known. With that being said, one uses the following approximations for the investment model:

\[ C_{\text{CAPEX, local}}[\€] = N_{\text{RRH}} (C_{\text{proc}}[\€] + C_{\text{cab/cell}}[\€]) + C_{\text{id/cell}}[\€] + C_{\text{CW/cell}}[\€] + C_{\text{lic/cell}}[\€] \] (9)

\[ C_{\text{CAPEX, CRAN}}[\€] = \sum_{p=1}^{N_{\text{pools}}} N_{\text{RRH},p} \left( V_{\text{base}}[\€] \right) \sigma + C_{\text{id/cell}}[\€] + C_{\text{CW/cell}}[\€] + C_{\text{lic/cell}}[\€] + C_{\text{FHT/cell}}[\€] \] (10)

where:
- \( C_{\text{base/cell}} \) – Cost of baseline components in C-RAN normalised per cell (includes switches and controllers).
- \( \nu \) – Added cost of C-RAN auxiliary technology.
- \( \sigma \) – GPP cost factor.
- \( V_{\text{p}} \) – Virtualisation factor.

Regarding OPEX analysis, the following expression accounts for all the components that need to be considered:

\[ C_{\text{OPEX}}[\€] = C_{\text{rent}}[\€] + C_{\text{p}}[\€] + C_{\text{mte}}[\€] \] (11)

where:
- \( C_{\text{rent}} \) – Rents expenses for equipment housing (per year).
- \( C_{\text{p}} \) – Power consumption expenses in equipment and in air conditioning (per year).
- \( C_{\text{mte}} \) – Maintenance expenses for equipment and transmission network (per year).

For the renting costs, one assumes that the cost variations are due to the different total area in both implementations and different prices per square metre. The power consumption considered in this model is only the one required for digital processing of each cell (the power for the RRHs is not considered). Naturally, it also depends on the multiplexing gain. However, one also accounts for extra energy consumption due to less efficient equipment (generic purpose). One assumes that the maintenance corresponds to a fraction of the initial investment with different percentages to hardware and civil work components.

B. Model Implementation

In order to solve the problem under study, one has structured a simulator with three main layers. The first component of implementation is the proliferation and it is developed to add a temporal dimension to the simulator. Being the technology under discussion a futuristic concept, one uses the proliferation algorithm to study how the architecture can scale and adapt to future network demands. Taking as input the current RRHs position and the number of years to proliferate, this layer adds RRHs coordinates based on an adjustable proliferation factor. The second component of the program is the technical layer. This layer is developed to make the RRH-BBU pool assignments based on different metrics. One purposes different strategies to optimise the results in multiple metrics. The third and final component of the model is the cost layer. In this phase, one develops a model to compare the costs in a traditional implementation and in a centralised implementation. The comparison results in possible savings obtained with the adoption of C-RAN technology instead of local RAN (green field deployments) This layer also offers different factors to account for possible computational implementations of the data centres. A generic flowchart of the algorithm used in the
simulator is presented in Figure 2. In the next subsections, the RRH proliferation module and the assignment strategies developed are presented.

![Figure 2 - Model Implementation Algorithm.](image)

1) **RRH Proliferation**

The main goal of this algorithm is to take as input the existing BS locations, traffic profiles and the simulation time to introduce new RRHs with different positions and traffic profiles in the scenario under study. The newly deployed RRHs are low-powered nodes (small cells) to increase the overall capacity of the network. The annual proliferation factor is used to control the number of deployed cells.

The algorithm takes into consideration the different densities of RRHs (base stations per area) to assign more RRHs in the denser zones of the networks. This assignment strategy is chosen because RRH density is usually associated with more populated zones in which the traffic is expected to be higher. The algorithm is adapted from [7] and works with the Probability Distribution Function of the distances to a reference point in the centre of the scenario. Based on this function one can extract an element with the probability conditioned by the distance to the reference point. After selecting one RRH, the algorithm chooses the two nearest neighbours and deploys a new RRH in the centroid of these three points.

The newly instantiated RRH is then added to the list of deployed RRHs after the traffic behaviour of that cell has been estimated based on one of four different metrics that reflect four different implementations:

- **Traffic average**.
- **Weighted traffic average** (only average of cells with the same type of profile).
- **Specific traffic profiles based on neighbouring RRHs** (assign default profiles based on neighbourhood).
- **Specific traffic profiles assigned randomly**.

