# Delay Minimization in Dynamic and Scalable Multi-operator Wireless Backhauling

Dídac Surís, Adrian Agustin and Josep Vidal
Dept. Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
Emails: didac.suris@alu-etsetb.upc.edu, {adrian.agustin, josep.vidal}@upc.edu

***Abstract*—This paper focuses on the optimization of routing and interference-aware resource allocation for a wireless backhaul network, focusing on the end-to-end delay minimization. We adopt a convex optimization formulation, which allows the decomposition of the problem, separating the network-plane and the communications-plane, and also allows an easy interpretation of results. We integrate the solution in a backhaul multi-tenant scenario amenable to a 5G dense small cells access network.**

***Index Terms*—delay minimization; wireless backhaul optimization; graph coloring; interference avoidance.**

## I. Introduction

Both delay and interference are key issues in 5G, the former because of the Internet evolution towards a tactile Internet, and the latter because of the 5G tendency to increase area spectral density through dense deployment of base stations (BS). Nevertheless, this solution comes at the cost of requiring a reliable high capacity backhaul. Deploying a wireless backhaul is an economically interesting solution as compared to a wired backhaul, but it poses challenging problems in terms of channel state acquisition, interference management and allocation of radio resources [1].

The current resource management of wireless backhaul networks tends to be static, i.e. resources devoted to different links are designed for the peak traffic conditions and the interference is avoided by assigning orthogonal resources. Additionally, each mobile network operator (MNO) runs its own backhaul, so resource utilization cannot benefit from traffic aggregation. The two aspects negatively impact on the efficient use of wireless resources. To combat these issues we investigate the technical tradeoffs and solutions that an infrastructure provider (InP) can adopt to manage the backhaul network for multiple MNOs (see Fig. 1). We assume that a single InP connects the access networks of different MNOs with their respective core networks, being responsible for providing the adequate quality-of-service (QoS) for the different network slices[1].

From an economical perspective, the virtualization of the network enables the InP to abstract and share the infrastructure and radio spectrum resources, reducing significantly the overall deployment and operational expenses [2]. Resource allocation and routing can be adjusted in a centralized way as a function of the current traffic demands and momentary network conditions.

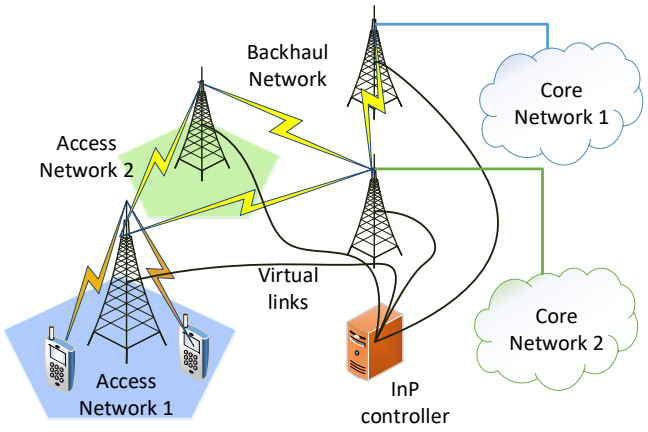



**Fig. 1:** The InP controller allocates physical resources to connect the access network to the core network of a MNO.

Several studies have separately addressed the optimization of the routing [3] and the resource allocation [4] for 5G networks. However, joint optimization can provide improved performance as shown in [5]. We elaborate on the approach found in [5] and extend it to introduce in the model key aspects of 5G [6], namely: a) Interference management, b) Delay minimization, and c) Virtualization and slicing for multi-operator management.

The network management is modeled as a convex optimization problem, from whose solution the proposed algorithms are obtained. The main contributions of this paper are:

- A method allowing the MNOs to control their communication resources in a virtualized way, while the end-to-end total delay is centrally managed at the InP, that controls the network flow resources.
- The adoption of mathematical constraints on the bandwidth allocation, based on graph theory [7], in order to implement inter-links interference avoidance by allocating bandwidth resources orthogonally.

This work has been supported by the Spanish Ministry of Economy and Competitiveness (Ministerio de Economa y Competitividad) through projects TEC2013-41315-R (DISNET) and TEC2016-77148-C2-1-R (RUNNER), and from the Catalan Government (AGAUR) through grant 2014 SGR 60.

[1]We refer to slicing as the physical resource splitting among MNOs.

## II. System Model

### A. Network model

The backhaul network is a wireless mesh network deployed by an InP, where each node is equipped with a router terminal. Nodes collect/distribute the traffic from/towards the access network of the MNO by means of BSs. The wireless backhaul connects the access network to the wired Core Network. For our purposes, the connections between nodes are assumed fixed, as well as the traffic each operator needs to send (or equivalently, the minimum traffic the network has to guarantee).

