Comparison of performance- and cost-optimal functional splits in 5G and beyond

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

Steffen Pundt
Chair of Communication Networks
Technical University of Munich
Munich, Germany
steffen.pundt@tum.de

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

***Abstract*—Centralization in 5G radio access networks brings two main benefits: reducing cost and improving performance. Although an ideal, fully-centralized architecture would provide minimum cost and maximum performance, actual deployments cannot simultaneously optimize both. Previous research focuses on how to select the functional split of a 5G network to either maximize performance or minimize cost on partially-centralized architectures, without exploring which approach is the most appropriate. In this work, we investigate the trade-off between cost and performance of both approaches, in order to figure out which one is more adequate for real network operators. We provide a comprehensive study under a wide range of network conditions and show that, in general, a performance-maximizing approach is more likely to produce a higher net revenue.**

***Index Terms*—5G, performance, cost, functional split**

I. Introduction

One of the key features of 5G radio access networks (RAN) is high cell density [1]. This is a necessary condition in order to provide high data rates over the air interface, since denser deployments of small base stations (gNodeBs, or simply gNBs, in 5G terminology) consume less power and provide a more uniform coverage than sparse deployments. The obvious drawback of dense deployments is, however, increased inter-cell interference, which needs to be countered with interference-mitigating techniques, such as coordinated scheduling and beamforming, or joint transmission and reception.

Implementing these interference-mitigating techniques requires, ideally, a fully centralized RAN architecture, in which the operation of all gNBs is moved to a single central location. This way, centralized functions can easily cooperate among one another to coordinate their transmissions [2]. Furthermore, a centralized RAN architecture has another major advantage: it may be substantially less costly to deploy and operate than distributed architectures [2]. The reason for this is twofold. First, centralization implies converting stiff hardware units into flexible software functions, which can be affordably deployed into general-purpose data centers. Second, the pooling of computational resources benefits from multiplexing gain, which translates into less required resources with respect to distributed architectures.

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).

Consequently, it is generally agreed that a fully centralized RAN would feature these two major advantages: high performance (in terms of user data rates) and low deployment and operating cost. Nonetheless, a centralized architecture requires high-capacity links connecting the remote cell sites with the central location [3], which renders such an architecture infeasible in many cases. As a result, the current 5G architecture is just *partially centralized*: a centralized unit (CU) hosts a subset of the software functions that make up each gNB, whereas the remaining functions are located at a distributed unit (DU). This division is called the 5G *functional split*.

Although feasible, a partially centralized architecture raises two new issues with respect to a fully centralized architecture. On the one hand, the optimal functional split of a partially centralized architecture is not static, but it varies with the instantaneous network conditions and user traffic. This motivates the adoption of a dynamically-adapting functional split to track the optimal operation point [4], which is even considered in the early description of 6G networks [5, p. 13]. On the other hand, whereas a fully centralized architecture is simultaneously optimal at both performance and cost (except for the cost of the high-capacity network), this may not be true for partial centralization. That is, the performance-optimal functional split may not be cost-optimal, and vice-versa.

To the best of our knowledge, this latter issue is not addressed in detail yet. Previous work mainly focuses on selecting the functional split which either maximizes performance [6] or minimizes operating cost [7], without exploring how big the performance-to-cost gap is for either approach. This is, nevertheless, an important issue, since both performance and operating cost contribute to the net revenue of a 5G/6G network [8]. In this paper, we investigate this topic by comparing performance-maximizing and cost-minimizing approaches and show that their performance and cost differ substantially. In summary, our contributions are mainly three: (i) we present a simple, unified formulation for directly comparing performance-maximizing and cost-minimizing approaches to select the optimal functional split, (ii) we provide comprehensive simulation results to estimate the cost and performance of both approaches under a wide range of network conditions, and (iii) we identify the network conditions at which each approach is superior.

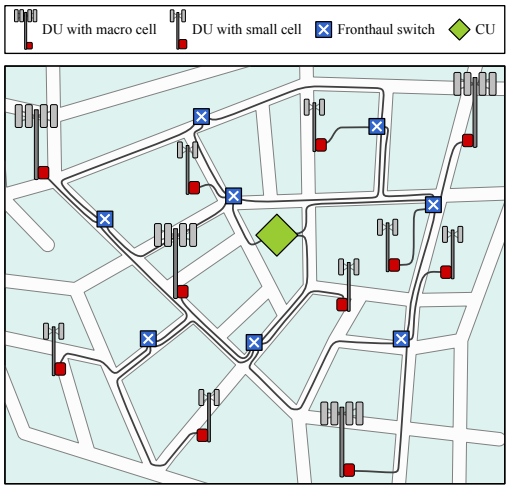


Figure 1: Example network with $G = 11$ gNBs (including macro and small cells) and eight fronthaul switches.

