Optimal Functional Splitting, Placement and Routing for Isolation-Aware Network Slicing in NG-RAN

Maria Mushtaq
mariamushtaq@uregina.ca

Morteza Golkarifard
mgolkari@uwaterloo.ca

Nashid Shahriar
nashid.shahriar@uregina.ca

Raouf Boutaba
rboutaba@uwaterloo.ca

Aladdin Saleh
aladdin.saleh@rci.rogers.com

***Abstract*—In the rapidly evolving landscape of 5G and its successor technologies, the Next Generation Radio Access Network (NG-RAN) stands out as a transformative pillar. Functional splitting, a core concept in NG-RAN, splits the traditional base station into distinct functional entities, notably the Distributed Unit (DU), Centralized Unit (CU) and Radio Unit (RU). With flexible functional splitting, Infrastructure Providers (InPs) can dynamically allocate RAN resources to cater to each network slice's distinct throughput and latency demand. However, the problem of optimally selecting functional splits, placement of RAN functions in DU/CU with constrained computational capacities and determining routing paths present an NP-hard challenge. The coexistence of multiple slices on shared infrastructure may necessitate slice isolation for security, performance, and operational reasons, adding another layer of complexity. To address this multifaceted problem, we formulate an Integer Linear Programming (ILP) model that seeks to maximize the InP profit considering computation, virtual machine instantiation and routing costs. Using Gurobi optimizer, we show that optimal slice admission solutions directly impact InP profit and that enhanced computational capacities can increase the number of slices admitted.**

***Index Terms*—Radio access network (RAN), functional splitting, network slicing, integer linear programming (ILP)**

I. Introduction

The paradigm shift in the design and implementation of network infrastructure, moving towards a more flexible, scalable, and programmable framework using Software Defined Networking (SDN) and Network Function Virtualization (NFV), has been a significant highlight in recent years. A critical development under this umbrella has been the emergence of Next Generation Radio Access Networks (NG-RAN) which remains a focal point of innovation and scrutiny. NG-RAN enables the disaggregation or 'split' of traditionally bundled base station functions into virtualized network functions (VNFs), hosted within RU, DU and CU on commoditized servers based on eight split options standardized by 3GPP.

5G introduces Network Slicing (NS), enabling multiple virtual networks to coexist on one physical infrastructure, catering to services like eMBB (enhanced Mobile Broadband), URLLC (Ultra-Reliable Low-Latency Communication), and mMTC (massive Machine Type Communication). Since each slice instance needs a complete set of RAN VNFs, the RAN's capabilities including computational and network resources, dictate which slice requests can be admitted. The placement and realization of these VNFs within the DU/CU are achieved by deploying virtual machines (VMs) on specific computing devices. By sharing VNFs among slices, fewer VNF instances are used, saving computing resources and instantiation costs. However, certain slice instances will mandate dedicated VNF isolation. This is due to varying security needs and the unique network protection measures they entail. Furthermore, functional splitting provides Infrastructure Providers (InPs) flexibility for VNF placement and traffic routing. For example, a URLLC slice may benefit from a distinct functional split compared to an eMBB slice. These decisions, contingent on available resources and prevailing network conditions, account for efficient resource utilization, which in turn results in a reduction in operational expenses. As a result, InPs can establish varied pricing for each slice, maximizing revenue opportunities.

As 5G and beyond networks materialize, realizing their full potential hinges on a trifecta of pivotal components: functional split selection, the strategic placement of RAN VNFs within the DU and CU, and the efficient routing for slice admission. Each of these elements individually holds transformative potential. Yet, when optimized collectively, they redefine how InP cater to diverse, stringent, and evolving service requirements. Therefore, our goal in addressing the existing research gap is to jointly consider the challenges of dynamic RAN functional splits, DU/CU placement, and routing for various types of slice requests subject to various network constraints.

Contributions: In this paper, we present the JointFSAPR, which focuses on the joint optimization of functional splits selection, slice request admission, CU/DU placement, and traffic routing. Our work centers on maximizing the profit of the InP by increasing the revenue generated from admitted requests, while simultaneously reducing the operational costs associated with these decisions. We formulate this optimization problem using ILP and solve it using off-the-shelf solver Gurobi. In contrast to prior research, we factor in the unique specifics of requests, particularly those related to isolation requirements that could potentially elevate costs.

