Optimization over time of reliable 5G-RAN with network function migrations

Optimization over Time of Reliable 5G-RAN with Network Function Migrations

Nicola Di Cicco[a], Federico Tonini[b], Valentina Cacchiani[c], Carla Raffaelli[c]

[a]*Department of Electronics, Information and Bioengineering (DEIB), Polytechnic University of Milan, Via Giuseppe Ponzio, 34, Milan, 20133, Italy*
[b]*Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden*
[c]*Department of Electrical and Information Engineering (DEI), University of Bologna, Viale Risorgimento, 2, Bologna, 40136, Italy*

Abstract

Resource optimization in 5G Radio Access Networks (5G-RAN) has to face the dynamics over time in networks with increasing numbers of nodes and virtual network functions. In this context, multiple objectives need to be jointly optimized, and key application requirements such as latency must be enforced. In addition, virtual network functions realizing baseband processing are subject to failures of the cloud infrastructure, requiring an additional level of reliability. Overall, this is a complex problem to solve, requiring fast algorithms to cope with dynamic networks while avoiding resource overprovisioning. This paper considers the problem of optimal virtual function placement in 5G-RAN with reliability against a single DU Hotel failure and proposes a solution that takes service dynamics into account. Firstly, the joint optimization of the total number of DU Hotels, of the RU-DU latency and of the backup DU sharing in a static traffic scenario is considered, and the DUOpt algorithm, based on Lexicographic Optimization, is proposed for solving efficiently this multi-objective problem. DUOpt splits the multi-objective problem into smaller Integer Linear Programming (ILP) subproblems that are sequentially solved, adopting for each one the most effective methodology to reduce the total execution time. The proposed DUOpt algorithm is extensively benchmarked to show its effectiveness in optimization of medium to large size networks: in particular, it is shown to greatly outperform an aggregate multi-objective approach, being able to compute optimal or close to optimal solutions for networks of several tens of nodes in computing times of a few seconds. Then, the problem is extended to a dynamic traffic scenario in which optimization is performed over time. In this context, in addition to the aforementioned objectives, the total number of network function migrations induced by multiple reoptimizations must be kept to the minimum. For solving efficiently this problem the DUMig algorithm is proposed, which extends and improves DUOpt. Reoptimization over a time horizon of one day in an illustrative dynamic traffic scenario is performed to evaluate the proposed DUMig algorithm against DUOpt, the latter being oblivious of the traffic dynamics. DUMig shows remarkable savings in the total number of migrations (above 86.1% for primary virtual functions and 83% for backup virtual functions) compared to DUOpt, while preserving near-optimal resource assignment.

Keywords: Reliable 5G-RAN, Lexicographic optimization, Network Function Migrations.

1. Introduction

Access networks are nowadays evolving towards a set of interconnected segments, possibly based on different communication technologies, spanning from the radio access to the high capacity core, through passive or active optical transport network solutions [1]. Virtual infrastructures are configured on top of these high-capacity networks with the aim of offering the flexibility required by the dynamic behaviour of the served applications in an efficient way.

The reconfiguration capability of virtual network functions placement in transport network nodes is provided by orchestration and management capabilities developed according to the Software Defined Networking (SDN) paradigm. Many degrees of freedom are offered by the possibility to configure the virtual infrastructure on top of the optical transport network. In fact, by exploiting the network connectivity and the availability of distributed processing power, optimized design can be obtained in relation to objectives such as total deployment cost or power efficiency. At the same time, constraints that arise from application performance requirements need to be met, such as the latency and reliability requirements for 5G and beyond scenarios [2, 3]. In addition, efficient strategies are needed to automatically optimize network resources during network operation. Therefore, a revolutionary change is expected in the management and orchestration capability to cope with this extremely complex and pervasive network and service scenario [4], towards the so-called zero-touch network management [5].

In this paper, the 5G Radio Access Networks (5G-RAN) segment is considered as the network segment that forwards the traffic to/from the antennas from/to the transport network nodes. 5G-RAN consists of three units, namely radio, distributed, and central units (RU, DU, CU, respectively) interconnected by RU-DU fronthaul and DU-CU midhaul network segments. Bandwidth and latency requirements are particularly critical for the fronthaul segment that needs to be properly designed; suitable placement of DU and CU functionalities allows to relax the main constraints on network capacity and maximum

latency. DU functionalities are usually located in a larger number of simpler distributed nodes with respect to fewer data centers where CUs are preferably located. The fronthaul segment between DU and RU is also the most constrained in terms of latency and demanding in terms of transmission capacity. Therefore, we focus here on DU functionalities.

Enhanced flexibility in 5G-RAN is represented by the introduction of SDN-controlled Network Function Virtualization (NFV), including DU RAN functions, which can be suitably located in virtualized Hotels to optimize cost and energy consumption in relation to reliability, latency, and bandwidth requirements [6]. Standardization bodies are actively working on different options for the functional split and in the definition of the related requirements, depending on network and service needs. For example, the O-RAN Alliance is currently working on an open interface for the RU/DU split that takes advantage of the NFV capabilities [7].

On the one hand, 5G-RAN virtualization offers several advantages in managing and operating mobile networks. On the other hand, virtualization makes the network prone to failures of the cloud infrastructure. Therefore, reliability represents a crucial aspect that involves also virtualized network functions and that is typically faced by adding extra resources for backup against failures (e.g., of a DU Hotel) with consequent additional costs and further constraints in the optimization model, making it computationally harder to solve.

Finally, the network function placement must be adapted to the dynamics of the traffic in the transport network. To this end, the optimization methods must attain computational times short enough to cope with the traffic change rate [8]. The obtained solutions can include potential migrations of network functions, that represents an additional cost that also needs to be accounted for along with the other objectives.

Overall, the optimal design of a virtual infrastructure for 5G-RAN is in general a multi-objective optimization problem which needs to be solved for medium to large size networks, namely in the order of several tens of nodes. This often leads to comprehensive but exceedingly complex problem formulations in which several, possibly competing objectives are optimized at the same time, typically by assigning them arbitrarily defined priority weights. However, specific application scenarios might implicitly require a well-defined priority ordering among objectives. Leveraging these practical considerations opens up the possibility of smarter optimization strategies.

In this paper, a time-efficient algorithm for the DU Hotel placement problem with reliability against single DU Hotel failure is proposed. Reliability is achieved by assigning RUs to primary and backup DUs hosted in distinct DU Hotels. We first consider, in a static traffic scenario, joint optimization of the total number of DU Hotels, of the total RU-DU distance and of the backup DU sharing, under maximum RU-DU distance and link capacity constraints. The DUOpt algorithm, which is based on Lexicographic Optimization, is proposed for efficiently solving this multi-objective problem. DUOpt splits the large multi-objective Integer Linear Programming (ILP) problem into several smaller single-objective ILP subproblems, which are solved sequentially according to their application-defined priority ordering. The bottlenecks in the optimization process are discussed, and a hybrid approach, partially based on local search, is described, which extends and improves previous contributions [9]. An extensive comparison with a classical aggregate multi-objective optimization approach is presented both in regular lattice networks and non-regular ones, showing the effectiveness of the lexicographic algorithm in solving the multi-objective problem in short computing times. Secondly, the problem is cast to a dynamic traffic scenario, for which the DUMig algorithm is proposed. DUMig, together with the aforementioned objectives, accounts for the cost of network functions migrations induced by multiple reconfigurations over time. Numerical evaluations show that DUMig is able to maintain near-optimal resource assignment with respect to DUOpt, with significant gains in terms of both network function migrations and computing times.

This paper is organized as follows: in Section 2 previous works on 5G-RAN optimization are presented and discussed in relation to the contribution of this paper. In Section 3 the 5G-RAN optimization problem to be solved is defined both in the static and in the dynamic traffic scenarios. In Section 4 the lexicographic algorithm to solve the optimal DU Hotel placement problem and the models of the corresponding subproblems are described. In Section 5 the extension of the algorithm to a dynamic traffic scenario, in order to deal with DU migrations, is introduced. In Section 6 results are reported and discussed in different scenarios to outline the effectiveness of the proposed algorithm, and to show its performance in a dynamic scenario over a time horizon of one day. Section 7 reports the conclusions of the work and addresses some open aspects. In Appendix Appendix A the aggregate model for the optimal DU Hotel placement problem proposed as a reference is reported.

2. Related works

The Cloudification of the RAN was firstly introduced in [10] with the term C-RAN. In C-RAN, baseband processing functions are centralized in selected locations (called hotels) and virtualized on general-purpose hardware, to achieve better performance and cost savings. However, this imposes extreme requirements on the transport network interconnecting radio and baseband units (called RRU and BBU, respectively). To cope with this, 3GPP recently proposed an evolution of the C-RAN concept where the different functions of the 5G new radio (NR) stack are divided into three parts: the Radio Unit (RU), the Distributed Unit (DU), and the Central Unit (CU) [11]. In this view, the DU and CU can be located (in Hotels) and virtualized to achieve better performance and cost savings. Several split options of the 5G NR functions are possible, resulting in different functions performed in the different units and, consequently, different requirements on the transport network [6]. The functional split between DU and CU relaxes the bandwidth and latency requirements over the transport network, allowing to use statistical multiplexing and to locate CUs in deep network nodes (e.g., a core data center), traversing high-capacity optical rings with inherent protection of the traffic. Conversely, the split of physical layer resources between RU

and DU imposes strict latency requirements ($\sim 100\mu s$), hence lower centralization, and usually requires the assignment of expensive dedicated high capacity resources to transport the data (e.g., dedicated wavelengths in a circuit transfer mode). Due to these requirements, the DUs are usually located in the access network, close to the RUs, with limited centralization gains. Therefore, optimizing the assignment of network resources and the placement of the DUs is crucial to contain the network cost. The focus of the paper is thus on the RU-DU split.

5G-RAN resource optimization is a topic of great interest, well-studied in the recent literature. In the following, several works considering network and processing resource optimization in static traffic conditions, i.e., with traffic not changing over time, are outlined. In this context, optimization is regarded as a planning problem in which several objectives (e.g., network energy consumption, spectrum utilization, reliability indicators) are typically optimized.

