A Dynamic Functional Split in 5G Radio Access Networks

Alberto Martínez Alba
Chair of Communication Networks
Technical University of Munich
Munich, Germany
alberto.martinez-alba@tum.de

Wolfgang Kellerer
Chair of Communication Networks
Technical University of Munich
Munich, Germany
wolfgang.kellerer@tum.de

***Abstract*—The 3rd Generation Partnership Project (3GPP) proposes a centralized architecture for the 5G radio access network (RAN) in order to reduce costs and mitigate inter-cell interference, which helps to increase user data rates. However, the limited capacity of current fronthaul networks renders it impossible for many RANs to be fully centralized. Instead, the operators can opt for a partially centralized architecture, in which only some of the functions of the RAN's processing chain are centralized. Previous work has tackled the optimal selection of these functions in a static or semi-static manner. In this paper, we present a 5G RAN that is able to dynamically adapt the subset of centralized functions to maximize data rates at runtime. We analyze the dynamics of a dense 5G RAN to derive a maximum convergence time for the selection algorithms and show that a dynamic functional split significantly improves data rates with respect to statically centralized solutions.**

***Index Terms*—dynamic, functional split, flexible, 5G, eMBB**

I. Introduction

The fifth generation of mobile networks (5G) aims at substantial performance improvements with respect to 4G networks. This is reflected by the three 5G use cases: enhanced mobile broadband (eMBB), which envisions data rates ten times higher than those provided by 4G networks; ultra reliable low-latency communications (URLLC), which promises delays ten times lower; and massive machine-type communications (mMTC), which will support hundreds of thousands of connected devices [1]. In order to achieve these ambitious objectives, the usage of network resources must be performed in a highly efficient manner.

The increment of data rates envisioned in eMBB necessarily entails denser mobile networks to provide strong radio signals to the users and improve frequency reuse. This densification, however, leads to an increase in the uplink and downlink interferences, which may actually hinder the achievement of high data rates. As a consequence, interference management techniques need to be used to allow for high-density 5G networks. With the intention of reducing costs and providing the fast communication between base stations that these techniques require, the 3GPP proposes a partially centralized architecture for the radio access network (RAN). In this architecture, each base station (gNodeB) in the RAN is divided into a distributed unit (DU), located close to the radio equipment, and a centralized unit (CU), deployed in a central location. The functions of the processing chain of each gNodeB are split between these two units, leading to the so-called *functional split* [2]. Since centralized units can take over the functions of multiple gNodeBs, this architecture allows for fast communication between functions, thus enabling interference management. In addition, a centralized architecture may reduce infrastructure and operating costs, as it reduces power consumption and rental fees [3].

Given the aforementioned advantages, mobile network operators are motivated to centralize as many RAN functions as possible. Nevertheless, in realistic scenarios the amount of functions that can be centralized is limited by the capacity of the fronthaul network connecting the CU and DUs, as well as by the processing capacity of the CU [4]. Therefore, operators can centralize only a subset of the RAN functions, which introduces the problem of optimally selecting them. This problem is complicated by the continuous variation in the interference situation caused by the movement of the users, which may lead to frequent changes in the optimal solution. One way to address this challenge would be to consider only static parameters, such as average user density or number of interfered resource blocks. However, such an approach has two important disadvantages. On the one hand, it relies on an estimation of the traffic and mobility that the network will face, which may be inaccurate. On the other hand, even if the estimated parameters are accurate in the long run, any temporary deviation from them will lead to wasted resources.

In this work, we opt instead for dynamically adapting the functional split to the instantaneous interference situation. This approach maximizes the data rates of the users and avoids wasting resources, thus being suitable for eMBB. In order to do this, we formulate the *dynamic centralization problem* for the 5G RAN and propose multiple strategies to solve it. As the complexity of the problem makes its solving time comparable to the time it takes for the optimal solution to change, we derive an analytical model to estimate the *coherence time* of a 5G RAN, i. e., the time during which the solution of the centralization problem does not change significantly, and use it to select the most appropriate strategy. In summary, our contributions are threefold: (i) we formulate the dynamic centralization problem and propose different algorithms to tackle it, (ii) we analyze the dynamics of a dense RAN to derive the time available for the algorithms to converge, and

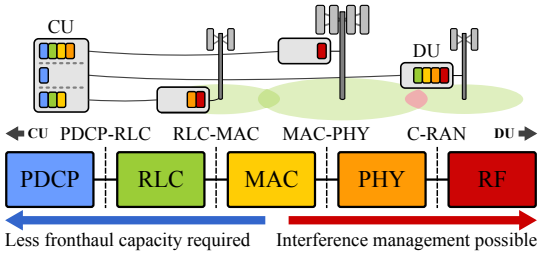

Fig. 1. Architecture of the 5G RAN, functional processing chain, and possible functional splits (depicted with dashed lines).