The rest of this paper is organized as follows. In Sec. II, we describe the system model. Sec. III presents the performance-maximizing and cost-minimizing approaches to find the optimal functional split. In Sec. IV, we simulate and compare the cost and performance of both approaches. Finally, Sec. V concludes the paper.

II. System Model

In this section, we present the network under consideration, including an overall description, the possible functional split options, a detailed description of the connection between DUs and CUs, and the modeling of the users.

A. Network description

We consider a network consisting of G gNBs, each of which is split into a CU and a DU. The DUs are deployed at remote locations, close to the radio equipment, whereas all CUs are deployed in a single, centralized data center. DUs are not necessarily collocated with their radio equipment. Instead, the radio equipment of a gNB may be deployed independently as an additional remote unit (RU) some distance away from the DU. Nonetheless, since we assume that each RU is connected to its DU by means of a dedicated link [3], they do not play a role when selecting the optimal functional split. CU and DUs are connected via a packet-switched fronthaul[1] network [9], whose modeling is addressed in Sec. II-C. A simple network consisting of $G = 11$ gNBs is depicted in Fig. 1.

The geographical distribution of DUs plays an important role in finding the optimal functional split, since it impacts the interference experienced by the UEs. In order to be as realistic as possible, we follow the recommendations for generating a dense urban scenario as specified in 3GPP TS38.193 [10]. Consequently, DUs are divided into two categories: macro and micro DUs. Macro DUs make up one fourth of all DUs and are located on a hexagonal layout 200 m away from one another. The remaining, micro DUs are randomly distributed over the area covered by the macro DUs. This results in an average cell density of ca. 115 DUs/km^2.

[1]When RUs are considered in the architecture, it is preferable to use the term *midhaul* for the network connecting DUs and CUs, reserving the term *fronthaul* for the links between RUs and DUs. Since RUs are not relevant in our system model, we just refer to this network as *fronthaul* network, as opposed to the *backhaul* network between CUs and the mobile core.

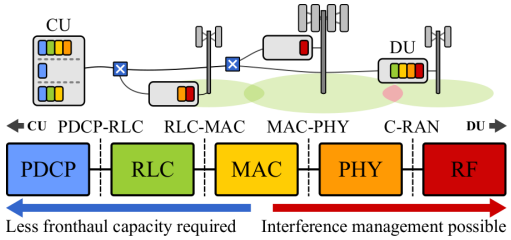


Figure 2: Scheme of the considered functional splits.

In this work, we focus on the downlink direction, that is, on communication originated at the gNB and terminated at the UE. Nonetheless, extrapolating the analysis and conclusions to the uplink is straightforward. In addition, we assume that all gNBs share the same spectrum, so that their main way of avoiding interference is to use interference-mitigating techniques at the CU.

B. Centralization levels

The full operation of a gNB can be decomposed into smaller *functions*. For instance, these functions are usually defined as the layers of the RAN protocol stack: SDAP, PDCP, RLC, MAC, PHY, RF, etc [3]. As mentioned before, we refer to each possible way of splitting gNB functions as a functional split. Since functions belong to the same processing chain, each functional split leads to a different *centralization level*. High centralization levels require high fronthaul capacity [3], but they allow for more functions to coordinate at the CU, thus enabling advanced interference-mitigating techniques. Conversely, low centralization levels require low fronthaul capacity, but only support basic interference management.

We denote by Q the number of functional splits that are available to the network. For example, in Fig. 2 there are $Q = 4$ centralization levels (PDCP-RLC, RLC-MAC, MAC-PHY, and C-RAN). At any time instant, the centralization level of a gNB g is denoted by $x_g \in \mathbb{Q}$, where $\mathbb{Q} \triangleq \{0, ..., Q-1\}$. We consider that $x_g = 0$ represents the lowest centralization level, i. e., the functional split with the lowest number of centralized functions. In contrast, $x_g = Q - 1$ represents the highest centralization level. In Fig. 2, $x_g = 0$ corresponds to the PDCP-RLC split, whereas $x_g = 3$ corresponds to the C-RAN split. The *centralization vector* containing the centralization levels of all gNBs is defined as $\mathbf{x} \triangleq [x_1, ..., x_G]$.

