Dynamics of the flexible functional split selection in 5G networks

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

Shakthivelu Janardhanan  
Chair of Communication Networks  
Technical University of Munich  
Munich, Germany  
shakthivelu.janardhanan@tum.de

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

Abstract—The architecture of the radio access network (RAN) in 5G features a functional split between centralized and distributed units, which can be leveraged to reduce inter-cell interference. Recent work proposes to dynamically adapt this split in accordance with the instantaneous interference situation experienced by all users. However, it is unclear whether performing this flexible adaptation is actually feasible, since the interference situation changes continuously as users move. In this work, we investigate the impact of mobility on the problem of dynamically selecting the optimal functional split. We employ a mobility simulator based on real street layouts and trace-derived traffic patterns to generate continuously varying interference situations. Then, we analyze how frequently and how much the performance of previous splits differs from the new, optimal one. The results allow us to estimate the time required for a viable flexible functional split adaptation.

Index Terms—5G, flexible, adaptation, functional, split

I. INTRODUCTION

The numerous improvements proposed for 5G mobile networks are commonly classified into three use cases: Ultra-Reliable Low-Latency Communications (URLLC), massive Machine-Type Communications (mMTC) and enhanced Mobile Broadband (eMBB). While the first two use cases aim at novel objectives for new markets, eMBB is directed towards a crucial selling point for current markets: increasing the user data rate. It is expected that the peak data rate in 5G will surpass 1 Gb/s, an order of magnitude higher than that of 4G.

There are three possible strategies to increase data rates in a mobile network: additional spectrum, improved spectral efficiency, or higher cell density. Since spectrum is a scarce and expensive resource, operators cannot rely on using new bands as their only strategy. Similarly, increasing spectral efficiency is limited by Shannon’s law, to which 5G modulation schemes are already very close [1]. Thus, operators have to increase cell density to improve user data rates.

The main disadvantage of increasing cell density is that users are exposed to stronger interference from neighbor cells. Therefore, a method to prevent or cancel these interference is required in order to leverage the benefits of high cell density. Several interference mitigation techniques have been proposed for 4G and 5G networks, such as Inter-cell Interference Coordination (ICIC) and Joint Transmission and Reception (JT and JR). These techniques require some level of information exchange between the interfering cells. For instance, in JT the whole physical signal needs to be shared by the coordinated cells at every transmission interval. This requirement hinders the application of these techniques in 4G deployments, since the architecture of its Radio Access Network (RAN) makes it difficult for the cells to coordinate quickly.

In 5G, base stations (gNodeBs) are divided into a Centralized Unit (CU), which hosts a subset of the RAN functions, and Distributed Units (DUs), which host the remaining functions. This division, known as the 5G functional split [2], allows for centralized functions to quickly coordinate with one another, thus enabling advanced interference mitigation techniques. Its main disadvantage, however, is the high capacity required for the fronthaul network connecting CU and DUs, as function centralization may require very high throughput [3].

Decreasing the number of centralized functions relaxes the fronthaul capacity requirements, but reduces the interference mitigation opportunities. Thus, for a fixed fronthaul capacity and a given interference situation of all users, we can define the functional split selection problem (FSSP) as the problem of optimally selecting the functional split for each gNodeB. Previous works have tackled the offline FSSP, whose solution is used in the RAN design. This provides a better average performance with respect to identical functional splits for all gNodeBs, but performance may be far from optimal when the instantaneous interference situation deviates from the mean.

A more advanced approach is to address the online FSSP, that is, the problem of dynamically adjusting the functional split to the instantaneous conditions of the network. However, solving the online FSSP entails an additional difficulty with respect to the offline FSSP: the mobility of the users and the changing traffic patterns lead to potentially short-lived solutions. The time during which these solutions are valid is hard to estimate, as the problem changes continuously. Nonetheless, it is crucial to know how much time is available to solve the FSSP and to implement the solution, as solutions may have expired by the time they are put into operation.

In this work, we present a comprehensive study on the validity of the solutions of the online FSSP, focusing on how frequently and how much its solutions change. For this, we model a dense urban scenario, formulate the FSSP to maximize the proportionally-fair user data rate, and present
an algorithm to solve it. Then, we use a simulator to produce realistic user behavior, based on available mobility models and traces, and solve the online FSSP in different mobility scenarios. We record the dynamics of the solution changes and the performance degradation of old solutions. The analysis and results presented here contribute to decide on the viability of implementing an adaptive functional split.