The next sections outline related work, transitions into our ILP model formulation, discuss the solution's evaluation, and concludes with potential future work.

II. Related Work

Previous work provides a thorough examination of various functional splits [1] [2]. Given the discrepancies, the authors suggest that the functional split used must be adjusted to the

978-3-903176-59-1 ©2023 IFIP

network's conditions. The authors in [3] laid the groundwork for the *functional split selection problem* (FSSP), defined as the problem of selecting the optimal functional split for each gNodeB. Hence, our goal is to In [4], authors introduce a Cloud RAN architecture that supports adaptable functional splits for each DU and incorporates additional computation and storage resources for offloading. FluidRAN [5] provides a framework targeting RAN cost reduction, presenting an algorithm to determine functional splits during deployment based on predicted average cell traffic. Notably, all these studies tackle offline FSSP issues. Online FSSP [6], on the other hand, is a more advanced approach that dynamically adjusts the functional split based on instantaneous network conditions. Flex5G [7] offers an algorithm for virtual network embedding (VNE) that allows operators to switch between functional splits upon new virtual network requests, though it's confined to virtualized RAN settings with only low-layer, intra-physical splits. In existing literature, the optimization of radio function placement in virtualized RAN, considering diverse functional split choices, has been tackled for cost reduction by authors such as [8] and [9]. Similar optimization studies [10] [11] also considers policies related to mappings for shareable VNFs and isolation constraints. On the contrary, PlaceRAN [12] addresses all disaggregated RAN elements in their problem formulation to optimize the aggregation of radio functions and reduce computing resources, with a focus on enhancing aggregation within the same layer amid diverse industry desegregation scenarios.

III. Joint Functional Splitting, Admission, Placement and Routing in NG-RAN

The RAN protocol stack, comprising a series of virtual network functions, can be hosted on VMs at either the DU or CU based on the functional split. This choice is influenced by slice request needs and network constraints, with certain scenarios further complicated by slice isolation. Additionally, placing DUs and CUs is challenging due to computational limits, and routing between them must satisfy stringent request requirements. However, these decisions are interrelated and should be made jointly for their mutual impact on the network performance. This section presents our JointFSAPR model, which jointly optimizes functional splits selection, slice request admission, CU/DU placement, and traffic routing to maximize InP profit.

A. System Model

We model the RAN using a graph $G = (\mathcal{V}, \mathcal{E})$ as illustrated in Fig. 1. $\mathcal{V}$ includes subsets of nodes that includes $\mathcal{N}$ DUs, $\mathcal{M}$ CUs and $\mathcal{H}$ forwarding elements (i.e. switches); $\mathcal{M}_0$ as the core node. We assume that $|\mathcal{N}| \gg |\mathcal{M}|$. These nodes are connected with a set of links represented by $\mathcal{E} = \{e_{i,j} : v_i, v_j \in \mathcal{V}\}$ with limited transmission capacity $B_{e_{i,j}}$ (Mbps). $\mathcal{P}$ is set of all the paths in the midhaul i.e. $\mathcal{P}_{nm}$ connects each DU$_n$ to CU$_m$. Each path p incurs an end-to-end delay D_p (ms). The RAN functions are virtualized and implemented on computing resources with limited capacities (MIPS) denoted by C_n and C_m for DU and CU respectively where $C_m > C_n$ $\forall m \in M, n \in N$. We consider various network slice requests $R = \{r_1, r_2, \dots\}$ each with specific throughput λ^r and latency η^r requirements. For each request $r \in \mathcal{R}$ a functional split $s \in \mathcal{S}$ must be selected. Each split option enforces rigorous bandwidth and delay constraints, as depicted in Table I. Our focus is primarily on four types of splits: D-RAN, PDCP/RLC, MAC/PHY, and C-RAN. In the uplink scenario, where each request involves a sequence of RAN functions, we examine this sequence following the f_0 (RF) - f_1 (PHY) - f_2 (MAC & RLC) - f_3 (PDCP) chain. Each function within this chain necessitates a specific degree of processing, denoted as ρ_0, ρ_1, ρ_2 and ρ_3 respectively. This processing demand directly impacts how the DUs/CUs are allocated to computational resources, which subsequently influences the routing path between them. Additionally, our system facilitates both the isolation and sharing of the VNF chain, thereby impacting overhead and resource utilization. Consequently, the decisions pertaining to the admission, placement of DU/CU based on the selected functional split, and routing introduce cost considerations for InPs. The notations used in this optimization model are detailed in Table II.