(iii) we evaluate the performance of the proposed algorithms for different network conditions.

The rest of the paper is organized as follows. Sec. II introduces related work on the topic. In Sec. III the model of a dense 5G RAN is presented. In Sec. IV we formulate the dynamic centralization problem, along with a quadratic relaxation. In Sec. V we derive the coherence time of a 5G RAN. Sec. VI presents a simulative evaluation of the proposed algorithms. Finally, Sec. VII concludes the paper.

II. Related work

The optimal selection of functional splits has been tackled, to a greater or lesser extent, by previous research. For instance, the authors in [5] envision a flexible 5G RAN that supports multiple functional splits to match the expected user traffic. Another example is [6], which tackles the optimal selection of the functions to be centralized as part of the network design. In FlexCRAN [7], a framework for a partially centralized RAN is presented, featuring on-the-fly changes in the functional split as a desired characteristic. The authors of [8] address the problem of selecting the optimal functional split when a new virtual network is added. Finally, in [9] the authors describe a real implementation that can switch between functional splits at runtime in a few milliseconds. This previous research sets the basis for our work, which is, to the best of our knowledge, the first addressing the problem of dynamically selecting the functional split.

III. Network model

In this section we provide details about the RAN configuration required to formulate the dynamic centralization problem.

A. General aspects

We consider a dense 5G scenario with G gNodeBs (gNBs), including macro and small cells, and U active user equipments (UEs). The use case considered is eMBB, which implies that the main objective of the network is to provide high data rates to the users. In our analysis, we focus on the downlink and assume that UEs are continuously receiving (full buffer assumption). These assumptions are not required for the conclusions to hold, but they simplify the analysis.

B. Functional split

We assume that each gNB in the RAN hosts a processing chain of RAN functions as depicted in Fig. 1. Each gNB can switch between two different functional splits: one enabling interference management (such as MAC-PHY or C-RAN), and one reducing the load on the fronthaul and on the CU (such as PDCP-RLC) [2]. We refer to the gNBs implementing the former split type as *centralized* gNBs and to the latter as *distributed* gNBs. The binary variable c_g is used to indicate whether gNB g is implementing a centralized ($c_g = 1$) or a distributed split ($c_g = 0$) at any given time. We assume that each gNB g can change the value of c_g at runtime without service disruption and with a latency lower or equal than 10 ms, as previous research shows [9]. The decision to change the split is taken by a central entity in the CU that has information from all gNBs and UEs. Finally, we model the maximum number of centralized gNodeBs that the RAN can support at any given time with the variable C. The value of C reflects the limited fronthaul capacity and the limited computing resources at the CU [4].

C. Interference management

We assume that centralized MAC, PHY, or RF functions can communicate to one another so as to prevent or cancel their mutual interference. This is accomplished either by coordinating transmissions or by jointly processing signals [10]. Formally, we state that the interference between gNBs g and g' is canceled if and only if $c_g = c_{g'} = 1$.

IV. Dynamic centralization problem

In this section, we present the dynamic centralization problem for a dense 5G RAN.

A. Full problem

The instantaneous data rate $r_u(t)$ of UE u at time t is proportional to the allocated bandwidth $B_u(t)$ and to the instantaneous spectral efficiency $\gamma_u(t)$. As $B_u(t)$ depends on the scheduler decisions, we take $\gamma_u(t)$ as our main performance indicator. In the following, we drop the dependence on time to simplify the notation. According to Shannon's law, the spectral efficiency of UE u can be expressed as:

$$\gamma_u(\mathbf{c}) = \log_2\left(1 + \frac{P_{u,s(u)}}{N_u + \sum_{g \neq s(u)} (1 - c_g c_{s(u)}) P_{u,g}}\right), \quad (1)$$

where N_u is the instantaneous noise power, $s(u)$ denotes the index of the gNB serving UE u, $P_{u,g}$ is the power received by UE u from gNB g, c_g is the indicator of centralization of gNB g, and $\mathbf{c} = \langle c_1, ..., c_G \rangle^{\mathsf{T}}$ is the centralization vector for all G gNBs. Note that, as anticipated in Sec. III-C, in (1) the interference received by UE u from gNB g is canceled if $c_{s(u)} = c_g = 1$, that is, if both the interfering and the serving gNBs are centralized.