2) **Minimise Delay**

The algorithm takes as input the locations of the deployed RRHs and possible locations of BBU pools. To define the list of possible pools to which each RRH can be connected to, a maximum distance is configured based on the maximum delay tolerated by the architecture. If any of the cells is too far of possible pools, it is marked as unserved. One also has the possibility to limit the capacity of each pool. In the latter case, the algorithm will perform better if it is divided into two iterations and if one sorts the RRHs by the distance to any pool in order to give preference to the cells that are closer to any data centre. This is valuable mostly if one is also aiming at clustering the neighbouring cells together for CoMP gains. With this approach, one can expect the lowest values of latency.

3) **Load Balancing**

The algorithm takes the same inputs as the minimise delay with the addition of traffic information of each RRH. One can select to balance the number of RRHs, the traffic capacity or the processing capacity. In the two last cases, one should select the time instance in which it intends to balance the load but still take into consideration the capacity limits in each hour of the day.

The algorithm starts by defining a set of possibilities of connections for each RRH based on latency and capacity constraints. If one RRH just has one possibility of connection, the algorithm assigns the RRH-BBU link and marks it as served with centralisation. All the RRHs that have multiple possibilities are studied in the second iteration. In this phase of the algorithm, one assigns each of the RRHs to the pool that has less load allocated to it. In the virtualisation case, one can also sort the RRHs that need to be assigned based on their capacity in order to prioritise the ones with higher demands that serve more loaded cells. With this approach, a more balanced distribution of load is expected forming a more fault resilient network and lower values of capacity required in a single pool.

4) **Maximise Multiplexing Gain**

The algorithm works with the same inputs as the previous ones. The idea behind this algorithm is to use all the possible pools in the scenario to serve the RRHs while ensuring that the overall traffic curve of each one of them is as flat as possible. With this approach, one attempts to correct unbalanced traffic profiles of the pools with the RRHs that have a complementary behaviour in time. As a result, the network centralised processing power will be lower than in a decentralised network and lower investment and operational costs are achievable. The algorithm starts by assigning the RRHs that only have one pool to connect to. A list of RRHs possibilities to each pool is created and updated after each assignment. Then a set of iterations is done in the workflow. In each iteration, one should work with the MG of each set of BBUs. The idea is to assign the RRH that improves the MG to the pool which has a lower gain. If at any instance there are none RRHs improving any of the MGs, one should assign the RRH that has less impact on a pool and starts a new iteration.

5) **Minimise Number of Pools**

With this assignment strategy, the idea is to minimise the number of required pools in the scenario. The goal is to maximise the centralisation with the assumption that there is no infinite capacity in the pools. The problem is solved based on a heuristic of a known mathematical problem named as vector bin packing with capacity awareness. The mathematical solution described in [13] for parallel processing is adapted to combine the maximum number of RRHs according to their load requirements in multiple time instances. One uses the sorted Permutation Pack solution proposed by the authors. The algorithm firstly works by assigning the RRHs that only have one possibility. Then an attempt is made to assign more RRHs to the pools that are already in use. In order to respect the time instances selected, one sorts the load requirements in the pool by hour and searches for an RRH with an inverse sorted list. If the
inverse order is not found in the possible RRHs, one relaxes the matching criteria by searching the same order with one less instance. The latter procedure is repeated until one RRH fulfilling the criteria is found and assigned.

IV. RESULTS ANALYSIS

A. Scenario

The geographical region used is the city of Porto, particularly the most populated area. In the reference scenario, one only considers the dense urban zone of Porto spanning from a radius of 20 km from the city centre, in Boavista. The scenario includes 614 RRHs and 19 possible pools of a Portuguese mobile operator.