C. Fronthaul network

We model the fronthaul network as a directed graph $\mathcal{D} = (\mathbb{N}, \mathbb{E})$, where $\mathbb{N}$ is the set of network nodes (including CUs, DUs, and network switches) and $\mathbb{E}$ is the set of links. The node corresponding to the CU is referred to as n_0, whereas those modeling the DUs are denoted by n_g, $\forall g \in \mathbb{G}$, where $\mathbb{G} \triangleq \{1, ..., G\}$. Each link $e \in \mathbb{E}$ has a fixed capacity ϕ_e. We model downlink communication from the CU to DU g as a

data flow between nodes n_0 and n_g. The fraction of this flow carried over link e is denoted as f_e^g. We define the vector of all flows as $\mathbf{f} \triangleq [f_1^1, ..., f_{|\mathbb{E}|}^1, f_1^2, ..., f_{|\mathbb{E}|}^G]$. Each centralization level x is related to a maximum data flow value $r(x)$, which is the result of the highest possible user data rate plus the overhead required by the functional split [3].

It is shown by previous research that fronthaul networks can be very heterogeneous, exhibiting a large variability in the number of links and nodes. In order to take into account this variability into our experiments, we define the *fronthaul network degree* Ψ, defined as the ratio of number of links to nodes (CU, DUs, and switches) in the network. The minimum fronthaul network degree is $\Psi = 2$, which corresponds to a tree network. We consider a maximum network degree of $\Psi = 5$ based on the data provided in [9]. In our simulations, we generate the fronthaul network by laying out the nodes, computing the minimum spanning tree to connect the CU with all DUs, and finally adding redundant links according to Waxman model [11] until the desired value of Ψ is achieved.

D. User distribution

For any considered interval, the network serves U simultaneously active user equipments (UEs). In our simulations, we set $U = 10G$ as recommended in 3GPP TS38.193 [10] for a dense urban scenario. The geographical distribution of UEs has to be taken into account when comparing the performance or cost achieved by any approach, as user clustering influences the interference distribution. This motivates the definition of a metric to quantify the level of concentration or dispersion of UEs. We use the metric proposed in [6], so that our results can be directly compared to those of previous work. We refer to this metric as the *UE concentration index* Θ, which is defined as the Gini coefficient of the 2-dimensional distribution of the number of UEs in a 50×50 m square grid. Thus, a value of $\Theta = 1$ corresponds to all UEs being located in a single 50×50 m square, whereas $\Theta = 0$ corresponds to a perfectly homogeneous distribution. In practice, however, the lowest observed value is usually $\Theta \approx 0.5$, which corresponds to a uniformly random distribution of UEs.

III. Problem formulation

As mentioned before, when selecting the functional split there are two possible objectives to optimize: cost and performance. Ideally, selecting the highest centralization level for every gNB would optimize both of them. Nonetheless, in an actual RAN the problem is constrained by the fronthaul network, which forces us to choose a combination of cost and performance as the optimization objective. It is not possible to define a single combination of cost and performance that maximizes revenue for all networks, since there are many factors influencing how performance is converted into revenue that are specific to each network. However, we can formulate the performance-maximizing and cost-minimizing functions in a homogeneous manner, so that they can be straightforwardly combined when this specific combination is known. In this section, we present such formulations.

A. Performance-maximizing formulation

Based on the analysis of the functional split selection problem shown in [6], we take the *geometric mean* of the spectral efficiency over all UEs as our performance indicator. Maximizing this indicator is equivalent to maximizing the data rate of all UEs in a proportionally fair manner [12]. This is due to the fact that performing proportionally-fair rate maximization translates into maximizing the sum of the *logarithm* of the rates [13]. From a performance perspective, the utilities are the user data rates, which are calculated as the product of $\eta_u(\mathbf{x})$, the spectral efficiency achieved by UE u given a centralization vector $\mathbf{x}$, and B_u, the bandwidth allocated to UE u. As a result, we are interested in maximizing:

$$\sum_{u=1}^{U} \log\left(B_u \eta_u(\mathbf{x})\right) = U \log(\widetilde{\eta}(\mathbf{x})) + \sum_{u=1}^{U} \log\left(B_u\right), \quad (1)$$

where $\widetilde{\eta}(\mathbf{x}) = \left(\prod_{u=1}^{U} \eta_u(\mathbf{x})\right)^{\frac{1}{U}}$ is the geometric mean of the spectral efficiency over all UEs. It is clear that any $\mathbf{x}$ maximizing $\widetilde{\eta}(\mathbf{x})$ also maximizes $\sum_{u=1}^{U} \log\left(B_u \eta_u(\mathbf{x})\right)$.