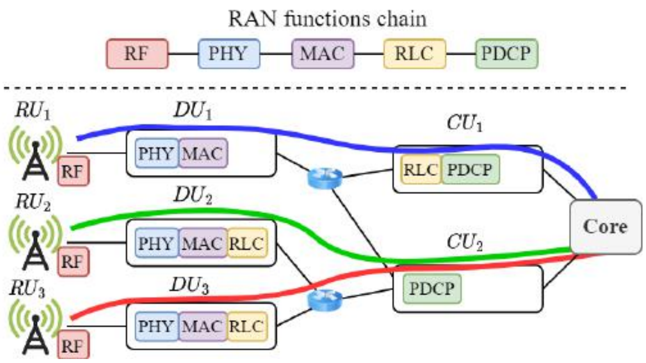


Fig. 1: **System Model:** Illustrative scenario detailing the placement and chosen traffic route for three admitted slices. One slice request demands isolation (highlighted in blue) and uses a MAC/RLC split, positioning the PHY and MAC in DU_1 and the RLC and PDCP in CU_1. The remaining two slices (highlighted in green and red), without isolation constraints, share the PDCP located in CU_2 while having PHY, MAC, and RLC functions hosted in DU_2 and DU_3, both of which have limited capacity.

B. Problem Formulation

This section introduces JointFSAPR, the joint optimization problem of admission, functional splitting, placement of DU/CU, and routing. This problem is formulated as an ILP model that takes into account the system constraints, the objective, and the decision variables we define.

Placement. Constraint 1 and 2 enforce that for any given request r, if the traffic of request r routes through path p, then a placement must exist on DU n and CU m, which act as the source $S(p)$ and destination $D(p)$ node of path p, respectively. Furthermore, these constraints prevent the placement of identical function chains in both DUs and CUs.

$$Z_p^{r,s} \leq X_n^{r,s} \cdot h_n^u, \quad \forall r \in \mathcal{R}, s \in \mathcal{S}, n \in \mathcal{N}_{DU}, u \in \mathcal{U}_r, \\ p \in P : S(p) = n, \quad (1)$$

$$Z_p^{r,s} \le Y_m^{r,s}, \;\; \forall r \in \mathcal{R}, s \in \mathcal{S}, m \in \mathcal{N}_{CU}, p \in \mathcal{P} : D(p) = m. \tag{2}$$

TABLE I: Data And delay requirements Of different splits, when traffic load is λ Mbps [9]

Split	Traffic (Mbps)	Delay (ms)	DU Functions	CU Functions
0	λ	30	f_0, f_1, f_2, f_3	-
1	λ	30	f_0, f_1, f_2	f_3
2	$1.02\lambda + 1.5$	2	f_0	f_1, f_2, f_3
3	2500	0.25	-	f_0, f_1, f_2, f_3

TABLE II: Notation (sets, parameters and variables)

	Notation	Definition
Sets	$\mathcal{N}, \mathcal{M}, \mathcal{M}_\circ$	Set of DUs, CUs and switches
	$\mathcal{U}_r$	Set of RUs that request r is under its coverage
	$\mathcal{P}$	Set of all paths from any DU to any CU
	$\mathcal{E}$	Set of edges
	$\mathcal{R}$	Set of requests
	$\mathcal{S}$	Set of split options
Parameters	λ^r	Traffic demand of request r
	O^r	Whether request r requires an isolated placement
	R^r	Revenue from serving one traffic unit of request r
	ω_n	Cost of instantiating VM in DU n or CU m
	$\alpha_{e_{i,j}}$	Cost of traffic transmission on link $e_{i,j}$
	β_n, β_m	Cost of computation on DU n and CU m
	$I^{DU}(n), I^{CU}(n)$	Number of instantiated VMs in DU and CU
	N_n	Maximum VM instances permitted on DU and CU
	h_n^u	Whether RU u is connected to DU n
	$B_{e_{i,j}}$	Link bandwidth between nodes i and j
	D_p	End-to-end delay of path p
	C_n, C_m	Computational capacity of DU n and CU m
	$\rho_1, \rho_2, \rho_3, \rho_4$	Processing requirements for function f_0, f_1, f_2 and f_3
	$\Pi^{\text{DU}}(s)$, $\Pi^{\text{CU}}(s)$	Computation requires at DU and CU for split s
	$L_{e_{i,j}}$	Traffic on link $e_{i,j}$
	η^r	Target delay of request
	μ^s	Latency requirement of split s
Variables	A^r	Whether request r is admitted
	$X_n^{r,s}$	Whether request r with split s is placed in DU n
	$Y_m^{r,s}$	Whether request r with split s is placed in CU m
	$Z_p^{r,s}$	Whether request r with split s selects path p

