Placement of Logical Functionalities in 5G/B5G Networks

Junior Ziazet, Brigitte Jaumard, Adel Larabi, Nicolas Huin

Placement of Logical Functionalities in 5G/B5G Networks

Junior Momo Ziazet
Computer Science and Software Engineering
Concordia University
Montreal (Qc) Canada

Brigitte Jaumard and Adel Larabi
GAIA - Global Artificial Intelligence Accelerator
Ericsson
Montreal (Qc) Canada

Nicolas Huin
ADOPNET
IMT Atlantique
Rennes, France

***Abstract*—5G technology has brought tremendous growth in connectivity, mobile traffic capacity, and enhanced performance with greater throughput, lower latency, ultra-high reliability, higher connectivity, and an expanded range of mobility. We present here a unified E2E logical functionality placement with the joint placement of distributed units (DU), centralized Units (CU) and user plane functions (UPF).**

The problem is modeled as a large-scale integer linear program, solved using decomposition techniques. It includes all key network and IT resources and optimizes DU, CU and UPF placements and their numbers.

Numerical results are presented on an open Montreal traffic dataset. Sensitivity analysis is performed, investigating the impact of the number of locations hosting DU, CU and UPF functionalities on the delays of the request flows.

***Index Terms*—5G provisioning, logical functionalities, UPF placement, DU and CU placement, latency requirements.**

I. Introduction

As 5G continues to evolve and we begin planning for 6G, difficult questions remain related to logical functionality, considering cost, energy and latency requirements. On the one hand, it is good to host the logical functionalities as close as possible to the radio access network, in order to satisfy the latency requirements, and in particular the latency requirements of the Ultra-Reliable Low Latency Communications (uRLCC) 5G traffic class. Conversely, sharing computing resources is easier when moving away from the radio access network. The question then arises as to how much computing resources we should schedule close to the RAN, and how much we can move further away to be able to meet user needs at any time, taking into account traffic fluctuations.

Moreover, considering the increase in quality of service (QoS) of 5G applications and the increase in overall 5G traffic, we have to deal with the huge power consumption of data centers involved in the different clouds hosting the growing software part of 5G networks. Again energy consumption needs to be kept in mind when taking care of the logical functionalities as to favor the sharing of their compute resources.

We therefore study here the placement of logical functionalities, DU, CU and UPF from an E2E network point of view. We propose a comprehensive mathematical model which integrates the key constraints and solve it using large-scale optimization techniques. We next use that model to explore the compromise between delays and number of locations for hosting the logical functionalities.

Several studies have been devoted to the placement of Virtual Network Functions and Service Function Chains, e.g., [1], [2], and much less on the placement of 5G logical functionalities. On the one hand, there are studies on the placement of DU and CU ([3], [4], [5]), then on UPF [6] on the other hand, but none on all the logical functionalities of 5G.

II. Mathematical Model: How to Optimize Optical Node Placement

A. Notation and Problem Statement

We consider a 5G Software-Defined Network (SDN) that is represented by a graph $G = (V, L)$ where V represents the set of nodes and L the set of physical links. We assume that only a subset of nodes $V^{\triangle} \subseteq V, \triangle \in \{\text{DU}, \text{CU}, \text{UPF}\}$ have computational resources available to host 5G logical functionalities, respectively for the DU, CU and UPF functionalities [7]. Following ITU-R, 5G traffic covers three traffic categories: Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (URLLC), and Massive Machine Type Communications (mMTC). On the user plane, each traffic flow request must follow a logical functionality chain made of an ordered set of logical functionalities as described in Figure 1, where we only present the dominant part of the traffic. Demand is then defined by a set of flow requests,

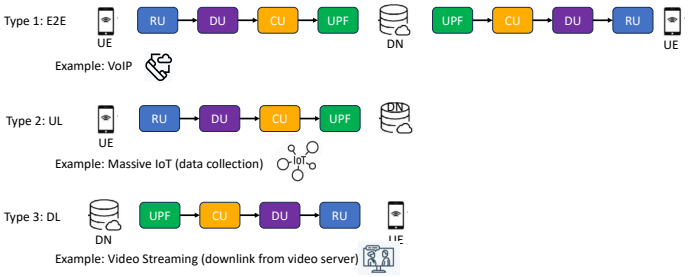


Fig. 1: 5G E2E logical functionality chain (user plane)

denoted by K, where each flow request k is characterized by a source SRC_k, a destination DST_k, a delay requirement δ_{CoS_k} and a logical functionality chain c_{CoS_k}, which both depend on the class of service of k, and a bandwidth requirement b_k (Gbs).

Each E2E logical functionality chain (LFC) c is defined as an ordered sequence of 5G logical functionalities. We denote by f_i the ith logical functionality in any given logical functionality chain, where $i \in I_{CoS_k} = \{1, \ldots, |c_{CoS_k}|\}$. Each logical functionality chain is mapped to a request depending on its service category. Let C be the set of all logical functionality chains.

