MOBILE COMMUNICATIONS AND NETWORKS

Impact of Network Densification on Joint Slicing and Functional Splitting in 5G

Behnam Ojaghi, Ferran Adelantado, Angelos Antonopoulos, and Christos Verikoukis

A discussion on he challenges and open issues of joint slicing and functional splitting is posed, taking into account the additional complexity resulted from instantiating multiple slices per DU, each one with a different functional split. The authors show that the additional complexity, which is translated into higher cost, is worthwhile depending on the RAN deployment density.

ABSTRACT

The virtualization of the fifth generation radio access network (RAN), which distributes the next generation Node B functions between a central unit (CU) and a distributed unit (DU), along with the emergence of slicing as a cornerstone of mobile communications networks, pose new challenges. One of these challenges is the management of the joint slicing and allocation of appropriate distribution of functions between CU and DU, known as functional split. In this work, a discussion on the challenges and open issues of joint slicing and functional splitting is presented, taking into account the additional complexity resulting from instantiating multiple slices per DU, each one with a different functional split. It is shown that the additional complexity, which is translated into higher cost, is worthwhile depending on the RAN deployment density.

INTRODUCTION

The fifth generation (5G) has been designed to handle the current and expected increase in mobile data traffic demand generated by a wide range of verticals and characterized by diverse requirements. According to the Ericsson Mobility Report released in June 2021, the number of 5G subscriptions worldwide will be 3.5 billion by 2026, and the global mobile data traffic will reach 236 exabytes/month [1]. In this context, the diversity of requirements and scenarios has led to the definition of three general use cases, namely enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC) [2]. eMBB corresponds to traffic with high data rate requirements, with up to 20 Gb/s peak data rate and 100 Mb/s everywhere (e.g. virtual and augmented reality), whereas URLLC services are mainly characterized by two key performance indicators (KPIs): low latency around 5 ms and very high reliability of 99.9999 percent (e.g., automated driving). Finally, mMTC defines services with massive deployment (up to 1 million devices/km^2) of low cost, low energy consumption devices (10 years of battery life) with relaxed data rate and latency requirements (e.g., smart city applications). The support of such diverse requirements simultaneously is a challenge. For instance, the high data rates demanded by eMBB services require tight coordination between neighboring cells to exploit advanced techniques such as coordinated multipoint (CoMP). This can be achieved by centralizing network functions of different cells in a single location. Instead, moving computation close to the users with the implementation of edge computing is required to achieve URLLC. The traditional architecture of the radio access network (RAN), designed as a *one-size-fits-all* approach, is unable to meet the extremely diverse requirements of the different use cases simultaneously, since it cannot be reconfigured to be adapted to each service. To overcome these limitations, the 5G network is designed on the basis of flexibility and reconfigurability to enable the adaptation of the network to the traffic demand [3]. Thus, the network can be dynamically set up on the fly to fit the service requirements.

Flexibility and reconfigurability are achieved by leveraging the network with network function virtualization (NFV) and software defined networking (SDN) technologies. Both technologies are enablers for the dynamic creation of network slices, defined as logical networks that provide specific network capabilities and network characteristics aimed at supporting a specific service on top [3]. Thus, slices tailored to guarantee the requirements of each service can be instantiated when needed.

NFV opens up the opportunity to overcome the traditional monolithic network function (NF) implementation and lead to a distributed allocation of virtual network functions (VNFs) among different nodes. This is relevant in 5G, where the RAN node, known as next generation Node B (gNB), is split up into two logical nodes: centralized unit (CU) and distributed unit (DU). A gNB has one CU and one or several DUs. The protocol stack is partially allocated in the CU and in the DU [4], thus providing the RAN with the flexibility to implement both gNB logical nodes in either a distributed or monolithic manner. Although the monolithic gNB implementation, with CU and DU in the same physical node, is not precluded, the distributed allocation of CU and DU enables the RAN to exploit the whole potential of CU/DU, which is detailed in subsequent sections. CU and DU are connected through an integrated fronthaul/backhaul network, often referred to as *X-haul* [5].

Ferran Adelantado and Behnam Ojaghi are with the Universitat Oberta de Catalunya (UOC), Spain; Angelos Antonopoulos is at Nearby Computing S.L., Spain; Christos Verikoukis is at University of Patras, Greece; ATHENA – Industrial Systems Institute, Greece; and Iquadrat Informatica S.L., Spain.