In [12] an Integer Linear Programming (ILP) model for 5G-RAN cost optimization is proposed and solved, aimed at minimizing the total network cost. Authors develop a multi-objective optimization problem taking into account the cost of baseband as well as electronic switches placement, and the total fiber utilization. The achieved centralization is evaluated on different transport options. Authors conclude that independently placing electronic switches allows for more efficient wavelength usage and higher centralization.

In [3] an ILP model for BBU Hotel placement with reliability against single link failures is proposed. Authors propose three different approaches, respectively based on dedicated backup links, on dedicated backup BBU Hotels and partial link sharing between RUs. The objective is the minimization of the total number BBU Hotels and of the wavelength utilization. Authors evaluate the proposed approaches on a reference network, and discuss the trade-off between the cost of resource redundancy and achieved reliability.

In [13] an architecture for C-RAN using Time-Wavelength Division Multiplexing Passive Optical Networks (TWDM-PON) as fronthaul is proposed. The objective is the minimization of the total energy consumption in a dynamic traffic scenario, given the possibility of deactivating virtual DU resources when not anymore needed. Authors propose an ILP formulation and develop a heuristic algorithm, motivated by poor ILP scalability to large networks. Results show that the proposed heuristic is able to achieve near-optimal solutions in a static traffic scenario and suboptimal solutions in a dynamic traffic scenario, with significant savings compared to peak-based dimensioning.

In [14] a robust optimization model for probabilistic protection in a cloud provider against multiple types of failures. Namely, the authors consider three survivability parameters pertaining to CPU, memory and the entire cloud provider considering both CPU and memory. The probabilistic protection formulation is transformed into a MILP problem via Robust Optimization, and solved via a Lexicographic method. Numerical results show that this approach can allocate backup resources in a more efficient way with respect to traditional approaches.

Several works in the literature extended the 5G-RAN optimization problem to a dynamic traffic scenario, i.e., with traffic evolving over time. In this context, one needs to perform periodic optimization over time, as old solutions might either become unfeasible due to a traffic spike or resource-wasteful in case of a traffic decrease. Moreover, network function migrations might have to be performed, whose cost needs to be accounted for in the optimization objectives. In the following, we outline several works on network and resource optimization in dynamic traffic scenarios.

In [15] a distributed heuristic for BBU Hotel placement with support for single BBU Hotel failure reliability was developed, since the proposed ILP centralized approach would not scale beyond networks of 30 nodes. The heuristic assumes that transport nodes have information regarding their neighbors only, therefore requiring exchange of information between transport nodes. The optimality gap with respect to an ILP model is thoroughly assessed showing some degree of suboptimality, especially for larger networks. However, the proposed distributed heuristic outperforms ILP in terms of number of network function migrations in dynamic scenarios.

In [16] a heuristic algorithm, based on Branch-and-Bound and Simulated Annealing, to solve the BBU Hotel placement problem, minimizing the overall link delays, given by the sum of propagation and processing delays. The development of the heuristic is motivated by poor ILP scalability to large networks. Moreover, in a dynamic traffic scenario, authors fine-tune the developed heuristic in order to find the optimal trade-off between migration delays and link delays.

In [8] an algorithm for dynamic slice provisioning based on traffic predictions is proposed. The objective is to minimize the migrated traffic and the slice degradation, the latter proportional to the slice priorities. Authors develop an algorithm that adjusts an overprovisioning resource margin based on a traffic prediction model, therefore reducing the need for network functions migrations when the traffic increases. Authors evaluate the trade-off between prediction-based resource overprovisioning and slice degradation, the latter induced by resource competition between different slices.

In [17] a heuristic algorithm is developed for solving the BBU Hotel migration and wavelength reassignment in Cloud-Fog RAN. In particular, the possibility of migrating BBU Hotels from Cloud to Fog nodes in case of increasing traffic, and vice versa, is investigated. The developed heuristic is based on the linear relaxation of the ILP, building a feasible solution based on the largest relaxed integer variables. The authors discuss the trade-offs between the overall network energy consumption and the service interruption probability.

In [18] an algorithm for dynamic placement of CU/DU Hotels is developed. The proposed approach takes into account the power consumption of CU/DU Hotels and the network congestion, in order to decide whether to centralize CUs or distribute them in the network. Experimental results show that the developed algorithm exhibits the best trade-off between power consumption and blocking probability, with respect to the considered benchmarks.

Differently from previous literature, this paper addresses a multi-objective DU Hotel placement problem with resiliency against single DU Hotel failures. In particular, our approach

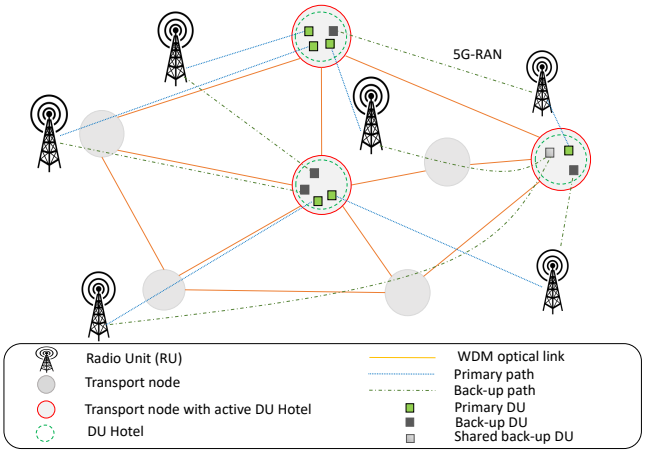

Figure 1: Scheme of a 5G-RAN with primary and backup DUs in DU Hotels.

aims to jointly optimize the following objectives: total number of DU Hotels, total RU-DU distance and the total number of backup DUs, under maximum RU-DU distance and link capacity constraints. Previous works either do not include the latency among the optimization objectives, or do not consider resilience against DU Hotel failures. To cope with the additional complexity introduced by jointly optimizing multiple competing objectives, a novel optimization approach based on Lexicographic Optimization is developed. To the best of the authors' knowledge, there are no other works in 5G-RAN optimization leveraging Lexicographic Optimization. The problem and the proposed methodology are extended to a dynamic traffic scenario, in which the number of network functions migrations is also optimized together with the aforementioned objectives.

This work extends [9] in several aspects. A fast and effective heuristic addressing the computational bottlenecks in the proposed methodology is developed. More extensive analyses of the advantages of the lexicographic approach in non-regular networks is presented. Above all, the problem is extended to a dynamic traffic scenario, performing optimization over time.

3. Problem statement for reliable 5G-RAN

The reference 5G-RAN architecture is illustrated in Fig. 1. A geographical area is covered by a set of RUs connected to a set S of transport nodes in a transport network. Each transport node is assumed to have computing capabilities (e.g., the one provided by an edge data center) to host virtualized DU functions according to service needs. In the following, a node that hosts at least one virtualized DU function will be interchangeably referred to as an active node or DU Hotel. Active nodes are modelled as binary variables A_j, either 1 or 0 if node $j \in S$ is active or not, respectively. Transport nodes are interconnected via lightpaths implemented as different wavelengths in Wavelength Division Multiplexing (WDM) optical fibers, forming the fronthaul network segment of the 5G-RAN. The set of optical links is denoted as L and each link is assumed to have up to M_W wavelengths. Due to the strict fronthaul latency requirements, circuit transfer mode is applied, therefore no queuing delay is present. In the following, it will be assumed that transport nodes can be equipped with wavelength converters.

DU Hotels are configured in the transport nodes in order to host virtual DU functions of different RUs. In case of failure of a DU Hotel (e.g., as a consequence of a power outage), all the DUs hosted in the failed DU Hotel cannot be reached by the RUs, leaving many users without service access. To achieve 5G-RAN reliability against single DU Hotel failure, one primary and one backup DU are assigned to each RU of the 5G-RAN and located in different DU Hotels at different transport nodes. DUs at nodes $j \in S$ for RUs hosted at nodes $i \in S$ are modelled as binary decision variables p_{ij} and b_{ij} for primary and backup DUs, respectively. In case of failure of the DU Hotel hosting a primary DU, the corresponding RU is assigned to its backup DU, which would in principle require to double the overall number of DUs. The number of backup DUs at node $j \in S$ is modelled as integer decision variables y_j. To achieve cost savings, the overall number of DUs can be reduced thanks to DU sharing, where a single backup DU can be shared by two (or more) RUs. In particular, a backup DU can be shared by two RUs if they are assigned to primary DUs hosted by distinct DU Hotel. In case of a DU Hotel failure, only one of the two RUs is affected and switches to the backup DU until the damage to the DU Hotel is repaired. In this case, only one backup DU is required instead of two. The assignment of backup DUs also needs to account for the backup wavelengths over optical links. An example of DU sharing is shown in Fig. 1. Note that backup DU sharing introduces a trade-off between total DU Hotels and total backup DUs. If DU Hotels are few, more RUs will share the same primary DU Hotel, therefore allowing for less backup DU sharing configurations.

5G-RAN optimization aims at minimizing the number of transport nodes that need to host DU functions by centralizing them in DU Hotels. This allows to achieve sizeable gains in terms of energy consumption and network management, since not all transport nodes need to host DU Hotels at the same time [6]. However, a higher centralization introduces higher delay, since the distance between DU Hotels and their assigned RUs increases. Therefore, 5G-RAN optimization must constrain the distance between DU Hotels and RUs (h_{ij}) to be lower than a target value (M_H), in order to limit the latency to the maximum allowed by the specific functional split [11]. In a dynamic traffic scenario, the number of RUs required varies over time depending on the traffic requirements, optimization needs to be over time during the network operation. Re-optimization may result in a new configuration of the DU Hotels, requiring virtual DU functions to migrate.

In the following, optimal DU Hotel placement both with and without migrations optimization will be considered. Optimally solving this problem without taking migrations costs into account will yield the optimal DU placement configuration for a given traffic state. Conversely, by optimizing also migrations, a trade-off needs to be made with respect to the optimal DU placement.