Let F be the set of 5G logical functionalities, indexed by f. Each functionality f requires some compute resources, i.e., CPU, RAM and storage for its processing. Let $\text{CPU}_f, \text{RAM}_f, \text{STO}_f$ be the amount of CPU, RAM and storage required by f for each unit of bandwidth (it is not necessarily a linear function). Therefore, assuming f is one occurrence of the LFC functions in request k, $\text{VM}_{fk}^{\square}$ represents the fraction of $\square \in \{\text{RAM}, \text{CPU}, \text{STO}\}$ of virtual machine (VM) (we assume that all VMs have the same configuration) that it needs for each of its occurrences. Each logical functionality has a limited number of replicas, denoted by REPLICA_f.

Note that each VM can serve several demands or functionalities, with each demand/functionality only using a fraction of the compute resources of each VM. Note that the complexity of a functionality may not always scale linearly with the unit of bandwidth of the functionality.

The **Placement of 5G/B5G Logical Functionalities** problem is critical for efficiently provisioning 5G requests and making sure they can go through their required logical functionality chain. We formally state it as follows. For a given set of request flows, where each flow k is characterized by a 5-tuple, $(s_k, d_k, c_{CoS_k}, b_k, \delta_{CoS_k})$, identify the best function locations in order to provision the set of LFCs and requests, while minimizing the overall network load subject to the transport and compute capacities (i.e., CAP_ℓ for the links and by $\text{CAP}^{\square}$ for $\square \in \{\text{CPU}, \text{RAM}, \text{STO}\}$ for the nodes), and optionally to a limit on the number of logical functionality replicas.

When provisioning a request k, its required logical functionalities are encountered in the same order as in c_{CoS_k}, with some functions possibly located at the same node.

TABLE I: Logical functionality chains (bandwidth and latency values adapted from [8])

Services	Class	Logical node sequences	Bandwidth per user or IoT system	E2E latency	Request bundles
Cloud gaming (CG)	eMBB	UPF_{CG} - CU - DU - RU	4 Mbps	80 ms	[40-55]
Augmented reality (AR)	eMBB	UPF_{AR} - CU - DU - RU	100 Mbps	10 ms	[1-4]
VoIP	eMBB	RU - DU - CU - UPF_{VOIP}- CU - DU - RU	64 Kbps	100 ms	[100-200]
Video streaming (VS)	eMBB	UPF_{VS} - CU - DU - RU	4 Mbps	100 ms	[50-100]
Massive IoT (MIOT)	mMTC	RU - DU - CU - UPF_{MIOT}	[1-50] Mbps	5 ms	[10-15]
Industry 4.0 (I4.0)	uRLLC	RU - DU - CU - $\text{UPF}_{\text{I4.0}}$ - CU - DU - RU	70 Mbps	8 ms	[1-4]

B. Layered Graph

In the context of Virtual/Container Network Function (VNF/CNF) placement, many authors have used the concept of layered graphs, see [9] for one of the early references, and then, e.g., [1], [2]. Here, we propose a revisited layered graph adapted to the logical functionalities of 5G, and in particular, to certain location restrictions of logical nodes, i.e., DU and CU in the access network, while UPF is either in transport or in the core network, depending on latency requirements.

TABLE II: CPU/RAM/Storage core usage for logical nodes

Logical nodes	vCPU	RAM (Gb) per 100 Mbps	Storage (Gb) per 100 Mbps	Logical node processing time per 0.01 msec unit
RU	1	4	7	12
DU	9	5	1	6
CU	11	15	2	22
UPF_{CG}	13	15	7	14
UPF_{AR}	5	2	5	16
UPF_{VOIP}	5	2	5	2
UPF_{CG}	5	11	10	4
UPF_{MIOT}	5	3	20	4
$\text{UPF}_{\text{I4.0}}$	5	4	11	4

A layered graph G^{L} is defined for each request flow k and its associated E2E logical functionality chain as follows. For a complete E2E chain RU $\rightarrow$ DU $\rightarrow$ CU $\rightarrow$ UPF $\rightarrow$ CU $\rightarrow$ DU $\rightarrow$ RU, the layered graph has 5 layers, and them 3 layers if the request is an uplink/downlink one, see Figure 3 for an illustration of the layered graph for an E2E request.

The initial network graph G is transformed into a *layered graph* G^{L} (not counting the RU_{SRC} and RU_{DST} nodes), where each layer is either associated with the transport network graph (or a subgraph of it) or the combination of the transport and core networks.

For every node $v \in V$, let v^i be the corresponding node in the ith layer $i \in I_{CoS_k}$. Every $(i-1, i)$ layer pair is connected by cross-layer links from v^{i-1} to v^i, if the function f^c_{i-1} can be installed and run on node v, see Figure 4 for an illustration. Therein, request k first goes through its RU_{SRC} following the base station to which it connects, and then to the access network in order to go through a DU and then a CU. Next, it needs to identify a UPF, either in the transport or in the core network, and then again a CU and then a DU, before reaching a RU and then the destination. Note that we do not need to consider the overall transport network, e.g., in Figure 2, we only need to consider TN_k^{SRC} for the DU-UL/CU-UL layers, and similarly only TN_k^{DST} for the CU-DL/DU-DL layers. Finding a path and a chain placement for a request $(s_k, d_k, c_{CoS_k}, b_k, \delta_{CoS_k})$ consists in finding a path on the layered graph G^{L} from node RU_{SRC} to node RU_{DST}. Note that each layer represents the progression of the chain, e.g., being on the second layer means that the placement of the first function of the chain is already decided. The placement of the logical functionality associated with layer i is given by the cross-layer link used to switch between layers i and $i+1$.