The remaining of this paper is structured as follows. Sec. II summarizes the related work on solving the FSSP. In Sec. III we introduce the system model and the FSSP formulation. Sec. IV describes the experimental setup. In Sec. V we present the simulation results. Finally, Sec. VI concludes the paper.

II. RELATED WORK

The problem of adapting the functional split to the static or dynamic conditions of a 5G network has attracted considerable attention in the recent years. The foundations of the FSSP are established in [2], whose authors envision a 5G RAN supporting multiple functional splits. Similarly, in [4], the authors propose a 5G RAN architecture that simultaneously supports different functional splits for each DU in the network.

The offline FSSP is tackled in [5], where an algorithm to select functional split during the deployment phase is presented, according to the average traffic expected in each cell. The authors of [6] also address the optimal selection of the functional split, as part of the network design.

The online FSSP is suggested in [7], where a platform for a partially centralized RAN is presented, although the problem is not fully addressed. In [8] the online FSSP is tackled within the scope of a virtualized RAN scenario, as the optimal functional split is recalculated whenever a new virtual network request arrives. However, none of these works consider the possibility to adapt the functional split to the instantaneous interference situation. This is first mentioned in [9], where the authors present a novel gNodeB implementation that is able to change its functional splits at runtime. Based on this platform, the FSSP is fully addressed in [10], where the authors describe multiple algorithms to dynamically adapt the functional splits of a 5G RAN according to the instantaneous user position. Moreover, this work includes a preliminary study on the time that is available to solve the problem.

III. SYSTEM MODEL

In this section we describe the general characteristics of the network under consideration, for which the FSSP is defined.

A. Network description

We focus on highly populated scenarios such as city centers or entertainment areas, as the variable activity in these zones motivates the interest in dynamic adaptations. Our area of interest is the coverage area of all DUs (including macro and small cells) that are connected to the same CU. DUs and CU are connected by means of a fronthaul network, whose limited capacity prevents full centralization for all cells. We denote the number of gNodeBs, and hence DUs, as $G$, and the number of active user equipments (UE) within the area of interest as $U$.

The total number of connected users may be larger by $U$, but since inactive users do not influence the interference situation, they can be ignored.

The considered use case is eMBB. The goal of the network operator is to maximize the data rate of the users, while also ensuring proportional fairness. We focus on the downlink and assume that active UEs are always receiving.

B. Functional split

Each gNodeB in the network hosts a chain of RAN functions as depicted in Fig. 1, where each function corresponds to a processing layer: PHY, MAC, RLC, and PDCP layers. An alternative, more granular definition of layers can also be used, e.g., by dividing each layer into two sublayers, although this does not affect the overall formulation of the problem. For simplicity, we assume that each gNodeB can select between two functional splits: one requiring low fronthaul capacity (such as PDCP-RLC) and one capable of performing interference mitigation (such as MAC-PHY or C-RAN). In addition, we consider that those gNodeBs implementing the latter split (centralized split) are able to fully cancel the interference they cause to one another, whereas this is not possible with the former (distributed) split (distributed split).

C. FSSP formulation

Given a network whose DUs can switch between the aforementioned splits, the objective of the online FSSP is to find the split option for each gNodeB such that the proportionally-fair data rate of all UEs is maximized. If there were no restrictions, this would be accomplished by using the centralized split for all gNodeBs. However, the number of gNodeBs implementing the centralized split is limited by the fronthaul network.

We use the binary variable $c_g$ to indicate whether gNodeB $g$ is implementing the distributed split ($c_g = 0$) or centralized split ($c_g = 1$) at any given time. Therefore, the objective is to find the optimal vector $c = \{c_1, ..., c_G\}$. We borrow the formulation of the online FSSP from [10]:

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

Fig. 1: Chain of gNodeB functions and possible functional splits (depicted with dashed lines).
where, $C$ is the maximum number of centralized gNodeBs that the fronthaul can support and $\gamma_u(c)$ is the downlink spectral efficiency achieved by UE $u$ with the split vector $c$. In this case, maximizing for spectral efficiency is equivalent to maximize for data rate, as the allocated bandwidth does not depend on $c$. Note also that the objective in (1) is to maximize the sum of the logarithm of the spectral efficiencies, which is done to ensure proportional fairness [10].

We can express $\gamma_u(c)$ as a function of the received signal and interference powers by means of the Shannon’s law:

$$\gamma_u(c) = \log_2 \left( 1 + \frac{P_{u,g}}{N + \sum_{u \neq s_u} (1 - c_g s_u) P_{u,g}} \right),$$

(2)

where, $P_{u,g}$ denotes the power received by UE $u$ from gNodeB $g$, $s_u$ is the index of the gNodeB serving UE $u$, and $N$ is the noise power (assumed constant for simplicity).