Dynamic centralization problem (DCP): The objective of the DCP is to find the centralization vector $\mathbf{c}$ that maximizes the sum of the logarithm of the spectral efficiency (to ensure proportional fairness [11]) for all UEs:

$$\begin{aligned} &\max_{\mathbf{c}} \; R(\mathbf{c}) = \sum_{u=1}^{U} \log(\gamma_u(\mathbf{c})), \\ &\text{s. t.} \sum_{g=1}^{G} c_g \leq C, \quad c_g \in \{0, 1\}. \end{aligned} \quad (2)$$

Solving (2) in real time leads to the instantaneous optimal configuration $\mathbf{c}^*$ of the RAN, which guarantees operation at maximum spectral efficiency. Nonetheless, the DCP is an integer non-linear optimization problem, and hence NP-Hard, which jeopardizes real-time solving. In this work, we consider two strategies to allow for fast near-optimal solutions to the DCP. One is to reformulate the problem into a relaxed version, which is addressed in the next section. Alternatively, we can employ an off-the-shelf heuristic such as the genetic algorithm, owing to its good applicability to integer optimization [12]. In a nutshell, the genetic algorithm works as follows. First, an initial population $\langle \mathbf{c}_1, ..., \mathbf{c}_J \rangle^\mathsf{T}$ of J solutions is randomly generated. The objective function is evaluated for all solutions and the best ones are kept for the next generation. In addition, new solutions are created by crossing and randomly mutating selected solutions. This process is repeated until the improvement achieved by new generations is low. The specific parameters used in our implementation are those recommended in [12], and thus we skip the details here.

B. Quadratic relaxation

Let us consider a simplified scenario in which the UEs only report the list of interfering gNBs, but not their signal power. In that case, we can assume that all interferences are received with the same power P and rewrite (1) as:

$$\gamma_u(\mathbf{c}) = \log_2 \left(1 + \frac{P_{s(u)}}{N_u + P \cdot \sum_{g=1}^{G} (1 - c_g c_{s(u)}) k_{u,g}} \right), \quad (3)$$

where $k_{u,g} = 1$ if and only if UE u is being interfered by gNB g. In (3), the summation $x_u(\mathbf{c}) = \sum_{g=1}^{G} (1 - c_g c_{s(u)}) k_{u,g}$ simply yields the number of gNBs that are interfering UE u. Therefore, we want to maximize the sum of this function:

$$\log(\gamma_u(\mathbf{c})) = \log \left(\log_2 \left(1 + \frac{P_{s(u)}}{N_u + P \cdot x_u(\mathbf{c})} \right) \right). \quad (4)$$

In a well planned network, we can assume that there are no more than, say, 10 simultaneous interferers. Thus we can take advantage of the limited domain $x_u(\mathbf{c}) \in \{0, ..., 10\}$ and the slow growth of (4) to approximate it as a linear function:

$$\log(\gamma_u(\mathbf{c})) \approx \cdot \left(1 - \beta \cdot \sum_{g=1}^{G} k_{u,g} \cdot (1 - c_g c_{s(u)}) \right). \quad (5)$$

The coefficients α and β can be obtained from linear regression of (4) at the desired points. Since these coefficients depend only on $P_{s(u)}$ and P, which should be similar among all UEs in a dense RAN, they do not influence the optimization problem. Now we can exploit the linear nature of (5) to express the new objective function in matricial notation:

$$\sum_{u=1}^{U} \log(\gamma_u(\mathbf{c})) \approx \alpha \cdot \left(U - \beta \cdot (\mathbf{1}^\mathsf{T} \mathbf{K} \mathbf{1} - \mathbf{c}^\mathsf{T} \mathbf{H} \mathbf{c}) \right), \quad (6)$$

where $\mathbf{H} = \frac{1}{2} \left(\mathbf{S}\mathbf{K} + \mathbf{K}^\mathsf{T} \mathbf{S}^\mathsf{T} \right)$ is the symmetric coefficient matrix, $\mathbf{K} = [k_{u,g}] \in \{0,1\}^{U \times G}$, and $\mathbf{S} = [s_{g,u}] \in \{0,1\}^{G \times U}$, where $s_{g,u} = 1$ if and only if $g = s(u)$. This new objective function can be reduced to remove the factors not depending on $\mathbf{c}$, which results in a new, simplified optimization problem.