The spectral efficiency that a UE may achieve is influenced by the centralization level of its serving gNB and all other gNBs. If the gNB is highly centralized, the interference received from other gNBs with the same or a higher centralization level would be small, owing to the interference-mitigation techniques that can be applied. Based again on [6], we model the ability of a centralization level to cancel interference by means of the function $c(x) : \mathbb{Q} \mapsto [0, 1]$, which represents the maximum cancellation factor that a gNB with centralization level x may apply to reduce the interference experienced by its served UEs. Using this function and Shannon's formula, we can formulate the spectral efficiency of UE u as [12]:

$$\eta_u(\mathbf{x}) = \log_2\left(\frac{s_u}{\varsigma + \sum_{g=1}^{G} i_{u,g} \cdot c(\min(x_{h_u}, x_g))}\right), \quad (2)$$

where ς is thermal noise power, s_u is the signal power received from its serving gNB, $h_u \in \mathbb{G}$ is the index of its serving gNB, and $i_{u,g}$ is the interference power received from gNB g. Note how this interference is multiplied by $c(\min(x_{h_u}, x_g))$, as the cancellation factor that can be applied is limited by the least centralized gNB. From (2), we can formulate our proportionally-fair performance-maximizing problem as [6]:

$$\max_{\mathbf{x},\mathbf{f}} \sum_{u=1}^{U} \log\left(\eta_u(\mathbf{x})\right), \quad \text{(P0a)}$$

subject to

$$\sum_{e \in \mathbb{E}^+(n)} f_e^g - \sum_{e \in \mathbb{E}^-(n)} f_e^g = \begin{cases} 0 & \forall n \in \mathbb{N} \backslash \{n_0, n_g\} \\ r(x_g) & \text{for } n = n_0 \\ -r(x_g) & \text{for } n = n_g \end{cases} \quad \forall g \in \mathbb{G}, \quad \text{(P0b)}$$

$$\sum_{g=1}^{G} f_e^g \leq \phi_e \qquad \forall e \in \mathbb{E}, \quad \text{(P0c)}$$

$$f_e^g \geq 0 \qquad \forall e \in \mathbb{E}, \forall g \in \mathbb{G}, \quad \text{(P0d)}$$

where $\mathbb{E}^+(n)$ is the set of edges leaving node n, and $\mathbb{E}^-(n)$ is the set of edges entering node n. Constraint (P0b) is the *flow conservation* constraint, which guarantees that the flow entering the DU and leaving the DU is $r(x_g)$ for all gNBs. Constraint (P0c) is the *link capacity* constraint for all links.

Problem (P0) is a mixed-integer non-linear problem (MINLP) which does not follow a standard formulation. In [6], a simpler, approximate reformulation into a mixed-integer linear problem (MILP) is presented. In this formulation, the vector of auxiliary variables $\mathbf{z} = [z_1^1, ..., z_G^1, z_1^2, ..., z_G^Q]$ is added and $\mathbf{x}$ variables are replaced by $\mathbf{y} = [y_1^1, ..., y_G^1, y_1^2, ..., y_G^Q]$ variables by means of the following variable change:

$$x_g = \sum_{q=1}^{Q-1} y_g^q, \quad y_g^q \in \{0,1\}, \quad y_g^q \geq y_g^{q'} \Leftrightarrow q \leq q'. \tag{3}$$

This leads to the following performance-maximizing problem formulation (refer to [6] for the derivation):

$$\max_{\mathbf{y},\mathbf{z},\mathbf{f}} \sum_{q=1}^{Q-1} \sum_{g=1}^{G} z_g^q \tag{P1a}$$

subject to

$$0 \leq z_g^q \leq Z_g^q y_g^q \qquad \forall q \in \mathbb{Q}, \forall g \in \mathbb{G}, \tag{P1b}$$

$$z_g^q \geq \sum_{k=1}^{G} \epsilon_{g,k}^q y_g^q - (1 - y_g^q) Z_g^q \qquad \forall q \in \mathbb{Q}, \forall g \in \mathbb{G}, \tag{P1c}$$

$$z_g^q \leq \sum_{k} \epsilon_{g,k}^q y_g^q \qquad \forall q \in \mathbb{Q}, \forall g \in \mathbb{G}, \tag{P1d}$$

$$y_g^1 \geq y_g^2 \geq ... \geq y_g^{Q-1} \qquad \forall g \in \mathbb{G} \tag{P1e}$$

$$\mathbf{y} \in \{0,1\}^G, \tag{P1f}$$

and (P0b) – (P0d),

where

$$\epsilon_{g,k}^q = \begin{cases} (c(q-1) - c(q)) \left(\sum_{u \in \mathbb{H}_g} \frac{i_{u,k}}{s_u} + \sum_{u \in \mathbb{H}_k} \frac{i_{u,g}}{s_u} \right) & \text{if } g \neq k, \\ 0 & \text{if } g = k, \end{cases} \tag{4}$$

$\mathbb{H}_g = \{u \,|\, h_u = g\}$ is the set of UE indices served by gNB g, and $Z_g^q = \sum_{k=1}^{G} \epsilon_{g,k}^q$. Formulation (P1) is an MILP can be solved quickly by off-the-shelf solvers, which is suitable for dynamic adaptation of the functional split [6].