Admission and Routing. In case where request r is admitted, it's important to ensure that the traffic of request r is routed solely through a single path p, signifying that only one path should exist defined by constraint 3. Also, there must be no more than one placement of DU and CU to serve that request represented by constraint 4 and 5.

$$\sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P}} Z_p^{r,s} = A^r, \quad \forall r \in \mathcal{R}, \tag{3}$$

$$\sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} X_n^{r,s} = A^r, \quad \forall r \in \mathcal{R}, \tag{4}$$

$$\sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{S}} Y_m^{r,s} = A^r, \quad \forall r \in \mathcal{R}. \tag{5}$$

Computational Capacity. We denote $\Pi^{\text{DU}}(s)$ as the processing required to execute one traffic unit of a portion of the RAN functions chain based on the selected split. Constraints 6 and 7 ensure that the processing requirements do not exceed the capacities of the DU and CU where the RAN function splits are placed.

$$\sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} X_n^{r,s} \cdot \lambda^r \cdot \Pi^{\text{DU}}(s) \le C_n, \quad \forall n \in \mathcal{N}_{DU}, \tag{6}$$

where,

$$\Pi^{\text{DU}}(s) = \begin{cases} \rho_1 + \rho_2 + \rho_3 + \rho_4, & \text{if } s = 0, \\ \rho_1 + \rho_2 + \rho_3, & \text{if } s = 1, \\ \rho_1, & \text{if } s = 2, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Similarly, for CU as:

$$\sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} Y_m^{r,s} \cdot \lambda^r \cdot \Pi^{\text{CU}}(s) \le C_m, \quad \forall m \in \mathcal{N}_{CU}, \tag{8}$$

where,

$$\Pi^{\text{CU}}(s) = \begin{cases} 0, & \text{if } s = 0, \\ \rho_4, & \text{if } s = 1, \\ \rho_2 + \rho_3 + \rho_4, & \text{if } s = 2, \\ \rho_1 + \rho_2 + \rho_3 + \rho_4, & \text{otherwise.} \end{cases} \tag{9}$$

Link Capacity. The traffic on any physical link must not exceed the maximum link capacity, $B_{e_{i,j}}$ as specified in constraint 12. $L_{e_{i,j}}$ defines the traffic load on link $e_{i,j}$ calculated using the traffic requirements of request r while using split s denoted by $\sigma(\lambda^r, s)$.

$$L_{e_{i,j}} = \sum_{r \in R} \sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P} : e_{i,j} \in p} Z_p^{r,s} \cdot \sigma(\lambda^r, s) \tag{10}$$

where,

$$\sigma(\lambda^r, s) = \begin{cases} \lambda^r, & \text{if } s = 0, \\ \lambda^r, & \text{if } s = 1, \\ 1.02\lambda^r + 1.5, & \text{if } s = 2, \\ 2500, & \text{otherwise.} \end{cases} \tag{11}$$

$$L_{e_{i,j}} \le B_{e_{i,j}}, \quad \forall e_{i,j} \in \mathcal{E} \tag{12}$$

Latency. Each request with a target delay of η^r must satisfy the end-to-end latency of the path, denoted by D_p, as well as meet the respective split latency requirement, denoted by $\mu(s)$. This is expressed as the minimum of the two values using the following constraint:

$$Z_p^{r,s} \cdot D_p < \min\{\eta^r, \mu(s)\}, \quad \forall r \in \mathcal{R}, s \in \mathcal{S}, p \in P, \tag{13}$$

where,

$$\mu(s) = \begin{cases} 30, & \text{if } s = 0, \\ 30, & \text{if } s = 1, \\ 2, & \text{if } s = 2, \\ 0.25, & \text{otherwise.} \end{cases} \tag{14}$$