C. Mathematical Model

We propose a model, called LFP_CG model, which relies on the concept of configurations. Therein, a configuration is defined by a potential path provisioning, called *service path* in the sequel, which is next described formally below for a given request.

A *Service Path* p for request flow k is a path, i.e., an ordered set of nodes from the source to the destination node of the request, going through node hosting the logical functionalities

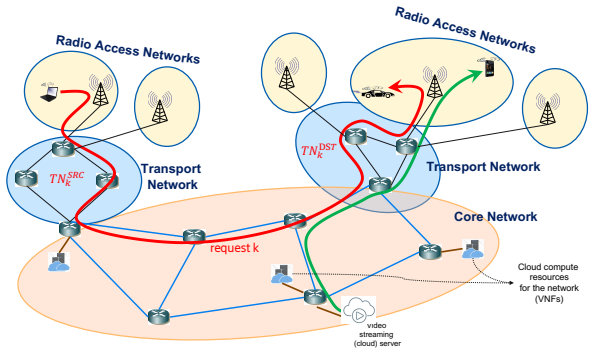


Fig. 2: Overview of the various networks

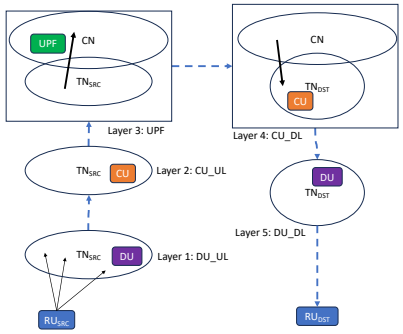


Fig. 3: E2E layered graph

of the service chain c_{COS_k} associated with the class of service (CoS) of request k, in the node order of c_{COS_k}. Let P_k denote the set of paths from SRC_k to DST_k for service chain c_{COS_k}. Note that the same DU and CU are shared by the request flows originating from the base station or going through the same RU location.

Notations

- $V^{\triangle}$ the set of nodes able to host logical functionality $\triangle \in \{\text{DU}, \text{CU}, \text{UPF}\}$
- V^{RU} is the set of RU nodes (we do not look at the placement of the RU functionality).
- K_{RU} is the set of requests originating from or destined to the UEs associated with RU
- $K^{\text{UL}}_{\text{RU}}$ (resp. $K^{\text{DL}}_{\text{RU}}$) is the set of requests originating from (resp. destined to) to the UEs associated with RU

Parameters

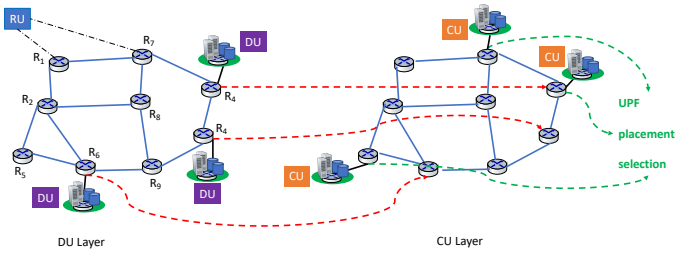


Fig. 4: Details of E2E layered graph

- $a^p_{iv} \in \{0,1\}$, where $a^p_{iv} = 1$ if $f_i^{c_{\text{COS}_k}}$ is executed on node v for *Service Path* $p \in P_k$.
- $a^p_{i,\text{UL},v}$ (resp. $a^p_{i,\text{DL},v}$) $\in \{0,1\}$, where $a^p_{i,\text{UL},v} = 1$ (resp. $a^p_{i,\text{DL},v} = 1$) if $f_i^{c_{\text{COS}_k}} = \text{DU}_{\text{UL}}$ or CU_{UL} (resp. $f_i^{c_{\text{COS}_k}} = \text{DU}_{\text{DL}}$ or CU_{DL}) is executed on node v for *Service Path* $p \in P_k$.
- $\text{NUM}^p_\ell \in \mathbb{N}$ denotes the number of occurrences of link ℓ in path p: $\sum_{i \in I_{\text{COS}_k}} \varphi^i_\ell$, where $\varphi^i_\ell = 1$ if path p goes through ℓ in layer i.
- $\text{PLACE}_{vf} \in \{0,1\}$, where $\text{PLACE}_{vf} = 1$ if function f can be hosted on node v and 0 otherwise.

Variables

- $y^k_p \in \{0,1\}$, where $y^k_p = 1$ if request k uses service path p, 0 otherwise.
- $z_{vf} \in \{0,1\}$, where $z_{vf} = 1$ if function f is placed on node v and 0 otherwise.
- $x^{\triangle}_{\text{RU},v} \in \{0,1\}$, where $x^{\triangle}_{\text{RU},v} = 1$ if at least one request $k \in K^{\text{UL}}_{\text{RU}} \cup K^{\text{DL}}_{\text{RU}}$ executes function $\triangle \in \{\text{DU}, \text{CU}\}$ on node v and 0 otherwise.
- $z^{\text{UPF}}_v \in \{0,1\}$, where $z^{\text{UPF}}_v = 1$ if v hosts at least one $\text{UPF}_\circ$ logical functionality for a given service $\circ \in \{\text{CG}, \text{AR}, \text{VoIP}, \text{VS}, \text{MIoT}, \text{I4.0}\}$

Objective

$$\min \sum_{k \in K} \sum_{p \in P_k} b_k \sum_{\ell \in L} \text{NUM}^p_\ell \, y^k_p \tag{1}$$

The objective is to minimize the network workload. For a given request provisioning and its associated routing (path), the network workload can be calculated by multiplying its hop count by the request's bandwidth requirement.