Regarding the RRHs configuration selected, one assumes that all the radio equipment is the same to reflect a scenario that is entirely urban. A bandwidth of 20 MHz is used with a MIMO configuration of 2x2. No capacity fronthaul constraints are assumed in this thesis. A value of 2x108 m/s is used for the propagation delay in the fibre. The maximum distance considered is 15 km corresponding to a latency requirement of 150 µs (round trip time). One also assumes that technological implementations with resources sharing are available, although the virtualisation and pooling in C-RAN are still an area under study. One uses the traffic measures in GBph and GOPS in order to do load balancing and to analyse the multiplexing gain. In the load balancing case, the reference hour is 10 p.m. due to the peak traffic of the network. Downlink LTE traffic is used as this direction produces more traffic than uplink, as expected. In reference simulations as well, the capacity used for the pools is infinite, as more processing cards can be added to the pools if needed. For costs, one has used real values for hardware and licences. The energy cost was taken as the average value paid in Portugal during busy hours. The price of renting is assumed based on the values usually paid for square metre in the city of Porto. One uses a smaller value for central pools location because their positioning is more flexible and can be made according to price considerations. The maintenance factors are derived due to the depreciation of computers (or boards), for the eventuality of civil repairs and regular hardware depreciation (interfaces, switches and power units). The virtualisation factor used as a reference is the inverse of the multiplexing gain measured for traffic curves in GBph. The added complexity and GPP factor are used as one due to technological uncertainties.

The scenario under studied contains cells classified based on their traffic profile. The classification is done automatically by comparing the traffic periods 9h-16h. and 17h-24h. If the difference is not bigger than 15%, the cells are classified as mixed, otherwise the dominance of the time periods defines the classification type.

B. Reference Scenario

The first parameter studied is the latency. Although the reference scenario has a dimension comparable to the maximum link distance (20 and 15 km, respectively) and the 19 pools are spread along the area, one should measure the delay values and use them as a reference for future technologies. Figure 3 presents the distance results obtained. All the RRHs can be centralised under the delay restriction. As one can see, all of the RRHs have a link with distance below 15 km. In terms of delay, these values are easily convertible with equation (2). For minimise delay, 23.11 µs is the average two-way delay of fronthaul connection.

As it is noticeable, the Minimise Delay algorithm has shorter fronthaul distances than the remaining algorithms, as expected. This is explained by the load balancing and multiplexing gain strategies making use of the maximum allowed fronthaul to achieve different metrics other than delay.

Next parameter under study is the average capacity based on the average values of traffic per hour. Figure 4 shows the results of the three algorithms developed.

The Load Balancing algorithm is run at 10 p.m. Naturally, this one is aiming at a more balanced distribution of capacity as the time sample under study is one of the most loaded during the day. With the Maximise MG algorithm, one actually achieves a balancing that is close to the one obtained with Load Balancing. In fact, with Maximising MG algorithm one is actually using a balanced approach to obtain higher gains by assigning cells over all the possibilities. As the values of traffic profiles under study are already a result of an average, one should notice that this actual capacity of the network should be higher to handle traffic fluctuations during the days. The average capacity is about 15.2 GBph.

Considering the load analysis based on GOPS, one also gets the results of computational processing power. The difference between the peak processing power in a day and the lowest value of processing power is not as significant as the traffic variation. This analysis supports the low values obtained for MG. The effectiveness of the balancing approach is also noticeable as all the pools process around 1.9 Tera Operations Per Second (TOPS).
One has also compute the three types of gain identified. The results are presented in Table 1 only for the Maximise MG algorithm. The MG value of the maximise algorithm is 3% higher than the one obtained with load balancing. Besides proving that the algorithm achieves higher gains, the latest fact also supports that both produce comparable results in what concerns balancing and gain.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.308</td>
<td>1.033</td>
<td>1.278</td>
</tr>
</tbody>
</table>

One also presents the results obtained with the GOPS scenario applied in order to present a very relevant result for current applications of LTE technologies as the MG application affects mostly the computational processing power. The first thing to notice is that the MG measured in processing capacity is quite low compared to the traffic one. When considering an approach with under-provisioning, one can see much higher gains. Although this value is close to the multiplexing gain with traffic, one should note that only a small portion is due to centralisation of resources. The under-provisioning is not specific of C-RAN but it is certainly facilitated due to the efficient scalability of data centres. The mobile operator can then prepare their processing power with a certain margin and upgrade their data centre when required.

Finally, the analysis of TCO is presented in Figure 5 using Maximise MG algorithm. As it expected, there is a considerable cost reduction in the CAPEX and in the OPEX. With centralisation, one can see that there is about 63% of savings in CAPEX and 31% of savings in OPEX.