Our system model, borrowed from [5], can be cast into a convex problem and split in two separated parts: the network flow model, and the communications model.

The network flow model is based on a directed graph (which is assumed to be connected), where each edge represents a link ($l = 1, \dots, L$), and each vertex, a node ($n = 1, \dots, N$). A link is an ordered pair $(i, j)$ of distinct nodes, and its presence means that the network is able to send data from the start node $i$ to the end node $j$. The network topology is represented by the *node-link incidence matrix* $\mathbf{A} \in \mathbb{R}^{N \times L}$, whose entry $A_{nl}$ is associated with the node $n$ and link $l$ via

$$A_{nl} = \begin{cases} 1 & \text{if } n \text{ is the start node of link } l \\ -1 & \text{if } n \text{ is the end node of link } l \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We define $\mathcal{I}(n)$ and $\mathcal{O}(n)$ as the set of links that are incoming and outgoing from node $n$, respectively.

In our model, each node can send different data flows to many destinations and receive data flows from many sources. Data flows are identified by their destinations, so flows in a link are treated separately depending on its destination, being its source irrelevant, and all flows going to the same destination node are treated as one. In the uplink direction, we can interpret each destination $d$ as the access to the core network of a specific MNO, so separating the traffic by destinations will imply separating it by MNO.

We assume that the destination flows are labeled $d = 1, \dots, D$, where $D \leq N$. For each destination $d$, we define a *source-sink vector* $s^{(d)} \in \mathbb{R}^N$, whose $n$-th entry ($n \neq d$) $s_n^{(d)}$ denotes the nonnegative flow (data rate in bits/s) injected into the network at node $n$ (the source) and destined to node $d$ (the sink). In order to accomplish the flow conservation law, the sink flow (in bits/s) at the destination is given by $s_d^{(d)} = -\sum_{n, n \neq d} s_n^{(d)}$, which is in fact the total flow destined to $d$. The network has to guarantee a minimum value $s_{\min}$.

On each link $l$ , we let $x_l^{(d)} \geq 0$ be the amount of flow (in bits/s) destined for node $d$, and we call $\mathbf{x}^{(d)} \in \mathbb{R}^L$ the flow vector for destination $d$. At each node $n$, components of the flow vector and the source-sink vector with the same destination satisfy the flow conservation law:

$$\sum_{l \in \mathcal{O}(n)} x_l^{(d)} - \sum_{l \in \mathcal{I}(n)} x_l^{(d)} = s_n^{(d)} \rightarrow \mathbf{A}\mathbf{x}^{(d)} = \mathbf{s}^{(d)}, \forall d \quad (2)$$

where $\mathbf{A}$ is the node-link incidence matrix previously defined. Finally, we impose capacity constraints on the individual links.

Let $c_l$ be the capacity of link $l$ and $y_l^{(d)}$ the fraction of $c_l$ given to each operator $d$, such that $\sum_d y_l^{(d)} = c_l$. We then require that $x_l^{(d)} \leq y_l^{(d)}$. We suppose a separate queue and server for each destination in each router.

In summary, our network flow model imposes the following constraints on the network flow variables $x^{(d)}$, $s^{(d)}$ and $y^{(d)}$:

$$\begin{aligned} \mathbf{A}\mathbf{x}^{(d)} = \mathbf{s}^{(d)}, \mathbf{x}^{(d)} \succeq 0, d = 1, \dots, D \\ \mathbf{s}^{(d)} \succeq_d \mathbf{s}_{\text{min}}^{(d)}, \mathbf{x}^{(d)} \preceq \mathbf{y}^{(d)}, d = 1, \dots, D \\ \sum_d y_l^{(d)} \leq c_l, l = 1, \dots, L \end{aligned} \quad (3)$$

where $\succeq$ means component-wise inequality, and $\succeq_d$ means component-wise inequality except for the $d$-th component (the sink flow $s_d^{(d)}$ is always negative).