Digital Object Identifier: 10.1109/MCOM.001.2100680

The distributed implementation of the gNB poses strict connection requirements between CU and DU, since maximum allowed latency and minimum data rate are imposed by the distribution of VNFs.

The 3rd Generation Partnership Project (3GPP) proposes eight possible VNF separation options between CU and DU denoted as *functional splits* [4]. As a rule of thumb, centralized functional splits (i.e., more VNFs allocated in the CU) result in high data rate and low latency requirements in the X-haul, whereas decentralized functional splits lead to relaxed requirements [6]. In terms of service, the requirements of the service determine the suitability of each functional split. For instance, for services demanding low latency and high reliability (i.e., URLLC), decentralized functional splits are required since the allocations of the medium access control layer (MAC) in the DU enables it to meet hybrid automatic repeat request (HARQ) maximum latency. Conversely, for services requiring higher gNB coordination (i.e., eMBB), centralized functional splits provide better coordination, resource sharing, and so on. There is not necessarily a one-to-one mapping between a service and a functional split, although in general there are subsets of functional splits that fit the requirements of each service.

There are proposals for optimal functional split selection per DU to get the network adapted to its characteristics (data rate and latency in the CU-DU connection) and the services' needs [7, 8]. However, existing works do not consider functional splitting based on the slice, thus allowing the slicing of a DU and the implementation of different functional splits per slice in a shared DU. In this context, joint slicing and functional split selection arises as a critical aspect enabling mobile network operators to build virtualized networks tailored to meet a variety of demands with diverse quality of service requirements. Instantiating multiple slices over a shared gNB (CU and DU), each with the most suitable functional split, is a promising solution, although it adds complexity. It may impact the design of protocols and requires programmable hardware to potentially support different functional splits per slice in a shared gNB and a common protocol architecture in the NFV environment addressing multiple slices controlled by an SDN controller.

In our previous work [9], a dynamic joint functional split and RAN slicing algorithm was proposed. It was apparent that instantiating each slice with the most suitable functional split can enhance the performance of the network. However, the complexity of instantiating multiple slices in a shared gNB, each with a different functional split, was not discussed. Here, we propose a joint functional splitting and slicing solution, discuss the open research issues, and analyze and discuss the performance of the dynamic joint functional split and RAN slicing algorithm in scenarios with different RAN node density. This work explores its potential gain and discusses how RAN density impacts the performance, thus resulting in conclusions on when the gain is worth the added complexity and cost. It is shown that in dense scenarios, the gain achieved is compromised, while the network performance benefits from it in sparsely deployed scenarios. The contributions of this work are summarized as:

- Analysis of the joint functional splitting and slicing performance as a function of the RAN density
- Discussion of the open issues and challenges

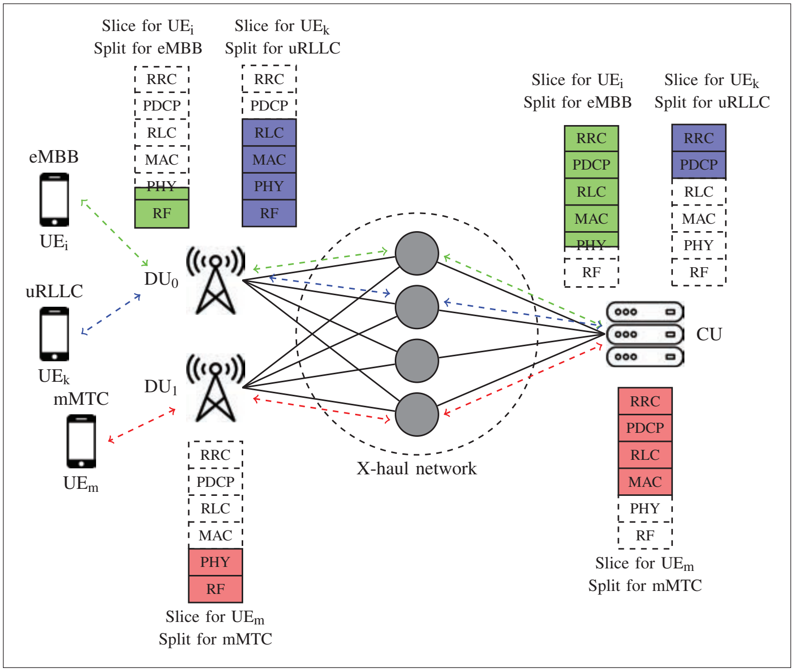

FIGURE 1. Example of joint slicing and functional splitting. Each color denotes a slice. Protocol layers are colored where implemented.