Quadratic Dynamic Centralization Problem (QDCP): The objective of the QDCP is to find the vector $\mathbf{c}$ that maximizes the product $\mathbf{c}^\mathsf{T} \mathbf{H} \mathbf{c}$:

$$\begin{aligned} &\max_{\mathbf{c}} \ \mathbf{c}^\mathsf{T} \mathbf{H} \mathbf{c}, \\ &\text{s. t.} \ \sum_{g=1}^{G} c_g \leq C, \quad c_g \in \{0,1\}. \end{aligned} \quad (7)$$

The QDCP approximates the DCP and allows for faster resolution. Although still NP-Hard, the QDCP is a special case of the quadratic 0-1 knapsack problem, in which the weights of all elements are the same. Thus, we can use algorithms in the state of the art to tackle it. In this work, we select three algorithms to solve the QDCP, based on their short convergence times. The first is Quadknap, a branch-and-bound algorithm that yields exact solutions reportedly faster than other exact approaches by applying a Lagrangian relaxation in the calculation of upper bounds [13]. The second is the greedy algorithm, which sets the non-zero elements in $\mathbf{c}$ as the first C unique indices associated with the greatest coefficients in $\mathbf{H}$. Finally, we evaluate an even simpler, faster greedy algorithm, which simply selects the C gNBs with the highest number of covered (interfered or served) UEs.

V. Coherence time of a dense 5G RAN

The motivation behind formulating and solving the DCP and QDCP is to dynamically adapt the 5G RAN to environment changes so as to maximize data rates and usage of resources at every instant. The ever-changing nature of the environment implies, however, that the problem itself is also continuously changing, which forces us to carefully consider the time it takes for a solution to be found and implemented, as it may have become useless by the time it is put into operation.

We can guarantee that the optimal solution $\mathbf{c}^*$ of the DCP/QDCP at a given time will perform as expected if the problem has not changed by the time $\mathbf{c}^*$ is implemented. Owing to this, in this section we derive an analytical expression to predict the time between changes in the problem. Nevertheless, the presence of changes in the problem does not necessarily mean that the solution $\mathbf{c}^*$ has varied. That is, the time between changes in the problem is not equivalent to the time between changes in the optimal solution, and the latter is actually more relevant when evaluating whether the convergence time of an algorithm is low enough. We call the latter time the *coherence time* of the RAN, as it is the time during which we can assume that the RAN exhibits an stable behavior. In this section, we also provide an estimation of this coherence time.

A. Time between changes in the problem

Let us define $\mathcal{T}_p$ as the random variable modeling the time between changes in the input parameters of the optimization problem. In the DCP these parameters are continuous variables, thus the problem is continuously changing and $\mathcal{T}_p = 0$.

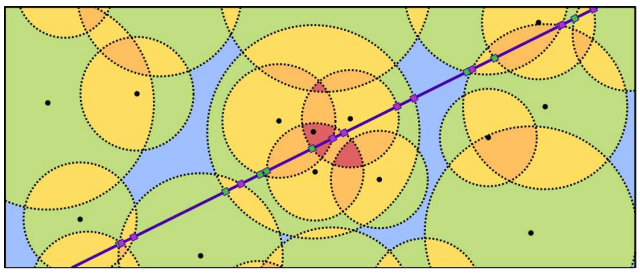

Fig. 2. Depiction of cells, interference regions, and a trajectory line. Green (purple) squares are left-transitions (right-transitions) on the trajectory line.

Conversely, the input parameters of the QDCP (matrices $\mathbf{S}$ and $\mathbf{K}$) have a discrete (binary) range, which allows for further analysis. We can relate changes in $\mathbf{S}$ and $\mathbf{K}$ to changes in the positions of the UEs. In order to see this, let us picture a simplified, geometrical model of a 5G RAN. We define *cell* as the area covered by one gNB, i. e., the area within which a UE may be served or interfered by a given gNB. For simplicity, we model these cells as circles[1] centered on the position of their radio equipment. The cells divide the area into *interference regions*, defined as those points covered by the same number of gNBs. These regions are shown in Fig. 2 as the areas with different colors. In this scenario, every change in $\mathbf{S}$ or $\mathbf{K}$ implies one or more UEs *transitioning* from one region to another. Therefore we can model the dynamics of $\mathbf{S}$ and $\mathbf{K}$ by analyzing the movement of UEs.

