Isolation-Aware 5G RAN Slice Mapping Over WDM Metro-Aggregation Networks

Hao Yu, *Student Member, IEEE,* Francesco Musumeci, *Member, IEEE,* Jiawei Zhang, Massimo Tornatore, *Senior Member, IEEE,* and Yuefeng Ji, *Senior Member, IEEE*

***Abstract*—To accommodate the heterogeneous requirements of emerging 5G services, 5G Radio Access Network (RAN) will support the rising paradigm of network slicing, where a single physical RAN infrastructure is divided into multiple logical networks characterized by diverse requirements in terms of performance, cost, latency, and security, etc. The optimality of RAN slicing depends on various factors, such as slice requirements, resource availability and adopted network architecture. In 4G, RAN architecture evolved from a distributed architecture towards a 2-layer architecture that separates remote radio heads (RRHs) from baseband processing units (BBUs). However, this architecture has soon shown scalability limitations. To address this problem, the concept of flexible 5G functional split has been proposed, where RAN functions are distributed on a 3-layer architecture featuring a radio unit (RU), a distributed unit (DU), and a central unit (CU). Besides RAN architecture, isolation among different slices, which is necessary for slicing security, has a significant influence on the RAN slicing. In this work, we consider the isolation-aware RAN slice mapping problem considering advanced functional splits and a 3-layer RAN architecture. We propose a dual-objective heuristic algorithm for RAN slice mapping over a physical infrastructure represented by a wavelength division multiplexing (WDM) metro-aggregation networks. The proposed heuristic targets the minimization of 1) the number of active COs (i.e, hosting DUs and/or CUs) and 2) the number of established wavelength channels under constraints of network capacity and latency requirements. As our algorithm is also designed to map the RAN slices with least amount of physical resources for a given level of isolation, we also investigate the impact of slice isolation on resource utilization. Results show how higher isolation results in higher network cost.**

***Index Terms*—RAN Slicing, Flexible Functional Split, Slice Isolation, Slice Mapping, WDM Metro-Aggregation Network.**

I. Introduction

A large number of new mobile services is emerging in 5G mobile networks, such as Augmented/Virtual Reality (AR/VR), automonous driving, industrial automation, etc. These new services are characterized by heterogeneous requirements in terms of data rate, end-to-end latency and reliability [1] and can be broadly classified into three categories: enhanced Mobile Broadband (eMBB), ultra-Reliable Low-Latency Communication (uRLLC), massive Machine Type Communication (mMTC). eMBB services represent the high-bandwidth demanding applications, e.g., AR/VR and live video. uRLLC services are characterized by a strict requirement on real-time interaction, e.g., as in autonomous driving. mMTC services include both long-range MTC and broadband MTC with low-cost and low-power, e.g., smart services for metering, environment monitoring and traffic control in the urban and rural areas, which result in very high device density. Satisfying these diverse and often conflicting service requirements over the legacy 4G system will be economically unviable as dedicating network resources to each service will result in high capital and operational expenditures (CapEx and OpEx). Operators, in order to provide customized network services while keeping CapEx and OpEx under control, are considering the new concept of network slicing (NS) [2], which divides a physical network into several logical virtual networks, or "slices". Each slice is customized and dedicated to a certain type of service or to a specific application in the context of a service type. On the one hand, network slicing can help reduce CapEx (i.e., costs including equipment investment, civil work, and installation & commissioning) by sharing resources among multiple slices. On the other hand, flexible resource management enabled by software defined network (SDN) and network function virtualization (NFV) can also increase the resource efficiency while reducing OpEx (i.e., costs including operation & maintenance (OM), energy consumption, etc). [3].

Nowadays, the 5G RAN architecture is still evolving to accommodate the concept of functional splits together with network slicing (5G RAN slicing). In 4G, cloud-RAN (C-RAN) transformed the traditional distributed RAN (D-RAN) into a 2-layer architecture by centralizing the baseband function units (BBUs) into a common pool while leaving remote radio heads (RRHs) in the cell sites (CSs). It becomes soon clear that high bandwidth demands of the fronthaul network interconnecting RRHs and BBUs would make it economically unsustainable to deploy C-RAN everywhere, especially in fiber-poor rural or suburban areas [4]. Hence, in 5G, new functional splits have been proposed to flexibly divide RAN functions among centralized and distributed units depending on factors as the quality of service (QoS), latency, reliability, and network cost. Thanks to the increased flexibility in locating functions, these advanced functional splits can facilitate effective RAN slicing. 3GPP [5] has proposed a flexible 3-layer RAN architecture, where three new functional blocks are created: the radio unit (RU), distributed unit (DU) and centralized unit (CU). The transport network is correspondingly divided into three parts: fronthaul, midhaul, and backhaul.

5G communications impose high requirements on bearer

Hao Yu, Jiawei Zhang, and Yuefeng Ji are with the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, 100876, China (E-mail: yuhao92@bupt.edu.cn; zjw@bupt.edu.cn)

Francesco Musumeci and Massimo Tornatore are with Politecnico Di Milano, Milan, Italy (E-mail: francesco.musumeci@polimi.it; massimo.tornatore@polimi.it)

Corresponding author: Yuefeng Ji (jyf@bupt.edu.cn)

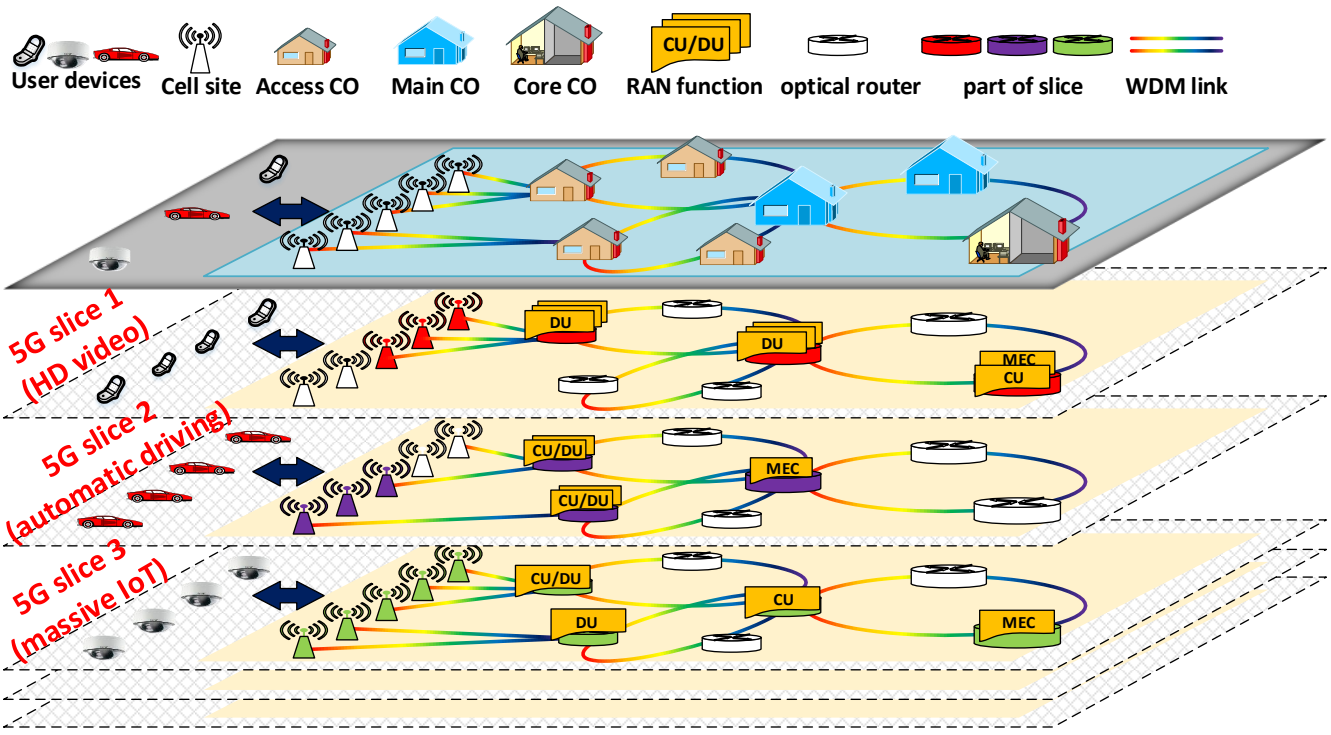


Fig. 1: 5G RAN slicing over WDM metro-aggregation network

network bandwidth, latency, flexibility, and cost. For example, in 5G, peak data rate (Gbit/s) will be 20 times higher and latency (ms) will be 10 times lower than the ones in 4G networks [6], which imposes great challenges to transport network. For fronthaul, wavelength division multiplexing-passive optical networks (WDM-PON) is the promising solution which might be an "easy and cheap" option [7]. For midhaul and backhaul portions of the 5G transport network, optical transport network (OTN) is the "best fit", as it is already widely deployed and can support the bandwidth, latency, and flexibility requirements of 5G services. Recently, China Telecom and ZTE have cooperated on the standardization of the Mobile-optimized OTN (M-OTN) technology [8] targeting at requirements of low latency, low power consumption, and low cost. In addition, optimal placement of DU/CU is crucial to improve resource efficiency while satisfying 5G requirements. The 3-layer architecture investigated in this paper significantly enhances the flexibility in DU/CU placement: CU and DU can be integrated or separated. In the early stage of 5G, DU/CU integration is adopted to promote low latency, low complexity operation and maintenance. However, in the mid-to-long term, the network will be upgraded to accommodate more challenging uRLLC and mMTC services, and DU/CU separation will become more beneficial [9] [10].