RAN ARCHITECTURE

Initial implementations of legacy C-RAN used point-to-point CU-DU connections, but integrated fronthaul/backhaul networks, namely X-haul, were shown to provide more flexibility and cost reduction [5]. The X-haul connects CU and DU, and carries different types of packets, such as IP packets or IQ samples encapsulated with enhanced Common Public Radio Interface. On top of the network, slices are created and managed. Figure 1 shows a simple example of the joint functional split and slicing concept. Three user equipments (UEs), UE$_i$, UE$_k$, and UE$_m$, are served by two DUs. UEs require eMBB, URLLC, and mMTC services, respectively. To adapt the network to the requirements of each service, the network creates two slices on the shared D$_0$ for UE$_i$ and UE$_k$, and a slice on DU$_1$ for UE$_m$. Slices are created across the corresponding DU, X-haul, and CU. In the slice created for UE$_i$, lower layers of the gNB are run in the DU (radio frequency [RF] and lower physical layer [PHY] functions), whereas the remaining layers are run in the CU. This distribution allows the implementation of joint processing and advanced receivers in the CU, needed to provide high data rate. The slice created for UE$_k$ moves most layers down to the DU (RF, PHY, MAC, radio link control [RLC]) and leaves the radio resource control (RRC) layer and Packet Data Convergence Protocol (PDCP) layer at the CU. This functional split reduces HARQ process delay, implemented in the DU, and allows the provision of URLLC services. The slice created for UE$_m$ implements an intermediate functional split to serve mMTC traffic. In Fig. 1, the X-haul is a simple network, although more complex network architectures are not precluded.

The isolation of slices is key to guarantee the requirements of services running on top, and accurate estimates of resources (computing, memory, and network capacity) are required. Moreover, the complexity of isolating different slices running on shared hardware increases when the functional split of each slice is different.

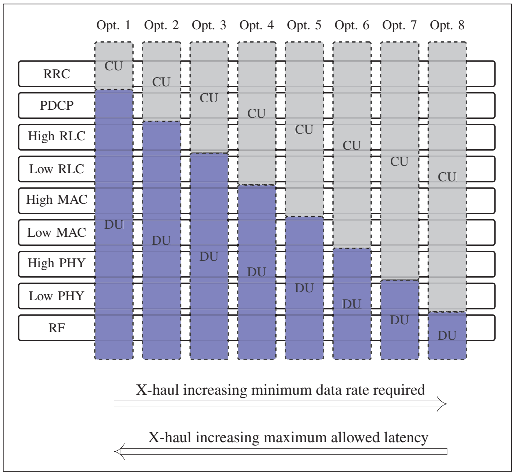

FIGURE 2. Functional splits/options defined by 3GPP and allocation of protocol stack layers in the DU and CU.

RAN Slicing and Functional Splitting

Functional Split Options

The operation of the gNB is modeled as a chain of functions, where functional splits define the placement of the functions [4]. 3GPP defines eight functional split options [6], each with the distribution of functions between CU and DU shown in Fig. 2.

The higher the centralization degree, the higher the resources management efficiency, the lower the DU complexity and cost. Conversely, X-haul data rate and latency requirements become more restrictive as centralization increases. For instance, option 8 requires X-haul data rate of around 2.5 Gb/s for 20 MHz bandwidth and 250 μs maximum latency. Instead, option 2 requires 150 Mb/s in the downlink and 50 Mb/s in the uplink, and a latency of tens of milliseconds, as shown in [4, Annex A].

The allocation of each layer determines the characteristics of the functional split. When PDCP is located in the DU (option 1), there is control and user planes separation. It is suitable for edge computing and consequently for URLLC. When moving PDCP toward the CU (options 2 to 8), the centralized aggregation of traffic from 5G and Long Term Evolution — Advanced (LTE-A) is enabled.