### B. Communications model

We define as *communication variables* the critical parameters on which the capacities of individual links depend, such as the transmit powers, bandwidths, or time-slot fractions. In this paper, the study is done considering bandwidth divisions, but an equivalent study can be done using time division. We denote the vector of transmitted powers by $\mathbf{p}$, and the vector of allocated bandwidths $\mathbf{w}$, where $p_l$ and $w_l$ are the power and bandwidth associated with each link $l$, respectively. In general, the capacity $c_l$ depends not only on $p_l$ and $w_l$, but also on communications resources allocated to other links in the network (due to interferences). However, we consider the case where the link capacity is only a function of the local resource allocation, $c_l = \Phi_l(p_l, w_l)$, and in section II-D we derive a series of conditions on $\mathbf{w}$ in order to avoid interferences. The functions $\Phi_l$ are concave and monotone increasing in $p_l$ and $w_l$, such as the classical Shannon capacity formula:

$$\Phi_l(p_l, w_l) = w_l \log_2 \left(1 + \frac{p_l}{\sigma_l w_l}\right) \quad (4)$$

The communication variables are themselves limited by various resource constraints, such as limits in the total transmit power at each node, modeled by $(\mathbf{A}_+)\mathbf{p} \succeq \mathbf{P}_{\text{max}}$, where $(\mathbf{A}_+)_{nl} = \max(0, A_{nl})$, only identifying the outgoing links at each node. For the bandwidth limits, we assume there are $W_{\text{total}}$ bandwidth resources.

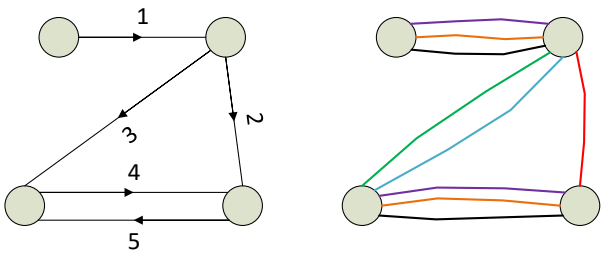
### C. Objective function to minimize delay

A common cost function used in the communication network literature is the total delay function [8]:

$$f(\mathbf{x}, \mathbf{y}) = \sum_l \frac{x_l}{y_l - x_l} \quad (5)$$

This function accounts for the delay incurred in the links, which is closely related to the occupation or utilization of the links (defined as $\rho_l = \frac{x_l}{y_l}$). It is convex with respect to the network variable $\mathbf{x}$, but it is not jointly convex in $\mathbf{y}$ and $\mathbf{x}$, so it is preferable to use another cost function with similar qualitative properties, which is the *maximum link utilization* [8]:

$$f(\mathbf{x}, \mathbf{y}) = \max_l \frac{x_l}{y_l} \quad (6)$$




**(a)** Initial graph **(b)** Extended graph

**Fig. 2:** Example of a network modeled using graphs

This function is quasiconvex [5], and therefore can be solved through a series of convex feasible problems [9, Section 4.2]. As we want to separate the problem into several MNOs, we consider a problem with independent queues for each operator, so as the final objective function to be minimized is:

$$\sum_d \mu_d \max_l \frac{x_l^{(d)}}{y_l^{(d)}} = \sum_d \mu_d f(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}) \tag{7}$$

where $\mu_d$ is a weight associated to each destination (or MNO) that accounts on the importance of delay: the larger the weight, the lower the delay that MNO will bear in its transmissions.

### *D. Interference avoidance conditions on the bandwidth* $\boldsymbol{w}$

We consider a model whereby interference among links is avoided by allocating orthogonal resources to close nodes, considering the interference among distant nodes negligible.

Assume the following interference management model: if two nodes are neighbors, they can not share the same resource. If both node A and node B want to transmit to C, in C the transmissions will collide so they should use different bands. And in the opposite case, if B transmits both to A and C, we also consider that B cannot use the same band to transmit the two signals, as they would also interfere with each other. We consider that in the rest of the nodes the interference is negligible. Hence, we need to fulfil the following conditions:

- All the links entering or leaving a node have to have different associated bands (each link having from 0 to $W_{\text{total}}$ associated bands).
- The links need to allocate the same resources at transmitter and receiver ends.

In order to incoroporate these constraints in our optimization problem we resort to graph coloring principles. First, let us start from a graph $G_1$, where each BS is a node, and each radio link is an edge. We illustrate the concept with the example in Fig. 2a.

We can think of $W_{\text{total}}$ as an integer representing the number of available frequency channels, and $w_l$ an integer representing the number of frequency channels used in link $l$, as well as the number of colors associated to link $l$. Note that some quantization of $w_l$ is needed.

Then, we create a new graph, which we call *extended graph* $G$, from the first one, where each edge $l$ (representing the link $l$) is replicated $w_l$ times (and hence is transformed in a set of $w_l$ edges), creating a multigraph. In the example of Fig. 2b links 2 and 4 are assigned only one channel (or color), link 3 and 5 two channels, and the link 1, three channels. The colors represent a possible channel allocation.

