Mobility aware Dynamic Resource management in 5G Systems and Beyond

Anna Tzanakaki
National and Kapodistrian University of Athens
Athens, Greece

Markos Anastasopoulos
National and Kapodistrian University of Athens
Athens, Greece

Alexandros Manolopoulos
National and Kapodistrian University of Athens
Athens, Greece

Dimitra Simeonidou
University of Bristol
Bristiol, UK

Abstract— **One of the fundamental challenges that needs to be addressed in 5G systems and beyond relates with delivery of demanding uninterrupted services to mobile users. In this context, optical transport networks play a key role as they facilitate disaggregation of the Radio Access and Core Network resources and the associated network functions. They can also enable disaggregation and dynamic allocation of compute resources and the relevant virtualised functions to address fast user mobility through solutions such as live VM migration. This paper presents relevant architectures and provide an evaluation analysis on the benefits of the proposed approach.**

Keywords—5G, Mobility, MEC, UPF

I. Introduction

Digital technologies have been identified as key in addressing fundamental challenges associated with societal and economic objectives, such as improved quality of living for citizens, sustainable development and economic growth. In this context, 5G and beyond infrastructures will play a fundamental role in bringing these technologies to society transforming their every day's life in the way services are provided, and businesses are run. However, this transformation requires new service capabilities that networks need to support including: i) connectivity for a growing number of very diverse devices, ii) ubiquitous access with varying degrees of mobility in heterogeneous environments and, iii) mission critical services, supporting highly variable performance attributes in a cost and energy-efficient manner. These demanding and diverse requirements bring the need of a paradigm shift migrating from closed purposely developed infrastructures into open elastic ecosystems able to support a variety of very diverse services and end-users. These ecosystems will rely on flexible architectural models adopting convergence and integration of a variety of network and compute technologies, network softwarisation, hardware programmability and disaggregation of compute/storage and network resources [1].

In this context it is important to identify suitable architectural approaches taking into consideration relevant ongoing standardisation activities (ETSI, 3GPP, IEEE [2]) as well as the most recent trends such as these defined by the Open-Radio Access Network (O-RAN) alliance [3]. Inspired by these activities we propose an architectural solution that efficiently integrates optical network and compute resources (Figure 1) in support of very demanding 5G services. The architectural principles adopted exploit the benefits of softwarisation migrating from the notion of network elements to network functions, the separation of user and control plane functions and Radio Access Network (RAN) disaggregation. RAN disaggregation refers to functional decomposition of the RAN BaseBand (BB) processing, to a set of functions that can be processed independently at the Remote, the Distributed and Central Units (RUs, DUs, Cus respectively). These can be placed either at one or more locations supporting a variety of BB processing functional splits.

Another architectural principle that we propose relates with the adoption of cloud computing in support of the processing requirements of the various Fronthaul (FH) and Backhaul (BH) services. The proposed approach introduces flexibility in the way compute resources are allocated across the 5G infrastructure as it allows both the integration of a central cloud solution as well as more distributed approaches where smaller scale compute and storage resources are placed at the network edge closer to the end user in accordance to the Mobile Edge Computing (MEC) paradigm. MEC will play a key role in order to further guarantee the capability of the 5G solution to support demanding requirements associated with reduced end-to-end latency and transport network capacity. This paper describes the detailes of the proposed 5G and beyond architecture and provides an evaluation analysis taking into consideration user mobility with the aim to quantify the benefits of the proposed approach.

II. Proposed Architecture

An optical transport network interconnecting a variety of general and specific purpose compute/storage and network elements is proposing adopting the concepts of hardware programmability and network softwarisation to support a variety of 5G-RAN deployment options. To achieve this, the optical transport network needs to provide the necessary interfaces to enable: i) disaggregation of Base Station nodes and, ii) separation of control plane (CP) and the user plane (UP) entities. Through Base Station disaggregation, the RAN functions (including RU, DU, CU) can be physically separated and hosted at different locations. Connectivity between the different protocols can be provided over the optical transport through interfaces such as the O-RAN FH interconnecting the