RLC is responsible for performing ARQ. When ARQ is centralized in the CU, the reliability of the X-haul is improved, and buffering and computational requirements in DU are reduced. It is suitable for non-ideal X-haul and corresponds to options from 3-1 (sub-option of 3) to 8.

The allocation of high MAC determines the centralization/decentralization of the scheduler. Options 5 to 8 implement centralized scheduling in the CU, which is suitable for inter-cell coordination. However, it means tight X-haul latency requirements.

Lower-layer splits are defined by options 7 and 8. Three different splits of PHY are defined, denoted by sub-options 7-1, 7-2, and 7-3.

Option 8 centralizes all layers in the CU except for RF. The main advantages are the isolation of RF components, which facilitates PHY upgrades, reuse of RF components, and more efficient management resources.

Despite the range of functional splits, there is a subset of representative ones. 3GPP recommends option 2 for highly decentralized applications, where cell coordination is not required, and latency and bandwidth are limited in the X-haul [4]. Option 6 is pushed by the SCF [10] as the optimal split for low-cost, low-capacity deployments. The O-RAN Alliance [11] supports option 7-2 (specifically the 7-2x variant) for networks with high capacity and high reliability requirements. This work evaluates the network performance considering the representative options 2, 6, and 7-2. Although not addressed in this work, there can be other factors impacting the functional split selection. A good example of these factors can be found in [12], where an analysis of the best protocol layer to aggregate multi-connectivity flows is proposed.

RAN Slicing and Isolation

3GPP defines the management and orchestration for slicing in [13]. Although this work is focused on the RAN, the slice is created across the RAN and the core network to guarantee minimum requirements in terms of bandwidth, end-to-end latency, reliability, data rate, and security. In [14], all processes involved in 5G network slicing are detailed. The isolation of slices is key to guarantee the requirements of services running on top, and accurate estimates of resources (computing, memory, and network capacity) are required. Moreover, the complexity of isolating different slices running on shared hardware increases when the functional split of each slice is different. Aspects such as inter-slice coordination or sharing of common functions arise as implementation challenges. The increase in the complexity of the DU is translated into high deployment costs due to the massive number of DUs. Therefore, there is a clear trade-off between the complexity of the DU (i.e., cost) and the gain achieved.

Impact of Network Densification on Joint Slicing and Functional Splitting

Given the added complexity resulting from the joint slicing and functional splitting, the density of the RAN — defined as the number of gNBs per area unit — impacts the cost of deployment (larger number of complex RAN nodes). In other words, is the network performance improvement in terms of throughput and served users achieved by joint slicing and functional splitting per slice worth the added complexity and cost?

The advantage of instantiating multiple slices and functional splits on a shared DU lies in the reduction of the distance between users and serving DUs. As services can only be served with a subset of functional splits, only the subset of DUs with a proper functional split can accommodate the slice. Modelling this effect is equivalent to removing the subset of DUs with inappropriate functional split from the set of feasible DUs for instantiating the slice.

Following our previous work [9], where capacity and structure of the X-haul were analyzed, hereafter, the impact of the RAN density (i.e., the number of deployed DUs) on the joint slicing and functional splitting is investigated. Thus, the description of the joint functional splitting and slicing algorithm is described in the sequel.

The slice creation and management are subject to a set of constraints. Such constraints are imposed by DU, CU, X-haul network, or services.

The slice creation depends on these constraints:

C_1: Spectrum availability. Imposed by the spectrum allocated to a specific DU, the available spectrum (i.e., number of physical resource blocks) must be enough to serve the requirements of the UE.

C_2: Computational capacity. The functional split defines the set of VNFs placed in CU and DU. The processing of these VNFs has a cost and requires computing resources. Centralizing computational resources improves the resources' efficiency and reduces the cost. Before creating a slice, computational resources of CU and DU are checked to guarantee the accommodations of the slice.

C_3: Service requirements. Service requirements determine the suitable functional split. Not all functional splits are fitted to the service requirements.

C_4: X-haul data rate and latency. The X-haul traffic routing must meet the diverse requirements of functional splits in terms of data rate and latency.

