Cost Optimization of Cloud-RAN Planning and Provisioning for 5G Networks
Osama Arouk, Thierry Turletti, Navid Nikaein, Katia Obraczka

To cite this version:
Osama Arouk, Thierry Turletti, Navid Nikaein, Katia Obraczka. Cost Optimization of Cloud-RAN Planning and Provisioning for 5G Networks. ICC 2018 - IEEE International Conference on Communications, May 2018, Kansa City, MO, United States. hal-01707682

HAL Id: hal-01707682
https://inria.hal.science/hal-01707682
Submitted on 13 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Cost Optimization of Cloud-RAN Planning and Provisioning for 5G Networks

Osama Arouk†, Thierry Turletti‡ and Navid Nikaein† and Katia Obraczka‡
†INRIA, Université Côte d’Azur - 2004 Route des Lucioles
06902 Sophia Antipolis Cedex, France - Email: firstname.lastname@inria.fr
‡Communications Systems Department - EURECOM, Biot, France - Email: navid.nikaein@eurecom.fr
Department of Computer Engineering, UC Santa Cruz - Email: katia@soe.ucsc.edu

Abstract—In this paper, we propose a network planning and provisioning framework that optimizes deployment cost in C-RAN based 5G networks. Our framework is based on a Mixed Integer Quadratically Constrained Programming (MIQCP) model which optimizes “virtualized” 5G service chain deployment cost while performing adequate provisioning to address user demand and performance requirements. We use two realistic scenarios to showcase that our framework can be applied to different types of deployments and discuss the computational cost and scalability of our solution.

I. INTRODUCTION

Approximately every 10 years, a new generation wireless communication system is deployed in order to satisfy ever growing demand from users and applications: starting with first generation, or 1G systems around 1982, then 2G around 1992, 3G in the early 2000’s, and 4G around 2012. The next generation mobile communication network, or 5G, is scheduled to become commercially available in the early 2020’s and promises to support, among other things, significantly higher end-user data rates, considerably lower latency, and massive number of connected devices.

In order to meet these goals and still keep CAPEX/OPEX financially viable, 5G providers will rely heavily on virtualization of network functions by adopting a “cloudified” radio access network architecture, or C-RAN. According to the latest 3GPP technical reports [1], [2], next-generation RAN will be disaggregated into three main units: the Remote Radio Unit (RRU), the Distributed Unit (DU), and the Centralized Unit (CU). The RRU contains all the necessary components related to signal transmission/reception [3]. The DU may perform a set of physical layer (PHY) functions that could be shifted to the cloud, as well as some higher layer functions, while the rest of higher layer functions is aggregated in the CU.

Satisfying user demand, while maintaining adequate levels of resource utilization and thus minimizing the cost will require 5G providers to dedicate considerable effort and attention to adequately plan their deployments. This network planning phase includes: (1) deciding how many RRUs are needed, if new ones need to be deployed, and if so, in which location, (2) deciding which data centers (DCs) will be used to host DUs and CUs, and whether new DCs need to be brought online, and (3) deciding how to connect RRUs, DUs, and CUs, which may use existing communication links or require new ones.

We focus on the important and timely problem of optimizing cost of infrastructure deployment in C-RAN based 5G networks. Most efforts to-date have focused on minimizing network cost (e.g., cost of running a deployed infrastructure) by sharing the available resources (e.g., base stations’ resources) among multiple operators [4], [5], [6], [7], while only a few have tackled the problem of infrastructure’s deployment cost. In [8], an Integer Linear Programming (ILP) model was introduced for Passive Optical Networks (PONs) when fibers are sparsely deployed. The work reported in [9] and its variation [10] propose an ILP model to minimize the deployment cost of cell sites and links to the selected Access Points (APs) in the case of sparsely deployed fiber. An ILP model for joint cost optimization of the fronthaul and the Base Band Units (BBUs) was introduced in [11]. While these existing models focus on horizontal scaling for certain parts of the network, i.e., they assume partial presence of infrastructure, our model can also be applied to scenarios where the infrastructure does not exist.