3.1. DU Hotels placement optimization

The problem statement for optimal DU Hotel placement can be defined as follows: given the network topology and the number of RUs per transport node at the current time instant ($r_{i,t}$), find the optimal DU Hotel assignment minimizing the number of active nodes, the total distance between DU Hotels and their assigned RUs, and the number of backup DUs. The latter can be shared by RUs assigned to distinct DU Hotels. Constraints must impose that DU Hotels are properly deployed in a redundant way ensuring primary and backup DU Hotels for each RU, that a maximum distance between DU Hotels and RUs is not exceeded, and that the capacity of the fronthaul WDM links is not exceeded as well.

3.2. DU Hotels placement and Migrations optimization

A straightforward solution method to deal with a dynamic traffic scenario is to repeatedly call the optimal DU Hotels placement algorithm in order to reconfigure the network according to the traffic changes. On the other hand, this approach would not consider the network configuration previously determined, thus potentially causing high reconfiguration costs. Due to that, a cost proportional to the total number of network functions migrations, induced by the virtual infrastructure reconfigurations, needs to be accounted for. To do so, information on the resource assignment at each time instant is needed, i.e., the set of inactive and active nodes ($S_{0,t}$ and $S_{1,t}$, respectively) and the set of inactive primary and backup assignments ($P_{0,t}$ and $B_{0,t}$).

Therefore, the problem statement for the optimal DU placement with migrations optimization can be defined as follows: given the network topology, the number of active RUs per transport node at the current time instant, and the resource assignment at the current time instant, determine a new feasible resource assignment that minimizes the displacement with respect to the current resource assignment, while deactivating as many nodes as possible if no longer needed. The rationale is to keep under control the total number of migrations, while ensuring near-optimal resource assignments. Numerical evaluations will show the effectiveness of the developed algorithm in attaining such goal.

4. Lexicographic Optimization for DU Hotel Placement

A generic multi-objective optimization problem can be expressed as follows:

$$\begin{aligned} &\min \mathbf{f}(\mathbf{x}) = \min\left(f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\right) \\ &\text{subject to} \quad \mathbf{x} \in \mathbf{X} \end{aligned}$$

where $\mathbf{X}$ denotes the feasible set. Without domain knowledge, optimizing at the same time multiple objective functions can be a challenging task. However, one can take advantage from the fact that in realistic application scenarios objectives are not of equal importance. On that note, Lexicographic optimization has proven to be an effective tool for solving challenging optimization problems in communication systems [19, 20, 14].

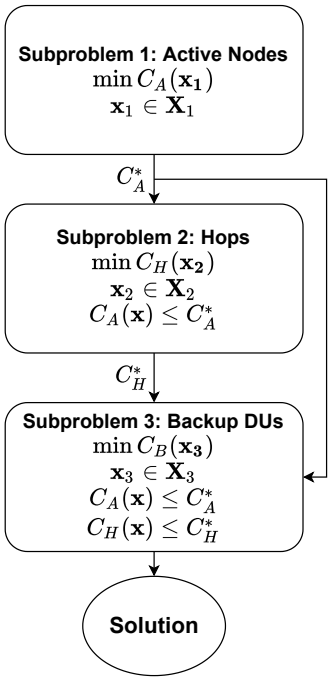

Figure 2: DUOpt flow diagram. Symbols $\mathbf{x}_i$ and $\mathbf{X}_i$ indicate the variables and the feasible set, respectively, for the i-th subproblem.

Assume that objectives are ranked in importance such that minimization of $f_i(\mathbf{x})$ is infinitely more important than minimization of $f_{i+1}(\mathbf{x})$, $i = 1, \ldots, n-1$. Objectives ranked in this way are said to be in lexicographic ordering. The Lexicographic method consists in solving a sequence of n single-objective subproblems in the following form:

$$\begin{aligned} &\min f_j(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{x} \in \mathbf{X} \\ &\qquad\qquad f_i(\mathbf{x}) \leq f_i^* \quad \forall i < j \end{aligned}$$

where f_i^* is the optimal solution value for the i-th single-objective problem. Objectives are sequentially optimized in a pre-defined priority order, and each single-objective subproblem is constrained so that the solution values found in the higher priority subproblems are not worsened. The final solution for the multi-objective problem is achieved when the objective with lowest priority is optimized.

The rationale is that, instead of having to solve one "difficult" problem, one can solve a sequence of potentially "easier" subproblems, with overall smaller computing times. Moreover, intermediate solutions can be used to provide a "warm start" (i.e., an initial integer feasible solution) for solving a subsequent subproblem, in order to further speed up the optimization process. In addition, more degrees of freedom are available on how to tackle each individual subproblem (e.g., by employing an efficient heuristic). Finally, this approach bypasses the use of large weights in the cost function, which may be cause of numerical issues [21].

The DU Hotel placement problem introduced in Section 3 is a multi-objective optimization problem, consisting in the minimization of the total active nodes, total RU-DU distance and

Table 1: Model parameters and variables.

	Parameters
S	Set of transport nodes.
L	Set of links.
T	Time horizon.
$S_{0,t}$	Set of inactive nodes at time t, $t \in [0, T]$.
$S_{1,t}$	Set of active nodes at time t, $t \in [0, T]$.
$P_{0,t}$	Set of inactive primary assignments at time t, $t \in [0, T]$.
$B_{0,t}$	Set of inactive backup assignments at time t, $t \in [0, T]$.
h_{ij}	Distance in hops between nodes i and j computed with the shortest path, $i, j \in S$
α	Weight for the active nodes in the cost function.
μ_P	Weight for primary DU function migrations in the cost function.
μ_B	Weight for backup DU function migrations in the cost function.
β	Weight for the distance in the cost function.
γ	Weight for the backup DU Hotels in the cost function.
$r_{i,t}$	Number of active RUs at site i at time t, $i \in S$, $t \in [0, T]$.
δ_{ij}^{l}	1 if the shortest path between i and j uses link l, 0 otherwise, i, $j \in S$, $l \in L$
M_W	Maximum number of wavelengths available in each link.
M_H	Maximum allowed distance in hops between RUs and DUs.
	Variables
A_j	1 if node j is active (i.e., hosts a DU Hotel), $j \in S$, 0 otherwise
p_{ij}	1 if the DU Hotel at node j is assigned as primary for RUs at node i, i, $j \in S$, 0 otherwise.
b_{ij}	1 if the DU Hotel at node j is assigned as backup for RUs at node i, i, $j \in S$, 0 otherwise.
y_j	Total backup DUs hosted at node j, $j \in S$
$c_{ijj'}$	1 if RRUs at node i are using the DU Hotel at node j as primary the DU Hotel at node j' as backup, i, j, $j' \in S$, 0 otherwise.

total backup DUs. In this model, the RU-DU distance is expressed in hops. This is because links in access networks span geographical distances similar enough not to have significant differences in propagation delays. Still, other distance metrics can be handled in the model.

To solve this problem, a priority ordering was identified in previous literature, related to the application scenario [15]. Firstly, energy consumption accounts for up to 40% of network operational costs (OPEX) and is projected to increase [22, 23], therefore minimizing the total number of active nodes is of primary importance. If hops minimization had higher priority than the active nodes, all transport nodes would host DU Hotels and the overall number of hops would be trivially minimized. Similarly, this would happen if backup DU sharing had higher priority than the minimization of the number of active nodes, since the number of shareable backup DUs increases with the number of active nodes. Therefore, minimization of the number of active nodes needs to have the highest priority. Minimization of the total hops is given higher priority than the backup DU sharing, since it allows to reduce the average delay and the wavelength utilization on the WDM fronthaul links. Therefore, the overall optimization algorithm, referred in the following as DUOpt, consists of three sub-problems to be solved in sequence, with results of each subproblem conditioning the optimization of the following ones. A flow diagram illustrating the lexicographic algorithm is illustrated in Fig. 2.

The parameters and decision variables illustrated in the following ILP formulations are reported in Table 1.

4.1. DUOpt-1: Minimization of total active nodes

This subproblem is used to determine the optimal number of active nodes in the transport network at time instant $t \in [0, T]$. The ILP model solved in this subproblem reads as follows:

$$\min C_A = \sum_{j \in S} A_j \tag{1}$$

$$\sum_{j \in S} p_{ij} = 1 \quad \forall i \in S \tag{2}$$

$$\sum_{j \in S} b_{ij} = 1 \quad \forall i \in S \tag{3}$$

$$p_{ij} + b_{ij} \leq A_j \quad \forall i, j \in S \tag{4}$$

$$(p_{ij} + b_{ij}) \cdot h_{ij} \leq M_H \quad \forall i, j \in S \tag{5}$$

$$\sum_{i \in S} \sum_{j \in S} (p_{ij} + b_{ij}) \cdot \delta_{ij}^{l} \cdot r_{i,t} \leq M_W \quad \forall l \in L \tag{6}$$

$$A_j \in \{0, 1\} \quad \forall j \in S \tag{7}$$

$$p_{ij} \in \{0, 1\} \quad \forall i \in S, j \in S \tag{8}$$

$$b_{ij} \in \{0, 1\} \quad \forall i \in S, j \in S \tag{9}$$

Objective function (1) minimizes the total number of active nodes, which is the objective of highest priority. Constraints (2) and (3) impose that each RU has a primary and a backup DU Hotel assigned. Constraints (4) are used both to count the active DU Hotels and to impose that RUs are assigned to distinct primary and backup DU Hotels, since the left-hand side of the inequality is restricted to be at most equal to 1. Constraints (5) impose that the distance between RUs and their assigned primary and backup DU Hotels does not exceed M_H. Constraints (6) impose that the maximum number of wavelengths M_W in each WDM fronthaul link is not exceeded. Finally, constraints (7)-(9) define the domain of decision variables A_j, p_{ij} and b_{ij}.

Notably, in this subproblem variables y_j and $c_{ijj'}$, related to backup DU sharing, are not present neither in the objective function or in any constraint, since they are not currently being optimized: this allows to reduce the size and the computational burden of the subproblem.