![Figure 5 - CAPEX and OPEX comparison (in 10 years).](image)

The values obtained are proving that there are very considerable advantages of C-RAN technology when compared to another green-field investment in RRH-BBU local technology. The main CAPEX reduction is not from the multiplexing gain but from the civil work involved (around 88% of total reduction). Naturally, the savings in interfaces, boards and civil work can justify a larger investment in the auxiliary technology and licences that increases 50%. However, the savings in hardware only are not enough pronounced to solely justify a C-RAN investment. Regarding the OPEX, renting is the main source of savings in operational expenditures with a reduction of 60% of the decentralised cost, although some important reduction is seen in both energy and maintenance (about 20% of savings in each one of them). This happens as expected, as there are less and cheaper areas to rent when considering the locations for data centres. The energy consumption has savings due to the multiplexing gain and one can see a large influence on the value obtained in this component. As the hardware does not have such a variation in C-RAN, the maintenance expenditures savings are not so pronounced. Still, one can see a reduction that reflects the easier maintenance of a few pool instead of hundreds of complete sites.

C. Latency Analysis

The values considered are all under the 15 km of reference, as this restriction is always expected to be more stringent with the adoption of new technologies and with the use of extra equipment introducing additional latency (switches, GPPs, etc.). As it can be seen in Figure 6, the share of traffic load being handled in a pool is not always 100% as some of the RRHs do not have a possible pool within the maximum fronthaul distance. With that being said, one actually is studying the effect of partial centralisation in the scenario under different delay constraints.

![Figure 6 - Share of traffic load being treated in a pool with maximum fronthaul distance.](image)

The partial centralisation affects the multiplexing gain. It starts increasing as more RRHs are served in a centralised pool. However, the overall gain stops increasing for fronthaul limits larger than 7 km. Above that distance, one can even note that not all of the pools are centralised (around 2%). The reason for this result is that the cells to serve are the peripheral ones and are mostly characterised by mixed traffic that do not improve MG.

To conclude the analysis on the centralisation achieved with different delay requirements, one should comment the cost evolution for the different distance values. Naturally, full centralisation implies higher cost savings. However, as one has seen for MG, for latency over 7 km one gets almost the same values for OPEX and CAPEX.

D. Traffic Analysis

Another approach that should be studied in this chapter is the influence of the traffic profiles. One uses reference values from other RATs such as 2G and 3G. The idea behind this problem is to consider all the traffic being handled by the network in a futuristick approach of combining all the technologies in centralised pools. Figure 7 represents the average load per pool obtained with all the traffic profiles. The algorithm used was the maximise MG, as one can see a balanced distribution of load throughout the day.

The traffic curves produce different multiplexing gains. Actually, the legacy technologies may achieve higher gains (around 1.4 in 3G and 1.6 in 2G) than LTE. However, as a final result, the immediate conclusion is that the observed gains do
not improve as one combines all the curves together (the MG decreases to 1.23). The reason for this result it that one is creating more mixed cells when analysing more traffic together (the share of mixed cells increased almost 12 %) while there are fewer cells being classified as office and residential ones. Finally, one can also state that the average traffic capacity that would be required in order to deal with the maximum traffic demand is almost twice of the reference scenario.

Figure 7 - Average load per pool with different traffic profiles.

Besides, one has also changed the LTE DL traffic input to the UL LTE traffic. Comparing with the previous reference results in downlink, one can see that values achieved with all the algorithms under consideration are significantly higher. For maximise MG, one gets a MG of 1.542. The latter sentence proves that the uplink profiles are more heterogeneous and higher gain is achieved. However, one should also note that the traffic being considered in this approach is lower than the downlink one. To proceed with the same comparison, one also studies the effects of the gain in computational power. The processing gains with and without provisioning are 1.23 and 1.76, respectively. The first conclusion is that the processing gains are higher. Besides the more specific traffic curves explained before, the uplink processing power is more load dependent and the variation will have more impact. Finally, the results of under-provisioning prove that much of the RRHs are currently processing significantly under the maximum processing capacity. The last conclusions support the initial approach of considering the multiplexing gain of the downlink as the one used in cost model, as the total traffic is quite above the opposite direction. However, uplink traffic is increasing each year due to multiple services such as streaming or social networks and the results for uplink should be taken into account. If one analyses the processing capacity, one can state that uplink processing power is higher than downlink processing power. This conclusion comes from the fact that uplink processing functions are more complex due to the radio characteristics and require more operation per second in each BBU. The relevance of this result implies that uplink should be considered in processing power gain analysis.