In this paper, we propose a network planning and provisioning model that optimizes deployment cost in C-RAN based 5G networks. To the best of our knowledge, our work is the first to propose a Mixed Integer Quadratically Constrained Programming (MIQCP) model that optimizes “virtualized” 5G deployment cost while performing adequate provisioning to address user demand and performance requirements. We showcase the generality of the proposed MIQCP model by employing it in two realistic deployment scenarios, namely: (1) a region with no existing networking infrastructure, and (2) a region that has partial network infrastructure coverage.

The rest of the paper is organized as follows. The C-RAN based 5G network deployment cost optimization problem is described in Section II along with our assumptions and network model. Section III models the problem using an Integer Linear Programming (ILP) formulation and derives the proposed MIQCP model. The performance of our MIQCP model and its computational cost and scalability are evaluated in Section IV and Section V, respectively. Section VI concludes the paper with some directions for future work.

†In the literature, this problem is also referred as horizontal scaling [12].
We consider the three-tier C-RAN architecture as envisioned by 3GPP [1] and illustrated in Fig. 1. Our problem can be stated as follows: given a specific geographic region, also known as Region of Interest, our goal is to minimize C-RAN deployment cost while still adequately provisioning resources to meet user demand. In particular, we need to consider C-RAN functional splits that satisfy bandwidth, latency, and processing requirements at RRU s, DUs, and CUs. In our cost minimization model, we assume that PHY functions are split between the RRU and DU, while higher network layer functionality is performed at the DU or CU. We also assume that resources such as CPUs and links have finite capacity, which make our model more realistic. We focus on two types of scenarios, namely: (1) a region with no existing infrastructure, and (2) a region with partially deployed infrastructure. Each one of these scenarios is described in more detail below.

1) The scenario with no existing infrastructure, dubbed as Abundance of Infrastructure, has the following features:
   - There are no RRU s installed.
   - The number and locations of the RRU s needed to satisfy user demand are known. Note that the cost of the cell site is not included in our model because it represents a fixed and mandatory cost.
   - Candidate locations of DCs for hosting the DU functions (i.e., DCDUs) and the CU functions (i.e., DCCUs) and their associated installation costs are known. Our model will select which locations to pick for hosting DCDUs and DCCUs.
   - No communication links have been installed in the Region of Interest and the connection between a RRU and a DCDU is either direct or will pass through another DCDU if decided by our cost minimization model.

   - We assume that it is possible to install a direct link between any DCDU and any DCCU. Connecting a DCDU and a DCCU is needed when DUs and CUs are hosted in the DCDUs and DCCUs, respectively, which are collaborating on functional splits.

2) The scenario with partial infrastructure features, named as Partial presence of Infrastructure:
   - Two types of RRU s: (1) Existing RRU s already connected to DCDUs, and (2) New RRU s that have to be installed at some known locations similarly to the first scenario.
   - Two types of DCDUs/DCCUs: (1) Existing DC- DUs/DCCUs possibly interconnected, and (2) New DCDUs/DCCUs that could be constructed if needed at some candidate locations.

III. PROBLEM FORMULATION

In this section, we formulate our network provisioning cost minimization problem as Mixed Integer Quadratically Constrained Programming (MIQCP) model [13], [14]. To the best of our knowledge, this is the first time this network planning cost optimization has been modeled using MIQCP. Besides yielding optimal cost, the proposed model can be extended to different deployment scenarios.

To showcase the generality of our model, we first consider scenarios with no existing network infrastructure and then discuss how to extend the model to scenarios with partially deployed infrastructure. The notation used in our derivation is summarized in Tables I and II.