4.2. DUOpt-2: Minimization of total hops

This subproblem is used to determine the optimal number of hops between DU Hotels and their assigned RUs at time instant $t \in [0, T]$. The ILP model solved in this subproblem reads as follows:

$$\min C_H = \sum_{i \in S} \sum_{j \in S} (p_{ij} + b_{ij}) \cdot h_{ij} \tag{10}$$

$$(2) - (9)$$

$$\sum_{j \in S} A_j \leq C_A^* \tag{11}$$

The objective function (10) minimizes the total number of hops between RUs and their assigned DU Hotels, which is the objective of second-highest priority. All the constraints defined for DUOpt-1 are imposed (i.e., constraints (2)-(9)), since a feasible assignment minimizing the total number of hops needs to

be found. Constraint (11) enforces the priority ordering among the objectives, imposing that the solution value C_A^* achieved at DUOpt-1, of higher priority, must not be worsened in solving DUOpt-2.

4.3. DUOpt-3: Minimization of backup DUs

This subproblem is used to determine the optimal number of backup DU at time instant $t \in [0, T]$. The ILP model solved in this subproblem reads as follows:

$$\min C_B = \sum_{j \in S} y_j \tag{12}$$

$$(2) - (9), \ (11)$$

$$c_{ijj'} \geq p_{ij} + b_{ij'} - 1 \quad \forall i, j, j' \in S, \ j \neq j' \tag{13}$$

$$y_{j'} \geq \sum_{i \in S} c_{ijj'} \cdot r_{i,t} \quad \forall j, j' \in S, \ j \neq j' \tag{14}$$

$$c_{ijj'} \in \{0, 1\} \quad \forall i \in S, j \in S, j' \in S, j \neq j' \tag{15}$$

$$y_j \geq 0, \ \textit{integer} \quad \forall j \in S \tag{16}$$

$$\sum_{i \in S} \sum_{j \in S} (p_{ij} + b_{ij}) \cdot h_{ij} \leq C_H^* \tag{17}$$

$$\sum_{j \in S} y_j \geq \frac{\sum_{i \in S} r_{i,t}}{C_A^* - 1} \tag{18}$$

Objective function (12) minimizes the total number of backup DUs, which is the objective of lowest priority. All the constraints defined for DUOpt-1 and DUOpt-2 are imposed, since a feasible assignment minimizing the total number of backup DUs needs to be found. Constraints (13) and (14) are used to count the backup DUs according to the sharing policy, that is, a backup DU can be shared between RUs that do not share the same primary DU Hotel. Constraints (15) and (16) define the variable domains $c_{ijj'}$ and y_j, i.e., binary and positive integer, respectively. Constraints (11) and (17) enforce the priority ordering, imposing that the solution values of DUOpt-1 and DUOpt-2 must not be worsened in solving DUOpt-3. Constraint (18) imposes a lower bound on the total number of backup DU Hotels, which is equal to the ratio between the total number of RUs and the total number of active DU Hotels minus one. Recall that RUs can share backup DU Hotels only if they have different primary DU Hotels. Therefore, the lower bound on number of backup DUs is obtained by considering the largest number of different primary nodes, which in the best case is equal to the number of active nodes. However, since a RU needs to have distinct primary and backup DU Hotels, one is subtracted from the denominator. Note that this constraint can be imposed in the ILP only within a lexicographic method, since it uses the solution from DUOpt-1. In fact, in an aggregate approach, this constraint would become nonlinear.

4.4. Simple Local Search Heuristic for DUOpt-3

From preliminary results, it was observed that solving the ILP model of DUOpt-3 in an exact way was the major bottleneck in the optimization procedure. Therefore, exploiting the separation into subproblems introduced by the lexicographic optimization, a simple heuristic was developed for solving DUOpt-3.

This heuristic algorithm consists of a local search procedure, in which a neighbourhood of the solution of DUOpt-2 is searched by solving a reduced ILP model for DUOpt-3.

Let A_j^* be the values of A_j, i.e., the DU Hotel placement, computed at the end of DUOpt-2. In the ILP model of DUOpt-3, variables A_j are fixed to A_j^*, i.e., $A_j = A_j^* \ \ \forall j \in S$ is imposed. Therefore, the model is solved optimizing only variables p_{ij}, b_{ij}, y_j and $c_{ijj'}$. In practice, the local search heuristic finds the DU assignment that allows the maximum backup DU sharing given a fixed DU Hotel placement. The core assumption behind this local search heuristic is that the DU Hotel placement found after DUOpt-2 is already optimal or near-optimal.

This simple local search heuristic performed remarkably well on the considered problem instances, both in terms of solution quality and computing times.

5. Lexicographic optimization of DU Hotel placement with minimal virtual network function migrations

In situations where traffic demand is dynamically changing over time, as expected in practice for 5G networks, an optimized solution at a given time instant might not be such later, with the possible need to reconfigure the placement of a few virtual network functions over time. The DUOpt lexicographic algorithm described so far (although appropriate for a static optimization) does not take into account the cost of virtual function migrations. An extension of the previous model is proposed to minimize the number of virtual function migrations within the DU Hotel optimization problem.

In each time interval of the considered time horizon (e.g., each minute of a day) the new DUMig algorithm applies a lexicographic approach in three steps, as in DUOpt, but accounting for the cost of virtual function migration: objectives are ranked according to their priority given by the application scenario (minimization of new node activations, and of migrations and hops), and optimized by solving sequentially three subproblems, in order to separate the decision variables in the objective function. The ILP models parameters and decision variables are reported in Table 1.

5.1. DUMig-1: Optimal node de-activation

This subproblem is used to determine the optimal number of active nodes in the transport network, while minimizing the displacement with the current resource assignment and allowing deactivation of nodes that are no longer needed, at time $t \in [0, T]$. The ILP model of this subproblem reads as follows:

$$\min C_D = \sum_{j \in S_{0,t}} A_j - \frac{1}{2} \sum_{j \in S_{1,t}} (1 - A_j) \tag{19}$$

$$(2) - (9)$$

The objective function (19) penalizes the activation of new nodes with respect to the current solution available at time t,

and favors the deactivation of redundant nodes, i.e., nodes that are no longer needed due to low traffic load. Note that the activation penalty is larger than the deactivation reward, so that the deactivation of a node followed by the activation of a different node results in a penalty. Indeed, deactivating a DU Hotel and activating a new one translates in having to migrate all hosted DUs causing migration costs. To limit migration costs, the activation of new nodes should be used only when the currently active nodes cannot satisfy the traffic demand, hence it is penalized. On the contrary, the deactivation of nodes that are not used gets a reward in order to avoid keeping active redundant nodes. Overall, the goal is, thus, to minimize the displacement in terms of active nodes with respect to the current solution available at time t while deactivating as many redundant nodes as possible.

5.2. DUMig-2: Minimization of hops and migrations

This subproblem is used to minimize the number of migrations and the total hops between RUs and DU Hotels, at time instant $t \in [0, T]$. The ILP model solved in this subproblem reads as follows:

$$\min C_M + C_H = \mu_P \sum_{(i,j)\in P_{0,t}} p_{ij} + \mu_B \sum_{(i,j)\in B_{0,t}} b_{ij} + \\ + \beta \sum_{i\in S} \sum_{j\in S} (p_{ij} + b_{ij}) \cdot h_{ij} \tag{20}$$

$$(2) - (9)$$

$$\sum_{j\in S_{0,t}} A_j - \frac{1}{2} \sum_{j\in S_{1,t}} (1 - A_j) \leq C_D^* \tag{21}$$

The weighted multi-objective function (20) minimizes the total primary and backup DU migrations, and the total number of hops. All the constraints defined in the ILP model of DUMig-1 are imposed, since the choice of active DU Hotels is not fixed from the previous subproblem. Constraint (21) imposes that the solution value C_D^* of the previous higher priority subproblem is not worsened in the final solution.

In the following, it will be assumed that $\mu_P \gg \mu_B \gg \beta$, in order to penalize more primary DU migrations with respect to backup DU migrations, and to give the hops less priority with respect to the migrations. Note that, the objective function could be indeed further decomposed via the Lexicographic method, optimizing in sequence primary migrations, backup migrations and total hops. From experimental results, it was observed that jointly optimizing migrations and hops was more time-efficient. This is because the Lexicographic method introduces a computational overhead due to instantiating and solving several models in sequence. Since the time for jointly optimizing the above objectives is already small, further decomposing the problem grants no benefit.

5.3. DUMig-3: Minimization of backup DUs

This subproblem is used to minimize the total number of backup DUs, at instant $t \in [0, T]$. The ILP model solved in this subproblem reads as follows:

$$\min C_B = \sum_{j\in S} y_j \tag{22}$$

$$(2) - (9),\ (13) - (16),\ (18),\ (21)$$

$$C_M + C_H \leq C_M^* + C_H^* \tag{23}$$

Objective function (22) minimizes the number of backup DUs, and constraint (23) imposes that the solution value achieved at DUMig-2 must not be worsened.

In order to mitigate the computing times of this last subproblem, a local search heuristic similar to the one developed for DUOpt-3 was employed. In particular, the values of A_j are fixed to the values A_j^* found at the end of DUMig-2. The major difference with respect to DUOpt lies in constraint (23) that constraints the migrations in addition to the hops, therefore leading to a smaller neighbourhood exploitable by the local search heuristic.

6. Numerical Results

In this section, the performance of the proposed lexicographic algorithm and its extension to dynamic traffic scenarios are extensively benchmarked in different case studies. First, we validate the proposed lexicographic approach for optimal DU placement DUOpt, by comparing solution quality and computing times against the aggregate approach, on several test networks. In addition, the results of the local search procedure for DUOpt-3 are reported to show its effectiveness in reducing computing times. Secondly, a dynamic traffic scenario in a practical 5G-RAN topology is considered to evaluate the proposed DUMig algorithm against DUOpt. The numerical results were obtained via the commercial solver CPLEX 12.10, running on an Intel Core i9-9900k@4.8GHz with 32GB RAM.