We can now translate the previous idea to the graph theory language: the solution we obtain from the problem solver has to allow the extended graph to be edge-colorable, as we do not want two links associated to the same node to share the same frequency channel (or color). It is important to remark that the optimization problem does not provide the specific colors or channels; it just ensures that we will be able to find them by later using a proper coloring algorithm.

Now we need to come up with the constraints in the problem for the extended graph to be edge-colorable. We will use the chromatic index concept[2] for the extended graph. Note that we need to color the extended graph, so we have to ensure that $W_{\text{total}} \geq \chi'(G)$, where $\chi'(G)$ is the chromatic index of $G$, as $W_{\text{total}}$ is the total number of colors.

The problem is that finding the chromatic index of a graph is an NP-Problem, and even more complicated when we actually do not know how $G$ (the extended graph) is, since it is created from the optimization results. Because of that, we will adopt an upper bound [7]:

**Theorem 1** (Shannon's Bound)**.** *Every colorable graph G satisfies:*

$$\chi'(G) \leq \left\lfloor \frac{3}{2}\Delta(G) \right\rfloor \tag{8}$$

where $\Delta(G)$ is the maximum degree of the graph G[3]. The Vizing and Extended Vizing bounds [7] can also be used with similar results. This theoretical bound can always be reached using the algorithm in the constructive proof of the theorem [10].

Using the Shannon's Bound we can transform the restrictions to:

$$W_{\text{total}} \geq \left\lfloor \frac{3}{2}\Delta(G) \right\rfloor \geq \chi'(G) \tag{9}$$

We can use a simpler upper bound, considering that in the cases where $\Delta(G)$ is odd we do not floor it:

$$W_{\text{total}} \geq \frac{3}{2}\Delta(G) \tag{10}$$

Now we transform $\Delta(G)$ into the parameters we actually are working with, and arrange the inequality:

$$\sum_{l \in \mathcal{O}(n), \mathcal{I}(n)} w_l \leq \frac{2}{3} W_{\text{total}} \; \forall n \in V(G_1) \tag{11}$$

[2]A *k-edge coloring* of a graph $G$ is an assignment of a color to each edge of $G$ in such a way that no two adjacent edges have the same color and at most $k$ different colors are used. The *chromatic index* or *edge chromatic number* $\chi'(G)$ of $G$ is the smallest integer $k0$ for which $G$ admits a $k$-edge coloring [7]

[3]The degree of a vertex of a graph is the number of edges incident to the vertex. The maximum degree of a graph is the maximum degree of its vertices

where $\sum_{l\in\mathcal{O}(n),\mathcal{I}(n)} w_l$ is $\delta(n)$ (degree of a node) and $V(G_1)$ are all the nodes of the initial graph. Using a compact notation, the expression on the left is simply abs($\mathbf{A}$)$\mathbf{w}$, where abs($\mathbf{A}$) represents the absolute value of each element in $\mathbf{A}$. This bound will provide fractional values that we propose to floor so as to get integer values that fit the restrictions.

## III. Network Optimization

All these models and constraints can be put together to define a convex problem that reflects how the link capacities depend on the allocation of communications resources, and how the overall optimal performance of the network can only be achieved by simultaneously optimizing routing and resource allocation:

$$\begin{aligned}
\underset{\mathbf{x},\mathbf{y},\mathbf{s},\mathbf{p},\mathbf{w}}{\text{minimize}} \quad & \sum_d \mu_d f(\mathbf{x}^{(d)},\mathbf{y}^{(d)}) \\
\text{subject to} \quad & \mathbf{A}\mathbf{x}^{(d)} = \mathbf{s}^{(d)}, \mathbf{x}^{(d)} \succeq 0, \mathbf{s}^{(d)} \succeq_d \mathbf{s}_{\min}^{(d)}, \forall d \\
& \mathbf{x}^{(d)} \preceq \mathbf{y}^{(d)}, \forall d \\
& \sum_d y_l^{(d)} \leq w_l \log_2\left(1+\frac{p_l}{\sigma_l w_l}\right), \forall l \\
& \mathbf{A}_+\mathbf{p} \preceq \mathbf{P}_{\max}, \mathbf{p} \succeq 0, \mathbf{w} \succeq 0 \\
& \text{abs}(\mathbf{A})\mathbf{w} \leq \frac{2}{3} W_{\text{total}}\mathbf{1}
\end{aligned} \tag{12}$$