Let us consider a RAN with cell centers distributed according to a 2-dimensional homogeneous Poisson point process (PPP) of density d cell centers per area unit. The PPP has been shown to accurately model real mobile deployments [14], and, as we show in the next section, it also provides valid results for a dense 5G RAN. We assume that each gNB g is associated with a circular cell of radius ρ_g and UEs move in a linear fashion throughout the area, which is accurate for short periods of time, thus following a *trajectory line*. We refer to the intersections between the trajectory line and the cell borders as *transition points*. In Fig. 2, these points are depicted as squares over the trajectory line. For every cell, there are two transition points, one when entering the cell and one when leaving it. Without loss of generality, we assume movement from left to right and thus refer to these points as *left-transition* and *right-transition* points, respectively.

From this layout, we can derive interesting properties of $\mathcal{T}_p$. We have structured the main statements that lead to them into four lemmas, whose proofs are included in Appendix A. The first two lemmas address the distribution of the distance between transition points, which are the points on the trajectory lines where $\mathcal{T}_p$ changes. Lemmas 3 and 4 tackle the conversion of distance between these points into time.

Lemma 1: The left- and right-transition points are 1-dimensional PPPs on the trajectory line of density $\lambda_l = \lambda_r = 2d\bar{\rho}$ (respectively), where $\bar{\rho}$ is the average cell radius.

[1]Note that this is also valid for a cell configuration of three 120° sectors, as their mutual interference can be canceled locally at their common DU.

Lemma 2: If $\bar{\rho} > \sqrt{\frac{\pi}{8d}}$, the superposition of the left- and right-transition processes can be approximated by a PPP with intensity $\lambda_t = 4d\bar{\rho}$.

Lemma 3: The distribution of the random variable $\mathcal{T}_p^u$ modeling the time spent by UE u in an interference region can be approximated by an exponential distribution of mean $\lambda_u = 4d\bar{\rho}v_u$.

Lemma 4: In a RAN with U independent UEs, the distribution of $\mathcal{T}_p$ can be approximated by an exponential distribution of mean $\lambda_p = 4d\bar{\rho}U\bar{v}$, where $\bar{v}$ is the average UE speed.

The condition $\bar{\rho} > \sqrt{\frac{\pi}{8d}}$ of Lemma 2 is easily met in dense scenarios [15]. As a result, we conclude that $\mathcal{T}_p \sim \mathrm{Exp}(4d\bar{\rho}U\bar{v})$, which implies that the average time between changes in the QDCP, $\bar{t}_p = \frac{1}{4d\bar{\rho}U\bar{v}}$, is inversely proportional to the density of the cells in the RAN d, the number of UEs U, the average cell radius $\bar{\rho}$, and the average UE speed $\bar{v}$.

B. Time between changes in the solution

Let us define $\mathcal{T}_s$ as the random variable modeling the time between changes in the optimal solution of the problem. Since the solutions of the DCP and QDCP cannot change faster than the input, we know that $\mathcal{T}_p$ is a lower bound (in distribution) of $\mathcal{T}_s$. However, this bound may not be tight, as the relationship between problem inputs and solutions is not injective. This implies that many inputs share the same solution, and it is likely that those inputs are similar to one another. As a result, we can conjecture that the more different two inputs are, the more likely it is that they do not map to the same solution. In the QDCP case, this means that the higher the ratio of transitioned UEs with respect to an initial setup, the more probable a change in the solution is.

In the absence of any other information, we can approximate the distribution of $\mathcal{T}_s$ by assuming a one-to-one relationship between the ratio of transitioned UEs and probability of a solution change. Thus we define $\mathcal{T}_s$ as the random variable modeling the time required for all UEs in the network to change its interference region, that is, $\mathcal{T}_s = \max\{\mathcal{T}_p^u\}$ for all u, where $\mathcal{T}_p^u$ is defined in Lemma 3 of Sec. V-A. Thus, its cumulative distribution function (CDF) can be expressed as:

$$F_{\mathcal{T}_s}(t) = \frac{1}{U}\sum_{u=1}^{U}\left(1 - e^{-4d\bar{\rho}v_u t}\right) = 1 - \sum_i p_{v_i} e^{-4d\bar{\rho}v_i t}, \quad (8)$$

where p_{v_i} reflects the proportion of UEs with velocity v_i. Note that $\mathcal{T}_s$ does not depend anymore on the number of UEs U.