B. Cost-minimizing formulation

In order to select a centralization vector $\mathbf{x}$ that minimizes cost, we base on the model and approach presented in [7]. There, the cost of operating a 5G RAN featuring configurable functional splits is divided into three components: (i) the cost of instantiating functions at the DUs and CU, (ii) the computational costs of running these functions, and (iii) the cost of routing the resulting flows. The first component κ_{inst} can be calculated as:

$$\kappa_{\text{inst}} = (\delta_{\text{CU}} + \delta_{\text{DU}})\, G, \tag{5}$$

where δ_{CU} (δ_{DU}) is the cost of instantiating the gNB functions at the CU (DU). According to [7], reasonable values for these components are $\delta_{\text{CU}} = 1$ ncu and $\delta_{\text{DU}} = 0.5$ ncu, where ncu stands for "normalized cost units", since absolute cost values are difficult to provide, as operators rarely disclose them.

The second component, the computational cost $\kappa_{\text{comp}}(\mathbf{x})$, can be expressed as a function of the centralization vector $\mathbf{x}$ as follows:

$$\kappa_{\text{comp}}(\mathbf{x}) = \sum_{g=1}^{G} \left(\alpha_{\text{CU}}(x_g)\gamma_{\text{CU}} + \alpha_{\text{DU}}(x_g)\gamma_{\text{DU}} \right) \rho_g, \tag{6}$$

where $\alpha_{\text{CU}}(x_g)$ ($\alpha_{\text{DU}}(x_g)$) is the CPU cycles per Gb/s required to deal with traffic at the CU (DU) with centralization level x_g, γ_{CU} (γ_{DU}) is the cost in ncu per CPU cycle at the CU (DU), and ρ_g is the downlink traffic at gNB g. In our simulation, we use those reference values provided in [7, Table I]. We can apply (3) again to replace $\mathbf{x}$ with $\mathbf{y}$ and yield a linear function of $\mathbf{y}$:

$$\kappa_{\text{comp}}^y(\mathbf{y}) = \beta^0 + \sum_{g=1}^{G} \sum_{q=1}^{Q} \left(\beta_{\text{CU}}(q)\gamma_{\text{CU}} + \beta_{\text{DU}}(q)\gamma_{\text{DU}} \right) \rho_g y_g^q, \tag{7}$$

where

$$\beta^0 = \sum_{g=1}^{G} \left(\alpha_{\text{CU}}(0)\gamma_{\text{CU}} + \alpha_{\text{DU}}(0)\gamma_{\text{DU}} \right) \rho_g, \tag{8}$$

and

$$\beta_{\text{CU}}(q) = \alpha_{\text{CU}}(q) - \alpha_{\text{CU}}(q-1), \;\; \beta_{\text{DU}}(q) = \alpha_{\text{DU}}(q) - \alpha_{\text{DU}}(q-1). \tag{9}$$

Finally, the third component, the routing cost $\kappa_{\text{rout}}(\mathbf{f})$ is calculated as:

$$\kappa_{\text{rout}}(\mathbf{f}) = \sum_{e \in \mathbb{E}^+(n_g)} \omega f_e^g \tag{10}$$

where ω is the average normalized cost per Gb/s over all links.

Therefore, the cost-minimizing functional split selection problem is:

$$\max_{\mathbf{y},\mathbf{f}} \;\; \kappa_{\text{inst}} + \kappa_{\text{comp}}^y(\mathbf{y}) + \kappa_{\text{rout}}(\mathbf{f}) \tag{P2}$$

subject to (P0b)–(P0d) and (P1e)–(P1f). Problem (P2) is an MILP that uses the same variables as (P1), except for the auxiliary $\mathbf{z}$ variables. As a result, it can be easily combined with (P1) if the relationship between performance and revenue is known.