Figure 3 shows the flowchart of the algorithm to create joint slicing and functional splitting in the network detailed earlier. The algorithm is initialized by creating an ordered list of candidate DUs per UE, from the DU with the best channel quality to the DU with the worst channel quality. For each service and UE demanding that service, conditions C_1, C_2, C_3, and C_4 are checked one by one with the DUs in the ordered candidate list of the UE. When all constraints are fulfilled for a DU, the UE is associated with that DU, and the slice is created or scaled. Also, routing constraint C_4 is set to guarantee that the X-haul supports the slice and functional split requirements. If constraints are not fulfilled by a UE, the UE is not served. The provision of the slices is on the fly, since it is subject to the dynamics of the traffic. As shown in Fig. 3, the algorithm allocates the slices in the following order: URLLC, mMTC, eMBB. This is done like this because services are *complementary* in terms of required resources. As aforementioned, URLLC is the first service to be allocated, since it needs large edge computing resources. Conversely, its X-haul requirements are substantially smaller than those required by mMTC and particularly by eMBB. Therefore, even though URLLC users hoard most DU computing resources, they just consume a small share of the X-haul resources. This complementarity facilitates the solution of the algorithm.

The proposed algorithm scales well with the number of UEs and DUs thanks to the characteristics of the problem. In cellular networks, the nature of the wireless medium limits the size of the candidates' list of the UEs, since only close DUs are included. Therefore, the increase in the number of DUs is not directly translated into larger candidates' list sizes. Moreover, the algorithm allocates first the URLLC slice (Fig. 3), with large DU computational requirements. This fact quickly limits the size of the candidate list as the algorithm progresses, thus improving the scalability. Finally, if the computational capacity in the CU is large enough, solutions become more geographically bounded, because optimal solutions depend mainly on its proximity. Therefore, large-scale scenarios can be divided into smaller scenarios and the computational runtime can be reduced.

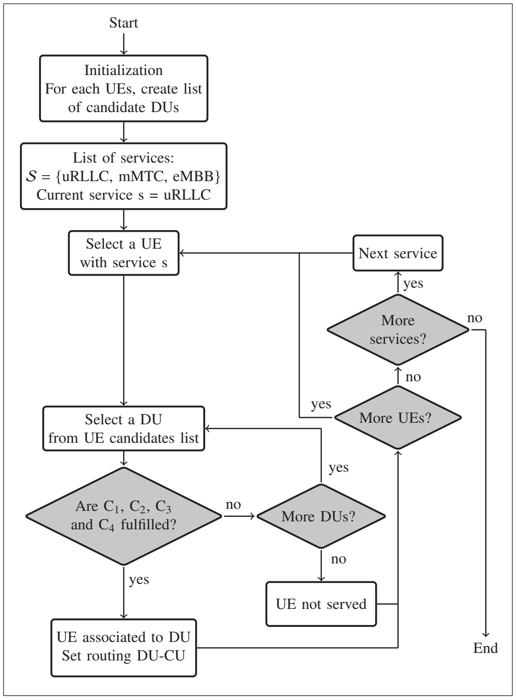

FIGURE 3. Flowchart of the algorithm to allocate UEs and create slices.

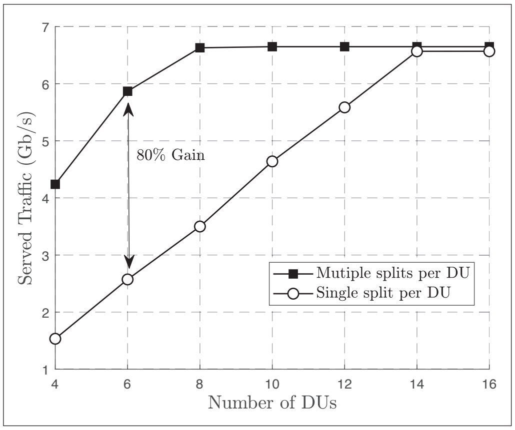

FIGURE 4. Served traffic for 4 to 16 DUs. The offered traffic is 6.65 Gb/s, and the capacity of the X-haul links is ∞. Results have 95 percent confidence level.