Now that we have a characterization of $\mathcal{T}_s$, we can use it to generate an estimation of the coherence time of the RAN. For instance, we could use its mean $\bar{t}_s = \frac{1}{4d\bar{\rho}\breve{v}}$ as our estimation, where $\breve{v}$ is the harmonic mean of all UEs' velocities. Nevertheless, $\bar{t}_s$ is not a robust estimator, since it tends to infinity when there are static users. We could use instead the median or the quartiles of $\mathcal{T}_s$ as coherence times, but (8) does not admit analytical expressions for them. Alternatively, we propose

$$\hat{t}_s = \frac{1}{4d\bar{\rho}\bar{v}} \quad (9)$$

Parameter	Scenario's density		
	Low	Medium	High
Layout	Macro cells: hex. grid + Small cells: 2-d PPP		
Small/macro ratio μ	3	6	9
Cell density d **(cells/km**^{2}**)**	110	200	290
Avg. cell radius $\bar{\rho}$ **(m)**	58.8	48.6	44.5
UE distribution	Uniform in macro + Clustered in small 10 users per cell		
UE speed distribution	80%: 3 km/h, 20%: 30 km/h		
Radio propagation model	COST Hata model		

Table 1. Attributes of a 5G dense urban scenario as specified by 3GPP [15].

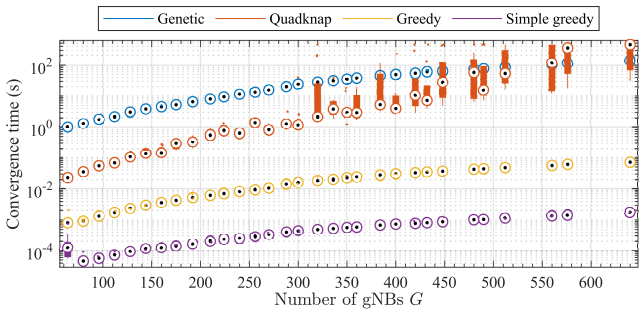

Fig. 3. Boxplots of the convergence time of the four proposed approaches with respect to the number of gNBs in the network.

as the coherence time of a dense 5G RAN, since it provides a conservative value (it can be trivially proven that $\hat{t}_s \leq \bar{t}_s$) and is more robust than $\bar{t}_s$.

VI. Evaluation

In this section, we assess the different strategies to operate a RAN that can dynamically change its functional split to maximize data rates. In order to simulate a realistic 5G RAN, we employ a custom MATLAB/C++ simulator implementing the recommended 3GPP parameters for a 5G dense urban scenario (see Table 1). The equipment used for the simulations is an 8-core Intel Core i7-6700 PC running at 3.40 GHz. In the following measurements, we assume a high-mobility scenario corresponding, for example, to the center of an active city. We consider linear movement of the UEs but still in a clustered manner, in order that the UE distribution remains the same. In addition, we always use $C = \frac{G}{2}$ to allow for consistent comparisons between measurements.

We first evaluate the convergence time of the four algorithms as a function of the number of G of gNBs in the RAN, since this is the length of the solution vector $\mathbf{c}$ (see Fig. 3). We observe a substantial difference in the convergence time of the four algorithms. The greedy and simple greedy algorithm converge in less than 100 ms in all cases, whereas the genetic and Quadknap algorithms take longer than 100 s if $G > 500$ gNBs. In addition, Quadknap exhibits a large increase in its convergence variance when the number of gNBs exceeds 300, which may degrade its performance in a real implementation.

Next, we calculate and simulate the time between changes in the problem $\mathcal{T}_p$ and the coherence time of the RAN $\mathcal{T}_s$ to see whether it clashes with the convergence times. In Fig. 4

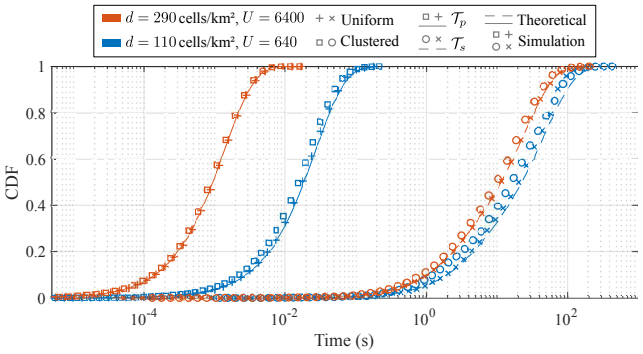

Fig. 4. Empirical and theoretical CDFs of $\mathcal{T}_p$ and $\mathcal{T}_s$ for low- and high-density scenarios with UEs distributed uniformly and in a clustered manner (according to the 5G dense urban model).