### A. Quasiconvex implementation

As the delay function is quasiconvex, the problem in (12) has to be solved through a series of convex feasible problems [9]. If we denote the delay function by $f_0(x,y)$, a family $g_t$ of convex functions has to be found such that:

$$\begin{aligned}
& f(\mathbf{x},\mathbf{y}) \leq t \leftrightarrow g_t(\mathbf{x},\mathbf{y}) \leq 0 \\
& t_1 \geq t_2 \rightarrow g_{t_1}(\mathbf{x},\mathbf{y}) \leq g_{t_2}(\mathbf{x},\mathbf{y})
\end{aligned} \tag{13}$$

for a $t \in \mathbb{R}$. In our case we used:

$$f^{(d)}(\mathbf{x},\mathbf{y}) = \max_l \frac{x_l^{(d)}}{y_l^{(d)}} \rightarrow g_t^{(d)}(\mathbf{x},\mathbf{y}) = \max_l \left(x_l^{(d)} - t y_l^{(d)}\right) \tag{14}$$

which fulfill the conditions in (13). In the convex feasibility problem, there are $D$ constraints $g^{(d)}(\mathbf{x},\mathbf{y}) \leq 0, \forall d$.

In (13) and (14), $t$ is the value of the objective function whose feasibility is being checked. If for an iteration $i$, with $t^{(i)}$, the problem is feasible, in the next iteration a $t^{(i+1)} < t^{(i)}$ will be chosen, and if it results to be infeasible, then $t^{(i+1)} > t^{(i)}$. The update of $t$ can be done using a simple bisection algorithm such as Algorithm 1 [9, p. 146]:

**Algorithm 1** Bisection method for quasiconvex optimization

$t_{\max} \leftarrow D$ ▷ The maximum value for (7) when $\mu = 1$
$t_{\min} \leftarrow 0$
tolerance $\epsilon > 0$
**repeat**
  $t = (t_{\max} - t_{\min})/2$
  Solve feasibility [9, p. 128] problem composed of the constraints in (12) along with the constraint $g_t(\mathbf{x},\mathbf{y}) \leq 0$ for $t$
  **if** they are feasible, $t_{\max} \leftarrow t$ **else** $t_{\min} \leftarrow t$
**until** $t_{\max} - t_{\min} < \epsilon$

### B. Decentralized network optimization

One of the key points of (12) is that it allows being decomposed in subproblems in different manners, still reaching the optimal solution, so that the centralized problem can be solved in a decentralized way. The decomposition in [5] consists in separating the problem in the communications subproblem and the network subproblem, using dual decomposition such that the final problem is:

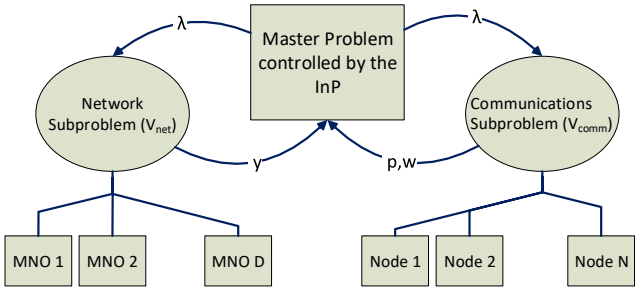



**Fig. 3:** The master problem can be decomposed into smaller subproblems, which at their turn can also be decomposed into smaller ones. The master problem is defined in (12), $V_{\text{net}}$ is defined in (16), $V_{\text{comm}}$ is defined in (17), and each MNO has to optimize (19)

$$\begin{aligned}
\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad & V(\boldsymbol{\lambda}) = V_{\text{net}}(\boldsymbol{\lambda}) + V_{\text{comm}}(\boldsymbol{\lambda}) \\
\text{subject to} \quad & \boldsymbol{\lambda} \succeq 0
\end{aligned} \tag{15}$$

where

$$V_{\text{net}}(\boldsymbol{\lambda}) = \inf_{\mathbf{s},\mathbf{x},\mathbf{y}} \left\{ \sum_d \mu_d f(\mathbf{x}^{(d)},\mathbf{y}^{(d)}) + \sum_l \lambda_l \sum_d y_l^{(d)} \;\middle|\; \mathbf{A}\mathbf{x}^{(d)} = \mathbf{s}^{(d)}, \quad \mathbf{x}^{(d)} \succeq 0, \mathbf{s}^{(d)} \succeq_d \mathbf{s}_{\min}^{(d)}, \; \mathbf{x}^{(d)} \preceq \mathbf{y}^{(d)} \; , \forall d \right\} \tag{16}$$