E. Number of Pools Analysis

Another approach that a mobile operator can elect to implement in its network is the scaling of pools during the day. The main goal of this strategy is to set a capacity limit for each pool as a way of balancing the load or as a consequence of physical limitations such as equipment capacity, physical area for hardware accommodation, energy supply or related problems. The results shown in Figure 8 were obtained with

Minimise the Number of Pools algorithm with three time instances used separately.

Figure 8 - Number of pools required with maximum capacity per pool.

One can see that the number of pools required during the night is quite under the number of pools required during busiest instances of traffic. In fact, 5 pools are enough to process the 6 a.m. traffic in the reference scenario. When the capacity limits are higher, the number of pools tends to be minimal. One can see for the 6 a.m. simulation that there is a limit of 2 pools in the scenario. This limitation is actually imposed by distance, as any of the possible pools is able to concentrate all the RRHs with a fronthaul constraint of 15 km. Finally, note that in the largest values of capacity there is not a relevant difference of pools as only three are enough to process the higher traffic demands. Naturally, this situation represents a waste of resources during the night as a major part of the capacity is unused. Further simulations have shown that a capacity of 245 GBph is required to enforce the minimal number of pools (2) at any time instance. Naturally, the algorithm also changes the results of the TCO. If all the RRHs are not served, the CAPEX and OPEX increase as in the latency analysis. The MG discussed previously is obtained with 6 pools under 55 GBph limit. Limitation higher than 55 GBph only have influence in OPEX, as less pools used means less renting costs. Savings of 48% in OPEX are possible with 125 GBph of limitation considering that no extra investments would be required.

F. Proliferation Analysis

One presents two scenarios in 5 years that reflect an operator forecast of traffic growth in a pessimistic and in an optimistic approach with proliferation factors of 54% and 75%, respectively. One also offers an intermediate scenario that works as a more realistic result with an increasing proliferation factor per year. Besides the proliferation, one assumes a significant growth of traffic in the already deployed macro BS considering they are still able to serve higher demands with other technologies such as CA. This growth is assumed as year 0 for simplicity of the model (growth factor) and has values of 0%, 25% and 75%. In Figure 9, the evolution of the network load in the peak hour is shown for the three cases considered. The proliferation was performed by simply averaging the traffic of the neighbouring RRHs of the new small cell. One can see the difference between the three in the two final years. When comparing the overall maximum in five years, one can see an increase compared to the reference case studied in previous sections of about 7.9 for the optimistic case, 4.8 for the realistic and 3 for the pessimistic.
The next analysis is important to understand the importance of the centralisation of small cells in the multiplexing gain. Although one has been using an average of neighbouring traffic for the newly deployed RRHs, one has proposed more deployment models of RRHs that represent other possible realistic scenarios. Table 2 shows the results obtained for multiplexing gain under the different traffic proliferation methods for the optimistic scenario. There is no significant difference between the optimistic and pessimistic scenario which implies that the proliferation time and factors have small influence on the value.

<table>
<thead>
<tr>
<th>Proliferation Method</th>
<th>Traffic Avg.</th>
<th>Pounded Traffic Avg.</th>
<th>Specific Traffic Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG</td>
<td>1.248</td>
<td>1.306</td>
<td>1.402</td>
</tr>
</tbody>
</table>

Other conclusion is that the deployment of new centralised RRHs will be beneficial only if they are installed in specific environments such as shopping centres, company’s facilities or spaces characterised by nocturnal activities. Note that the main difference of the specific traffic profiles (higher gain) is that one is using very pronounced traffic curves and not averaging the already existent ones. In the deployments of pounded traffic average, one avoids the deployment of mixed profiles but it is still averaging the office and residential cells, resulting in an almost constant gain. Note that with traffic average one gets more mixed cells as the combination of residential ones with office results in a mixed cell. The increase in the multiplexing gain is also translated to an increase in gain in processing power but with less relevance. Regarding the gains with under-provisioning, one gets for the same scenario more loaded cells with less spare capacity (growth factor applied to year 0), causing an overall gain close to the centralisation gain. For a mobile operator, this actually means that the gain of under provisioning is easily lost when the traffic growths and the cells start getting their maximum loads. However, this last value also depends on radio conditions of cells (spectral efficiency of the power model) and an operator can still have some profits in cells with very favourable conditions such as small cells.