Although network slicing can improve resource efficiency and decrease network cost, when different tenants create slices over the same infrastructure, inter-slice security becomes a significant issue. Therefore, inter-slice isolation should be performed, where RAN functions and traffic of different slices are isolated from each other for security and privacy reasons. GSMA [11] has recommended different levels of isolation among slices. Slice functions and network connectivity can be partly or fully isolated according to the isolation level. In contrast to network sharing, isolation makes each slice operating over dedicated physical resources, hence reducing resource efficiency.

In this work, we focus on a new solution to perform isolation-aware RAN slice mapping to improve the processing and bandwidth resource utilization. As shown in Fig. 1, we consider RAN slicing in a multi-layer OTN over WDM metro-aggregation network, where several CSs, collecting mobile traffic, are distributed and interconnected with central offices (COs) by optical fiber links. COs are divided into multiple stages, depending on their positions in the aggregation hierarchy. Specifically, at the lowest network stage (say stage 0) CSs are present, while stages 1 and 2 are composed by Access COs and Main COs, respectively. And a single Core CO, acting as point of presence (PoP), represents the interface toward the core network segment. These metro COs are organized in "ring-and-spur" topology and each CO supports optical-electronic-optical (OEO) signal conversion, which can be used to perform traffic grooming, but introducing additional latency due to the switching and signal conversion. Besides the switching ability, metro COs can also be equipped with processing capacity for hosting RAN functions. In this work, we solve the RAN function (e.g., DU, CU, and MEC[1]) placement in the metro COs, jointly considering routing and wavelength assignment (RWA) for the traffic flows between functions. We

[1]Mobile Edge Computing server, we assume that all the services will be served in the MEC server

design a heuristic algorithm to solve the problem with two objectives: minimizing the number of active COs and minimizing the number of wavelengths. Then we investigate the interaction between RAN function placement and RWA, and evaluate the impact of capacity/latency constraints, and different isolation levels on the network and processing resource efficiency.

The rest of the paper is organized as follows. Section II gives a brief overview of the evolution of RAN and flexible functional split RAN architecture. In Section III, we define the concept of 5G RAN and explain how slice isolation affects RAN slicing. In Section IV, we introduce the isolation-aware 5G RAN slicing problem in a WDM metro-aggregation networks and propose a heuristic algorithm to address this problem. Illustrative numerical results are shown in Section V. Section VI concludes the work.

II. Related Work

The 5G RAN architecture and the concept of network slicing are still under study and standardization by several organizations, e.g., 3GPP, ITU-T, IEEE, etc. Next, we present recent related works on 5G RAN and network slicing.

A. RAN evolution

In recent years, C-RAN architecture has been widely investigated. Some studies compared D-RAN and C-RAN in terms of cost and power consumption. Ref. [12] [13] emphasized the impact of optical communication as an enabling technology in 5G networks. Ref. [14] analyzed the total cost of ownership (TCO) for constructing a RAN infrastructure, showing that the optical infrastructure plays a dominating role in the migration from D-RAN to C-RAN. Ref. [15] investigated CapEx when deploying C-RAN, showing cost saving in terms of optical switches compared to a distributed architecture. Recently, using a WDM metro-aggregation network for C-RAN deployment is under investigation (see, e.g., overviews in [16] [17]), Refs. [18] [19] investigated the BBU placement problem over an optical aggregation network, and proposed ILP and heuristic solutions for BBU placement in a 2-layer C-RAN architecture. Moreover, always in the context of 2-layer C-RAN, Refs. [20] [21] [22] proposed solutions to dynamically associate RRUs and BBUs through reconfigurable lightpaths in fronthaul for enhancing wireless performance metrics, such as energy efficiency, radio coordination gains, etc. Ref. [23] focused on the mapping between radio resource blocks and optical bandwidth resource in the beam-forming scenario. Considering the high bandwidth requirements for fronthaul, more flexible functional split between RAN functions are becoming a key proposition in designing new 5G RAN architecture. Ref. [24] performed a techno-economic study on how to design a low-cost C-RAN, showing that the proposed Hybrid-RAN with functional split can achieve up to 15% cost saving compared to D-RAN. Both Refs. [25] and [26] studied baseband function placement considering different functional splits. Ref. [25] presented a graph-model-based functional splitting, then investigated the deployment cost under delay constraints. Ref. [26] discussed different RAN functional splits with a specific emphasis on multiplexing gain in computational resources. To the best of our knowledge, existing works have only considered baseband function placement in 2-layer RAN, and there are no works analyzing DU/CU placement in 3-layer RAN. Therefore, in our preliminary work [27], we modeled a DU/CU placement problem for C-RAN deployment with a 3-layer RAN architecture using ILP. The results in [27] has proved that 3-layer RAN outperforms than 2-layer RAN in the aspect of baseband function consolidation. In this work, we emphasize on the impact of isolation on RAN slicing in the 3-layer RAN architecture.

B. 5G network slice

Ref. [28] proposed a grooming graph for dynamic slicing to minimize service blocking by jointly considering radio, transport and cloud domains. As understanding user's behavior is highly significant for effective network slicing, a big-data-analytics-based slice admission policy was proposed in [29], where traffic-prediction-based dynamic slicing is shown to achieve lower service degradation. All existing works discussed above investigate network slicing in traditional C-RAN or D-RAN, however, functional split and 3-layer RAN architecture will also have a significant impact on the performance of RAN slicing and will be the subject of investigation in this work.

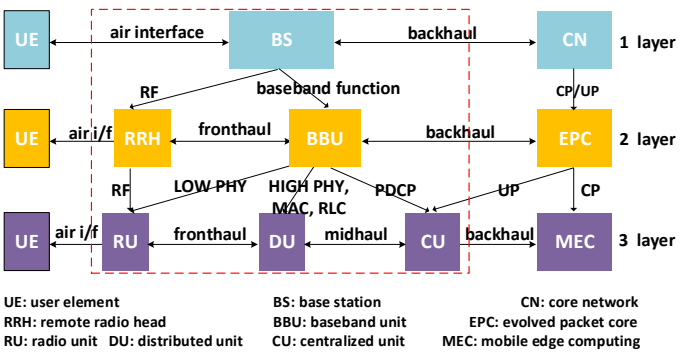


Fig. 2: RAN architecture evolution

III. Flexible RAN Architecture with Functional Split

As shown in Fig. 2, RAN architecture is evolving towards a more distributed functional decomposition. Following increased densification and coordination of small cells in 4G RANs, it has already been proposed to split traditional BS. Centralizing BBUs in a pool facilitates efficient resource management and cell coordination, which can reduce the total network cost. Because of high bandwidth demand between BBUs and RRHs, and emergence of diversified 5G applications, RAN is evolving towards a more flexible architecture, a 3-layer architecture has been proposed recently [5]. Part of physical layer functions are moved to the CS, combined with original RRH, called RU. Higher-layer PHY functions, MAC layer and RLC layer constitute the DU, while the CU includes PDCP layer functions and user plane functions of the core network. The MEC mainly includes control plane

functions from the original evolved packet core (EPC) and edge application server.

New network interfaces between RU, DU and CU are currently under standardization by 3GPP [5] and other alliances [30] [32], etc. As shown in Fig. 3, 3GPP has proposed multiple functional split options, typically listed from option 1 to option 8. Based on the proposal by 3GPP, ITU-T [33] recommends the split at Option 2 (interface F1) and Option 7 (interface Fx) as the standard split options between RU, DU and CU.

Transport bandwidth: The interface F1 at Option 2 (also known as midhaul interface) dynamically scales its bandwidth requirement with traffic load and requires only slightly more bandwidth than the backhaul interface. The bandwidth of fronthaul interface Fx at Option 7 also varies with the air interface traffic load but requires higher rates (typically by up to an order of magnitude compared to backhaul).