and

$$V_{\text{comm}}(\boldsymbol{\lambda}) = \sup_{\mathbf{p},\mathbf{w}} \left\{ \sum_l \lambda_l w_l \log_2\left(1+\frac{p_l}{\sigma_l w_l}\right) \;\middle|\; \begin{array}{l} \mathbf{A}_+\mathbf{p} \preceq \mathbf{P}_{\max}, \mathbf{p} \succeq 0 \\ \text{abs}(\mathbf{A})\mathbf{w} \leq \frac{2}{3} W_{\text{total}}\mathbf{1}, \mathbf{w} \succeq 0 \end{array} \right\} \tag{17}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^L$ is the dual variable linking the two subproblems, associated to the constraint $\sum_d y_l^{(d)} \leq w_l \log_2\left(1+\frac{p_l}{\sigma_l w_l}\right)$. More insight into the meaning and computation of $\boldsymbol{\lambda}$ will be provided in the following section III-C.

### C. Coordination among operators

In this section, we explain the decomposition of the Network Flow Subproblem into $D$ single-commodity flow problems,

so that each MNO can optimize its own network parameters according to some global conditions. The communications problem can be similarly decomposed. We start from the initial problem (16).

We can directly decompose the problem into $D$ subproblems, where the variable $\lambda_l$ links to the resources subproblem. Since in each iteration of the main problem this parameter is fixed, we can separate these problems as follows:

$$\sum_d \mu_d f(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}) + \sum_l \lambda_l \sum_d y_l^{(d)} = \sum_d \left( \mu_d \max_l \frac{x_l^{(d)}}{y_l^{(d)}} + \sum_l \lambda_l y_l^{(d)} \right) \quad (18)$$

Eventually we have $D$ independent problems $V_{\text{net}}^{(d)}$ like:

$$\begin{aligned} \underset{\mathbf{x}^{(d)}, \mathbf{y}^{(d)}, \mathbf{s}^{(d)}}{\text{minimize}} \quad & \mu_d f(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}) + \sum_l \lambda_l y_l^{(d)} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x}^{(d)} = \mathbf{s}^{(d)}, \mathbf{x}^{(d)} \succeq 0, \mathbf{s}^{(d)} \succeq_d \mathbf{s}_{\min}^{(d)} \\ & \mathbf{x}^{(d)} \preceq \mathbf{y}^{(d)} \end{aligned} \quad (19)$$

In this decomposition we still need the information of all the resources in every subproblem, but we can separate the $D$ destinations. We associate each destination to a MNO that wants its flow to go from one specific point to another, and every link has an associated price $\lambda_l$, meaning that using that link has a cost $\lambda_l$ per traffic unit. In practice, each operator decides, according to its cost function (in this case we used the delay function for all of them, but any other convex function can also be used), if it is worthwhile to send more or less traffic through a certain link, as the cost incurred is included in its objective function.

When all the operators have decided the amount of traffic to transmit, the master problem (controlled by the InP) updates the prices $\lambda_l$ (if a link is highly demanded, its price is increased, and the other way around, following a supply and demand rule), depending on the information arriving from the communications subproblem (as represented in Fig. 3), until the optimal is obtained. Note that in an iteration the total capacity is not fixed or imposed, there is just a cost for using it, so the different operators do not depend on each other when fixing their $y^{(d)}$. The choices of other operators affect in the calculation of $\lambda_l$ at the main problem for the next iteration: its value will be greater if the other users are also interested in using that link.

Mathematically, the update of $\boldsymbol{\lambda}$ by the master problem does *not* depend on whether or not the decomposition in (19) has been done for calculating $V_{\text{comm}}$, but understanding it this way provides brighter insights into the problem. The update of $\boldsymbol{\lambda}$ can be done using the subgradient method [11]. A possible $(k+1)$-th iteration $\boldsymbol{\lambda}$ value could be:

$$\boldsymbol{\lambda}^{(k+1)} = \max(\boldsymbol{\lambda}^{(k)} - \alpha_k \mathbf{h}^{(k)}, 0) \quad (20)$$

where $\mathbf{h}^{(k)} \in \mathbb{R}^{L \times 1}$ is the subgradient of the master problem (12) at $\boldsymbol{\lambda}$, and $\alpha_k$ is a positive scalar stepsize that has to guarantee convergence. For the subgradient, an expression obtained by deriving (15) with respect to $\boldsymbol{\lambda}$ is:

$$h_l^{(k)} = w_l \log_2 \left(1 + \frac{p_l}{\sigma_l w_l}\right) - \sum_d y_l^{(d)}, \; l = 1, \ldots, L \quad (21)$$

using the values obtained after solving $V_{\text{net}}(\boldsymbol{\lambda}^{(k)})$ and $V_{\text{comm}}(\boldsymbol{\lambda}^{(k)})$. And a simple choice for $\alpha_k$ that guarantees convergence is $\alpha_k = \beta/k$, where $\beta$ is a positive constant [5]. Then, the algorithm for solving the dual problem is as in Algorithm 2.