The set of constraints can then be expressed as follows.

Constraints

One path per demand

$$\sum_{p \in P_k} y^k_p = 1 \qquad k \in K \tag{2}$$

Link capacity

$$\sum_{k \in K} \sum_{p \in P_k} b_k \, \text{NUM}^p_\ell \, y^k_p \leq \text{CAP}_\ell \qquad \ell \in L \tag{3}$$

Compute node resource capacity

$$\sum_{k \in K} \sum_{i \in I_{\text{COS}_k}} \sum_{p \in P_k} \text{VM}^{\square}_{f_i k} b_k \, a^p_{iv} \, y^k_p \leq \text{CAP}^{\square}_v$$

$$\square \in \{\text{CPU}, \text{RAM}, \text{STO}\}, v \in V^{\triangle} \tag{4}$$

Limited # of logical functionality occurrences

$$\sum_{v \in V^{\triangle}} z_{vf} \leq \text{REPLICA}_f \qquad f \in F \tag{5}$$

Checking the existence of a service path, i.e., all its functions are hosted on a node along it

$$\sum_{p \in P_k} a^p_{iv}\, y^k_p \leq z_{vf_i} \qquad v \in V^{\triangle}, i \in I_{Cos_k}, k \in K \tag{6}$$

Logical functionality placement

$$z_{vf} \leq \text{PLACE}_{vf} \qquad f \in F, v \in V^{\triangle}. \tag{7}$$

Limited number of UPF nodes

$$\sum_{v \in V^{\triangle}} z^{\text{UPF}}_v \leq \text{UB} \qquad f \in F^{\text{UPF}} \tag{8}$$

$$z_{vf} \leq z^{\text{UPF}}_v \qquad f \in F^{\text{UPF}}, v \in V^{\triangle}. \tag{9}$$

For a given RU, outgoing and incoming flows must go through the same DU and CU

$$\sum_{k \in K^{\text{UL}}_{\text{RU}}} \sum_{p \in P_k} a^p_{\triangle,\text{UL},v} y^k_p + \sum_{k \in K^{\text{DL}}_{\text{RU}}} \sum_{p \in P_k} a^p_{\triangle,\text{DL},v} y^k_p \leq |K_{\text{RU}}| x^{\triangle}_{\text{RU},v}$$
$$v \in V^{\triangle}, \text{RU} \in V^{\text{RU}}, \triangle \in \{\text{DU}, \text{CU}\} \tag{10}$$

$$\sum_{v \in V^{\triangle}} x^{\triangle}_{\text{RU},v} = 1 \qquad \triangle \in \{\text{DU}, \text{CU}\}, \text{RU} \in V^{\text{RU}}. \tag{11}$$

D. Solution of the Mathematical Model

The mathematical model of the previous section has an exponential number of variables and therefore requires a decomposition solution scheme using column generation techniques [10] in order to scale. The latter scheme reuses the mathematical model of the previous section as a so-called Master Problem (MP) and the Restricted Master Problem (RMP), i.e., MP with a very small subset of configurations/columns, and the so-called Pricing Problem (PP), i.e., a configuration generator. Consequently, the Restricted Master Problem corresponds to (1) - (11) with a very limited number of variables. Its role is to select the best provisioning, one for each request, while the pricing problem generates improving configurations, i.e., configurations such that, if added to the current RMP, improve the value of its linear relaxation. Once the linear relaxation of MP is solved, an integer solution is sought. Reader who is not familiar with column generation and how to seek an integer solution is referred to, e.g., [11] and [12]. Figure 5 provides an illustration of the flowchart of the solution when combining column generation and integer linear programming techniques.

We now describe the pricing problem whose role is to generate a valid *Service Path* for a given request. Once again, the formulation relies on the layer graph (G^{L}) introduced in Section II-B. Its objective is defined by the so-called reduced cost (see [11] if not familiar with linear programming concepts).

- $u^{(j)}$ represents the vector of dual variables of constraints (j) in the RMP.

Variables:

- $\alpha_{iv} \in \{0,1\}$, where $\alpha_{iv} = 1$ if $f_i^{c_{Cos_k}}$ is installed on node v, 0 otherwise.