Processing complexity: The processing complexity of baseband functions can be measured in Giga Operation Per Second (GOPS) [34]. The processing complexity of low-PHY is related to carrier bandwidth and the number of antennas (called "cell-processing" functions, as shown in Fig. 3), whereas the processing complexity of MAC/RLC/PDCP layer and high-PHY layer functions are also related to the traffic load (called "user-processing" functions).

Transport latency: Transport latency requirements of backhaul and midhaul links are determined by service latency requirements, i.e., around 10 ms for eMBB, about 1 ms for uRLLC and ranging from 1 ms to several 10 ms for mMTC [35]. For the fronthaul link at Option 7, the latency is determined by the requirements of the RAN technology. For example, a latency budget of round-trip time RTT = 3 ms is available between a DU and its corresponding RU [36] due to Hybrid Automatic Repeat reQuest (HARQ) processing. The latency contributors modeled in this work include the propagation latency and switching latency: propagation latency depends by the distance between source and destination nodes of a connection, and it is assumed to be 5 $\mu s/km$; switching latency includes OEO and queueing time in switching nodes. The formulas below allow to calculate processing complexity and bandwidth requirement.

$$C_{RU} = k_R \cdot B \cdot A \tag{1}$$

$$C_{DU} = k_D \cdot B \cdot A \cdot L \tag{2}$$

$$C_{CU} = k_C \cdot A \tag{3}$$

$$C_{MEC} = k_M \cdot B \cdot L \tag{4}$$

$$D_f = k_{F1} \cdot B \cdot A \cdot L + k_{F2} \cdot A \tag{5}$$

$$D_m = k_M \cdot B \cdot L \tag{6}$$

$$D_b = k_B \cdot B \cdot L \tag{7}$$

where C is the processing complexity required by a given RAN function (DU, CU, etc.), D is the bandwidth required by the network connections, k is a coefficient related to the specific RAN function, B is the wireless bandwidth of the cell, and A is the MIMO number of the antennas, L is the traffic load under the cell, which is expressed in numbers of wireless resource blocks (RBs). The reference values for coefficient k for the various RAN functions are taken from [30].

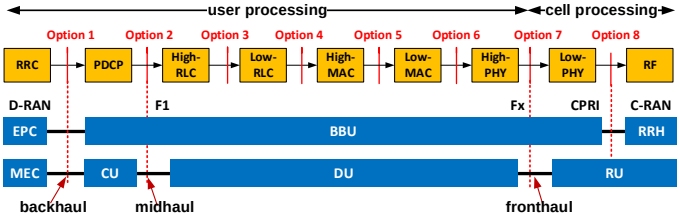


Fig. 3: Example of functional split options (adapted from [31])

IV. Isolation in RAN Slicing

A. Definition of 5G Network Slice

A network slice represents a collection of network functions and network connections dedicated to certain mobile services with similar requirements in bandwidth, processing, latency, reliability, etc. In this work, we assume that the mobile users within a slice are located in n adjacent CSs. For different types of slices, users' traffic are quite different and can be measured by the number of requested RBs. As an example, for eMBB slice, mobile users need higher data rate for transmitting a large amount of content, while for mMTC slice, users may not require high data rate, but need continuous data transmission during the whole lifecycle of slice. Similarly, for latency, an eMBB slice requires the user plane delay not to be larger than 4 ms, while fronthaul delay must be less than 100 μs, and midhaul delay must be less than 150 μs. Instead, a uRLLC slice requires stricter delay, namely, user plane delay must be less than 0.5 ms, while fronthaul delay must be less than 50 μs [37].

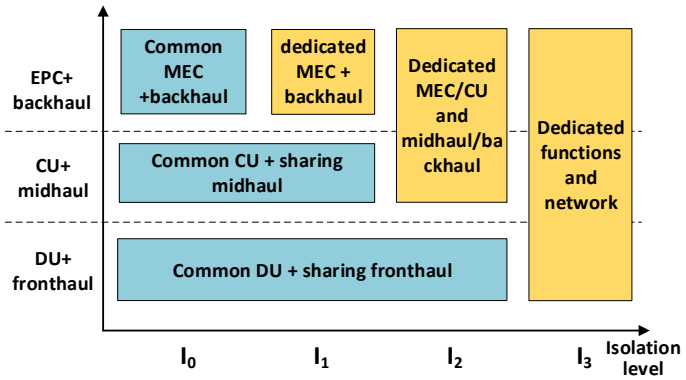


Fig. 4: Example of isolation levels

B. Isolation requirement

Isolation is regarded as one of the key features of network slicing. For security considerations, slice tenants may require a slice to be fully or partly isolated from other slices. Generally,

TABLE I: Network Requirement of Slice Request

Requirement	Description
coverage	$C_s = (c_1, c_2, ..., c_n)$
RB	$R_s = (r_1, r_2, ..., r_n)$
latency	$L_s = (l_f, l_m, l_b)$
isolation	$I_s = \{I : I_0, I_1, I_2, I_3\}$

RAN slice isolation is divided into two parts: processing isolation and transport isolation. Processing isolation refers to isolating functions of different slices into separate physical entities (e.g., by placing functions of different slices in different virtual machines (VMs) or different physical servers). Transport isolation means traffic should be transmitted in separated wavelengths, or even different physical links. Different RAN isolation levels have been defined [11], according to how many functions need to be isolated, as shown in Fig. 4. For example, slice tenants might call for a shared RAN with isolated core network functions, or both isolated RAN and core network functions. For isolation level I_0, all functions can share physical resource among multiple slices, including VM resources and transport resources. In I_3, all functions need to be isolated. As isolation level increases, more RAN functions need to be isolated. Higher isolation level implies higher network cost because each slice is provided with specialized VMs and transmission bandwidth. From operators' perspective[2], isolation requirements can be set in two different ways: i) slice tenants directly ask for isolation, or ii) tenants do not express any isolation requirement, but the operator takes decisions on isolation based on other requirements (e.g. service level agreements (SLA) or resource availability). We assume that i) within the same operator, isolation levels are different for different slices; ii) among different operators, slices must always be isolated. To achieve isolation between slices, NFV allows deploying network function instances to different VMs on general-purpose servers, while transport isolation can be implemented by OTN/WDM technology and FlexE technology which can provide independent wavelengths or time slots for different connections.

In summary, according to the requirements on processing complexity, transport bandwidth[3], latency, and isolation level, the properties of 5G RAN slice are summarized in Table I. Each generic slice request s is defined by its slice coverage, represented by a set of n cells $(c_1, c_2, ..., c_n)$, the number of RBs requested per cell, represented by $(r_1, r_2, ..., r_n)$, its latency requirement in fronthaul l_f, midhaul l_m and backhaul l_b, and its isolation level (to be chosen among I_0, I_1, I_2, I_3).

V. A Heuristic Algorithm for Isolation-Aware 5G RAN Slice Mapping Problem

A. Problem Statement

Since the metro-aggregation network has a hierarchical structure, it is beneficial to concentrate more network functions into fewer COs in higher network stages to achieve higher multiplexing gain. We define a metro CO as an "active" node if it hosts at least one active VM (hosting DU/CU) and then we define a metric called "consolidation ratio" $R = N_a/N_t$, where N_a represents the number of active nodes (COs), and N_t represents the number of total nodes in the networks. Lower values of R correspond to higher consolidation. Minimizing the number of active nodes (and hence R) by consolidating more DU/CU into fewer COs is beneficial to operators, especially for energy saving. However, more consolidation (i.e., lower R) comes at the cost of higher bandwidth consumption and higher latency, because traffic has to traverse longer paths and more COs. Moreover, the isolation requirements also affect processing/bandwidth utilization when mapping network slices into metro-aggregation networks.

TABLE II: Network Parameters

Parameter	Description
$G(N, E)$	Physical Topology
N	Set of physical nodes, $N = N_m \cup N_c$ N_m represents metro nodes, N_c represents CSs
E	Set of physical links, $e(i, j)$ represents the link with source i and destination j
C_i	Processing capacity of physical node i
$B_{i,j}$	Bandwidth of link $e(i, j)$ (number of wavelength)
$D_{i,j}$	Distance of link $e(i, j)$
V	Processing capacity of VM
W	Capacity of wavelength

Hence, in this study we investigate the ***isolation-aware 5G RAN network slice mapping problem*** in WDM metro-aggregation networks, that can be formally stated as follows. Given i) a hierarchical multistage metro-aggregation network topology, represented by a graph $G(N, E)$, where N is the set of nodes (including COs and CSs) and E is the set of optical fiber links, and ii) slice requests defined in Section IV.A, we decide how to map the RAN slice requests into a physical infrastructure (mapping includes functions placement and RWA for fronthaul/midhaul/backhaul traffic), with two objectives: i) minimizing the number of active nodes, constrained by network link capacity, latency and isolation requirement, ii) minimizing the established wavelengths, constrained by network processing capacity, latency and isolation requirement.