Number of Instantiated VMs. Requests may require isolation, indicated by a binary $O^r \in \{0, 1\}$, or permit shared RAN functions on a VM. The computational capacity sets a limit, N_n, on VMs instantiated per computational resource for DU or CU. Eq. 15 dictates that the total VMs, $I^{DU}(n)$, across

DUs for all network requests shouldn't surpass this limit, with the first term indicating isolated instances and the second for shared ones.

$$I^{DU}(n) = \sum_{\substack{s\in\mathcal{S}:\\ s\neq 3}} \sum_{r\in\mathcal{R}} X_n^{r,s}\cdot O^r + \sum_{\substack{s\in\mathcal{S}:\\ s\neq 3}} \mathbb{1}_{\{\sum_{r\in\mathcal{R}} X_n^{r,s}\cdot(1-O^r)>0\}} \tag{15}$$

$$I^{DU}(n) < N_n, \quad \forall n\in\mathcal{N} \tag{16}$$

Similarly for CUs as below:

$$I^{CU}(n) = \sum_{\substack{s\in\mathcal{S}:\\ s\neq 3}} \sum_{r\in\mathcal{R}} Y_n^{r,s}\cdot O^r + \sum_{\substack{s\in\mathcal{S}:\\ s\neq 3}} \mathbb{1}_{\{\sum_{r\in\mathcal{R}} Y_n^{r,s}\cdot(1-O^r)>0\}} \tag{17}$$

$$I^{CU}(n) < N_n, \quad \forall n\in\mathcal{M} \tag{18}$$

Objective. The goal of the optimization problem is to maximize the revenue derived from admitted requests while minimizing the total cost. The total cost is the sum of the transmission cost in physical links, instantiation and computational costs in VMs.

$$\begin{aligned}\max \sum_{r\in\mathcal{R}} R^r\cdot\lambda^r\cdot A^r - \Bigg(& \sum_{e_{i,j}\in E} \alpha_{e_{i,j}}\cdot L_{e_{i,j}} \\ & + \sum_{n\in\mathcal{N}} \omega_n\cdot I^{DU}(n) + \sum_{n\in\mathcal{M}} \omega_n\cdot I^{CU}(n) \\ & + \sum_{n\in N}\sum_{r\in R}\sum_{s\in S} \beta_n\cdot X_n^{r,s}\cdot\lambda^r\cdot\Pi^{\mathrm{DU}}(s) \\ & + \sum_{m\in M}\sum_{r\in R}\sum_{s\in S} \beta_m\cdot Y_m^{r,s}\cdot\lambda^r\cdot\Pi^{\mathrm{CU}}(s)\Bigg)\end{aligned} \tag{19}$$

The first term in the equation represents the revenue achieved by serving requests, while the remaining terms denote the total costs, which include the sum of the transmission cost on physical links, instantiation costs in DUs and CUs, and computational costs in DUs and CUs. R^r denotes the revenue achieved by serving one traffic unit of request r. $\alpha_{e_{i,j}}$ indicates the cost incurred by transmitting one unit of traffic on a physical link. ω_n represents the cost of instantiating a VM in a DU or CU. β_n and β_m are the costs of utilizing one unit of computing resource in a DU and CU, respectively.

IV. Evaluation

In this section, we evaluate the solutions of our ILP formulation in Eq. 19 utilizing the commercial mathematical optimization solver, Gurobi [13], satisfying all the constraints covered in section III-B.

A. Experimental Setup

In order to assess the effectiveness of our proposed JointFS-APR formulation, we employ the identical network topology depicted in Fig. 1, which comprises 3 DUs, 2 switches, and 2 CUs, each with confined processing capabilities (measured in MIPS). We take into consideration three distinct types of network slice requests, the specifications for which are detailed in Table III. Our study uses a variety of bandwidth and latency parameters for each request, which are selected randomly using a uniform distribution. The options for functional splits, along with their corresponding delay and available bandwidth for each request, are inferred from the data provided in Table I. By utilizing Eq. 9 and Eq. 11, we calculate the computational requirements for each request for the chosen split. In all plots, the results we present are derived from the average of 30 runs.