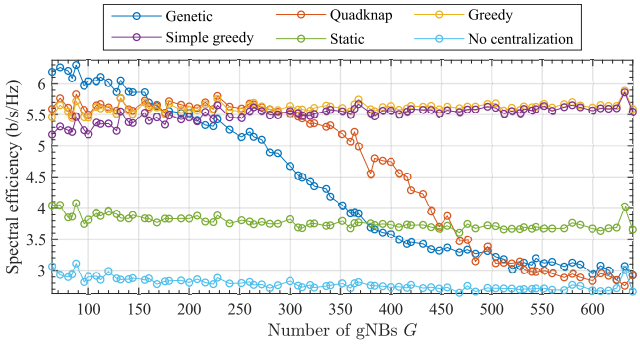

Fig. 5. Spectral efficiency achieved by four dynamic optimization approaches in a low-density RAN performing real-time adaptation. The performances of a static and a non-centralized solution are shown for comparison.

we show empirical and theoretical values of the CDFs of $\mathcal{T}_p$ and $\mathcal{T}_s$ for low- and high-density scenarios (see Table 1) and two types of UE distributions: uniform (as in the theoretical analysis), and clustered (from the 3GPP dense urban model). We observe that the theoretical model closely resembles the simulative data. The median values of $\mathcal{T}_p$ are around 20 and 1 ms for low- and high-density scenarios, respectively, which is shorter than the convergence time of all but the simple greedy algorithm and comparable to the implementation time of any solution, as mentioned in Sec. III-B. We conclude that, in general, it cannot be guaranteed that the optimization problem has not changed by the time a solution is put into operation. Regarding $\mathcal{T}_s$, we notice that is several orders of magnitude higher than $\mathcal{T}_p$ and that the differences between low- and high-density scenarios have been reduced with respect to $\mathcal{T}_p$. This can be seen via the coherence time: $\hat{t}_s = 8.9$ s and 17.5 s according to (9), respectively.

Finally, we simulate the spectral efficiency achieved by the four proposed algorithms right after the calculation of their solutions to observe the performance degradation due to their convergence time. For comparison, we also simulate the performance of a static solution, obtained from solving (2) with the average parameters of the RAN. The average spectral efficiencies after 30 repetitions for a low-density network are shown in Fig. 5, from which we can distinguish two types of behaviors. The greedy and simple greedy algorithms yield a constant spectral efficiency, since their convergence time is always below the coherence time of the RAN. By employing

any of them, the RAN can increase user data rates by a factor of 150%, compared to static solutions. Conversely, the genetic and the Quadknap algorithms, although adequate when the RAN is small, clearly exhibit a maximum G beyond which their performance degrades due to their long convergence times. These values of G can be accurately predicted by searching the points where the coherence time $\hat{t}_s = 17.5$ s equals their convergence time: $G \approx 280$ for the genetic algorithm and $G \approx 400$ for Quadknap. At these points, their relative performance improvements with respect to the static solution when $G = 64$ have decreased by half. We therefore conclude that these algorithms are not suitable for real-time optimization of large 5G RANs.

VII. Conclusion

The 3GPP currently proposes a partially centralized architecture for the 5G RAN, in which a subset RAN functions are deployed in a central location. This architecture reduces costs and enables interference mitigation, which helps to improve data rates. In this work, we tackle the problem of dynamically centralizing these functions to maximize data rates, according to the state of the network. We show that this can be accomplished in real time by using simple optimization algorithms. Indeed, we foresee a substantial increase in the achievable data rates with respect to static optimization approaches. In addition, we provide a theoretical analysis of the dynamics of a 5G RAN that can be used to select more suitable algorithms for real time operation.

Acknowledgment

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647158 - FlexNets). The authors alone are responsible for the content of the paper.