Equation (2) reflects the spectral efficiency for constant parameters $P_{u,g}$ and $s_u$. Nevertheless, for the online FSSP these parameters change over time. Hence, we can consider them as time-dependent functions: $P_{u,g}(t)$ and $s_u(t)$. As a result, this time dependence is transmitted to the spectral efficiency, which we denote as $\gamma_u(c, t)$.

### D. FSSP solving

After combining (1) and (2), we can formulate an instance of the FSSP as a function of the powers received by each UE in the network. As shown in [10], the FSSP is, in general, NP-Hard, which implies that there is not an efficient (polynomial) exact algorithm to solve it. Four algorithms are proposed in [10] to provide near-optimal solutions to the FSSP: a genetic algorithm, a branch-and-bound algorithm, a greedy algorithm that centralizes those pairs of gNodeBs producing the highest interference, and simpler greedy algorithm that centralizes the gNodeBs with the highest number of active connected users. It is shown that the genetic algorithm yields the best solutions out of them. Thus, we select it for our experiments. The summarized operation of our genetic algorithm is as follows:

1) An initial generation $\{c_1, ..., c_{100}\}$ of 100 candidate solutions is generated from random 0 − 1 vectors, the solutions provided by the aforementioned greedy algorithms, and the solution found for the previous instance.

2) The value of each solution in the population is evaluated according to the objective function in (1) and a penalization value for constraint violation.

3) The best solutions are kept and mutated in order to create the next generation. Then, the process is repeated until no new best solution is found in 50 generations.

The population size, mutation rate, and elite count, and other parameters are experimentally adjusted and/or set to those values recommended in [11] for each problem instance.

Although the solutions provided by this genetic algorithm may not be strictly optimal, in order to simplify the notation we henceforth refer to them as optimal and denote them as $c^*$. Indeed, given the complexity of the FSSP and the limited time to solve its online version, aiming to find strictly optimal solutions would be unrealistic.

### IV. Experimental design

The objective of this work is to evaluate whether dynamically selecting the functional split is feasible in a realistic 5G network or, conversely, whether the scenario from which the FSSP is defined changes too abruptly. In simple scenarios, an analytical evaluation may be considered. However, the NP-Hardness of the FSSP, as well as the complexity of modeling the UE mobility and the radio propagation effects, render the analysis intractable for all but the simplest scenarios. Owing to this, we utilize a simulator to generate the FSSP parameters according to a predefined scenario, solve the problem, and record the performance and validity of the solutions.

#### A. Simulator description

In order to generate the time-dependent parameters $P_{u,g}(t)$ and $s_u(t)$ for a realistic 5G eMBB scenario, we use a MATLAB simulator with the following characteristics:

1) Simulated area: The coverage area is simulated via a street layout that is fed to the simulator as an undirected graph. These graphs can be generated either manually or automatically from services like OpenStreetMap. For our experiments, we use the city centers of Munich, London, Madrid, and Chennai, which yielded areas ranging from 1.25 to 1.46 km².

2) Cell layout: We use parameters provided in the 3GPP dense urban model [12] to generate the cell sites. Macro cells are arranged in a hexagonal grid with an inter-site distance of 200 meters, whereas small cells are randomly positioned.

3) Path loss model: We use the log-distance path loss model for urban scenarios as described in [13].

4) User distribution and mobility: We follow the recommendations of the 3GPP dense urban model [12] with 50% mobile users (80% pedestrians and 20% vehicles) and 50% static users. With the intention of simulating mobility as realistically as possible, mobile users can only move within streets. This creates random, temporary clusters in streets and intersections, which may affect the problem solving.

5) Zones of interest: In order to model the popularity of certain places, which may cause the appearance of UE clusters, we implement a slightly modified version of the opportunities model [14], in which the probability of a UE being attracted to a zone of interest (ZOI) is not only affected by the amount of population but also by a manually-defined value. This is done to allow the manual creation of zones of interest.

#### B. Validity indicators

The aforementioned simulator is able to provide a time sequence of power parameters $P_{u,g}(t)$ and $s_u(t)$, which in turn can be translated to a time sequence of FSSP instances. We are interested in knowing for how long a solution to the instance defined at time $t$ remains valid. The concept of time of validity of a solution is subjective, but it can be intuitively understood as the time during which the performance of the solution does not differ substantially from the optimal performance. We propose the following two indicators to reflect this intuition.

- **Validity of time:** This is the amount of time a solution remains valid. It is calculated as the difference between the current time and the time when the solution was found to be optimal. A valid solution is one that remains valid for a long period of time.