IV. Approach comparison

A. Metrics

We define $(\mathbf{x}_p^*, \mathbf{f}_p^*)$ as the performance-maximizing centralization and flow vectors, respectively, obtained after solving (P1). The cost κ_p^* and spectral efficiency η_p^* of the performance-maximizing approach are defined as $\kappa_p^* \triangleq \kappa_{\text{inst}} + \kappa_{\text{comp}}^y(\mathbf{x}_p^*) + \kappa_{\text{rout}}(\mathbf{f}_p^*)$ and $\eta_p^* \triangleq \widetilde{\eta}(\mathbf{x}_p^*)$, respectively. Similarly, we define is $(\mathbf{x}_c^*, \mathbf{f}_c^*)$ as the cost-minimizing centralization and flow vectors, respectively, obtained after solving (P2). The cost κ_c^* and spectral efficiency η_c^* of the cost-minimizing approach

are defined as $\kappa_c^* \triangleq \kappa_{\text{inst}} + \kappa_{\text{comp}}^y(\mathbf{x}_c^*) + \kappa_{\text{rout}}(\mathbf{f}_c^*)$ and $\eta_c^* \triangleq \widetilde{\eta}(\mathbf{x}_c^*)$, respectively.

We refer to $\xi_\eta \triangleq \eta_p^*/\eta_c^*$ and $\xi_\kappa \triangleq \kappa_p^*/\kappa_c^*$, as the performance and cost ratios, respectively. They quantify how much better the performance of the performance-maximizing approach is when compared to the cost-minimizing approach, and how costly the performance-maximizing solution is with respect to the cost-minimizing approach. Since η_p^* is the maximum spectral efficiency and κ_c^* is the minimum operating cost, it is clear that $\xi_\eta \geq 1$ and $\xi_\kappa \geq 1$.

Finally, we define $\xi_{\eta,\kappa} \triangleq \xi_\eta/\xi_\kappa$ as the performance-to-cost ratio over both approaches. This metric is useful since it quantifies how good the performance improvement of the performance-maximizing approach is when compared to the cost reduction of the cost-minimizing approach. In fact, under mild assumptions this ratio can be used to conclude which approach leads to a superior revenue. Namely, if we consider that the gross network revenue increases linearly[2] with the achieved spectral efficiency, we can express the net revenue R_c (R_p) of the cost-minimizing (performing-maximizing) approach as

$$R_c \approx a\eta_c^* + b - \kappa_c^*, \quad R_p \approx a\eta_p^* + b - \kappa_p^*, \tag{11}$$

for some arbitrary a and b. Then, $R_c < R_p$ is equivalent to $a\eta_c^* - \kappa_c^* < a\eta_p^* - \kappa_p^*$. If $a\eta_c^* - \kappa_c^* > 0$ and $\xi_{\eta,\kappa} > 1$, the following holds:

$$a\eta_c^* - \kappa_c^* < a\xi_\kappa\eta_c^* - \xi_\kappa\kappa_c^* + a(\xi_\eta - \xi_\kappa)\eta_c^* \tag{12}$$

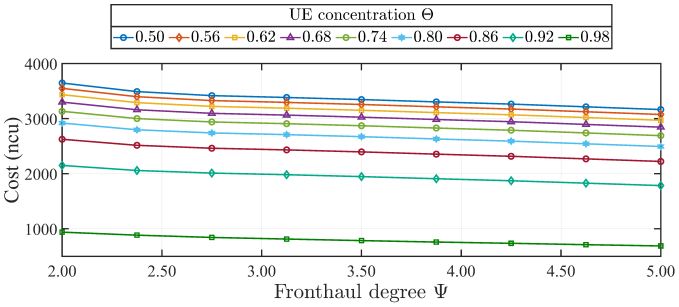
$$= a\eta_p^* - \kappa_p^* \tag{13}$$

Thus, in that case, $\xi_{\eta,\kappa} > 1$ ($\xi_{\eta,\kappa} < 1$) is a sufficient (necessary) condition for a performance-maximizing (cost-minimizing) network to be more profitable than a cost-minimizing (performance-maximizing) approach. The same reasoning can be applied for the case $a\eta_c^* - \kappa_c^* < 0$, which yields reciprocal conclusions. As a result, the value of $\xi_{\eta,\kappa}$ can be used as a strong indicator of the potential superiority of either approach.