Note that, since the RU performs radio frequency and some physical layer functions, it always needs to be deployed on a dedicated processing platform at the cell sites. Hence, placement is only devised for DU, CU, and MEC.

B. Algorithm Description

Fig. 5 illustrates an example of RAN slice mapping. We assume mobile users are of two types, uRLLC users and eMBB users, which belong to different operators (colored by red, yellow, blue and green). Fig. 5(a) shows how the processing isolation works when mapping RAN functions. Due to low latency of a uRLLC service, all its functions will be placed in COs near the cell sites, see, e.g., the placement of blue and green users in Fig. 5(a). For example, note that, in the case of the DUs for blue users, due to latency constraints, all the functions can only be placed in access CO 1. For the eMBB users, the functions can be deployed at higher-stage

[2]In this work, we assume that there is only one physical infrastructure owned by infrastructure provider (InP), several operators buy/lease the physical resource and rent the resource to the slice tenants

[3]We use the requested RBs to represent the processing complexity and transport bandwidth, and refer to [30] for how to calculate processing complexity and transport bandwidth according to the number of RBs

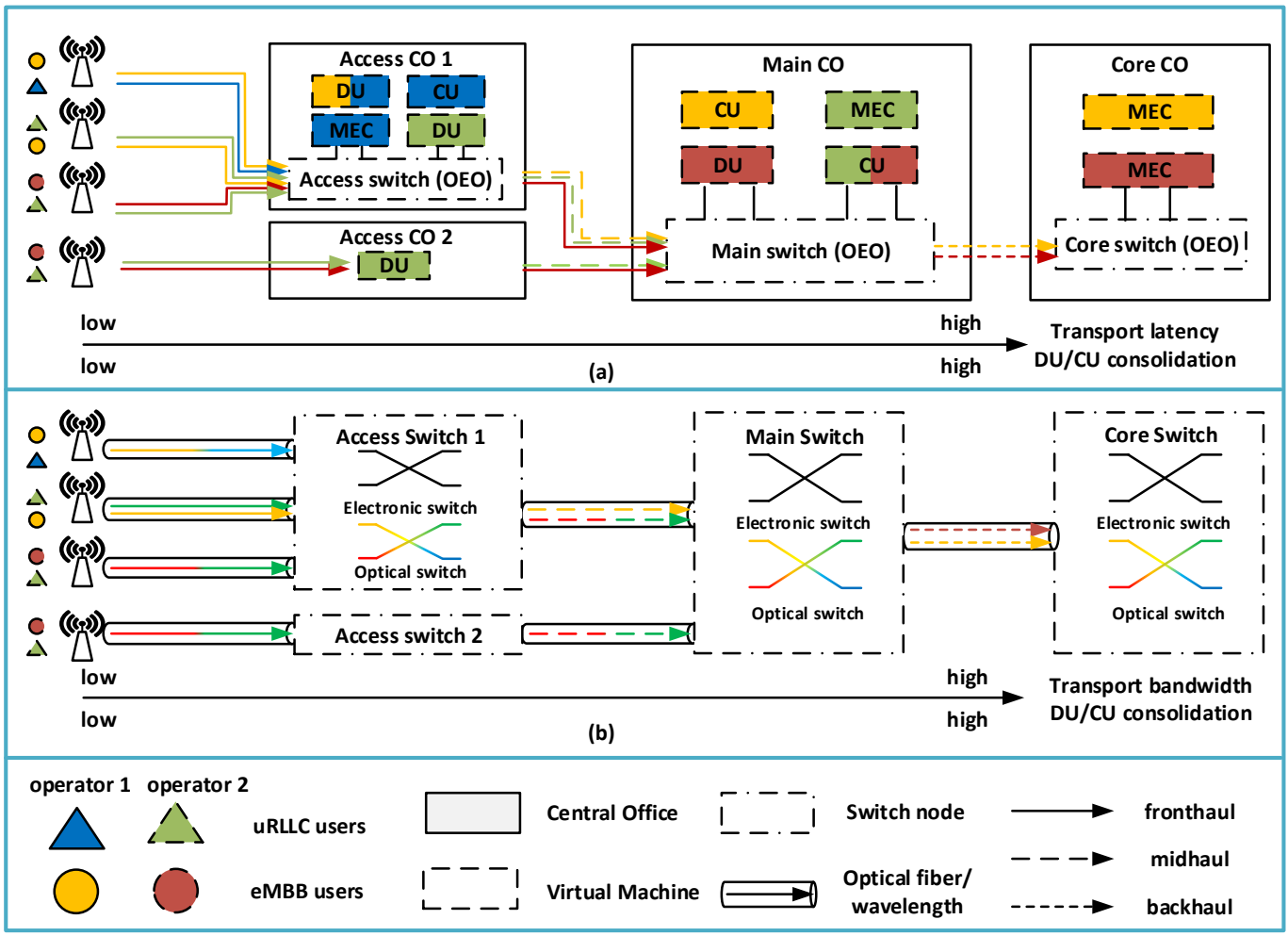


Fig. 5: Illustration of 5G RAN slice mapping

COs to achieve higher processing consolidation as latency constraint is looser. Placing functions in core CO enables higher DU/CU consolidation, but also higher transport latency. Concerning isolation, the blue DU and yellow DU are located in the same VM (dashed square) in access CO 1, however, the green DU is located in another VM, because green users belong to the different operators from blue and yellow users. Fig. 5 (b) shows the illustration of bandwidth isolation which also implies the relationship between DU/CU consolidation and bandwidth consumption, more consolidation comes at the cost of more bandwidth consumption.

The proposed algorithm addresses the 5G RAN slice mapping problem in a static scenario, i.e., when all the slice requests are given, the algorithm aims at accommodating all the slice requests and targets at the minimization of network and computational resource. According to problem statement, the network slicing problem in this work can be regarded as a service function chain (SFC) mapping problem which can be divided into node mapping and link mapping. As for node mapping, we solve it using bin-packing methodologies. Then after determining the locations of RAN functions, the connections between functions need to be mapped. This problem is a routing and wavelength assignment problem in a multilayer OTN over WDM aggregation network. As we know, SFC problem is an NP-Hard problem [38], hence, to find an scalable solution, we divide it into two objectives: 1) minimize the number of active nodes and 2) minimize the established wavelengths in the network. When considering objective 1, we set bandwidth capacity in each link as limited and the computational resource as always enough, so we can target at the minimization of active nodes. On the contrary, when considering objective 2, we set the computational resource in each CO as limited and the bandwidth resource as always enough, so we can aim at the minimization of established wavelengths in the network.

The execution of the algorithm is detailed in Algorithm 1. The characteristics of slice requests are given by the parameters shown in Table I, and the network parameters are summarized in Table II.

1) Node Ranking: For function mapping, the problem is solved by using existing bin-packing methodologies. All the physical nodes are ranked in a certain order considering the residual processing resource of nodes, the residual bandwidth of optical ports of nodes, the distance to the source node, and the node level (Access CO, Main CO, and Core CO) in the metro-aggregation network, then the RAN functions are

placed into the physical nodes with highest ranking value. To this end, we propose the following node ranking method:

$$F_n = \alpha U_n + \beta B_n + \lambda H_n + \gamma L_n \tag{8}$$

where F_n is the comprehensive evaluation index of the physical node n, U_n is the used processing resource of node n, and B_n is the used bandwidth resource by all the ports of node n, $B_n = \sum_{p \in P(n)} BW(p)$, $BW(p)$ is used bandwidth of each port of node n. H_n is the average distance between the node n and all the source node in slice request, $H_n = \frac{\sum_{r \in R} dis(r)}{|R|}$, $dis(r)$ is the shortest distance between node n and each cell site r. R is the set of source nodes in slice request. L_n is the stage level of node n in the metro networks. $\alpha, \beta, \lambda, \gamma$ are coefficients and $\alpha + \beta + \lambda + \gamma = 1$.

The coefficient $\alpha, \beta, \lambda, \gamma$ can be adjusted according to different slice types and algorithm objectives. For example, for the eMBB slice, the value of α can be larger, so that the nodes with more remaining processing resource are more likely to be chosen. For uRLLC type slice, the value of β should be larger compared to other coefficients, so that the nodes closer to the cell sites are more likely to be selected.