Given a set of \( R \) RRU s, a set of \( D \) and \( C \) candidate data centers (DCs) to host the DUs and CUs, respectively, the overall cost of the network can be expressed as follows:

\[
\min \sum_{i=1}^{R} \sum_{j=1}^{D} \lambda_{i,j}^d + \sum_{i=1}^{D} \sum_{j=1}^{D} (\Gamma_{i,j}^d + \Psi_{i,j}^d) \\
\times e_{i,j}^d + \sum_{i=1}^{D} (\Delta_{i} + z_{i}) x_{i} + \sum_{i=1}^{D} \sum_{j=1}^{C} (\Gamma_{i,j}^c + \Psi_{i,j}^c) x_{i} \\
+ \sum_{i=1}^{C} (\Delta_{i} + z_{i}) x_{i}
\]  

(1)

The first term of the expression in (1) represents the cost of the links between the cell sites where the RRU s will be located and the DCs that will host the DUs functions (i.e., DCDUs). The second term of this expression is related to the cost of the links among DCDUs, while the fourth term represents the cost of the links between DCDUs and DCCUs. The cost of DCDUs and DCCUs is specified by the third and fifth terms of (1). In order to meet the requirements of the functions hosted in DUs and CUs, the following conditions should be satisfied for \( i \in \{1, ..., R\}, j \in \{1, ..., D\}, k \in \{1, ..., C\} \):

1) RRU-DCDU links
   - A RRU can be connected to only one DCDU:

\[
\sum_{j=1}^{D} v_{i,j}^d = 1, \quad x_{i,j} \geq v_{i,j}^d \quad \text{and} \quad x_{i,j} \geq \omega_{i,j}
\]  

\[
\text{for } i \in \{1, ..., R\}, j \in \{1, ..., D\}, k \in \{1, ..., C\}
\]
2) DCDU-DCDU links

- Links exist only between selected DCDUs:
  \[ x_d(j_1) \geq v_d^d(j_1, j_2) \quad \text{and} \quad x_d(j_2) \geq v_d^d(j_1, j_2); \quad j_1 \neq j_2 \]  
  \]  
  (2)

- A link between two DCDUs should exist when placing a DU function on a DCDU not connected directly to its related RRU:
  \[ v_d^d(j_1, j_2) \geq v_d^d(i, j_1)\omega_d(i, j_2) \]  
  (3)

- The capacity of selected links should not exceed a certain predefined threshold, \( \phi_d^d(j_1, j_2) \leq \Omega_d \), where:
  \[ \phi_d^d(j_1, j_2) = \sum_{i=1}^{R} v_d^d(i, j_1)\omega_d(i, j_2)\beta_d^cu(i) \]

Note that the variable \( v_d^d(j_1, j_2) \) can be omitted from the equation above as it is guaranteed by other conditions like the inequalities 2 and 3. It is also worth noting that the maximum allowed capacity for each link can also be modified by changing the value \( \Omega_d \). After determining the capacity of the links, their cost can then be calculated by:

\[ \Psi_d^d(j_1, j_2) = \phi_d^d(j_1, j_2)\alpha_d^d \]

- The delay of the links should respect the latency requirements of the functions to be deployed.

\[ \sum_{j_1=1}^{D} v_d^d(i, j_1)(t_d^d(i, j_1) + \sum_{j_2=1}^{D} t_d^d(j_1, j_2)v_d^d(j_1, j_2) \times \omega_d(i, j_2)) + \rho_d^c(i) \leq \rho_d^c(i) \]

It is also possible here to omit the variable \( v_d^d(j_1, j_2) \) as it is guaranteed by inequalities 2 and 3.

3) DCDUs

- The capacity of the selected DCDUs should not exceed a given predefined threshold:
  \[ \sum_{i=1}^{R} \rho_d^c(i)\omega_d(i, j) \leq \delta_d^c(j) \]

The cost of the DCs can be calculated by:

\[ z_d(j) = \gamma_{dcdu} \sum_{i=1}^{R} \rho_d^c(i)\omega_d(i, j) \]

4) DCDU-DCCU links

- Link delay should satisfy the latency requirements of the functions hosted at the DCs. We consider the latency requirement of a DU function (when placed at a DC) to be equal to the maximum latency the function can tolerate.

\[ \omega_d(i, j)\omega_d(i, k)(\rho_d^c(i) - \rho_d^c(i)) \geq t_d^d(j, k) + \rho_d^c(i) \]

- The capacity of the selected links is limited by

\[ \sum_{i=1}^{R} \omega_d(i, j)\omega_d(i, k)\beta_d^cu(i) \leq \Omega_d \]

The cost of the links can be determined by:

\[ \Psi_d^d(j, k) = \alpha_d^d \sum_{i=1}^{R} \omega_d(i, j)\omega_d(i, k)\beta_d^cu(i) \]

5) DCCEUs

- The capacity of selected DCCEUs is limited by a given threshold:

\[ \sum_{i=1}^{R} \rho_c^c(i)\omega_c(i, k) \leq \delta_c^c(k) \]

The cost of DCCEUs can be determined by \( z_c(k) = \gamma_{dcce} \sum_{i=1}^{R} \rho_c^c(i)\omega_c(i, k) \).