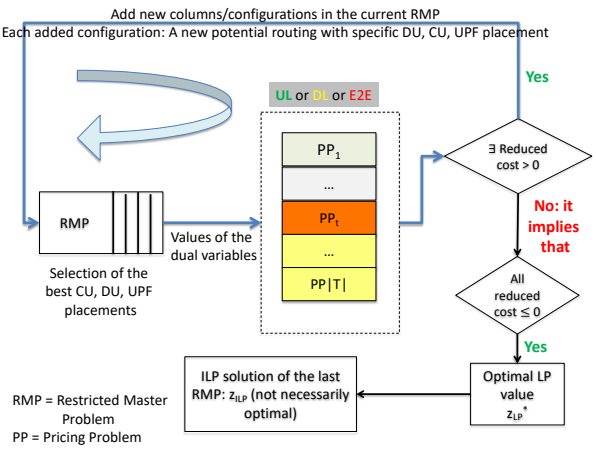


Fig. 5: CG ILP flowchart

- $\varphi^i_\ell \in \{0,1\}$, where $\varphi^i_\ell = 1$ if the flow is forwarded on link ℓ on layer i, i.e., links in each layer in graph G^{L}, 0 otherwise.
- $\beta_{vf} \in \{0,1\}$, where $\beta_{vf} = 1$ if function f is installed in node v. Note that chains can contain multiple occurrences of the same function.
- $\alpha_{vf_i^{c_{Cos_k}}} \in \{0,1\}$, where $\alpha_{vf_i^{c_{Cos_k}}} = 1$ if function $f_i^{c_{Cos_k}}$ is installed in node v, allowing a function to intervene several times (e.g., firewall application) in a service chain, and to be run in different locations, i.e., potentially a different one for each occurrence.

The service path generator (so-called pricing problem in mathematical programming and decomposition models) is written for each request k. We describe it below for the case of an uplink request, and it can be written in a similar way for the two other cases, i.e., for a downlink or a bidirectional request. Its objective function, so-called reduced cost, can be expressed as follows.

$$\min \sum_{\ell \in L} \sum_{i=1}^{|c_{Cos_k}|} \varphi^i_\ell \times \left(b_k - b_k u^{(3)}_\ell\right)$$
$$- u^{(2)}_k + \sum_{\square \in \{\text{CPU},\text{RAM},\text{STO}\}} b_k \sum_{v \in V^{\triangle}} u^{(4)}_{v,\square} \sum_{i=1}^{|c_{Cos_k}|} \text{VM}^{\square}_{f_i^{c_{Cos_k}}} \alpha_{vf_i^{c_{Cos_k}}}$$
$$+ \sum_{v \in V^{\triangle}} \sum_{i=1}^{|c_{Cos_k}|} \alpha_{vf_i^{c_{Cos_k}}} u^{(6)}_{k,v,i} + \text{COST}_{\text{UL}} + \text{COST}_{\text{DL}}, \tag{12}$$

where

- $\text{COST}_{\text{UL}} = \sum_{\triangle \in \{\text{DU},\text{CU}\}} \sum_{v \in V^{\triangle}} \alpha_{\triangle,\text{UL},v} u^{(10)}_{\text{RU}_k^{UL},\triangle,v}$ for uplink and E2E requests and 0 for downlink requests.
- $\text{COST}_{\text{DL}} = \sum_{\triangle \in \{\text{DU},\text{CU}\}} \sum_{v \in V^{\triangle}} \alpha_{\triangle,\text{DL},v} u^{(10)}_{\text{RU}_k^{DL},\triangle,v}$ for downlink and E2E requests and 0 for uplink requests.

Constraints are written as follows.

Flow conservation: they correspond to flow constraints (i.e.,

route) from one logical functionality to the next one in the 5G E2E logical functionality chain for the $\text{SRC}_k \rightsquigarrow \text{DST}_k$ request for which the pricing problem is solved (constraints (13)), and then flow constraints from the source node to the location of the DU function of the service chain (constraints (14)), and similarly from the location of the last function, i.e., UPF for an uplink request, of the service chain to the destination node (constraints (15)). Note that $a_v^i = 0$ for all nodes that do not host any logical node/functionality and that we take care of the possibility that several logical functionalities can be located on the same node, including on the source or destination nodes. Let $k \in K^{\text{UL}}$. Denote by TN_k the transport network associated with k as in Figure 2.

Selecting the CU placement

$$\sum_{\ell \in \omega^+(v)} \varphi_\ell^{\text{TN}_k^{\text{CU}}} - \sum_{\ell \in \omega^-(v)} \varphi_\ell^{\text{TN}_k^{\text{CU}}} + \alpha_{\text{CU,UL},v} - \alpha_{\text{DU,UL},v} = 0 \qquad v \in V^{\text{TN}_k^{\text{CU}}} \tag{13}$$

Selecting the DU placement

$$\sum_{\ell \in \omega^+(v)} \varphi_\ell^{\text{TN}_k^{\text{DU}}} - \sum_{\ell \in \omega^-(v)} \varphi_\ell^{\text{TN}_k^{\text{DU}}} + \alpha_{\text{DU,UL},v} = \begin{cases} 1 \text{ if } v = \text{SRC}_k \\ 0 \text{ else} \end{cases} \qquad v \in \text{TN}_k^{\text{DU}} \tag{14}$$

Selecting the UPF placement

$$\sum_{\ell \in \omega^+(v)} \varphi_\ell^{\text{TN}_k^{\text{UPF}} \cup \text{CN}} - \sum_{\ell \in \omega^-(v)} \varphi_\ell^{\text{TN}_k^{\text{UPF}} \cup \text{CN}} - \alpha_{\text{UPF},v} = \begin{cases} -1 \text{ if } v = \text{DST}_k \\ 0 \text{ else} \end{cases} \qquad v \in \text{TN}_k^{\text{UPF}} \cup \text{CN}. \tag{15}$$

Link capacity.