**Algorithm 2** Algorithm for solving the dual problem (15)

$\boldsymbol{\lambda} > 0$, $\beta > 0$
**repeat**
  Solve $V_{\text{net}}(\boldsymbol{\lambda})$ and $V_{\text{comm}}(\boldsymbol{\lambda})$
  Communicate results to the master problem
  Calculate $\mathbf{h}^{(k)}$ with (21) and update $\alpha_k = \beta/k$
  Update $\boldsymbol{\lambda}$ with (20)
  Communicate the new $\boldsymbol{\lambda}$ to the subproblems
**until** convergence

## IV. Results

The results in Fig. 4 have been simulated considering a 250m x 250m area, with 10 randomly situated nodes, 3 of which considered as destinations (3 different MNOs with their respective core networks as destinations). Carrier frequency is 6 GHz, considering free space path loss and a noise level of -174 dBm/Hz. The maximum power has been fixed at 10 mW in all nodes and the total BW is 8 MHz. $s_{\min}$ (minimum traffic guaranteed) from each node to each destination is a random value in a uniform distribution from 0 to 1 Mbps.

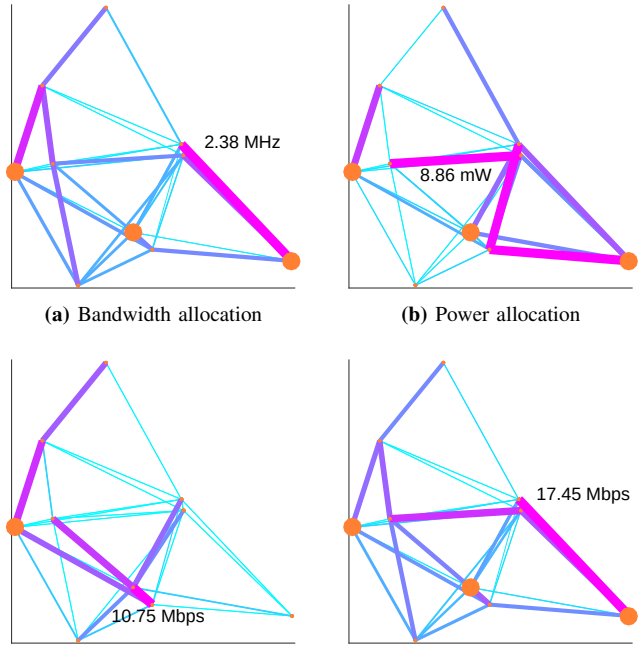
The previous simulation has been repeated for different numbers of nodes, and the mean value delay for 5 different simulations is shown in Fig. 5. The weighted end-to-end delay is calculated from the obtained results using (5), weighting with the actual flow passing through each link and normalizing with respect to the total flow through the network:

$$\text{delay} = \frac{\sum_d \sum_l \frac{x_l^2}{y_l - x_l}}{\sum_d \sum_{n \neq d} s} = \frac{\sum_d \sum_l \frac{\rho_l}{1-\rho_l} x_l}{\sum_d \sum_{n \neq d} s} \quad (22)$$

From the previous simulations we can observe two main results. The first one, seen in Fig. 4, is that the optimal resource and traffic distribution (both among links and among MNOs through slicing) is not uniform, not even close, and that makes the resource and routing optimization very relevant.

The second important result, easily seen in Fig. 5, is that the delay does not worsen significantly with the increasing number of nodes, so the performance does not degrade significantly with an increasing number of nodes. The uniform distribution implies $w_l = W_{\text{total}}/L \; \forall l$, and the power of each node $n$ is uniformly distributed among all $l \in \mathcal{O}(n)$.

The algorithm converges pretty fast with few iterations, each one representing the solving of a convex problem. Fig. 6 shows the value of the calculated delay in each iteration minus the




**(a)** Bandwidth allocation **(b)** Power allocation

**(c)** Flow to the marked destination **(d)** Total flow to all destinations

**Fig. 4:** Optimal routing and resource allocation results. The value is proportional to the line thickness. As a reference, the maximum value is shown. Big dots denote destinations.