6.1. Comparing DUOpt with aggregate multi-objective optimization in static conditions

The performance of the proposed optimal DU placement lexicographic algorithm (DUOpt) is shown, compared against a traditional aggregate approach. Four regular Lattice networks of 36, 49, 64 and 100 nodes are considered, with $r_{i,t} = 10$ RUs per node and a maximum of $M_W = 80$ wavelengths per link. Sample non-regular topologies are also considered obtained by removing 10 random links from the 49 and 64 Lattice networks, and 30 random links from the 100 nodes Lattice network. In the numerical evaluations, $\alpha = 10^6$, $\beta = 10^3$, and, $\gamma = 1$ are used as weights in the multi-objective function of the aggregate model as in [15], thus imposing the same objective ordering as in the lexicographic algorithm. Note that the lexicographic approach bypasses the use of such large weights, avoiding potential numerical issues during execution. Finally, time limit for execution is set to 1 hour. In the lexicographic approach, the time limit is set to 200 seconds for DUMig-1, 200 seconds for DUMig-2, and 3200 seconds for DUMig-3.

Since the lexicographic and aggregate approaches have different objective functions, the following formulas are defined

Table 2: Results of DUOpt and aggregate approaches for lattice networks.

Approach	$\|S\|$	$\|L\|$	M_H	C_A	C_H	C_B	Gap%
DUOpt	36	60	5	4	156	180	0
Aggregate	36	60	5	4	156	180	19.16
DUOpt	36	60	6	3	194	180	0
Aggregate	36	60	6	3	194	180	0
DUOpt	49	84	5	4	259	250	0.0022
Aggregate	49	84	5	4	259	250	0
DUOpt	49	84	6	4	259	250	0.0022
Aggregate	49	84	6	4	259	250	11.21
DUOpt	64	112	5	5	348	300	0.0026
Aggregate	64	112	5	5	348	290	0.0002
DUOpt	64	112	6	5	344	270	0.002
Aggregate	64	112	6	5	356	220	18.16
DUOpt	100	180	5	8	506	500	0.0042
Aggregate	100	180	5	8	535	470	15.37
DUOpt	100	180	6	8	506	540	23.5
Aggregate	100	180	6	18	607	280	68.9

in order to compute an equivalent optimality gap (G_{eq}) for the two approaches:

$$C_{eq} = \alpha \cdot C_A^* + \beta \cdot C_H^* + \gamma \cdot C_B^* \quad (24)$$

$$LB_{eq} = \alpha \cdot LB(C_A) + \beta \cdot LB(C_H) + \gamma \cdot LB(C_B) \quad (25)$$

$$G_{eq} = \frac{C_{eq} - LB_{eq}}{C_{eq}} \quad (26)$$

where $LB(\cdot)$ is the best lower bound achieved in the execution time limit for each subproblem, and C^* is the objective value of the best solution found in that subproblem. In this way, it is possible to compare the quality of the solutions produced by the two approaches.

In Table 2 the objective values and relative percentage gaps obtained by DUOpt and aggregate approaches are reported. Performance of DUOpt approach is shown either similar or significantly better than the aggregate approach (which obtains suboptimal solutions within the given time limit). In particular, DUOpt is able to find either optimal or near-optimal solutions for all the considered instances but the largest one. On the other hand, the aggregate model shows large relative gaps for the largest considered problem instances (namely, 64 nodes and $M_H = 6$, and 100 nodes), therefore providing less information on the actual solution quality. For the largest considered problem instance (i.e., 100 nodes and $M_H = 6$) DUOpt shows a non-negligible optimality gap, however the computed solution is significantly better with respect to the aggregate, with 8 active nodes and 506 hops against 18 active nodes and 607 hops.

In Fig. 3 the computing times required by DUOpt-1 and DUOpt-2, varying the number of nodes in a regular Lattice network, are reported. For all instances but the largest (i.e., 100 nodes and $M_H = 6$), DUOpt is able to both compute the optimal solutions and prove their optimality for both DUOpt-1 and DUOpt-2 within the time limit. For the largest instance,

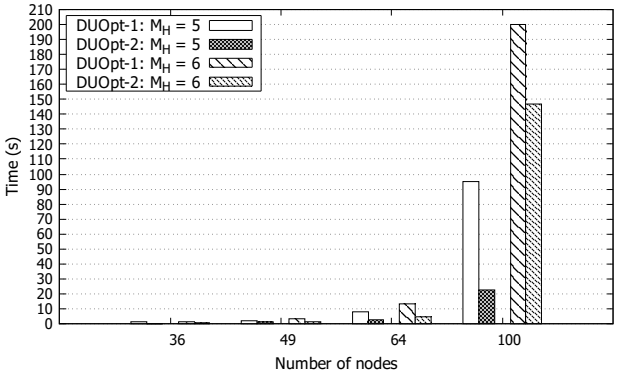

Figure 3: Execution times of the first two subproblems of DUOpt varying the number of nodes in a regular Lattice network and the maximum distance M_H.

Table 3: Results of DUOpt and aggregate approaches for non-regular networks.

Approach	$\|S\|$	$\|L\|$	M_H	C_A	C_H	C_B	Gap%
DUOpt	49	74	5	5	226	230	0
Aggregate	49	74	5	5	229	210	17.98
DUOpt	49	74	6	4	269	220	0.0013
Aggregate	49	74	6	4	269	250	9.21
DUOpt	64	102	5	6	305	320	0.0030
Aggregate	64	102	5	6	306	380	14.86
DUOpt	64	102	6	5	363	340	0.0034
Aggregate	64	102	6	6	331	230	30.43
DUOpt	100	150	5	9	493	530	10.5
Aggregate	100	150	5	9	530	480	18.8
DUOpt	100	150	6	9	492	440	31.6
Aggregate	100	150	6	21	604	330	69.2

DUOpt-1 only reaches the time limit, and the computed solution is of much better quality with respect to the aggregate.

Conversely, the aggregate approach, as shown in Table 2, reaches the time limit of 3600 seconds for most of the considered problem instances, being able to prove optimality of the computed solutions in only two cases (namely, 36 nodes with $M_H = 6$ and 49 nodes with $M_H = 5$).

In Table 3 the results for the reference 49, 64 and 100 nodes non-regular topologies are reported. Similarly to what was observed in Table 2, DUOpt consistently attains smaller optimality gaps with respect to the aggregate method in all of the considered problem instances, and also finds better solutions for the largest network.

Since in solving the ILP model of DUOpt-3 the $3200s$ timeout is often reached without having found the optimal solution, a specific local search for this heuristic was developed. In Table 4 the results of the local search heuristic for DUOpt-3 are reported. Columns C_B^{Heur} and T^{Heur} indicate the solution values found by the local search algorithm and the respective computing times, while columns $C_B^{\text{DUOpt-3}}$ and $T^{\text{DUOpt-3}}$ indicate the best solution values found by solving DUOpt-3 in an exact way and the respective computing times. The local search heuristic is able to find solutions similar to those obtained by solving

Table 4: Results of the heuristic algorithm for DUOpt-3.

$\|S\|$	$\|L\|$	M_H	C_B^{Heur}	$T^{\text{Heur}}(s)$	$C_B^{\text{DUOpt-3}}$	$T^{\text{DUOpt-3}}(s)$
36	60	5	180	0.17	180	183.3
36	60	6	180	0.20	180	47.11
49	84	5	250	0.38	250	3200
49	84	6	250	0.42	250	3200
49	74	5	240	0.53	230	1416
49	74	6	210	0.47	210	3200
64	112	5	290	0.86	300	3200
64	112	6	260	0.87	270	3200
64	102	5	320	0.83	320	3200
64	102	6	300	1.31	340	3200
100	180	5	510	2.94	500	3200
100	180	6	540	2.88	540	3200
100	150	5	490	2.92	530	3200
100	150	6	400	2.91	440	3200

model (12)-(18), in computing times at most equal to 2.94s for the largest problem instances. In particular, it can be seen that for some instances the heuristic is able to find better solutions with respect to the exact method, the latter remaining stuck in suboptimal solutions when reaching the time limit. In the worst case, the heuristics finds solutions that are only slightly worse with respect to solving DUOpt-3 in an exact way. Therefore, the heuristic proves to be a remarkably good solution method for mitigating the exceedingly long computing times required by solving the ILP model of DUOpt-3. Thus, in the following, minimization of the total backup ports for both DUOpt and DUMig will be solved via the developed local search heuristic.

6.2. *Optimization over time with DUOpt and DUMig*

To evaluate the algorithm performance in a dynamic scenario, a realistic reference network with 38 nodes is considered [24], as shown in Fig. 4. A sample traffic variation over a time horizon of one day, expressed in average number of active RUs per node [25, 26], is assumed, as shown in Fig. 5. Such traffic behaviour is representative of both low and high traffic periods, and of both positive, negative and null gradients with respect to time. In the numerical evaluation a random contribution of active RUs uniformly distributed between $[-2, 2]$ is added for each time sample instant to the mean values of Fig. 5 to take into account small deviations that can happen in practice with respect to the mean.

The described DUMig algorithm is applied to maintain the virtual DU function assignment optimized in relation to the sample traffic profile of Fig. 5. The time instants of execution can be set periodically with a given time granularity, either constant or variable during the observation period, providing that enough time is available for the execution of the optimization algorithm and the network reconfiguration. For this reason, in order to apply the procedure with fine time granularity, the optimization algorithm must converge accordingly. The considered numerical evaluations refer to a constant time granularity,

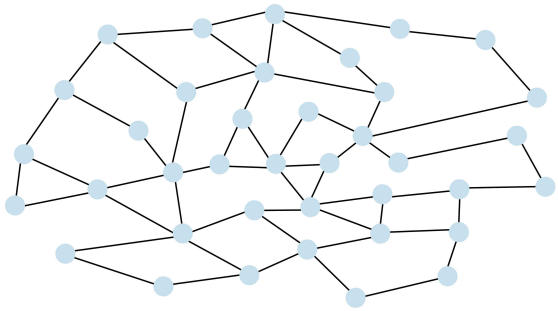

Figure 4: Sample 38-node 5G-RAN topology used for numerical evaluations for the optimization over time.