PERFORMANCE EVALUATION

Simulations have been conducted with the RAN architecture shown in Fig. 1 and a variable number of DUs. The X-haul is composed of four forwarding nodes, each connected point-to-point to the CU and to all DUs, creating a one-hop connection between the CU and each DU. The processing time of forwarding nodes is assumed negligible, and the capacity of the links is varied in different simulations. Each DU has a bandwidth of 20 MHz and a processing capacity equal to 1 CPU reference core per gigabit per second. As for the CU, the computing capacity is set to 100 CPU reference core per gigabit per second. The processing capacity required by the RLC and MAC functions is set to 0.75 CPU reference core per gigabit per second, and the processing capacity required by high PHY functions is set to 3.25 CPU reference core per gigabit per second. Three services are assumed, one for each traffic class, to showcase a complex 5G scenario. For URLLC class, a medical application with a data rate equal to 120 kb/s is considered. This traffic requires a decentralized functional split to guarantee high reliability and low latency (i.e., functional split option 2).

The proposed algorithm scales well with the number of UEs and DUs thanks to the characteristics of the problem. In cellular networks, the nature of the wireless medium limits the size of the candidates' list of the UEs, since only close DUs are included.

The described functional split selection is based on the RAN characteristics (architecture, resources, etc.). However, the creation and management of end-to-end slices across the RAN and the core network is more complex and requires further research.

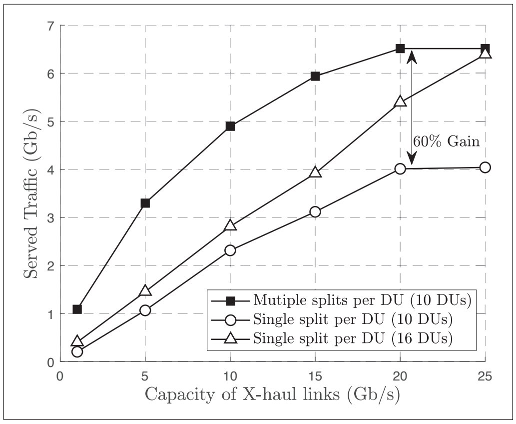

FIGURE 5. Traffic served by the network when the offered traffic is 6.65 Gb/s and the X-haul links' capacity ranges from 100 Mb/s to 25 Gb/s. Results have 95 percent confidence level.

For mMTC class, an Internet of Things (IoT) service is considered with a data rate of 30 kb/s. This service requires functional split option 6. Finally, a video streaming application with 30 Mb/s data rate is simulated for eMBB class. As explained, this service requires a centralized functional split such as option 7-2. The percentage of devices of each class are 15, 80, and 5 percent for URLLC, mMTC, and eMBB, respectively, and each device demands a single service. In terms of required data rate between CU and DU, for functional split 2 the data rate is equal to the served traffic data rate; for functional split 6 the required data rate is equal to 1.5 Mb/s plus 1.02 times the served traffic data rate (due to signalling, etc.); and for functional split 7-2 the required data rate is 2.5 Gb/s for 20 MHz bandwidth (IQ samples are exchanged). Both UEs and DUs are distributed uniformly in a square-shaped scenario of 30,000 m^2. Simulations have been conducted on a CPU processor of Core i7-8550U with 16 GB of RAM.

Results obtained with the algorithm detailed earlier are labeled *Multiple splits per DU*. In order to observe the gain achieved by instantiating different slices with different splits on a shared DU, results obtained with the additional constraint of using a single functional split per DU, as proposed in [15], are also plotted and labeled *Single split per DU*. Results have been compared to the optimal solution, and the gap is less than or equal to 2 percent.

Figure 4 shows the throughput of the network when the offered traffic is equal to 6.65 Gb/s. In this figure, the capacity of the X-haul links is assumed to be ∞ to analyze the effect of the number of DUs without X-haul constraints. As observed, the performance achieved with *Multiple splits per DU* is significantly better than with *Single split per DU*, reaching up to an 80 percent increase when 6 DUs are simulated. The reason for this behavior lies in the mean distance between the UE and the serving DU. With *Multiple splits per DU* the probability of having a nearby DU with a suitable functional split increases, thus improving the spectral efficiency. Moreover, the number of alternative DUs to get a UE served with proper split increases, and so does the flexibility of the RAN to associate UEs and DUs. However, when the density of the RAN (i.e., the number of DUs) increases, the gap between the two alternatives is reduced. Note that the increase in the density of DUs reduces the mean distance between UEs and serving DUs. Therefore, the gain achieved from instantiating multiple splits with different functional splits per DU vanishes.