2) RAN slice mapping: Firstly, the algorithm sorts the slice requests in S in descending order according to the total amount of RBs requested by each slice, put the sorted slice requests into S', then iterates the slice requests in the set S' (line 1). Next, we perform node ranking (line 2) for current slice request s, then sort the CSs of slice s in descending order of requested RBs in each CS, then put the sorted CSs into set C' (line 4). Next, we pop up the CS c in C', calculate the processing requirement of DU, CU and MEC and the bandwidth requirement of fronthaul, midhaul and backhaul using formula (1) $\sim$ (7) (line 6).

- **Objective 1: Minimize the active nodes**. For the function f in $\{DU, CU, MEC\}$, check whether there is at least one path between node n' in set N' and CS c whose distance is lower than $l_f/l_m/l_b$, if there exists such a path, check if there is enough available bandwidth in each link along the path (line 10). Then try to place function f in the node n' with $MapFunction(f, n')$ (line 11), otherwise, iterate next node in N'. During the placement of function f into the physical nodes, isolation is the main consideration when performing $MapFunction(f, n')$. First, we try to re-use the existing active VM v, since reusing active VMs can maximize utilization of processing resources, then we check if there is an isolation conflict between the function f and the functions already in the VM v. (1) Check if the function f belongs to the same operator with functions in VM v. (2) According to the function types, check the isolation level. For DU, if the already placed functions in v and function f are both with isolation level $I_0 \sim I_2$, then function f can be placed in VM v sharing the processing resource with other slices. If either function is with isolation level I_3 or their operators are different, the function f should be placed by creating a new VM (line 29-41). Link mapping is after the function f being placed, since there is already an available path satisfying the latency and bandwidth constraints, then the isolation is the only consideration when mapping connection l between function f and its former function into physical links. Just like function placement, transport isolation means two connections that need to be isolated should be mapped into two separated wavelengths. For fronthaul, if the connection l and connections in existing wavelengths are both with isolation level $I_0 \sim I_2$, then map this connection l, otherwise, establish a new wavelength.
- **Objective 2: Minimize the established wavelengths** Different from objective 1, the minimization of established wavelengths requires another solution for link mapping. In objective 2, the capacity of physical nodes is set to be limited and bandwidth capacity is set infinite. In the node mapping phase, after node ranking, all the ranked nodes are checked if satisfying with processing capacity and latency constraints. Then, among the nodes which satisfy the constraints, calculate the number of links of the path between candidate nodes and CS c, choose the node which has a minimum number of links to CS c in order to avoid establishing more wavelengths along the lightpath. Once the candidate node is decided, call the $MapFunction(f, n')$ and $MapLink(l, n^*)$ to finish the node mapping and link mapping.

Algorithm 1 5G RAN slice mapping scheme

Input: slice requests ***S***, physical topology ***G(N,E)***
Output: RAN functions placement, routing and wavelength allocation of each slice

1: Sort the slice requests in ***S*** in descending order according to the total amount of RBs requested by each slice, put the sorted slice requests into ***S'***
2: Set the parameter α, β, χ according to the type of slice request, calculate the node evaluation factor F_n, put the ranked nodes into ***N'***
3: **for** all slice request s in ***S'*** **do**,
4: sort the CSs belonging to s in descending order according to the requested RBs, put the sorted CSs into ***C'***
5: **for** all CS c in ***C'*** **do**
6: calculate the processing requirement of RU, DU, CU and MEC function and bandwidth requirement of fronthaul, midhaul and backhaul with formula (1) $\sim$ (7). For each CS, RU function is deployed locally in CS.
7: **for** each function f in DU, CU, MEC **do**
8: **Objective 1: Minimize the active nodes**
9: **for** all node n' in ***N'*** **do**
10: **if** (1) there is at least one lightpath between node n' and CS c whose latency is lower than $l_f/l_m/l_b$, (2) if the lightpath exists, whether there is enough bandwidth available in each link on this ligthtpath l **then**
11: MapFunction(f, n')
12: MapLink(l, physical links)
13: **end if**
14: **end for**
15: **Objective 2: Minimize the active wavelengths**
16: **for** all node n' in ***N'*** **do**

if (1) whether there is at least one lightpath between node n' and CS c whose latency is lower than $l_f/l_m/l_b$, (2) if there is enough processing capacity in the node n' **then**
put node n' into set N''
end if
for each node n'' in N'' **do**
calculate the number of links between node n'' and CS c and choose the lightpath l with minimum number of links between node n^* and CS c
end for
MapFunction(f, n^*)
MapLink(l, physical links)
end for
end for
end for
end for
function MAPFUNCTION(function, physical node)
check the isolation level of this slice request s
if the isolation level is I0 $\sim$ I2 **then**
find a existing active VM v with enough processing capacity
if the slice s' in VM v and slice s belong to the same operator and the isolation level of slice s' is also I0 $\sim$ I2 **then**
map the DU into this VM v
else
create a new VM to instantiate this DU
end if
else
create a new VM to instantiate this DU
end if
end function
function MAPLINK(lightpath, physical links)
for all candidate links n in the path L **do**
if the isolation level is I0 $\sim$ I2 **then**
find a existing active wavelength w with enough bandwidth capacity
if the slice s' in wavelength w and slice s belong to the same operator and the isolation level of slice s' is also I0 $\sim$ I2 **then**
map the lightpath into this wavelength w
else
create a new wavelength to map this lightpath
end if
else
create a new wavelength to map this lightpath
end if
end for
end function

VI. Illustrative Numerical Results

A. Case Study

We initially consider a small-scale RAN slice deployment scenario, with 9 metro nodes and 6 cell sites in the network, at most 2 wavelengths at 1 Gbps in each fiber link and VM capacity set to 100 GOPS. The radio configuration of each sector of a cell site is 20MHz carrier bandwidth, QPSK, and 2 $\times$ 2 MIMO and 100 RB under this configuration (1 RB for 180kHz and 7 OFDM symbols). All the slice requests belong to the same operator. Each slice request contains 5 cell sites, and the requested RBs follow a normal distribution with an average 10 and deviation 5. In addition, there is no isolation requirement in this case. Using this simplified case, we validate the performance of our heuristic against the ILP in [27].

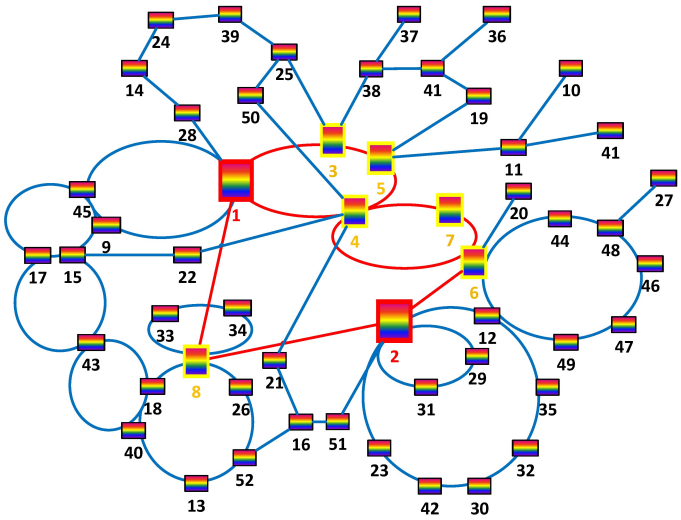


Fig. 6: Large-scale network topology

Then we consider the large-scale topology in Fig. 6, with 52 metro nodes, consisting of 2 core COs, 6 main COs, 44 access COs, each access CO is associated with 3 CSs, thus 132 CSs in this topology (not shown in the figure for the sake of clarity). At most 40 wavelengths at 10 Gbps can be established in each fiber link. VM capacity is 1000 GOPS and node capacity is 3/10/20/50 VMs for cell sites/access CO/main CO/core CO. The radio configuration is in line with the guidelines of a 5G urban aggregation network, as specified in [39], which is 100MHz, 256 QAM and 8 $\times$ 8 MIMO. Under this radio configuration, 500 RBs are available in each cell. Multiple slices can share these 500 RBs in each CS, which means that the number of requested RBs of all the slices in one CS can not exceed 500. We assume that the requested RBs of each CS within one slice are normally distributed, for the eMBB/uRLLC/mMTC type, the average number of requested RBs is 100/30/30. Moreover, the number of CSs in a slice is uniformly distributed between [5,10]/[5,10]/[20,30] for eMBB/uRLLC/mMTC type. Three operators are considered in this case. The maximum tolerated fronthaul/midhaul/backhaul latency for eMBB/uRLLC/mMTC type under the considered RAN split are shown in the Table III, derived from [40] [41].