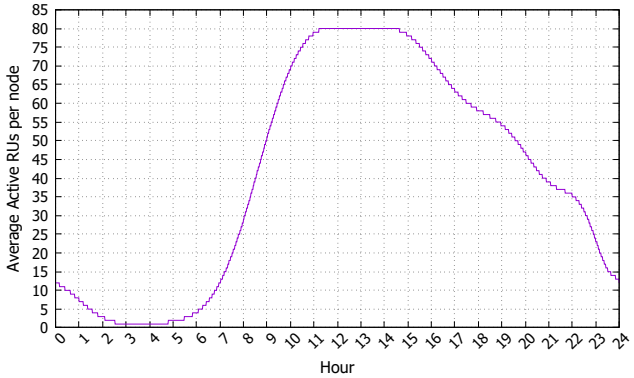

Figure 5: Average active RUs per node for a time horizon $T = 24$ hours.

which was set to be either fine (60 seconds) or coarse (30 minutes), in order to assess the sensitivity of DUMig to this design parameter. DUMig is compared with DUOpt (which neglects migrations and is designed for a static scenario) in order to show that the slight worsening in DUMig on resources optimization (number of nodes, hops and backup DUs) allows for a significant gain in the reduction of primary and backup migrations.

6.2.1. *Fine time granularity evaluations*

Firstly, DUMig was applied with a fine time granularity, which is every 60 seconds in the 24 hour range. The choice of this time granularity is to analyze the performance of the algorithm in a challenging scenario, where the RU profile needs to be followed closely and the time for the execution of the optimization is very limited. The time limit was set to 53 seconds, with at most 35 seconds for DUMig-1, 15 seconds for DUMig-1 and the residual time for the heuristic for DUMig-3. For DUOpt, at most 45 seconds were given to DUOpt-1, 15 seconds to DUOpt-1 and the residual time for the heuristic for DUOpt-1.

Fig. 6 shows the execution times of the instances calculated by DUMig and DUOpt, with time granularity equal to 1 minute. The execution time of DUMig depends on the traffic variations, but it is on average below 10s. In particular, DUMig obtains optimized results in less than a few seconds, with peaks of a few tens of seconds in the intervals e.g., 360-480 (corresponding to 6:00 and 8:00 AM, respectively) where the rate of traffic change

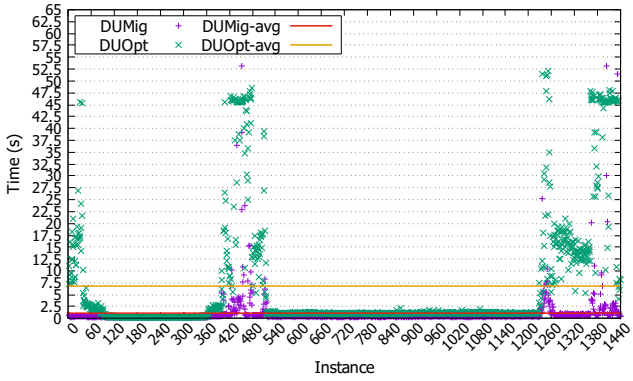

Figure 6: Execution time of each instance for DUMig and DUOpt, with time granularity equal to 1 minute over a time horizon $T = 24$ hours.

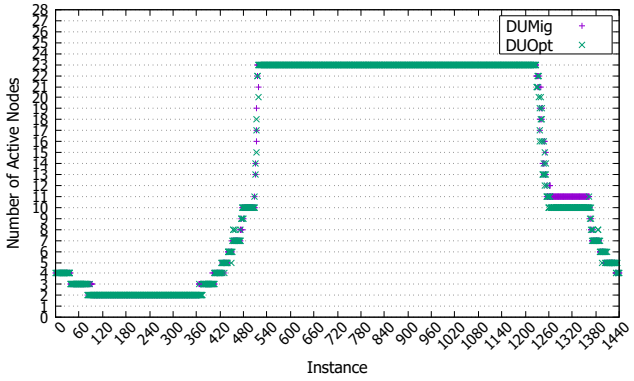

Figure 7: Number of active nodes A_j of each instance for DUMig and DUOpt, with time granularity equal to 1 minute over a time horizon $T = 24$ hours.

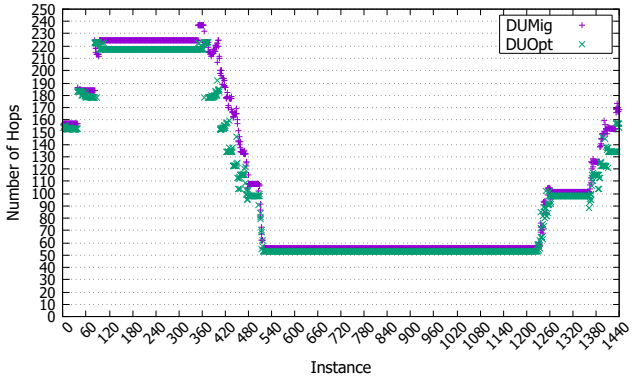

Figure 8: Total number of hops of each instance for DUMig and DUOpt, with time granularity equal to 1 minute over a time horizon $T = 24$ hours.

is high. This is due to the complexity of the reconfiguration required to achieve optimization.

DUMig reaches the time limit in a few instances out of 1440, for which however the optimal solution was found, lacking the proof of optimality only. On the other hand, DUOpt reached the timeout of $45s$ for active node optimization in several instances, for which in some cases it computed a worse solution than DUMig.

Fig. 7 shows the number of active nodes computed by DUMig in comparison with the number of active nodes obtained by DUOpt, with the same time granularity. DUOpt reached the time limit in several instances, nevertheless it computed solutions with the same or a better number of active nodes than DUMig in most cases. The reason is that DUOpt, myopic to migrations, completely neglects the solution computed in the previous time period: clearly, minimizing deviations with respect to the previous solution is instead very important in a dynamic setting to limit the network reconfiguration costs. In addition, one can observe that the two sets of results almost overlap, meaning that DUMig is able to attain near-optimal resource assignment with respect to DUOpt, which does not account for migrations costs. In particular, the average worsening of DUMig with respect to DUOpt throughout all instances is equal to 0.0882 nodes. Focusing on the time intervals in which the number of active nodes varied the most, between 6:00 and 9:00 the average worsening is equal to 0.0555 nodes, whereas between 20:00 and 24:00 is equal to 0.445 nodes. Therefore, the resource assignments computed by DUMig are either optimal or near optimal even if DUMig also accounts for migrations, and one can conclude that the minimization of the displacement from the previous time instant does not hinder the optimization of the number of active nodes.

In Fig. 8 the total number of hops computed by DUMig is compared with that obtained by DUOpt in the same time limit. The average worsening of DUMig with respect to DUOpt throughout all instances is equal to 7.16 hops, between 6:00 and 9:00 is equal to 22.1 hops, and between 20:00 and 24:00 is equal to 7.85 hops. Since the minimization of the total number of hops has lower priority with respect to the migrations, it is reasonable to observe a larger worsening with respect to optimal or near-optimal static traffic solutions. Nevertheless, the obtained solutions are of acceptable quality (6.00% relative gap with respect to DUOpt) and most importantly compliant with the maximum distance constraint, which is paramount for latency requirements.

In Fig. 9 the total number of backup DUs is shown for DUMig and DUOpt. As expected, the amount of primary DUs is the same for both algorithms, as they require one primary DU per RU. Conversely, DUOpt needs fewer backup DUs with respect to DUMig. This is because the heuristic for DUMig-3, constrained on the maximum number of migrations it can perform, explores a much smaller neighbourhood with respect to DUOpt. As a consequence, DUMig achieves less sharing of backup DUs with respect to DUOpt, which results in more backup DUs required. In any case, backup DU sharing allows to save a significant amount of resources with respect to a 1:1 protection scheme, where the number of backup DUs would be the same as the primary DUs.

In Fig. 10 and 11 the total number of primary migrations and backup migrations are shown, respectively. DUOpt requires significantly more DU migrations with respect to DUMig. It can be observed that number of migrations induced by DUOpt slightly drops for the instances in which the fronthaul links capacity approaches saturation, as the solutions found tend to be more similar. Overall, the number of migrations induced

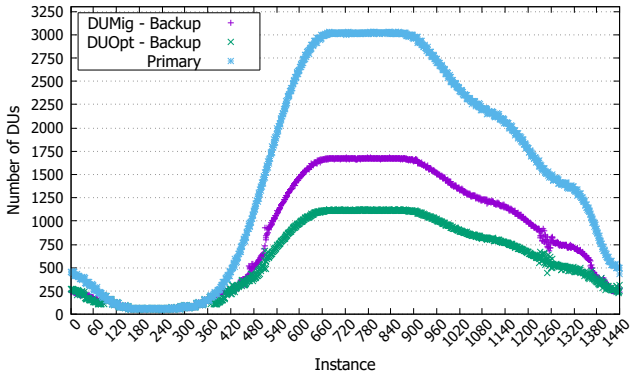

Figure 9: Total primary and backup DUs (p_{ij} and b_{ij}, respectively) of each instance for DUMig and DUOpt, with time granularity equal to 1 minute over a time horizon $T = 24$ hours.

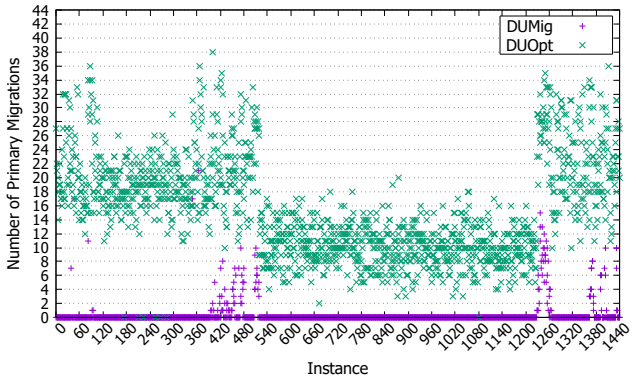

Figure 10: Primary DU migrations of each instance for DUMig and DUOpt, with time granularity equal to 1 minute over a time horizon $T = 24$ hours.

by DUOpt is completely out of control, being it not included in the optimization procedure.

In addition, the majority of migrations for DUMig happens in those time instants when nodes are either activated or deactivated, as can be expected. With reference to the traffic profile, it could be reasonable to execute the algorithm during the time intervals when the traffic is rising fast, so as to maximize responsiveness. However, when traffic is decreasing, it may be beneficial executing the algorithm over longer time intervals and performing nodes de-activations and consequent migrations only few times over the decreasing slope. This would limit the number of times the optimization is executed with potential positive impact on energy saving.