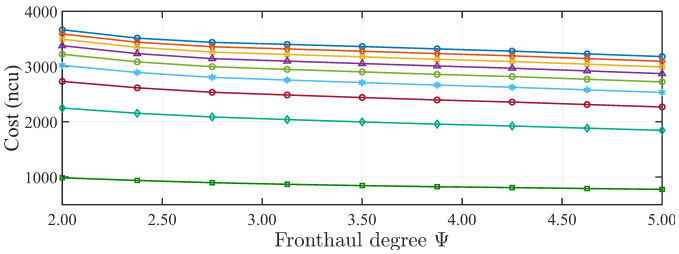
B. Simulation setup

In order to evaluate the performance and cost of both approaches under realistic conditions, simulations are performed for a network with $G = 300$ gNBs and $U = 3000$ UEs spread over an area of 2.6 km^2, corresponding to the center of medium-sized city. We choose $Q = 4$ centralization levels and $c(x) = \{1, 0.6, 0.2, 0.01\}$ for $x = \{0, 1, 2, 3\}$, respectively as suggested in [4]. The capacity of links in the fronthaul network is $\phi_e = 1$ Tb/s $\forall e \in \mathbb{E}$. Maximum required capacity of each split is $r(x) = \{4, 8, 80, 160\}$ for $x = \{0, 1, 2, 3\}$ as stated in [3]. The fronthaul network degree Ψ and UE concentration index Θ are kept as variables during the simulations, so as to evaluate their impact in the performance-to-cost ratio. For each (Ψ, Θ) pair, the simulation is repeated 300 times to ensure statistically tight results.

[2]This can be justified, for instance, by the relatively small range of spectral efficiencies that can be achieved by either approach.



(a) Mean operating cost κ_c^* of the cost-minimizing approach.



(b) Mean operating cost κ_p^* of the performance-maximizing approach.

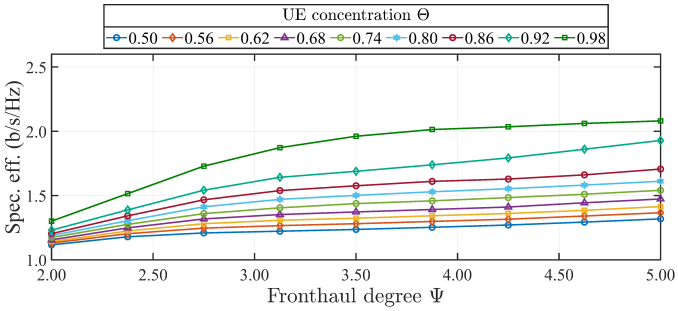
Figure 3: Mean operating cost of cost-minimizing and performance-maximizing approaches.

C. Cost comparison

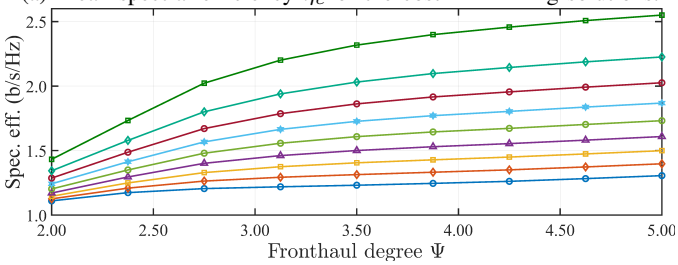
Fig. 3 shows the mean operating cost, in normalized cost units, of running a 5G RAN under cost-minimizing (Fig. 3a) and performance-minimizing (Fig. 3b) approaches, for fronthaul network degrees ranging from $\Psi = 2$ to $\Psi = 5$ and UE concentration indices ranging from $\Theta = 0.5$ to $\Theta = 0.98$. We observe that the operating cost decreases linearly as the fronthaul network degree increases. This is due to the fact that a denser fronthaul network allows for more RAN functions to be centralized, which reduces the operating cost. This trend holds for every UE concentration value and both approaches, resulting in an average operating cost reduction of 13% to 22% at $\Psi = 5$ with respect to $\Psi = 2$. In addition, the operating cost also decreases consistently as the UE concentration index increases. The explanation is that, when users are concentrated, the majority of the network's activity comes from the reduced subset of gNBs whose DUs are close to the user clusters. Thus, centralizing these gNBs leads to substantial cost reductions, which becomes more effective the more concentrated the users and, hence, the less gNBs are involved in serving these UEs. Namely, the operating cost is 74% to 78% lower when $\Theta = 0.98$ than when $\Theta = 0.5$. Finally, we observe that the operating cost of the performance-maximizing approach is only marginally worse than that of the cost-minimizing approach, as κ_p^* is less than 4.5% higher than κ_c^* in all cases.

D. Performance comparison

In Fig. 4 we show the geometric mean of the spectral efficiency achieved by the cost-minimizing (Fig. 4a) and performance-maximizing (Fig. 4b) approaches. We observe



(a) Mean spectral efficiency η_c^* of the cost-minimizing solutions.



(b) Mean spectral efficiency η_p^* of the performance-maximizing solutions.

Figure 4: Mean spectral efficiency of cost-minimizing and performance-maximizing approaches.