Node ranking is essential for the heuristic since it provides a criterion during the node mapping phase, which will affect the performance of the heuristic. Therefore, the parameters in (8)

TABLE III: Latency requirement of front/mid/backhaul for eMBB/uRLLC/mMTC Type

	user plane latency	fronthal	midhaul	backhaul
eMBB	$4ms$	$\leq 100\mu s$	$\leq 1ms$	$\leq 20ms$
uRLLC	$0.5ms$	$\leq 50\mu s$	$\leq 50\mu s$	$\leq 500\mu s$
mMTC	$3 \sim 10ms$	$\leq 250\mu s$	$\leq 10ms$	$\leq 20ms$

TABLE IV: Reference value of $\alpha, \beta, \lambda, \gamma$ for different objectives and different slice types

objective	slice type	α	β	λ	γ
1	eMBB	0	0.3	0	0.7
1	uRLLC	0	0	0.3	0.7
1	mMTC	0	0.1	0.1	0.8
2	eMBB	0.8	0	0.1	0.1
2	uRLLC	0.2	0.2	0.6	0
2	mMTC	0.4	0	0.3	0.3

should be set accordingly[4]. The value of parameter $\alpha, \beta, \lambda, \gamma$ for different slice types is presented in the Table IV. All the parameters used in the numerical evaluation are summarized in Table V.

B. Discussion

In the following numerical evaluation, we define the following metrics: (1) consolidation factor R; (2) number of created VMs; (3) number of established wavelengths, and we study the impact of slice demand, slice types, isolation level and network capacity on these metrics.

1) Small Topology: We solve the RAN slice mapping problem using ILP and heuristic with small topology. The details of ILP are shown in our previous work [27] and the heuristic only performs objective 1. In this case, no isolation is performed and no specific slice type is enforced. Fig. 7 shows the comparison between ILP and heuristic in terms of consolidation factor R. Lower value for R represents more consolidation, i.e., less COs are activated to host the RAN functions. As shown in Fig. 7(a), ILP achieves the best performance (smaller values) in terms of R, but heuristic results lay approximatively within 10% of the ILP. Leveraging its finer granularity in placing RAN function, 3-layer architecture achieves a better performance than in 2-layer architecture. In fact, when placing functions, the functions with less processing requirements are more likely to be mapped into the VMs without exceeding residual capacity.

Besides residual capacity, the RAN function placement also depends on other factors, e.g., latency requirement (in this case, we set the user plane latency requirement between RU and MEC varying from 100 μs to 1000 μs, and the fronthaul latency requirement is fixed value with 250 μs [42] in this case). As shown in Fig. 7(b), the value of R increases and reaches a maximum value at 500 slice requests, then decreases with the maximum latency requirement. When the latency requirement is strict, almost all the functions are placed in the cell sites (red bar). When the latency requirement is relatively loose, part of functions (e.g., CU and MEC) tend to be placed in the higher stage of networks, while few functions (e.g., DU) are still located in cell site due to the strict fronthual latency requirement, thus the total active COs increase. When the latency requirement is loose enough, most of the functions will be placed in the core COs remaining few DUs in the cell sites.

Fig. 7(c) shows the impact of the number of wavelengths on the consolidation factor R. Besides the latency requirement, the link bandwidth capacity also influences the RAN function placement. For instance, fronthaul connections are characterized by high bandwidth demand, shortage of bandwidth capacity in physical links will force the RAN functions to be placed in the COs near the CSs. Therefore, the more bandwidth capacity in physical links, the more consolidation in processing resource. In addition, 3-layer architecture is also proved with a better performance in function consolidation because of the flexible RAN functions placement and traffic routing than 2-layer architecture.

2) Large Topology: In this case, we solve the problem only by a heuristic as the ILP cannot solve this case in a reasonable time. In Fig. 8(a), we compare the performance of heuristic in terms of R for 3 slice types: eMBB, uRLLC and mMTC. It shows that uRLLC slices activate most COs, which is the lowest consolidation in processing resource since the latency requirements of uRLLC slices are much stricter than the other slice types. As the latency requirement of mMTC slices is the loosest among the three types, mMTC case achieves the highest resource consolidation. Note that, to clearly show the impact of latency requirement on processing resource consolidation, we set the overlay as a metro-aggregation network transport solution, where each traffic flow is allocated with dedicated wavelength without traffic grooming. Therefore, the slice type (also latency requirement) is the only impacting factor in this case.

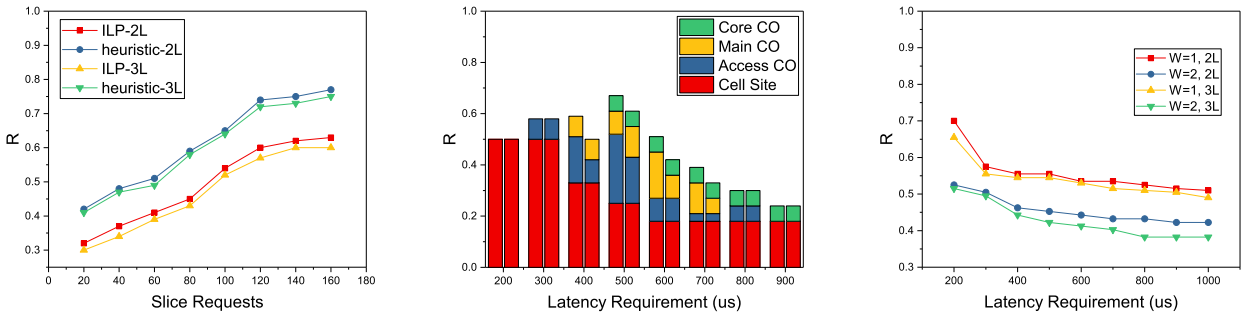
Next, the results show the impact of transport network capacity (i.e., number of wavelength W) on the consolidation factor R in Fig. 8(b). If $W = 1$, the number of active nodes increases rapidly with slice requests, when the number of slice requests is 100, almost all the CSs and part of COs are active for processing the RAN functions since there is not enough transport bandwidth, most of the functions must be placed locally. If $W \geqslant 10$, the curves rise smoothly and more processing consolidation can be achieved.

Then, the impact of slice isolation as a key requirement when mapping RAN slices is evaluated, and we also investigate the computational resource consolidation under VM isolation. To clearly show the impact of isolation on resource utilization, we change the metric from R to the number of created VMs. We investigate the number of created VMs with the increasing slice requests for different types of slices. As shown in Fig. 9(a), due to the huge traffic demand of eMBB service, the number of VMs created for eMBB slices is the highest among the three types. On the other hand, we can observe that the better performance of 3-layer architecture

[4]For the objective 1, minimization of active nodes is the main objective, the weight of node level should be set highest and weight of used processing resource is set to be 0, other parameters could be set according to the slice types. For example, the weight of bandwidth is supposed to be higher in order to increase the possibility of multiplexing gain in bandwidth for eMBB case, the weight of latency should be set higher in order to increase the possibility of mapping uRLLC slice.

TABLE V: Parameters used in the numerical evaluation

	Parameter	Values for small-scale network	Values for large-scale network
Wireless part	Carrier bandwidth	20	100
	Modulation Format	QPSK	256 QAM
	# of MIMO	2×2	8×8
	# of Resource Blocks	100	500
Metro part	# of COs	9	52
	# of CSs	6	132
	# of wavelengths per link	2	≤ 40
	Wavelength capacity	1 Gbit/s	10 Gbit/s
	VM capacity	100 GOPS	1000 GOPS
Slice part	# of operators	1	3
	Type of slices	NULL	eMMB/uRLLC/mMTC
	average # of requested RBs	10	100/30/30
	# of CSs	5	[5,10]/[5,10]/[20,30]
	Isolation Level	I_0	I_0, I_1, I_2, I_3
	Latency requirements	$[100\mu s \sim 1000\mu s]$	Table III
Others	$\alpha, \beta, \lambda, \gamma$	0.2,0.2,0,0.6	Table IV