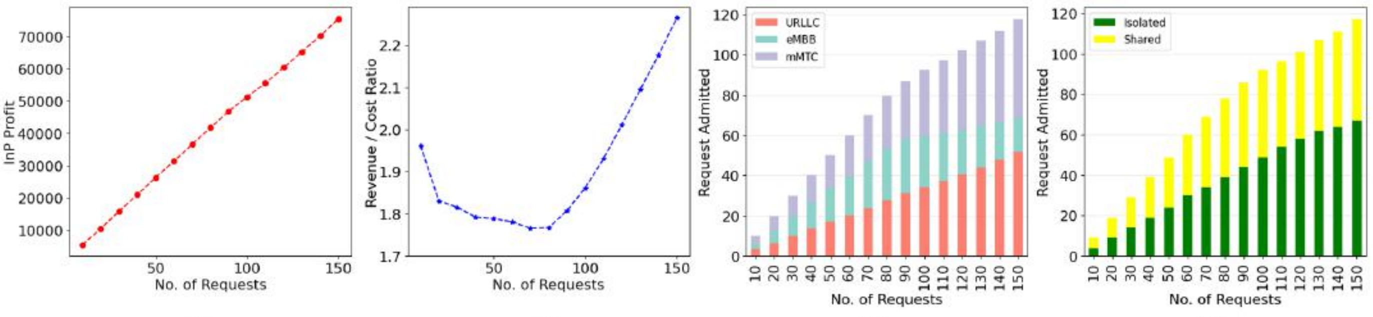
TABLE III: Network Slices Parameters

Slice Type	Throughput (Mbps)	Latency (ms)
URLLC	[10, 20]	[1, 2]
eMBB	[50, 100]	[4, 10]
mMTC	[15, 25]	[10, 30]

B. Results

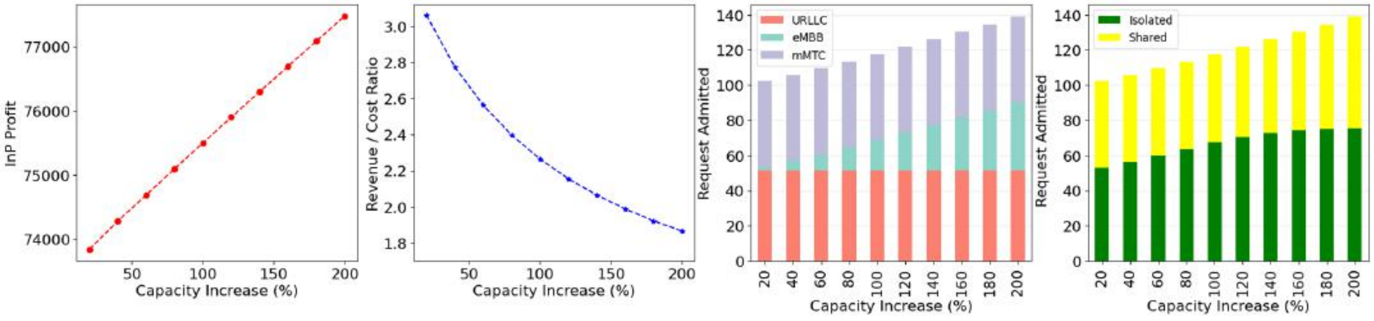
In our initial experiment, we progressively increased the number of slice requests to examine its effect on various parameters, as illustrated in Fig. 2. The InP profit, denoting the difference between revenue and cost, exhibited an upward trend with the increase in slice requests, as seen in Fig. 2(a). This is likely attributable to the rising count of admitted requests, corroborated by Fig. 2(c). However, an inverse trend was observed in the revenue-to-cost ratio as the number of requests escalated, until reaching a threshold near 90 requests as shown in Fig. 2(b). Post this point, the ratio sharply rose, indicating that the revenue from accepting requests outweighed the solution costs. This shift illustrates the correlation between admitting requests and incurring costs, which directly influences the generated revenue. Our model showed a preference towards admitting more URLLC and mMTC requests compared to eMBB, presumably due to the higher cost of placement and routing associated with eMBB and higher revenue associated with URLLC and mMTC in our setup as shown in Fig. 2(c). Interestingly, the number of isolated requests exceeds those that are shared as depicted in Fig. 2(d), indicating that there isn't an existing placement that can meet the requirements of other requests, necessitating the instantiation of new VMs. This highlights a clear trade-off between maintaining isolation and enhancing admission rates.