### TABLE I: List of notations related to RRU and DCDU

<table>
<thead>
<tr>
<th>Definition</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters related RRU-DCDU links</td>
<td></td>
</tr>
<tr>
<td>Unit function</td>
<td>( \Psi_d(i, j) )</td>
</tr>
<tr>
<td>Link delay between RRU and DCDU</td>
<td>( t_d^d(i, j) )</td>
</tr>
<tr>
<td>Cost of the link</td>
<td>( \Omega_d )</td>
</tr>
<tr>
<td>Parameters related DCDU-DCDU links</td>
<td></td>
</tr>
<tr>
<td>Unit function</td>
<td>( \Psi_d^d(j_1, j_2) )</td>
</tr>
<tr>
<td>Constant cost that needs to be paid when deploying the link</td>
<td>( \xi_d^d(i, j) )</td>
</tr>
<tr>
<td>Variable cost that needs to be paid based on the required capacity</td>
<td>( \rho_d^c(i) )</td>
</tr>
<tr>
<td>Cost per unit of resource for the links DCDU-DCDU</td>
<td>( \tau_d^d(i) )</td>
</tr>
<tr>
<td>Total capacity requirements on the link between the two DCDUs</td>
<td>( \rho_d^c(i) )</td>
</tr>
<tr>
<td>Maximum allowed capacity for a link between two DCDUs</td>
<td>( \Delta_d^d(i) )</td>
</tr>
<tr>
<td>Link delay between DCDU and DCDU</td>
<td>( \tau_d^d(i) )</td>
</tr>
<tr>
<td>Unit function</td>
<td>( \gamma_{dcdu} )</td>
</tr>
<tr>
<td>Required data rate for the data center</td>
<td>( \gamma_{dcdu} )</td>
</tr>
<tr>
<td>Latency constraint of the DU function of the 1st RRU</td>
<td>( \Delta_d )</td>
</tr>
</tbody>
</table>

### TABLE II: List of notations related to DCCEU

<table>
<thead>
<tr>
<th>Definition</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters related RRU-DCDU links</td>
<td></td>
</tr>
<tr>
<td>Unit function</td>
<td>( \Psi_d(i, j) )</td>
</tr>
<tr>
<td>Link delay between RRU and DCDU</td>
<td>( t_d^d(i, j) )</td>
</tr>
<tr>
<td>Cost per unit of resource for the links DCDU-DCDU</td>
<td>( \tau_d^d(i) )</td>
</tr>
<tr>
<td>Total capacity requirements on the link between the two DCDUs</td>
<td>( \rho_d^c(i) )</td>
</tr>
<tr>
<td>Maximum allowed capacity for a link between two DCDUs</td>
<td>( \Delta_d^d(i) )</td>
</tr>
<tr>
<td>Link delay between DCDU and DCDU</td>
<td>( \tau_d^d(i) )</td>
</tr>
<tr>
<td>Unit function</td>
<td>( \gamma_{dcde} )</td>
</tr>
<tr>
<td>Required data rate for the data center</td>
<td>( \gamma_{dcde} )</td>
</tr>
<tr>
<td>Latency constraint of the DU function of the 1st RRU</td>
<td>( \Delta_d )</td>
</tr>
</tbody>
</table>

As previously noted, the resulting cost optimization model is considered as a Mixed Integer Quadratically Constrained Programming (MIQCP) model [13], [14] since it includes: i) discrete (e.g., boolean) and continuous variables, ii) objective function with quadratic terms, and iii) at least one quadratic constraint. Furthermore, our model can be extended to other types of network deployment scenarios. For instance, it can handle regions with partial network infrastructure by simply
setting the corresponding boolean variables to “1”. The corresponding cost of these components will then be represented as a constant value added to the objective function. From a mathematical point of view, minimizing the objective \((f + \alpha)\) is the same as minimizing the objective \((f)\), where \(\alpha\) is constant and all the variables are non-negative.