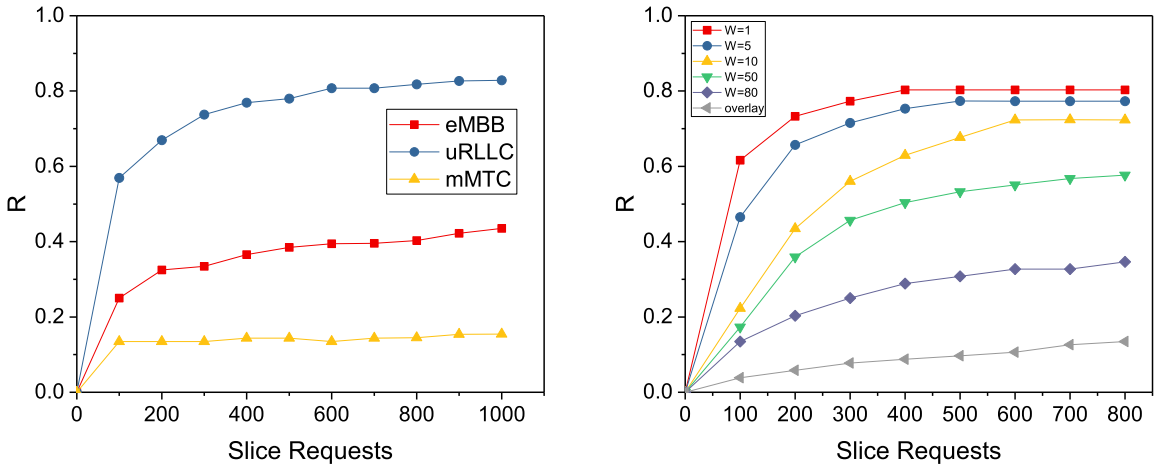
(a) (b) (c)

Fig. 7: (a) ILP vs heuristic in R with a small topology; (b) R values vs maximum service latency requirement (left: 2-layer, right: 3-layer); (c) R value vs maximum service latency requirement with different W

measured in the number of active VMs. Especially, processing consolidation of eMBB case with 3-layer architecture is the highest among three slice types, e.g., almost 28% less created VMs for 1200 slice requests.

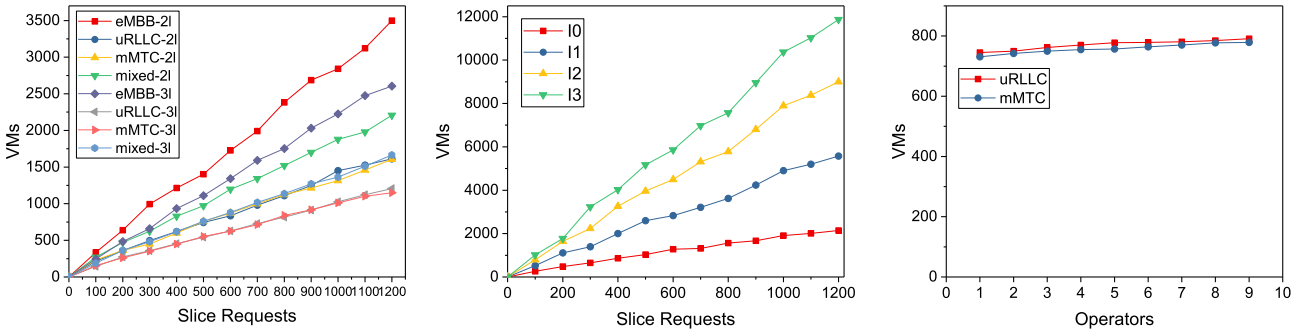
In the results above, the isolation level of all the slice requests is set to be I_0, which means no isolation is adopted, all the RAN functions share the whole physical infrastructure as long as there are enough available capacity in the VMs. In the following, the number of created VMs with different isolation levels are shown. In this case, we consider three operators and three slice types mixed randomly. As shown in Fig. 9(b), when the isolation level changes from I_0 to I_3, the number of created VMs increases by 6 times on I_3 with respect to I_0. In addition, we set the number of slice requests and isolation level fixed to show the impact of the number of operators on created VMs for both uRLLC and eMBB, as shown in Fig. 9(c). We find that the number of operators has not a prominent influence as much as the isolation level on the created VMs. We can envision that more operators (possibly, virtual operators) can co-exist in the same infrastructure sharing the whole network resource without excessive cost increase, even if all the virtual operators operate separately.

In Fig. 10, we perform the objective 2 to minimize the established wavelengths (the sum of active wavelengths in all the physical links) and evaluate the relationship between the number of established wavelengths and slice types. In this case, we set the isolation levels of slices as level I_3. Firstly, we can observe that the difference in established wavelengths between 2-layer and 3-layer seems not as obvious as in created VMs. Secondly, the number of wavelengths of different slice types shows different trends at different stages. For eMBB slices, because of the high bandwidth demand, a large amount of wavelengths are established in the beginning stage. However, for uRLLC slices, the established wavelengths increase rapidly and exceed the ones for eMBB slices when the traffic load is high. As we know, different slice types mean different latency requirements, because of the low latency property, the RAN functions of uRLLC slices are more likely to be distributed in the networks, there will be less possibility for the multiplexing of bandwidth in fewer wavelengths. On the contrary, more connections of eMBB slices can be multiplexed in the higher stages of metro networks. Although in the beginning, eMBB slices consume more wavelengths because of high bandwidth demand, the uRLLC slice will consume many more wavelengths than the eMBB case eventually.



(a) (b)

Fig. 8: (a) R value with increasing slice request for different slice types; (b) R value with increasing slice request with different W



(a) (b) (c)

Fig. 9: (a) Number of VMs vs slice requests for different slice types; (b) Number of VMs vs slice requests with different isolation levels; (c) Number of VMs vs number of operators for different slice types

Finally, we present the number of established wavelengths under different isolation levels. In this case, three types of slices are randomly mixed. As shown in Fig. 11, the result shows that higher isolation level, more wavelengths established. Note that the type of isolation level does not have a significant effect on the number of wavelengths (in fact, the four curves are very close to each other) at low traffic load. Because a large number of new wavelengths are established in the beginning, with the increasing of slice requests, the slice tends to be mapped in existing wavelengths, so the curves increase slowly after a certain point.

VII. Conclusion

In this paper, we investigate the problem of 5G RAN slicing considering a functional-split based 3-layer RAN architecture over WDM metro-aggregation networks, with dual-objective of 1) minimizing the number of active metro nodes, 2) minimizing the number of established wavelengths. To solve this problem, we propose a RAN slice mapping heuristic for the RAN functions placement and RWA of mobile traffic considering the capacity and latency constraints, as well as slice isolation requirements. Results show that the RAN slice mapping in 3-layer architecture outperforms the one in 2-layer architecture in terms of processing resource consolidation (e.g., created VMs decrease by 28% less with 3-layer architecture on eMBB case), due to the more flexible placement with a finer granularity of functions in 3-layer architecture. Finally, the trade-off between network sharing and slice isolation is investigated, and the results show that the slice isolation is achieved at the cost of higher resource occupation (6 times more VMs on I_3 case than I_0 case).

Acknowledgment

This work was supported by the National Key R&D Program of China (No. 2018YFB1800802), the National Nature Science Foundation of China Projects (No. 61871051), the

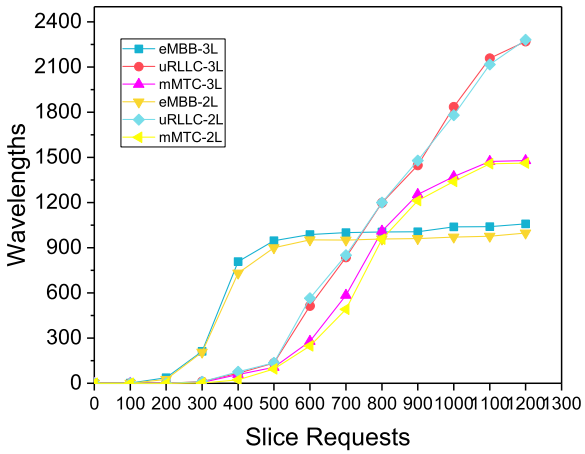


Fig. 10: Number of wavelengths vs slice requests for different slice types

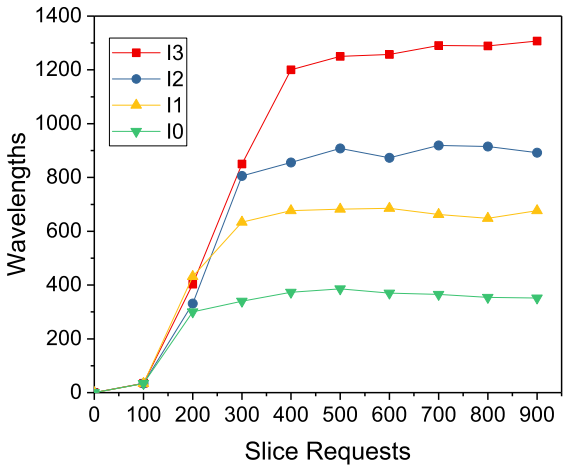


Fig. 11: Number of wavelengths vs slice requests with different isolation levels

Beijing Natural Science Foundation (No. 4192039), and the fund of State Key Laboratory of Advanced Optical Communication Systems and Networks, China, No. 2019GZKF5, China Scholarship Council Foundation, and the European Community under grant agreement No.761727 Metro-Haul project.