A final comment on the multiplexing gain is the combination of office and residential cells leading to higher gains. Although the ratios of deployment are not easily controllable by an operator, it still gives an analysis of what kind of cell types should be installed in an area in which offloading of traffic is needed. For instance, if a macro cell is overloaded one can decide if the offloading of traffic should be done in a residential hotspot or in an office centre. Figure 10 shows the results with traffic proliferation strategy being the specific traffic profile assignment based on probabilities.

![Figure 10 - Multiplexing Gain variation with share of deployed RRHs.](image)

Note that one assigns a probability of zero for the deployment of mixed cells, as one already proved that they are not good for the gain and they should not be considered to maximise MG. The results were obtained in the second year of proliferation in the optimistic case under maximise MG algorithm. As one can see, if one deploys 60% of residential cells in the scenario (40% of office ones), the maximum gain in achieved (more than 1.41). Note that this value corresponds to a ratio of 44.4% of global residential cells and 43.8% of office ones, as there were already some deployed RRHs. The conclusion is that mobile operator should try to keep a balanced deployment of both types of cells. The results come as a consequence of the office cell traffic and residential cell traffic having comparable peaks (on average).

Regarding TCO, centralised deployment of small cells is still not implemented. As SC equipment is already small, there is no expectable difference in renting and civil work. The extra savings achievable are only due to the higher multiplexing gains that can be achieved. Naturally, this result needs to be compared with the extra investment required to implement stringent fronthaul connections than backhaul to link the cells to the network.

V. CONCLUSIONS

The main goal of this work was to analyse the multiple performance parameters of the deployment of a C-RAN
architecture in an already existent LTE-A network from a Portuguese mobile operator. One has developed a model and a computational tool to receive the LTE network as input and to define the C-RAN configuration and export results regarding capacity in the pools (traffic per hour or processing power), latency, multiplexing gain and costs of investment and operational expenditures to characterise cost savings.

The reference scenario was the city of Porto. Simulations have confirmed that a fronthaul limit of 15 km is enough to centralise all the cells. In fact, using the minimise delay algorithm one can get even more satisfying values of latency for future technologies. The load balancing algorithm and maximising MG algorithm are effective but do not show significant differences in the results. An average capacity of 15.2 GBpH is obtained for the processing centres. An average of 1.9 TOPS is required for processing power. The MG obtained was 1.31 and slightly under the registered values in literature. However, mixed cells are considered in this realistic scenario. The value does not match with the processing gain when considering GOPS, as most of the functional processing are not load-dependent. The value of processing gain alone is surely not enough to justify the consideration in current technology but should also be added to the possibility of under-provisioning of cells due to IT scalable solutions. The cost savings achieved correspond to a 63% reduction in CAPEX and 31% in OPEX. The main cause of the reduction is the civil work investment and not the hardware. However, MG still has importance in the OPEX and causes 1% savings in HW with the assumptions taken.

Latency analysis has proven that stringent values of fronthaul distance can cause partial centralisation with noticeable decrease of MG and cost savings. However, a distance limit of 7 km is enough for the scenario under study. About capacity analysis, one has concluded that the 19 pools are not required depending on the capacity limits imposed by the operator. In the limit, 2 pools are required due to distance restriction. There is also a considerable difference on the number of pools required for 10 p.m. and 3 p.m. when compared to 6 a.m. demand which might justify the adoption of smart switching schemes to increase the resource utilisation efficiency. One has also studied the effect that the deployment of heterogeneous networks to handle new traffic demands would have with C-RAN. An increase of about 8-times in the traffic per hour is estimated for an optimistic scenario and corresponds to about 3 thousands of hotspots in the city centre. The processing power does not scale with the same number has small cells require a little less computational complexity (factor of 4). Overall MG obtained can be improved if the deployment of small cells is done in specific and known type of spots with a controlled ratio to match the share of office and residential cells. The savings introduced with the improved MG should be compared with the fronthaul (instead of backhaul) investment of centralised HetNets. Also for future work, other theoretical advantage in C-RAN related with cooperative transmission and signal processing should be characterised and quantified. In fact, even if the cost savings are considered to be not acceptable for new architecture deployment, the increase in data rates and coverage may support the introduction of C-RAN, as legacy technology has latency problems associated with this kind of implementations.

REFERENCES