As numerical results, the total number of primary and backup migrations for DUOpt are respectively equal to 22507 and 22163, whereas for DUMig they are equal to 613 and 1070, i.e., there are 97.3% less primary migrations and 95.2% less backup migrations than in DUOpt. This confirms that DUMig is able to preserve near-optimal resource assignment while keeping the total number of migrations as minimal. Moreover, given that minimizing primary migrations was given higher priority than backup migrations, DUMig performs 42.7% less primary migrations with respect to backup migrations.

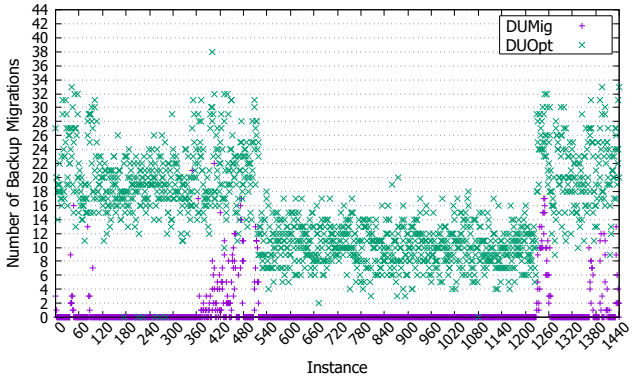

Figure 11: Backup DU migrations of each instance for DUMig and DUOpt, with time granularity equal to 1 minute over a time horizon $T = 24$ hours.

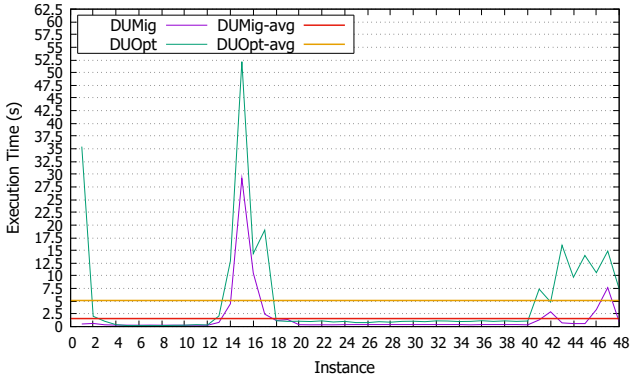

Figure 12: Execution time of each instance for DUMig and DUOpt, with time granularity equal to 30 minutes over a time horizon $T = 24$ hours.

6.2.2. *Coarse time granularity evaluations*

The algorithm was applied with a coarser time granularity, which is every thirty minutes. The time limit for both DUMig and DUOpt was set to 1700 seconds, with at most 1200 seconds for DUOpt-1 and DUMig-1, at most 500 seconds for DUOpt-2 and DUMig-2, and the residual time to the heuristics for DUOpt-3 and DUMig-3.

Fig. 12 shows the computing times of DUMig and DUOpt for each instance. One can observe that even though the time granularity is much coarser than in the previous case, and therefore traffic variations between two consecutive instances are much larger, the average computing times remain in the same order of magnitude as shown in Fig. 6. In particular, the average computing times are 1.66s and 5.21s for DUMig and DUOpt, respectively.

Fig. 13 shows the number of active nodes and total hops from DUMig versus DUOpt. With respect to the nodes, the average worsening of DUMig with respect to DUOpt over the entire day is equal to 0.041 nodes, between 6:00 and 9:00 is equal to 0.167 nodes and between 20:00 and 24:00 is equal to 0.125 nodes. Again, one can observe that the DUMig and DUOpt plots almost overlap, showing that DUMig is able to maintain near-optimal resource assignment. Moreover, even though a coarser time granularity was considered and therefore

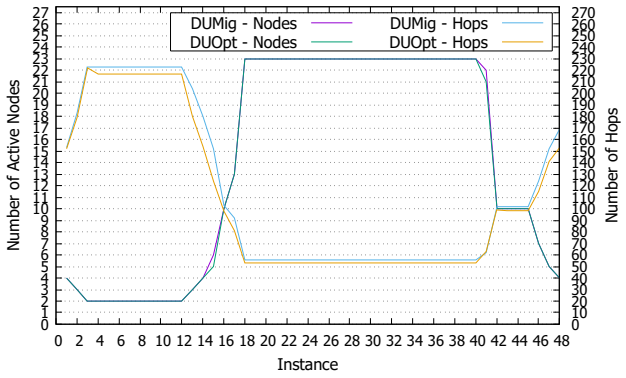

Figure 13: Active nodes and hops of each instance for DUMig and DUOpt, with time granularity equal to 30 minutes over a time horizon $T = 24$ hours.

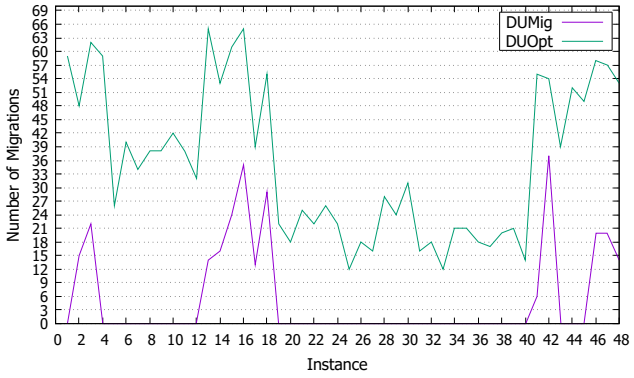

Figure 14: Total DU migrations of each instance for DUMig and DUOpt, with time granularity equal to 30 minutes over a time horizon $T = 24$ hours.

larger traffic variations, the minimization of the displacement from the previous solution does not hinder the optimization of the number of active nodes.

With respect to the hops, the average worsening of DUMig with respect to DUOpt over the entire day is equal to 5.69 hops, between 6:00 and 9:00 is equal to 16.2 hops and between 20:00 and 24:00 is equal to 6.25 hops. As in the scenario with finer time granularity, solutions of acceptable quality are achieved (5.19% relative gap with respect to DUOpt), even though minimization of the total number of hops is given lower priority with respect to the migrations.

Fig. 14 shows the total number of migrations performed by DUMig and DUOpt. Overall, DUMig performs 84.5% less migrations than DUOpt, again showing its capability to produce near-optimal resource assignments while keeping under control the total number of migrations. Fig. 15 shows the migration savings from DUMig with respect to DUOpt. In particular, DUMig obtains 86.1% and 83.0% less primary and backup migrations, respectively, than DUOpt. This is because primary migrations were given a higher cost with respect to backup migrations, and are therefore discouraged.

The total number of DUMig migrations over the considered time period is much smaller with respect to the 60 seconds time granularity: this can be of particular interest in case of decreasing traffic, where the timeliness of VNFs reconfiguration is not crucial, leading to further savings in computing power. Since with decreasing traffic the computed solution maintains feasibility, the advantage of adopting a coarser time granularity is twofold: firstly, the total number of migrations to be performed is much smaller, avoiding redundant re-configurations; secondly, the algorithm can be called fewer times, thus saving computational power. On the other hand, when the traffic is increasing, it is paramount to reconfigure the network as fast as possible, therefore a finer time granularity should be adopted.

In Fig. 16 the active nodes at 8:00 are reported after a traffic increase with respect to 7:30. In particular, previously active nodes kept active are highlighted in red, previously active nodes that have been turned off are circled in red, and new active nodes are highlighted in green. As reported in Fig. 13, at 8:00 there are 5 new active nodes with respect to 7:30, for a

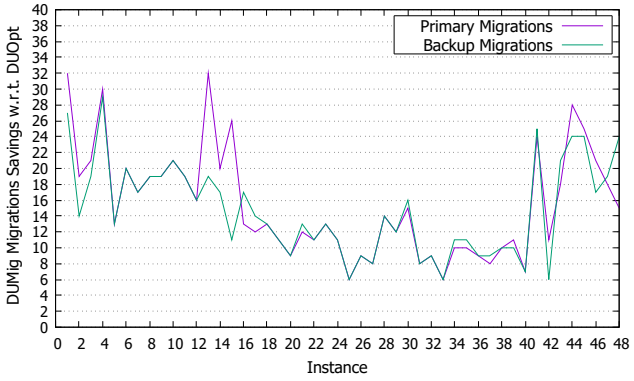

Figure 15: DUMig migrations savings of each instance w.r.t. DUOpt, with time granularity equal to 30 minutes over a time horizon $T = 24$ hours.

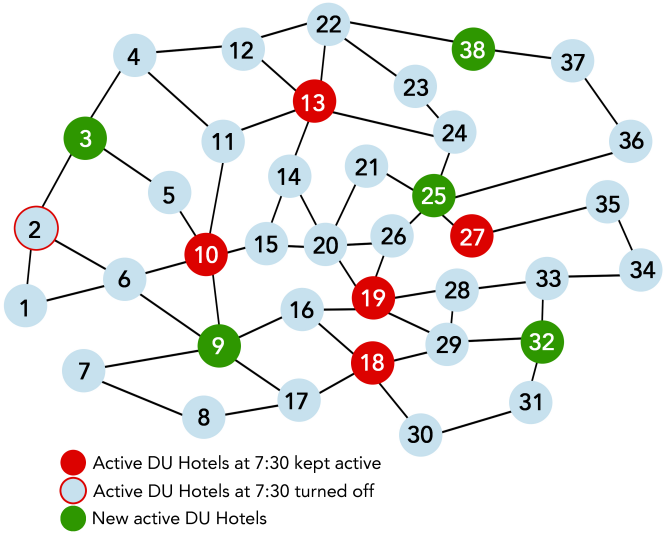

Figure 16: Active DU Hotels placement at 8:00 after a traffic increase w.r.t. 7:30, with time granularity of 30 minutes over a time horizon $T = 24$ hours.

total of 10 active nodes. Five new nodes (3, 9, 25, 32, 38) have been activated; with the previous solution fixed, that would have led to 11 active nodes. In fact, since the objective function of DUMig rewards the deactivation of nodes that are no longer needed, node 2 has been turned off. This results in DUMig reaching the optimal DUOpt solution (10 active nodes) both in shorter computing time and minimizing the total number of DU Hotel migrations.