In our second experiment, we analyzed the effects of enhancing DU/CU computational capacities up to 200% for a total of 150 requests. As illustrated Fig. 3(a), an increasing trend in the objective value was noted. On a positive note, it was observed that the admission of eMBB requests experienced a beneficial impact with the increased capacities as shown in Fig. 3(c). This explains the decreasing revenue-to-cost ratio in Fig. 3(b) as increasing capacity allows for balancing the admission of all three request types at the expense of increased costs. This trend also indicates that simply escalating capacities may not necessarily contribute to reducing costs associated with the placement and routing of requests. Conversely, we note that requests permitting shared placements tend to benefit from increased capacities. As illustrated in Fig. 3(d), there's a nearly equal number of admissions for both isolated and shared requests.



(a) (b) (c) (d)

Fig. 2: Impact of increasing the number of requests: (a) InP profit, (b) Revenue-to-cost ratio, (c) Admitted requests by slice type, (d) Admitted requests by isolation type



(a) (b) (c) (d)

Fig. 3: Impact of increasing DU and CU capacities on (a) InP profit, (b) Revenue-to-cost ratio, (c) Admitted requests by slice type, (d) Admitted requests by isolation type

V. Conclusion and Future Work

This paper demonstrates the profound significance and potential of a joint optimization approach to address the complexities inherent in functional splits, RAN VNF placement in DUs and CUs, and traffic routing in a NG-RAN setting. Our proposed JointFSAPR model efficiently balances the intricate interdependencies and trade-offs among these critical aspects factoring in the operational cost determined by chosen functional splits and slice isolation requirements. The preliminary findings of this research offers significant insights for InPs. We show how optimal choices in functional splits, RAN VNF placement in DU/CU, and routing for admissible slices can boost InP profit. Moreover, increasing computational capacities improve slice admission and revenue to cost ratio. While off-the-shelf solvers like Gurobi can find optimal solutions, they are best suited for smaller-scale networks. For larger networks, Gurobi's execution time escalates, making it impractical for real-world settings due to the NP-hard nature of the problem. In our future work, we aim to develop a more efficient heuristic approach and will offer a detailed comparison of the results.

Acknowledgement

This work was supported in part by Rogers Communications Canada Inc. and in part by a Mitacs Accelerate Grant.

References

[1] P. Rost *et al.*, "Cloud technologies for flexible 5g radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, 2014.

[2] J. Bartelt *et al.*, "Fronthaul and backhaul requirements of flexibly centralized radio access networks," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 105–111, 2015.

[3] A. Maeder *et al.*, "Towards a flexible functional split for cloud-ran networks," in *2014 European Conference on Networks and Communications (EuCNC)*. IEEE, 2014, pp. 1–5.

[4] O. Chabbouh *et al.*, "Cloud ran architecture model based upon flexible ran functionalities split for 5g networks," in *International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 2017, pp. 184–188.

[5] A. Garcia-Saavedra *et al.*, "Fluidran: Optimized vran/mec orchestration," in *IEEE INFOCOM*, 2018, pp. 2366–2374.

[6] C.-Y. Chang *et al.*, "Flexcran: A flexible functional split framework over ethernet fronthaul in cloud-ran," in *IEEE International Conference on Communications (ICC)*, 2017, pp. 1–7.

[7] D. Harutyunyan *et al.*, "Flex5g: Flexible functional split in 5g networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 961–975, 2018.

[8] L. M. Moreira Zorello *et al.*, "Power-aware optimization of baseband-function placement in cloud radio access networks," in *International Conference on Optical Network Design and Modeling*, 2020, pp. 1–6.

[9] F. W. Murti *et al.*, "An optimal deployment framework for multi-cloud virtualized radio access networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2251–2265, 2021.

[10] C. Mei *et al.*, "5g network slices embedding with sharable virtual network functions," *Journal of Communications and Networks*, vol. 22, no. 5, pp. 415–427, 2020.

[11] W. da Silva Coelho *et al.*, "Function splitting, isolation, and placement trade-offs in network slicing," *IEEE Transactions on Network and Service Management*, vol. 19, no. 2, pp. 1920–1936, 2021.

[12] F. Z. Morais *et al.*, "Placeran: Optimal placement of virtualized network functions in the next-generation radio access networks," *IEEE Transactions on Mobile Computing*, 2022.

[13] L. G. Optimization, "Gurobi optimizer reference manual," 2020.