Appendix A

Proof of Lemma 1: The feet of the perpendiculars between the cell centers and the trajectory line is a 1-d PPP, given the symmetry of a 2-d PPP. The left- and right-transition points are the result of a random displacement of this 1-d PPP. By the displacement theorem, they are also 1-d PPPs. The cells with radius ρ generate a left-transition (right-transition) if and only if their centers are at distance ρ or less from the trajectory line. If, w. l. o. g., we assume that the trajectory line is horizontal, the area containing these centers would be a rectangle whose height and width are 2ρ and W, respectively. Thus, the density of points in this rectangle is $\frac{2d\rho W}{W} = 2d\rho$, and hence this is also the intensity of the PPP produced by cells of radius ρ. Assuming that all centers and radii are independent, the resulting PPP for all radii is the continuous superposition of the PPPs for each radius, whose intensity is $\lambda_l = \lambda_r = \int_0^\infty 2d\rho f_P(\rho)\,\mathrm{d}\rho = 2d\bar{\rho}$, where $f_P(\rho)$ is the PDF of the radii of all cells. □

Proof of Lemma 2: Let us define $\mathcal{Q}$ as the random variable modeling the distance between the left- and the right-transition point belonging to the same cell. This variable models the chord length distribution of a circle of radius ρ, and thus its mean is $\bar{q} = \mathrm{E}\{Q\} = \frac{4\bar{\rho}}{\pi}$ [16]. When the average chord length between left- and right- transition points is greater than the average length between two left- or right-transition points, the processes become loosely correlated and their superposition behaves like a PPP with intensity $\lambda_t = \lambda_l + \lambda_r$. This happens when $\frac{4\bar{\rho}}{\pi} > \frac{1}{2d\bar{\rho}}$, which leads to the condition in Lemma 2. □

Proof of Lemma 3: For each user u with velocity v_u, the time between transitions can be obtained by scaling distance between transitions by the factor $\frac{1}{v_u}$, which results in another PPP of intensity $\lambda_u = \lambda_t v_u = 4d\bar{\rho}v_u$. □

Proof of Lemma 4: The process resulting from the superposition of U independent PPPs is another PPP with intensity $\lambda_p = \sum_{u=1}^{U} 4d\bar{\rho}v_u = 4d\bar{\rho}U\bar{v}$. □

References

[1] R. ITU-R, "R m. 2083-0,," *IMT vision–framework and overall objectives of the future development of IMT for*, vol. 2020, 2015.

[2] 3GPP, "Study on new radio access technology: Radio access architecture and interfaces," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.801, 03 2017, version 14.0.0.

[3] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networksa technology overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[4] U. Dötsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for lte," *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, 2013.

[5] A. Maeder, M. Lalam, A. De Domenico, E. Pateromichelakis, D. Wübben, J. Bartelt, R. Fritzsche, and P. Rost, "Towards a flexible functional split for cloud-ran networks," in *2014 European Conference on Networks and Communications (EuCNC)*. IEEE, 2014, pp. 1–5.

[6] X. Wang, A. Alabbasi, and C. Cavdar, "Interplay of energy and bandwidth consumption in cran with optimal function split," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[7] C.-Y. Chang, N. Nikaein, R. Knopp, T. Spyropoulos, and S. S. Kumar, "Flexcran: A flexible functional split framework over ethernet fronthaul in cloud-ran," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–7.

[8] D. Harutyunyan and R. Riggio, "Flex5g: Flexible functional split in 5g networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 961–975, 2018.

[9] A. Martínez Alba, J. H. Gómez Velásquez, and W. Kellerer, "An adaptive functional split in 5G networks," in *2019 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019.

[10] A. Łukowa and V. Venkatasubramanian, "Centralized ul/dl resource allocation for flexible tdd systems with interference cancellation," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2443–2458, 2019.

[11] F. Kelly, "Charging and rate control for elastic traffic," *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[12] K. Deep, K. P. Singh, M. L. Kansal, and C. Mohan, "A real coded genetic algorithm for solving integer and mixed integer optimization problems," *Applied Mathematics and Computation*, vol. 212, no. 2, pp. 505–518, 2009.

[13] A. Caprara, D. Pisinger, and P. Toth, "Exact solution of the quadratic knapsack problem," *INFORMS Journal on Computing*, vol. 11, no. 2, pp. 125–137, 1999.

[14] C.-H. Lee, C.-Y. Shih, and Y.-S. Chen, "Stochastic geometry based models for modeling cellular networks in urban areas," *Wireless networks*, vol. 19, no. 6, pp. 1063–1072, 2013.

[15] 3GPP, "Study on scenarios and requirements for next generation access technologies," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.913, 07 2018, version 15.0.0.

[16] P. Sidiropoulos, "N-sphere chord length distribution," *arXiv preprint arXiv:1411.5639*, 2014.