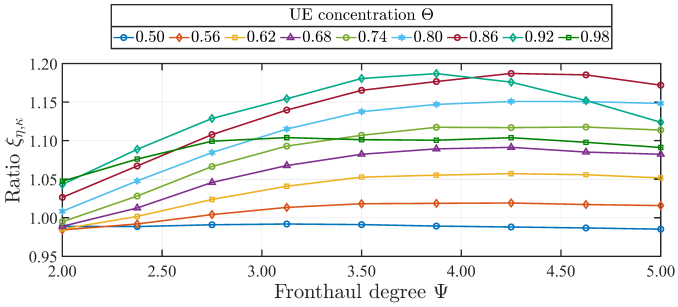


Figure 5: Mean performance-to-cost ratio $\xi_{\eta,\kappa}$.

that the spectral efficiency increases steadily with the fronthaul network degree, since a denser networks allows for more centralized RAN functions and thus better interference management. In addition, it is clear that the higher the UE concentration, the higher achieved the spectral efficiency, owing to the lower number of involved gNBs. Finally, we observe that the spectral efficiency of the cost-minimizing approach is clearly smaller than that of the performance-maximizing approach. This is specially noticeable for dense fronthaul networks and concentrated users. For example, at $\Psi = 2.75$ and $\Theta = 0.62$, $\eta_c^* = 1.29$ b/s/Hz and $\eta_p^* = 1.33$ b/s/Hz (a 3.5% improvement), but at $\Psi = 5$ and $\Theta = 0.98$, $\eta_c^* = 2.08$ b/s/Hz and $\eta_p^* = 2.54$ b/s/Hz (a 22.4% improvement).

E. Performance-to-cost ratio comparison

In Fig. 5 we show the performance-to-cost ratio $\xi_{\eta,\kappa}$ for $\Psi = 2$ to $\Psi = 5$ and $\Theta = 0.5$ to $\Theta = 0.98$. We observe that $\xi_{\eta,\kappa} > 1$ for most points, except when the UEs are uniformly distributed over the coverage area ($\Theta = 0.5$). This implies that the performance improvement achieved by the performance-maximizing approach outmatches its suboptimal cost when the UEs are only slightly clustered. In addition, even at $\Theta = 0.5$ the cost-minimizing approach is only marginally better than the, whereas when UEs are moderately concentrated the performance-to-cost ratio reaches $\xi_{\eta,\kappa} \approx 1.19$. This trend does not continue, however, for highly concentrated UEs if the fronthaul network is sufficiently dense, although the performance-maximizing approach remains always superior.

V. Conclusion

The ability to adapt its centralization level to the network conditions is regarded as a highly desired feature of a 5G or 6G RAN. However, partial centralization can be separately motivated by either cost reduction or performance improvement, and the advantages of selecting one goal over the other may be unclear. In this work, we provide a comprehensive comparison between both approaches. We present them with a unified formulation and perform simulations over a wide range of network conditions. We observe that the performance-maximizing approach offers a performance improvement that is proportionally higher to the optimality gap of its cost for almost all network conditions.

References

[1] NGMN Alliance, "5G white paper," Next Generation Mobile Networks, White paper, 2015.

[2] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen, "Evaluating C-RAN fronthaul functional splits in terms of network level energy and cost savings," *IEEE Journal of Communications and Networks*, vol. 18, no. 2, pp. 162–172, 2016.

[3] 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.

[4] A. Martínez Alba and W. Kellerer, "A Dynamic Functional Split in 5G Radio Access Networks," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[5] T. Taleb, R. L. Aguiar, I. Grida Ben Yahia, B. Chatras, G. Christensen, U. Chunduri, A. Clemm, X. Costa, L. Dong, J. Elmirghani *et al.*, "White paper on 6G networking," 6G Research Visions no. 6, University of Oulu, White paper, 2020.

[6] A. Martínez Alba, S. Janardhanan, and W. Kellerer, "Enabling dynamically centralized RAN architectures in 5G and beyond," *IEEE Transactions on Network and Service Management*, 2021, early access.

[7] A. Garcia-Saavedra, G. Iosifidis, X. Costa-Perez, and D. J. Leith, "Joint optimization of edge computing architectures and radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2433–2443, 2018.

[8] A. Martínez Alba, P. Babarczi, A. Blenk, M. He, P. Krämer, J. Zerwas, and W. Kellerer, "Modeling the Cost of Flexibility in Communication Networks," in *IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2021.

[9] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith, and G. Iosifidis, "FluidRAN: Optimized vRAN/MEC orchestration," in *IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2018, pp. 2366–2374.

[10] 3GPP, "Study on scenarios and requirements for next generation access technologies," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.913, 07 2020, version 16.0.0.

[11] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, 1988.

[12] A. Martínez Alba, S. Janardhanan, and W. Kellerer, "Dynamics of the flexible functional split selection in 5G networks," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2020.

[13] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Communications letters*, vol. 9, no. 3, pp. 210–212, 2005.