$$b_k \sum_{i=0}^{|c_{\text{Cos}_k}|} \varphi_\ell^i \leq \text{CAP}_\ell. \qquad \ell \in L. \tag{16}$$

Compute node capacity. For $v \in V^{\text{LN}}$,

$$\sum_{i=0}^{|c_{\text{Cos}_k}|-1} (\text{VM}^\square{}_{f_i^{c_{\text{Cos}_k}}} b_k) \times \alpha_{iv} \leq \text{CAP}_v^{\text{VM}^\square} \qquad \text{VM}^\square \in \{\text{CPU}, \text{RAM}, \text{STO}\}. \tag{17}$$

Delay constraint for request k: it includes a link delay (propagation and average queuing delays: it corresponds to the contribution of k to the overall queuing delay on link ℓ, with the assumption that queuing delay depends linearly on the bandwidth , and a node delay (function processing delay).

$$\sum_{\ell \in L} \sum_{i=0}^{|c_{\text{Cos}_k}|} \delta_{\ell,k} \varphi_\ell^i + \sum_{v \in V^{\text{LN}}} \sum_{i=0}^{|c_{\text{Cos}_k}|-1} \delta_{v, f_i^{c_{\text{Cos}_k}}} \alpha_{iv} \leq \delta_k \tag{18}$$

III. Numerical Results

We now report on the numerical results. First, we describe the data set we used and present the placement produced by the proposed LFP_CG model as well as the function's logical connectivity. Next, we examine the trade-off between the number of logical functionality locations and the request's path delays. Finally, we study the impact of reducing the number of logical functionality locations on the overall capacity utilization. The proposed LFP_CG model was implemented on an AMD Ryzen Threadripper 2990WX-32-Core processor @ 2.95 GHz with 128 GB of RAM in a Python environment with the GUROBI library.

To assess the performance of the proposed LFP_CG model, we considered four scenarios with different numbers of CU, DU, and UPF in the network. The details of each scenario are represented in Table III. In scenarios 1 and 2, the numbers of DU and CU are the same, respectively, 18 and 9, as well as in scenarios 3 and 4, respectively, 12 and 6. In scenarios 1 and 3, UPF_{AR}, UPF_{MIOT}, and $\text{UPF}_{\text{I4.0}}$ can be placed in the core network, and two replicas are allowed per UPF type, while in scenarios 2 and 4, they cannot be placed in the core, and only one replica is allowed. As UPFs of different services can be collocated, the parameter UB indicates the overall number of admitted UPFs locations.

TABLE III: Input configurations per scenario: blue, green and yellow colors indicate that UPF can be located only in MEC, only in core and in both MEC and Core, respectively

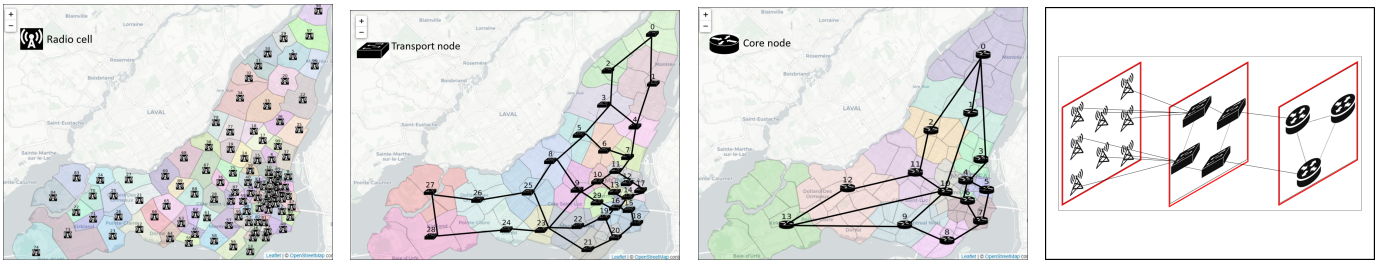
				REPLICA$_f$						
Scenario	RU	DU	CU	UPF_{CG}	UPF_{AR}	UPF_{VOIP}	UPF_{VS}	UPF_{MIOT}	$\text{UPF}_{\text{I4.0}}$	UB
1	100	18	9	2	2	2	2	2	2	7
2	100	18	9	1	1	1	1	1	1	5
3	100	12	6	2	2	2	2	2	2	7
4	100	12	6	1	1	1	1	1	1	5

A. Dataset

The data set used in this work is an extended version of the data set proposed by Ziazet *et al.* [13]. It contains E2E traffic generated by refactoring the urban data of the city of Montreal. The 5G cell locations are inspired by the real cell locations of a mobile operator in the city of Montreal. 5G transport and core nodes are mapped by aggregating cells in the same neighbourhood, and network connectivity is done as in Figure 6. The algorithm will select nodes where we can place processing points for logical functionalities. We considered six types of services, as illustrated in Table I with different bandwidths and delay requirements. Table II provides the compute resource modeling that we used with the required amount of CPU (in terms of percentage of CPU per user), RAM, and storage for each logical node. These values do not come from real use cases (due to lack of access to real data) and only provide a certain order of magnitude. More details about the data set can be found in [13].