- **Validity of performance:** This is the difference between the performance of the solution and the optimal performance. A valid solution is one that performs as well as or better than the optimal solution for a long period of time.
implementations. With the intention of providing meaningful performance of “old” optimal solutions for every new problem instance, in order to measure the impact of delayed solution implementations, we use the following notation.

\[ \tilde{\gamma}(c, t) \]

where \( \tilde{\gamma}(c, t) \) is evaluated by the simulator, whereas the squares represent the values of \( \tilde{\gamma}(c, t) \) at those times.

Fig. 2: Variation of \( \tilde{\gamma}(c, t) \) and \( \tilde{\gamma}^*(t) \) in a simple scenario.

1) Delayed spectral efficiency: We evaluate the performance of “old” optimal solutions for every new problem instance, in order to measure the impact of delayed solution implementations. With the intention of providing meaningful performance units, we reformulate (1a) as follows:

\[
\sum_{u=1}^{U} \log(\gamma_u(c, t)) = \log \left( \prod_{u=1}^{U} \gamma_u(c, t) \right) = U \log(\tilde{\gamma}(c, t)),
\]

(3)

where \( \tilde{\gamma}(c, t) \) is the geometric mean of the spectral efficiency \( \gamma_u(c, t) \) over all UEs. Therefore, problem (1) is equivalent to maximizing \( \tilde{\gamma}(c, t) \) at a given instant \( t \). For the optimal spectral efficiency function, we use the following notation.

\[
\tilde{\gamma}^*(t) = \tilde{\gamma}(c^*(t), t).
\]

(4)

The usage of these functions in an example scenario is depicted in Fig. 2. Fig. 2a shows the movement of the UEs from time \( t = 0 \) to \( t = 8 \) minutes, as well as the position of the DUs. In Fig. 2b, the value of \( \tilde{\gamma}(c, t) \) is shown over time, with the points registered by the simulator marked as dots. It can be seen that the spectral efficiency achieved by the blue and red solutions decay over time, as the green solution raises.

In order to denote the spectral efficiency at time \( t + \tau \) of the solution obtained for time \( t \), we use the following function:

\[
\tilde{\gamma}_t(\tau) = \tilde{\gamma}(c^*(t), t + \tau).
\]

(5)

Note that \( \tilde{\gamma}_t(\tau) = \tilde{\gamma}^*(t) \) when \( \tau = 0 \) and whenever \( c^*(t) = c^*(t + \tau) \). From these formulations, we propose the delayed spectral efficiency \( \Gamma_t(\tau) \) as:

\[
\Gamma_t(\tau) = \frac{\tilde{\gamma}_t(\tau)}{\tilde{\gamma}_t(0)} = \frac{\tilde{\gamma}(c^*(t + \tau), t + \tau)}{\tilde{\gamma}(c^*(t), t)},
\]

(6)

where, \( \tilde{\gamma}^* = \mathbb{E}\{\tilde{\gamma}^*(t)\}_t \) is used as a scaling factor, so that \( \Gamma_t(\tau) \) reaches \( \tilde{\gamma}^* \) when \( \Gamma_t(\tau) \) yields the optimal value. This is done to prevent the compression of all indicators to the range \([0, 1]\), which would hinder the comparison of scenarios. Within a single scenario, \( \Gamma_t(\tau) \) can be treated as a random process of index \( t \) for each value of \( \tau \), and thus its mean, variance, range, and other statistics can be computed.

2) Time between solution changes: Apart from \( \Gamma_t(\tau) \), which measures how much a solution degrades with respect to the optimal value, we also measure the time between changes in the solution. That is, for each \( t \) such that \( c^*(t + \Delta \tau) \neq c^*(t) \), \( \Delta \tau \) being the time increment of the simulator, we store the following value as a realization of the random variable \( \Theta \):

\[
\Theta = \inf_{\theta} \{ \theta \mid c^*(t + \theta \cdot \Delta \tau) = c^*(t), \ \theta > 0 \}.
\]

(7)

V. SIMULATION RESULTS

In this section we evaluate the performance and time between changes of the solutions of the FSSP in three types of scenarios. First, we consider a stationary scenario, in which the UE density remains constant during the whole simulation. Second, we simulate a change of UE density in accordance to several possible region types. Finally, we study the case in which the overall UE density remains constant but their distribution changes between two ZOIs. The time increment of the simulations is always \( \Delta \tau = 1 \) second.