IV. PERFORMANCE EVALUATION

In our experimental evaluation, we use the network topology proposed in [15]. While our network planning framework can be applied to multiple Region of Interests, in our experiments, we consider a single one as shown in Fig. 2. We implemented the proposed MIQCP model in the CPLEX Optimizer\(^2\). We ran our experiments on an OpenStack platform using a VM with 64GB of RAM and 28 virtual CPUs.

For the functional split, the RF-PHY split [1] is used for the DU, while higher layer functions are grouped at the CU. The requirements of DU and CU functions are obtained from OpenAirInterface [16], while data rate requirements are based on [2], [17]. Also, we consider relative normalized costs as shown in Table III, allowing us to derive real costs given the cost of links and DCs. Note that those costs are derived from [11], [18], [19]. More specifically, the cost of a DC is in the order of $10,000,000, and thus we use \(\alpha10^6\) to represent the DC cost, where \(\alpha\) is a constant value, while the cost of optical fiber is equal to $210/m. We normalize the two costs by dividing them by 210, and obtain 1 (unit cost/m) for links and 4.76\(\alpha10^3\) (unit cost) for DCs. We then use 4.76\(\alpha = 15\) for DCDUs and 3 \(+ 4.76\alpha = 45\) for DCCUs. In our experiments, the value \(\text{ref}\) in Table III is set to 1. As for the maximum allowed CPU capacity per DC node, we use 4 CPU cores for DCDUs and 8 CPU cores for DCCUs.

<table>
<thead>
<tr>
<th>TABLE III: Relative link and DC costs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
</tr>
<tr>
<td>RRU-DCDU link</td>
</tr>
<tr>
<td>DCCU-DCDU constant cost</td>
</tr>
<tr>
<td>DCU-DCDU variable cost</td>
</tr>
<tr>
<td>DDU-DCDU variable cost</td>
</tr>
<tr>
<td><strong>Parameters related to the links</strong></td>
</tr>
<tr>
<td>RRU-DCDU constant cost</td>
</tr>
<tr>
<td>RRU-DCDU variable cost</td>
</tr>
</tbody>
</table>

In order to show the effect of the density of the Region of Interest’s infrastructure, we vary the number of RRRUs in the selected region between 3 and 12, the number of candidate DCDU locations from 2 to 6, and use 3 DCCU candidate locations. These values were selected in order to run realistic experiments in reasonable time. To this end, we fix the region’s size and subsample the number of RRUs and DCDUs. Note that the computational cost and scalability of our model is discussed in Section V.

1) Scenario With No Infrastructure: Figs. 3a and 3b show the cost of DCs for 6 and 3 candidate DCDUs, respectively, while the cost of links is shown in Figs. 4a and 4b, also for 6 and 3 candidate DCDUs, respectively. From these figures, many important observations can be made. First, the dominant cost of the network is the cost of the X-haul, since the cost of DCs is relatively low compared to the cost of the links. More specifically, the cost of DCDU-DCCU links is higher than the cost of RRU-DCDU links as DCCUs are usually located farther away from DCDUs (< 185 km), when compared to the distance between RRUs and DCDUs (< 15 km). Thus, higher cost is required to interconnect DCDUs with DCCUs than to interconnect DCDUs with RRUs. The second reason for cheaper RRU-DCDU links is that the target topology considers a relatively small number of RRUs. It is expected that the cost of RRU-DCDU links will become higher when increasing the number of RRUs, and this cost may even exceed the cost of DCDU-DCCU links.

Generally, the cost of RRU-DCDU links increases faster than the one of DCDU-DCCU links. Indeed, longer RRU-DCDU links may need to be installed when increasing the number of RRUs, while the increased cost of the DCDU-DCCU links is only related to the cost paid for increasing the capacity of the links (i.e., links with higher capacity have to be installed). Moreover, the data rate on the links RRU-DCDU is much higher than the one for DCDU-DCCU links.