REFERENCES

[1] NGMN, "NGMN 5G White Paper," NGMN Alliance, 2015.
[2] NGMN, "Description of Network Slicing Concept," NGMN Alliance, Jan. 2016.
[3] M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska and P. Monti, "Demonstration of dynamic resource sharing benefits in an optical C-RAN," in IEEE/OSA Journal of Optical Communications and Networking, vol. 8, no. 8, pp. 621-632, Aug. 2016.
[4] "Assessment of candidate transport network architectures for structural convergence" COnvergence of fixed and Mobile BrOadband access/ aggregation networks- COMBO Project, Tech. Rep., June, 2016. [Online]. Available: https://www.ict-combo.eu.
[5] 3GPP TR 38.801 V14.0.0 (2017-03), "Radio access architecture and interfaces," (Release 14).
[6] "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond." 2015.
[7] N. J. Gomes, P. Chanclou, P. Turnbull, A. Magee, and V. Jungnickel, "Fronthaul evolution: from CPRI to Ethernet," Optical Fiber Technology 26 (2015): 50-58.
[8] "China Telecom's Requirements on 5G Transport," China Telecom, 2017. Available: https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20171016/Documents/3.Rick.pdf.
[9] "China Telecom 5G Technology White Paper," June 2018. Available: http://www.chinatelecom.com.cn/2018/ct5g/201806/P020180626325685163826.pdf.
[10] "The transport network considerations for 5G in CMCC," Available: https://www.opennetworking.org/wp-content/uploads/2018/12/The-transport-network-consi-deration-for-5G-in-CMCC.pdf.
[11] GSMA, "Network Slicing Use Case Requirements Whitepaper," April, 2018.
[12] Y. Ji, J. Zhang, X. Wang and H. Yu, "Towards converged, collaborative and co-automatic (3C) optical networks," in Science China Information Science, vol. 61, no. 12, pp. 11301, 2018.
[13] Y. Ji, J. Zhang, Y. Xiao and Z. Liu, "5G flexible optical transport networks with large-capacity, low-latency and high-efficiency," in China Communications, vol. 16, no. 5, pp. 19-32, May 2019.
[14] S. S. Lisi, A. Alabbasi, M. Tornatore and C. Cavdar, "Cost-effective migration towards C-RAN with optimal fronthaul design," 2017 IEEE International Conference on Communications (ICC), Paris, 2017, pp. 1-7.
[15] L. Velasco, A. Castro, A. Asensio, M. Ruiz, G. Liu, C. Qin, R. Proietti, and S. J. B. Yoo, "Meeting the Requirements to Deploy Cloud RAN Over Optical Networks," in IEEE/OSA Journal of Optical Communications and Networking, vol. 9, no. 3, pp. B22-B32, March 2017.
[16] T. Pfeiffer, "Next generation mobile fronthaul and midhaul architectures," in IEEE/OSA Journal of Optical Communications and Networking, vol. 7, no. 11, pp. B38-B45, 1 November 2015.
[17] X. Liu and F. Effenberger, "Emerging optical access network technologies for 5G wireless," in IEEE/OSA Journal of Optical Communications and Networking, vol. 8, no. 12, pp. B70-B79, December 2016.
[18] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, and S. Gosselin, "Optimal BBU Placement for 5G C-RAN Deployment Over WDM Aggregation Networks," in Journal of Lightwave Technology, vol. 34, no. 8, pp. 1963-1970, 15 April15, 2016.
[19] F. Musumeci, O. Ayoub, M. Magoni and M. Tornatore, "Latency-aware CU placement/handover in dynamic WDM access-aggregation networks," in IEEE/OSA Journal of Optical Communications and Networking, vol. 11, no. 4, pp. B71-B82, April 2019.
[20] J. Zhang, Y. Ji, X. Xu, H. Li, Y. Zhao and J. Zhang, "Energy efficient baseband unit aggregation in cloud radio and optical access networks," in IEEE/OSA Journal of Optical Communications and Networking, vol. 8, no. 11, pp. 893-901, Nov. 2016.
[21] J. Zhang, Y. Ji, S. Jia, H. Li, X. Yu, and X. Wang, "Reconfigurable Optical Mobile Fronthaul Networks for Coordinated Multipoint Transmission and Reception in 5G," in IEEE/OSA Journal of Optical Communications and Networking, vol.9,no.6, pp. 489-497, 2017.
[22] H. Yu, J. Zhang, Y. Ji and M. Tornatore, "Energy-efficient dynamic lightpath adjustment in a decomposed AWGR-based passive WDM fronthaul," in IEEE/OSA Journal of Optical Communications and Networking, vol. 10, no. 9, pp. 749-759, Sept. 2018.
[23] J. Zhang, Y. Xiao, D. Song, L. Bai, and Y. Ji, "Joint Wavelength, Antenna, and Radio Resource Block Allocation for Massive MIMO Enabled Beamforming in a TWDM-PON Based Fronthaul," in IEEE/OSA Journal of Lightwave Technology, vol.37, no.4, pp. 1396-1407, 2019.
[24] X. Wang, L. Wang, S. E. Elayoubi, A. Conte, B. Mukherjee and C. Cavdar, "Centralize or distribute? A techno-economic study to design a low-cost cloud radio access network," 2017 IEEE International Conference on Communications (ICC), Paris, 2017, pp. 1-7.
[25] J. Liu, S. Zhou, J. Gong, Z. Niu and S. Xu, "Graph-based framework for flexible baseband function splitting and placement in C-RAN," 2015 IEEE International Conference on Communications (ICC), London, 2015, pp. 1958-1963.
[26] M. Shehata, A. Elbanna, F. Musumeci and M. Tornatore, "Multiplexing Gain and Processing Savings of 5G Radio-Access-Network Functional Splits," in IEEE Transactions on Green Communications and Networking, vol. 2, no. 4, pp. 982-991, Dec. 2018.
[27] H. Yu, F. Musumeci, Y. Xiao, J. Zhang, M. Tornatore and Y. Ji, "DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks," 2019 IEEE Optical Networking Design and Modelling, Athens, 2019.

[28] M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska and P. Monti, "Dynamic slicing approach for multi-tenant 5G transport networks [invited]," in IEEE/OSA Journal of Optical Communications and Networking, vol. 10, no. 1, pp. A77-A90, Jan. 2018.
[29] M. R. Raza, A. Rostami, L. Wosinska and P. Monti, "A Slice Admission Policy Based on Big Data Analytics for Multi-Tenant 5G Networks," in Journal of Lightwave Technology, vol. 37, no. 7, pp. 1690-1697, 1 April1, 2019.
[30] "Small Cell Virtualization: Functional Splits and Use Cases," Small Cell Forum White Paper, rel. 7, July 2016.
[31] ITU-R Technical Report, "Transport network support of IMT-2020/5G, GSTR-TN5G," May 2018.
[32] Next generation Mobile Network (NGMN) Alliance, "Project RAN evolution: Further study on critical C-RAN technologies," Mar. 2015.
[33] G.sup.5GP, "5G Wireless Fronthaul Requirements in a PON Context," ITU-T (expected to be released by October 2018).
[34] B. Debaillie, C. Desset and F. Louagie, "A Flexible and Future-Proof Power Model for Cellular Base Stations," 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), Glasgow, 2015, pp. 1-7.
[35] Nokia, "White Paper 5G new radio network," Document code SR1803023634EN (April).
[36] 3GPP TS-36.213 (Physical layer procedures). (2015). [Online]. Available: http://www.3gpp.org.
[37] FUJITSU, "New-Transport-Network-Architectures-for-5G-RAN," 2018.
[38] S. Khebbache, M. Hadji, and D. Zeghlache, "Virtualized network functions chaining and routing algorithms," Computer Networks 114 (2017): 95-110.
[39] "Deliverable 3.3: Analysis of Transport Network Architectures for Structural Convergence," COnvergence of fixed and Mobile BrOadband access/ aggregation networks- COMBO Project, Tech. Rep., Jul. 2015. [Online]. Available: https://www.ict-combo.eu.
[40] IEEE Standards Association. "Dimensioning Challenges of xhaul," Reference document for discussion in meeting[S].
[41] R3-161813, "Transport requirement for CU&DU functional splits options," CMCC.
[42] Netmanias, "Fronthaul Size: Calculation of maximum distance between RRH (at cell site) and BBU (at CO)," Tech-Blog.