B. Placement analysis

We provide placement results in Figure 7 along with logical functionality connectivity for Scenario 1, for UPF placement



(a) Radio Network (b) Transport network (c) Core network (d) Network hierarchy

Fig. 6: 5G network where the placement will be done: radio, transport and core infrastructures and their hierarchy

in the core network. As expected, UPFs of services that we associated with URLLC, i.e., MIoT, AR, and I4.0 in our use case, are placed in the MEC by the proposed approach. This is in line with what is done in real life, as these services are latency-sensitive and their UPFs have to be located near the end user. Some processing points co-locate DU, CU, and UPF and are located in areas with dense populations (downtown Montreal). This is surely to accommodate the peak of traffic that can be encountered in such areas.

From Figure 6d, while in general RU is connected to the closest DU, we can see that the model does not necessarily connect the RU with its closest DU. This might be because of the capacity constraints of that processing point. Another reason might be that the traffic pattern in that area is such that the selected connectivity is the one aligned with the goal of minimizing network workload while maintaining acceptable delay. It is counter-intuitive, as we might think RU should be connected to the closest DU. However, the benefit of our approach is that it will perform the connectivity that benefits the goal and that sometimes is not trivial to find by heuristics or other rule-based approaches. The same interpretation can be made for the connectivity of DU and CU. The proposed LFP_CG model concentrates the backhaul network in the downtown part of the city, where we have more traffic, and it positions the UPFs in the area. Back-haul logical links show that a CU might be logically connected to multiple UPFs. This is because, based on the type of requested service and the network status, physical paths connecting a CU and any UPFs might be created on the core network.

C. Delay vs. number functions

To examine the trade-off between the number of logical functionality locations and the path delay, we considered the four scenarios previously described and compared their spare delays. Figure 8 presents the resulted spare delay for each scenario for the augmented reality service. We can see that for all the cases, the LFP_CG model can find solutions so that the overall sparse delay is more than 50% of the requirement. This is beneficial as it guarantees delay QoS is met, and this spare delay converted into energy will represent an important reduction in energy consumption.

As we reduce the number of replicas of logical functionalities (i.e., DU from 18 to 12, CU from 9 to 6, and UPF from 2 to 1), we observe an increase in the path delay. This is because, with fewer available logical functionalities, requests would need longer paths to access the available ones and reach the destination. This increase is not too important based on the results indicating the capacity of the model to find an optimized placement given the conditions.

Similar results are obtained for the five other services, i.e., cloud gaming, video streaming, VoIP, MIoT, and Industry 4.0.

D. Capacity utilization

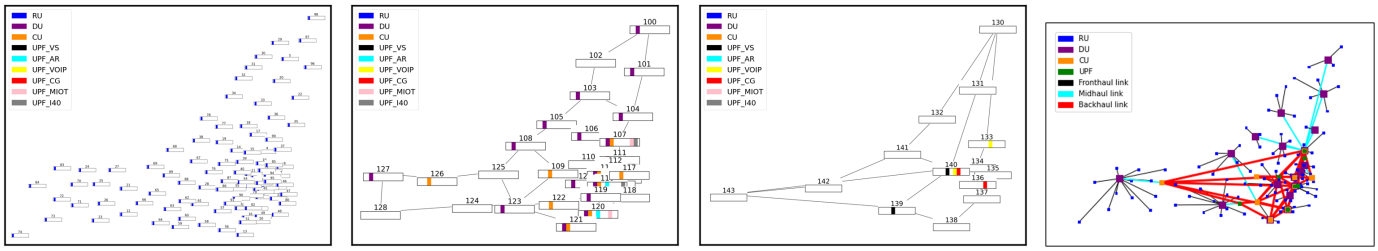
Figure 9 shows the link capacity utilized in all four scenarios. Overall, the obtained results are in line with those of the delay comparison in the previous section. Indeed, for the same reasons as the delay, reducing the number of replicas of logical functionalities will result in an increase in link capacity utilization. Comparing scenarios 1 and 3 which differ from the number of DU and CU replicas and scenarios 1 and 2 which differ from the number of UPFs, we see that in our use case, reducing the number of UPFs has less negative impact on the link capacity utilization than reducing the number of DU and CU. While being influenced by the characteristics of the traffic generated, the placement of DUs and CUs is also very dependent on the dimensioning of the transport network and the availability of computing resources.

IV. Conclusions

This paper formulates the DU, CU, and UPF placement problem as a large-scale integer linear program and solves it using decomposition techniques. The proposed solution does not only find the placement of logical functionalities but also the paths to serve the requests, which are later used to generate the logical connectivity of the 5G functions. Experiments with Montreal traffic data showed that the placement is done so as to reduce link capacity utilization, and the requests are provisioned with relatively small delays. The next step would be to quantify the gain, i.e., spare capacity and spare delay, in terms of energy for a more energy-efficient network.