Fig. 4a shows that the cost of RRU-DCDU links sometimes decreases when increasing the number of RRUs, e.g., from 6RRUs to 7RRUs. This is because there may not be enough capacity in already chosen DCDUs to support the additional costs.

and 3 candidate DCCUs (shown as red stars). In addition, when a component (i.e., RRU, DCDU, or DCCU) is selected, the corresponding symbol is filled by the same color. An important observation from this figure is that there is no link among the DCDUs. Usually, this could happen for two main reasons: i) lack of resources, ii) latency between two DCDUs does not allow to connect a RRU to a DCDU through another DCDU. Fig. 5b illustrates the case when the RRU-DCDU links are not constrained by latency. In this figure, there are two important observations: i) presence of links among the DCDUs, and ii) lower number of DCs. Again from Fig. 5a, it can be seen that there are two DCDUs connected to one of the DCCUs, and only one DCDU connected to the second DCCU (the one in the top right corner of the figure). This can be explained as follows. One or multiple RRRUs connected to the DCDU which is connected to the second DCCU (the one in the top right corner of the figure) cannot be connected to another available DCDU due to latency constraints as described above. Thus, a new DCDU needs to be used. In addition, the same DCDU is connected to a different DCCU (top right) than the one the two other DCDUs are connected although the distance between them is less than 140 km. The reason is that the cost to bring online this new DCCU (on the top right) is less than the cost of establishing a link to the DCCU in the bottom of the figure.

2) Scenario with Partially Deployed Infrastructure: Figs. 6a and 6b show the cost of DCs, and Figs. 7a and 7b illustrate the cost of links for 6 and 3 candidate DCDUs, respectively. The hexagon in Fig. 8 is used to indicate that the component is already constructed/installed. In this experiment, there is only one installed DCDU and one installed DCCU. As expected, when compared to the previous scenario, i.e., where there is no existing infrastructure (Figs. 3a, 3b, 4a, and 4b), a significant cost reduction can be observed: in these experiments, 45% for DC cost and 27% for link cost for the case of 3 candidate DCDUs. This cost reduction is due to already provisioned infrastructure. An important observation from Figs. 6a, 6b, 7a, and 7b is that both DC and link costs exhibit similar trend as in the first scenario, even though they are considerably reduced in this scenario.

V. Computational Cost and Scalability
Solving our MIQCP model as described in Expression (1) in the CPLEX Optimizer can be divided in two phases, namely: building the CPLEX object and effectively solving the model. Figs. 9a and 9b show the average execution time for each
of these two phases as a function of the number of C-RAN elements. We observe that there is no significant difference between the time to build the problem and the time to solve it for small number of RRUs and DCDUs. As the number of RRUs and DCDUs increases, both times increase exponentially. However, the build time exhibits a much more significant increase which, in these experiments, is up to 100 times longer than the time to solve the problem. Therefore, the bottleneck is the time to construct the CPLEX object, which could be highly reduced by using parallel distributed computing techniques. It is also worth noting that network planning and provisioning is usually done "offline", i.e, as network providers are in the planning stage of deployment. As such, we argue that longer computational times can be tolerated especially if they result in finding optimal deployments that offer substantial cost savings.

VI. CONCLUSION

In this paper we proposed a novel cost optimization framework for planning and provisioning of 5G three-tier C-RANs based on a Mixed Integer Quadratically Constrained Programming (MIQCP) model. We showcase the generality of our approach through simulations with realistic cost and C-RAN functional requirements. Our simulation results confirm that the overall cost is dominated by the X-haul, and also that the more candidate locations for DCs, the higher the chance to minimize the overall network cost. The advantage of the presented model is that it is general and flexible one. As future work, we plan to consider the problem of dynamic RAN function scaling and placement based on spatio-temporal multi-user traffic variability.

ACKNOWLEDGMENTS

This work was partly funded by the French Government through the Investments for the Future Programs reference UCN@Sophia Labex ANR-11-LABX-0031-01.

REFERENCES