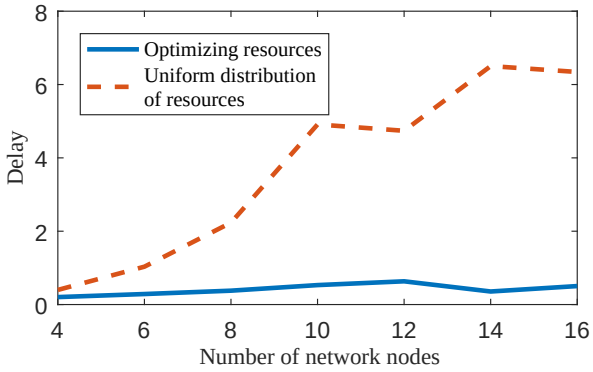



**Fig. 5:** Average end-to-end delay for an increasing number of nodes.

**TABLE I:** Computing Time of (12)

| Number of network nodes | 4 | 8 | 12 | 16 |
|---|---|---|---|---|
| Computing time (s) | 7.87 | 15.37 | 19.94 | 34.12 |

final optimal result. Negative values are associated to non-feasible solutions in certain iterations. The graph shows that the number of iterations does not depend on $N$.

The computing time of (12) grows linearly with the number of nodes (see Table I), where the mean calculation of the 5 previous simulations is shown. The simulations have been done using CVX [12] in Matlab. Note that for these simulations not all the benefits of the convex formulation have been exploited, as very simple and generic algorithms have been used.

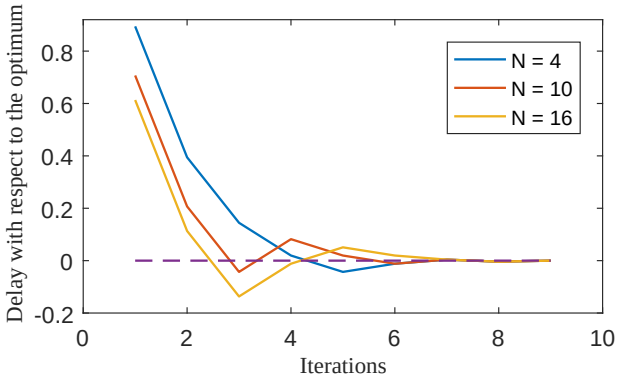



**Fig. 6:** Optimized value through the iterations of Algorithm 2.

## V. Conclusions

We have considered the problem of minimizing the delay in a backhaul network, jointly optimizing its routing and resource allocation. The delay minimization is a crucial aspect of 5G, which implies exploiting efficiently bandwidth resources and interference, due to the foreseen scenario of increasingly dense base station deployments. This is why we presented interference avoidance conditions and thus improve links capacity.

The obtained results show remarkable gains with respect to the non-optimized case, and show that the resource optimization is critical when the network grows in size and traffic.

Future work may include the study of the control plane and the management of the overhead it introduces in order to determine the network state, as well as a practical implementation in a real case scenario.

## References

[1] U. Siddique *et al.*, "Wireless backhauling of 5G small cells: challenges and solution approaches," *IEEE Wireless Communications*, vol. 22, pp. 22–31, Oct. 2015.

[2] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Communication Surveys & Tutorials*, vol. 17, no. 1, pp. 358–380, First Quarter 2015.

[3] H. Yao *et al.*, "An optimal routing algorithm in service customized 5G networks," *Mobile Information Systems*, 2016.

[4] V. Venkatasubramanian *et al.*, "On the performance gain of flexible UL/DL TDD with centralized and decentralized resource allocation in dense 5G deployments," in *Personal, Indoor, and Mobile Radio Communication (PIMRC), 2014 IEEE 25th Annual International Symposium on*, Sep. 2–5, 2014.

[5] L. Xiao, M. Johansson, and S. P. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Transactions On Communications*, vol. 52, no. 7, pp. 1136 – 1144, July 2004.

[6] J. G. Andrews *et al.*, "What will 5G be?" *IEEE Journal On Selected Areas In Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[7] D. Scheide and M. Stiebitz, "On Vizings bound for the chromatic index of a multigraph," *Discrete Mathematics*, vol. 309, p. 49204925, 2009.

[8] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs: NJ: Prentice Hall, 1992.

[9] S. Boyd and L. Vandenberghe, *Convex Optimitzation*. New York: Cambridge University Press, 2004.

[10] J. M. Gilbert, "Strategies for multigraph edge coloring," *John Hopkins APL Technical Digest*, vol. 23, no. 2 and 3, pp. 187 – 201, 2002.

[11] S. Boyd *et al.*, "Notes on decomposition methods," 2008, available at https://stanford.edu/class/ee364b/lectures/decomposition_notes.pdf.

[12] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.