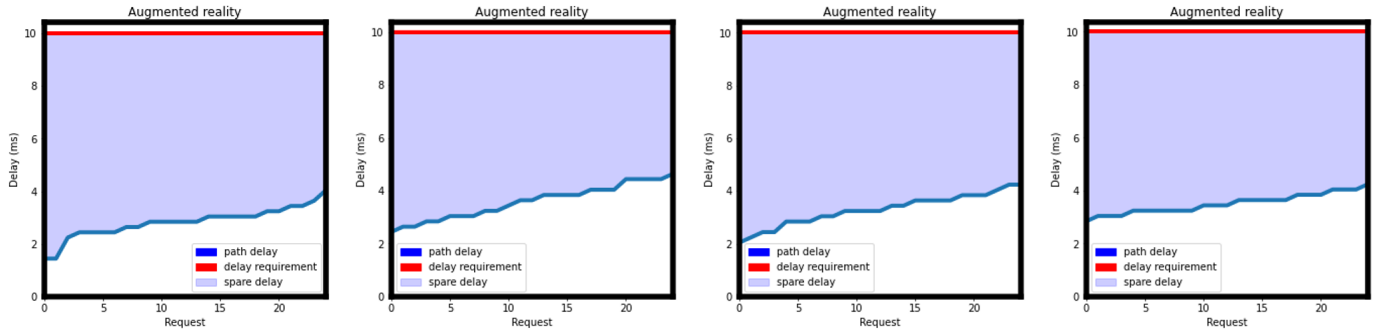
Acknowledgment

First author was supported by a NSERC/INNOVEE internship in collaboration with Ericsson, GAIA Montreal. A special thanks to Jennie Diem Vo for clarifying some concepts of the current 5G technology.



(a) RU placement (b) DU, CU, MEC UPF placement (c) Core UPF placement (d) 5G logical connectivity

Fig. 7: Placement of logical functionalities in radio (a), transport (b) and core (c) network. Obtained logical connectivity associated to each logical functionality (d)



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3 (d) Scenario 4

Fig. 8: sparse delay per scenario

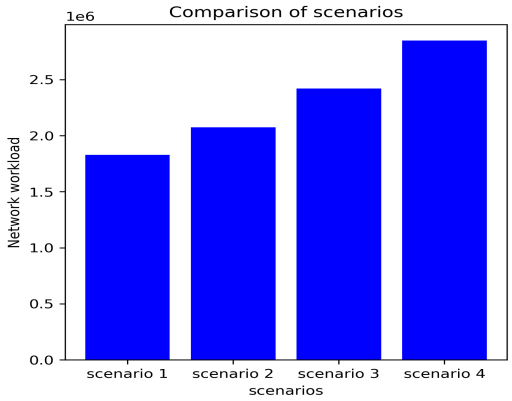


Fig. 9: Network workload comparison

REFERENCES

[1] N. Huin, B. Jaumard, and F. Giroire, "Optimal network service chain provisioning," *IEEE/ACM Transactions on Networking*, vol. 26, no. 3, pp. 1320–1333, June 2018.

[2] N. Hyodo, T. Sato, R. Shinkuma, and E. Oki, "Virtual network function placement for service chaining by relaxing visit order and non-loop constraints," *IEEE Access*, vol. 7, pp. 165 399–165 410, 2019.

[3] J. Liu, B. Zhao, M. Shao, Q. Yang, and G. Simon, "Provisioning optimization for determining and embedding 5G end-to-end information centric network slice," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 273–285, 2021.

[4] M. Klinkowski, "Latency-aware DU/CU placement in convergent packet-based 5G fronthaul transport networks," *Applied Sciences*, vol. 10, no. 21, p. 7429, 2020.

[5] H. Yu, F. Musumeci, J. Zhang, Y. Xiao, M. Tornatore, and Y. Ji, "DU/CU placement for C-RAN over optical metro-aggregation networks," in *Optical Network Design and Modeling (ONDM)*, ser. Lecture Notes in Computer Science, A. Tzanakaki *et al.*, Eds., vol. 11616. Springer, Cham, 2019, pp. 1291–1306.

[6] I. Leyva-Pupo, C. Cervelló-Pastor, and A. Llorens-Carrodeguas, "Optimal placement of user plane functions in 5G networks," in *IFIP International Conference (WWIC)*, Bologna, Italy, 2019, pp. 105 – 117.

[7] 3GPP, "System architecture for the 5G system (5GS)," ThreeGPP, Technical Specification (TS) 23.501, 12 2021, TS 23.501, Version 17.3.0.

[8] L. Askari, A. Hmaity, F. Musumeci, and M. Tornatore, "Virtual-network-function placement for dynamic service chaining in metro-area networks," in *International Conference on Optical Network Design and Modeling (ONDM)*, Dublin, Ireland, 2018, pp. 136 – 141.

[9] A. Dwaraki and T. Wolf, "Adaptive service-chain routing for virtual network functions in software-defined networks," in *Workshop on Hot topics in Middleboxes and Network Function Virtualization (HotMIddlebox)*, 2016, pp. 32–37.

[10] M. Lübbecke and J. Desrosiers, "Selected topics in column generation," *Operations Research*, vol. 53, pp. 1007–1023, 2005.

[11] V. Chvatal, *Linear Programming*. Freeman, 1983.

[12] C. Barnhart, E. Johnson, G. Nemhauser, M. Savelsbergh, and P. Vance, "Branch-and-price: Column generation for solving huge integer programs," *Operations Research*, vol. 46, no. 3, pp. 316–329, May-June 1998.

[13] J. Ziazet, B. Jaumard, H. Duong, P. Khoshabi, and E. Janulewicz, "A dynamic traffic generator for elastic 5G network slicing," in *IEEE International Symposium on Measurements & Networking (M&N)*, 2022, pp. 1–6.