7. Conclusions

In this paper, 5G-RAN multi-objective optimization is considered for jointly addressing the following points for the first time, in a scalable, effective and generalizable way: 1) cost of virtual network functions migrations in a dynamic traffic scenario, 2) deployment of redundant resources for resiliency against failures, 3) constraints on the maximum latency. The multi-objective optimization problem is divided into three subproblems, one per objective, which are solved in sequence according to the priority of the objectives using a lexicographic method. A bottleneck in the optimization procedure is tackled via a novel local search heuristic able to compute in negligible computing times solutions of similar quality with respect to an exact method. Computing times short enough to allow dynamic network optimization over time, with time granularities down to a few seconds, are achieved. In addition, the method is extended to optimize virtual function migrations to achieve optimization over time in a dynamic traffic context. To that regard, remarkable savings in the total number of migrations (86.1% for primary virtual functions and 83% for backup virtual functions) are gained while maintaining near-optimal resource assignment, with a worst-case performance of less than 1 additional active DU hotel required, on average, over a time horizon of a day. In future work, we are planning to extend the model to account also for failures of the transport network domain (e.g., in case of fiber cuts) considering different access network topologies.

Acknowledgements

Valentina Cacchiani acknowledges the support by the Air Force Office of Scientific Research under award number FA8655-20-1-7019.

Appendix A. Aggregate Multi-objective Optimization

The multi-objective problem introduced in 3.1 can be solved by scalarization: each objective function is weighted by a positive scalar, which magnitude is proportional to the objective priority. The weighted objective functions are summed, and optimized at the same time in an aggregate way by solving a single ILP model. Let α, β and γ be positive scalars, such that $\alpha \gg \beta \gg \gamma$. Then, the ILP model for reliable 5G-RAN optimization solved in an aggregate way reads as follows:

$$\min C = \alpha C_A + \beta C_H + \gamma C_B \tag{A.1}$$

$$\sum_{j \in S} p_{ij} = 1 \quad \forall i \in S \tag{A.2}$$

$$\sum_{j \in S} b_{ij} = 1 \quad \forall i \in S \tag{A.3}$$

$$p_{ij} + b_{ij} \leq A_j \quad \forall i, j \in S \tag{A.4}$$

$$(p_{ij} + b_{ij}) \cdot h_{ij} \leq M_H \quad \forall i, j \in S \tag{A.5}$$

$$\sum_{i \in S} \sum_{j \in S} (p_{ij} + b_{ij}) \cdot \delta^l_{ij} \cdot r_i \leq M_W \quad \forall l \in L \tag{A.6}$$

$$c_{ijj'} \geq p_{ij} + b_{ij'} - 1 \quad \forall i, j, j' \in S,\ j \neq j' \tag{A.7}$$

$$y_{j'} \geq \sum_{i \in S} c_{ijj'} \cdot r_{i,t} \quad \forall j, j' \in S,\ j \neq j' \tag{A.8}$$

$$A_j \in \{0, 1\} \quad \forall j \in S \tag{A.9}$$

$$p_{ij} \in \{0, 1\} \quad \forall i \in S, j \in S \tag{A.10}$$

$$b_{ij} \in \{0, 1\} \quad \forall i \in S, j \in S \tag{A.11}$$

$$c_{ijj'} \in \{0, 1\} \quad \forall i \in S, j \in S, j' \in S, j \neq j' \tag{A.12}$$

$$y_j \geq 0,\ \textit{integer} \quad \forall j \in S \tag{A.13}$$

Since $\alpha \gg \beta \gg \gamma$, solving this ILP model to optimality yields the same objective functions values of the Lexicographic algorithm presented in Section 4.

This problem is NP-Hard, since it generalizes the Uncapacitated Facility Location Problem [27]. In particular, the considered problem additionally includes the minimization of the number of backup DUs, and requires the assignment of primary and backup DU Hotels for each node where RUs are present.

References

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, L. Dittmann, Cloud RAN for mobile networks—a technology overview, IEEE Communications Surveys & Tutorials 17 (1) (2015) 405–426. doi:10.1109/COMST.2014.2355255.

[2] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, M. Liyanage, Survey on 6G frontiers: Trends, applications, requirements, technologies and future research, IEEE Open Journal of the Communications Society 2 (2021) 836–886. doi:10.1109/OJCOMS.2021.3071496.

[3] C. Colman-Meixner, G. B. Figueiredo, M. Fiorani, M. Tornatore, B. Mukherjee, Resilient cloud network mapping with virtualized BBU placement for cloud-RAN, in: 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2016, pp. 1–3. doi:10.1109/ANTS.2016.7947790.

[4] C. Benzaid, T. Taleb, AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions, IEEE Network 34 (2) (2020) 186–194. doi:10.1109/MNET.001.1900252.

[5] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, M. Zorzi, Toward 6G networks: Use cases and technologies, IEEE Communications Magazine 58 (3) (2020) 55–61. doi:10.1109/MCOM.001.1900411.

[6] L. M. P. Larsen, A. Checko, H. L. Christiansen, A survey of the functional splits proposed for 5G mobile crosshaul networks, IEEE Communications Surveys & Tutorials 21 (1) (2019) 146–172. doi:10.1109/COMST.2018.2868805.

[7] O-RAN use cases and deployment scenarios towards open and smart RAN, White paper, O-RAN Alliance (February 2020).

[8] H. Yu, F. Musumeci, J. Zhang, M. Tornatore, L. Bai, Y. Ji, Dynamic 5G RAN slice adjustment and migration based on traffic prediction in

wdm metro-aggregation networks, Journal of Optical Communications and Networking 12 (12) (2020) 403–413. doi:10.1364/JOCN.403829.

[9] N. Di Cicco, V. Cacchiani, C. Raffaelli, Scalable multi-objective optimization of reliable latency-constrained optical transport networks, in: 2021 17th International Conference on the Design of Reliable Communication Networks (DRCN), 2021, pp. 1–6. doi:10.1109/DRCN51631.2021.9477394.

[10] K. Chen, et al., C-RAN: The road towards green RAN, White paper, Mobile China Research Institute (12 2013).

[11] TR 38.801 v.14.0.0 Radio access architecture and interfaces (Release 14), Tech. rep., 3GPP (March 2017).

[12] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, S. Gosselin, Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks, Journal of Lightwave Technology 34 (8) (2016) 1963–1970. doi:10.1109/JLT.2015.2513101.

[13] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park, B. Mukherjee, Energy-efficient virtual base station formation in optical-access-enabled cloud-RAN, IEEE Journal on Selected Areas in Communications 34 (5) (2016) 1130–1139. doi:10.1109/JSAC.2016.2520247.

[14] M. Ito, F. He, E. Oki, Robust optimization model for probabilistic protection with multiple types of resources, IEEE Transactions on Network and Service Management (2021). doi:10.1109/TNSM.2021.3093066.

[15] B. M Khorsandi, F. Tonini, C. Raffaelli, Centralized vs. distributed algorithms for resilient 5G access networks, Photonic Network Communications 37 (06 2019). doi:10.1007/s11107-018-00819-7.

[16] D. Bhamare, A. Erbad, R. Jain, M. Zolanvari, M. Samaka, Efficient virtual network function placement strategies for cloud radio access networks, Computer Communications 127 (2018) 50–60. doi:10.1016/j.comcom.2018.05.004.
URL http://dx.doi.org/10.1016/j.comcom.2018.05.004

[17] R. I. Tinini, D. M. Batista, G. B. Figueiredo, M. Tornatore, B. Mukherjee, Energy-efficient vBBU migration and wavelength reassignment in cloud-fog RAN, IEEE Transactions on Green Communications and Networking 5 (1) (2021) 18–28. doi:10.1109/TGCN.2020.3035546.

[18] L. Askari, F. Musumeci, L. Salerno, O. Ayoub, M. Tornatore, Dynamic DU/CU placement for 3-layer C-RANs in optical metro-access networks, in: 2020 22nd International Conference on Transparent Optical Networks (ICTON), 2020, pp. 1–4. doi:10.1109/ICTON51198.2020.9203072.

[19] D. Pinchera, S. Perna, M. Migliore, A lexicographic approach for multi-objective optimization in antenna array design, Progress In Electromagnetics Research M 59 (2017) 85–102. doi:10.2528/PIERM17042106.

[20] T. Gomes, L. Jorge, P. Melo, R. Girão-Silva, Maximally node and SRLG-disjoint path pair of min-sum cost in GMPLS networks: a lexicographic approach, Photonic Network Communications 31 (06 2015). doi:10.1007/s11107-015-0524-0.

[21] M. Ehrgott, Multicriteria Optimization, Springer Science & Business Media, 2005.

[22] GSMA, Energy efficiency: An overview, https://www.gsma.com/futurenetworks/wiki/energy-efficiency-2/ (2019).

[23] A. S. G. Andrae, T. Edler, On global electricity usage of communication technology: Trends to 2030, Challenges 6 (1) (2015) 117–157. doi:10.3390/challe6010117.
URL https://www.mdpi.com/2078-1547/6/1/117

[24] S. Zhang, M. Xia, S. Dahlfort, Fiber routing, wavelength assignment and multiplexing for DWDM-centric converged metro/aggregation networks, in: 39th European Conference and Exhibition on Optical Communication (ECOC 2013), 2013, pp. 1–3. doi:10.1049/cp.2013.1648.

[25] L. Chen, D. Yang, D. Zhang, C. Wang, J. Li, T.-M.-T. Nguyen, Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization, Journal of Network and Computer Applications 121 (2018) 59–69. doi:https://doi.org/10.1016/j.jnca.2018.07.015.

[26] N. Saxena, A. Roy, H. Kim, Traffic-aware cloud RAN: A key for green 5G networks, IEEE Journal on Selected Areas in Communications 34 (4) (2016) 1010–1021. doi:10.1109/JSAC.2016.2549438.

[27] G. Cornuéjols, G. Nemhauser, L. Wolsey, The uncapicitated facility location problem, Tech. rep., Cornell University Operations Research and Industrial Engineering (1983).