Results shown in Fig. 4 are obtained without X-haul capacity constraints. In Fig. 5, served traffic is plotted for *Multiple splits per DU* and *Single split per DU* when 10 DUs are deployed and for X-haul link capacities ranging from 100 Mb/s to 25 Gb/s. The X-haul capacity has a huge impact on the served traffic. In the case of *Multiple splits per DU*, served traffic only reaches the values observed in Fig. 4 when the capacity of the links is around 20 Gb/s. This is because eMBB devices generate most of the traffic despite being only 5 percent of the UEs, and these users require functional split option 2, which results in an extremely high data rate in the X-haul. By inspecting the X-haul network, it is also clear that the X-haul bottleneck resides in the links between the forwarding nodes and the CU, since they aggregate the traffic of all the DUs. However, the impact of the X-haul links capacity constraint is even higher for the *Single split per DU*. While the *Multiple splits per DU* approach reaches the value obtained in Fig. 4 with 20 Gb/s links capacity, the *Single split per DU* approach does not reach the values of Fig. 4.

Figure 5 also includes the *Single split per DU* approach with 16 DUs. It can be seen that even in the case of having constraints in the link capacity, the *Single split per DU* approach can reach the results of the *Multiple splits per DU* by increasing the number of DUs.

Results show that the proposed solution provides higher throughput than *Single Split per DU*, particularly for sparse deployments. Given that the improvement depends on the RAN density, the joint slicing and functional splitting can be seen as an alternative to increasing the number of DUs in sparse deployments.

OPEN ISSUES AND CHALLENGES

The joint slicing and functional splitting poses significant challenges and opens up research opportunities. In the following, the most relevant ones are discussed.

End-to-end functional split selection and slicing. The described functional split selection is based on the RAN characteristics (architecture, resources, etc.). However, the creation and management of end-to-end slices across the RAN and the core network is more complex and requires further research. As discussed for the RAN, the 5G Core (5GC) network will also go through a virtualization process. The end-to-end slicing will have to consider not only the functional split, but also the rest of the RAN VNFs, the core VNFs, and, if the far end of the service is another UE, the far end functional split. Multi-access edge computing (MEC) will also play a key role in the slice creation. The pool of virtualized resources offered by MEC will provide each slice with differentiated capabilities, ranging from edge computing capability for URLLC to caching resources for eMBB.

Accurate models. The optimal selection of the functional split, the placement of VNFs, and, in general, the creation of slices rely on the accurate estimate of resources required by each slice. In the literature, there are initial analytical and experimentation-based models to estimate the computational and latency requirements of the different VNFs. However, it has been shown that results depend on a wide range of factors, such as the particular platform on which VNFs are instantiated. The reliability of the models has a huge impact on the slice provision and the efficiency of the resource usage, since the over- or under-estimation of VNFs' resource requirements results in low resource usage efficiency and lack of slices' isolation, respectively. Artificial intelligence driven methods are expected to fill the existing gap.

Operation and isolation of slices with different functional splits in a shared DU. Isolation is one of the key aspects of network slicing. This isolation becomes more complex when multiple slices with different functional splits are instantiated in a shared DU. The main problem lies in determining which functions/operations can be shared between slices, the RF components sharing, and so on while maintaining isolation. In terms of inter-slice coordination, when slices have different functional splits (e.g., one slice with functional split option 2 and another slice with functional split option 7), information could be distributed between CU and DU, and coordination can require sub-frame time latency constraints. How these functions are shared when instantiated in the same physical node, and how they coordinate with each other when they are implemented in different physical nodes, is still an open research problem.

Scheduling of different slices through a common X-haul. The use of multiple functional splits in a DU results in the transmission of completely different data/signaling through the X-haul. For instance, centralized functional splits (e.g., options 8 or 7) require high data rate and low delay through the X-haul and must be prioritized.

Conversely, decentralized functional splits, such as option 1, have more relaxed latency and data rate requirements. Further than the routing of the traffic, how to schedule packets with such a heterogeneous nature over a shared link is still an open issue.

CONCLUSIONS

This work proposes a joint functional splitting and slicing solution as a step forward to enable the dynamic adaptation of complex RANs to traffic diversity. Also, the open issues and research challenges are identified. Simulation results show the existing trade-offs between the joint slicing and functional splitting — and the associated cost — and the performance in different RAN density scenarios. In sparse deployments, it is worth facing the added complexity and cost of DUs by deploying DUs able to instantiate multiple functional splits, one for each slice. With this, the MNOs reduce the number of required DUs. Conversely, in dense scenarios, the additional cost of the DUs could not compensate for the achieved gain.