A. Stationary scenario

In a stationary scenario, a constant number of UEs move uniformly, that is, without ZOIs, within the simulated area. This represents an area in which the population is approximately constant and uniformly distributed, such as residential or office regions during the morning or afternoon on a week day. This scenario serves as a baseline for the rest.

In Fig. 3 we show the delayed spectral efficiency\(^1\) for two different UE densities: 5 times and 20 times the number of gNodeBs in the network (according to the 3GPP dense urban model [12], the reference value is 10 times). From these figures, we can see that the performance of old solutions barely degrades over time, especially when the number \( C \) of gNodeBs that can be centralized is large. We can also see that, the larger the number of UEs, the less degradation in the old solutions. This is a consequence of the law of large numbers: the more UEs, the closer the scenario is to a static scenario of uniformly distributed UEs. In conclusion, in this scenario old solutions converge to an spectral efficiency that is between 85% and 95% that of the optimal ones.

In Fig. 4, the cumulative distribution functions (CDF) of the times between changes in the solutions for multiple uniform scenarios are shown. We can see that, although the delayed

\(^1\)The range of spectral efficiencies shown in these figures may seem lower than expected for 5G. This is because \( \Gamma_t(\tau) \) is a geometric average, thus it is biased towards low values.
$U = 5 \cdot G \text{ (5 UEs per gNodeB on average)}$

$U = 20 \cdot G \text{ (20 UEs per gNodeB on average)}$

**Fig. 3:** Evolution of $E\{\Gamma_t(\tau)\}$ for the stationary scenario. The lower and upper dashed lines represent the 2.5th- and 97.5th-percentiles, respectively, so that 95% of the values of $\Gamma_t(\tau)$ are contained in the shaded area.

In a tide scenario, the number of UEs increases from a valley to a peak value (or vice-versa) while roaming uniformly over the considered area. This reflects the tidal effect that is observed in residential, business, and transport regions in the early morning or late afternoon hours. Accordingly, we set the initial and final UE densities, along with the duration of the simulations, for four different region types as described in [15]. These parameters are shown in Table I.

**B. Tide scenario**

In Fig. 5, we show the delayed spectral efficiency for these regions. It can be seen that the spectral efficiency of all regions, decays in approximately the same way as in the stationary scenario. This suggests that the performance of solutions are not severely affected by changes in the UE density.

**Fig. 4:** Time between solution changes for a uniform scenario. The spectral efficiency seems to suggest that this scenario is mainly static, solutions do change frequently, being very unlikely that the same optimal solution is kept for more than 60 s.

**C. Event scenario**

In the event scenario, a constant number of UEs move from being distributed around an initial ZOI to being distributed around a final ZOI. This represents the internal movement of UEs within a coverage area that includes more than one region type (such as UEs moving from a residential to a business region), or the movement of UEs as they gather to attend an entertainment event.

In Fig. 7, we show the delayed spectral efficiency for the event scenario, given randomly positioned initial and a final ZOIs with an attraction radius of 100 m. We evaluate three different transition speeds: the attraction between ZOIs changes continuously within 30, 60, and 120 minutes. We can see that the delayed spectral efficiency stays high for a longer time compared to the previous scenarios, which may seem counter-intuitive. This is because the UEs tend to be clustered when they move between two ZOIs, which allows for a more effective interference cancellation. Conversely, the delayed spectral efficiency does not converge quickly to a low value, but it keeps decreasing even after 15 minutes, reaching, on average, less than 70% of the optimal geometric spectral efficiency at that point. This is indeed expected, as solutions tend to change rapidly. In over 90% of the cases, a solution change occurred in less than 15 seconds.

**VI. Conclusion**

Adapting the functional split in 5G networks is a promising strategy to improve interference mitigation and, hence,
increase user data rates. However, the variability of mobile scenarios may render a dynamic adaptation useless, as the performance of a solution may have changed too much by the time it is put into operation.

In this work, we study the influence of this variability on the dynamics of the optimal functional split. We show that the optimal functional split changes frequently over time, even when UEs are distributed and moving uniformly over the considered area. Indeed, the mean time between solution changes is below 10 seconds for all simulated scenarios.

 Nonetheless, the swiftness in the change of solutions does not imply that the performance of previous solutions decays rapidly. For the stationary and tide scenarios, it is observed that the delayed spectral efficiency drops quickly, reaching its lowest point in approximately 5 minutes, but the final spectral efficiency is between 85% and 95% of the optimal value. For the case of UEs moving between two clusters, the spectral efficiency achieved by old solutions decays more slowly than in the other scenarios, but it steadily decreases over time, reaching, on average, less than 70% of the optimal geometric spectral efficiency after 15 minutes.

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.