ACKNOWLEDGMENTS

This work has been partially supported by the SPOTS project (RTI 2018-095438-A-I00) funded by the Spanish Ministry of Science, Innovation, and Universities, the FEM-IOT (001-P-001662) funded by the European Union Regional Development Fund, and the EC-funded (Horizon 2020) research projects Affordable5G (GA 957317), 5GMED (GA 951947), 5G-ERA (GA 101016681), 5GMediaHUB (GA 101016714), and 5G-Epicentre (GA 101016521) and OPTIMIST (GA 872866).

REFERENCES

[1] P. Cerwall, "Ericsson Mobility Report," Tech. Rep., June 2017; https://www.ericsson.com/en/reports-and-papers/mobility-report.

[2] ITU-R, "ITU-R M.2410-0 Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)," Tech. Rep., Nov 2017; https://www.itu.int/pub/R-REP-M.2410.

[3] I. Afolabi *et al.*, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Commun. Surveys & Tutorials*, vol. 20, no. 3, 2018, pp. 2429–53.

[4] 3GPP, "3GPP TR 38.801 v14.0.0: Study on New Radio Access Technology: Radio Access Architecture and Interfaces (Release 14)," Tech. Rep., Mar 2017; https://www.3gpp.org/DynaReport/38-series.htm.

[5] A. D. La Oliva et al., "Xhaul: Toward an Integrated Fronthaul/Backhaul Architecture in 5G Networks," *IEEE Wireless Commun.*, vol. 22, no. 5, Oct. 2015, pp. 32–40.

[6] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Commun. Surveys & Tutorials*, vol. 21, no. 1, 2019, pp. 146–72.

[7] A. M. Alba, S. Janardhanan, and W. Kellerer, "Enabling Dynamically Centralized RAN Architectures in 5G and Beyond," *IEEE Trans. Network and Service Management*, vol. 18, no. 3, 2021, pp. 3509–26.

[8] I. Koutsopoulos, "The Impact of Baseband Functional Splits on Resource Allocation in 5G Radio Access Networks," *IEEE INFOCOM*, 2021, pp. 1–10.

[9] B. Ojaghi *et al.*, "SlicedRAN: Service-Aware Network Slicing Framework for 5G Radio Access Networks," *IEEE Systems J.*, 2021, pp. 1–12.

[10] Small Cell Forum, "159.07.02 Small Cell Virtualization Functional Splits and Use Cases (Release 7)," Tech. Rep., Jan 2016; https://www.scf.io/en/index.php.

[11] O-RAN Alliance, "White Paper: O-RAN Use Cases and Deployment Scenarios," Tech. Rep., Feb. 2020; https://www.o-ran.org/resources/

[12] A. Prasad *et al.*, "Enabling RAN Moderation and Dynamic Traffic Steering in 5G," *2016 IEEE VTC*, 2016.

[13] 3GPP, "3GPP TS 28.530 v17.1.0 Management and Orchestration; Concepts, Use Cases and Requirements (Release 16)," Tech. Rep., Mar 2021; https://www.3gpp.org/DynaReport/28-series.htm.

[14] R. Ferrus *et al.*, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, May 2018, pp. 184–92.

[15] A. Garcia-Saavedra *et al.*, "Fluidran: Optimized VRAN/MEC Orchestration," *IEEE INFOCOM 2018*, 2018, pp. 2366–74.

BIOGRAPHIES

BEHNAM OJAGHI (bojaghi@uoc.edu) is an early stage researcher at Universitat Oberta de Catalunya (UOC) and a Ph.D. candidate at Universitat Politecnica de Catalunya (UPC).

FERRAN ADELANTADO (ferranadelantado@uoc.edu) is an associate professor at UOC and an adjunct associate professor at Universitat de Barcelona.

ANGELOS ANTONOPOULOS (aantonopoulos@nearbycomputing.com) is research and innovation director at NearbyComputing S.L.

CHRISTOS VERIKOUKIS (chverik@gmail.com) received his Ph.D. degree from UPC in 2000. He is an associate professor at Patras University.