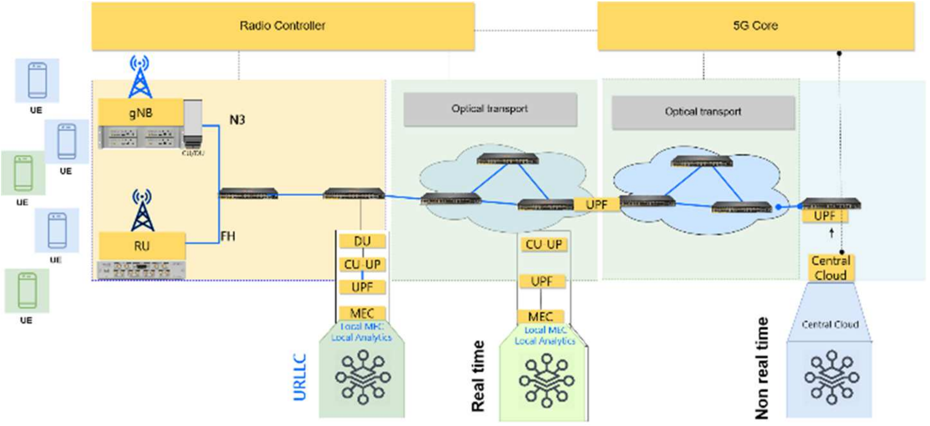


Fig 1: O-RAN 5G Architectural approach

978-3-903176-33-1 ©2021 IFIP

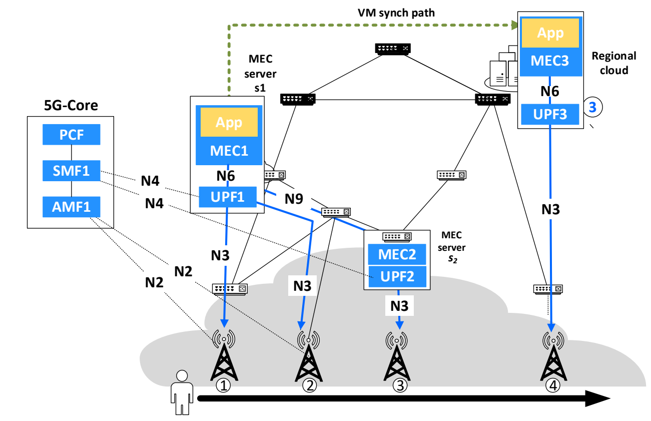


Fig. 2: The Joint user handover and VM migration problem to ensure

RU with the DU and the F1 interface interconnecting the DU with the CU. Separation of the CU Control Plane and User Plane enables flexibility in network deployment and operation, as well as cost efficient traffic management.

The user plane has the role to provide connectivity between the different network elements. This includes connectivity of the UE and the Access Network (AN) (NG-RAN in case of 5G) over the radio access technology, connectivity of the AN to the User Plane Function (UPF) over the N3 interface, connectivity between UPFs with different roles via the N9 interface, and finally connectivity from the UPF towards the external Data Network (DN) over the N6 interface [4]. User plane data that travel over N3 and N6 interfaces are carried over GPRS Tunneling Protocol User plane (GTP-U) tunnels. It should be noted that a big part of the user plane functionality in 5G Systems is handled by the UPF, which has to be designed to support challenging 5G services with very tight performance requirements. Part of the UPF's functionality is to set the data path between the UE and the Data Network and, as such, it is responsible for the PDU session establishment and the maintenance of the UE connectivity under user mobility. In this context, UPF is responsible to provide a broad range of functionalities including: i) Mapping of traffic to appropriate tunnels based on the QoS Flow Identifier (QFI) information, ii) Steering of packets to the appropriate output port and taking the necessary packet forwarding actions, iii) Packet counting for charging and policy control purposes, iv) Deep packet inspection for security and anomaly detection purposes, v) Buffering and queuing management for traffic service differentiation and assurance of end-to-end delays. On the other hand as shown in Fig 1 optical transport nodes need to support all user plane protocols and handle a large number of packet detection rules required to support policies at very high rates [5].

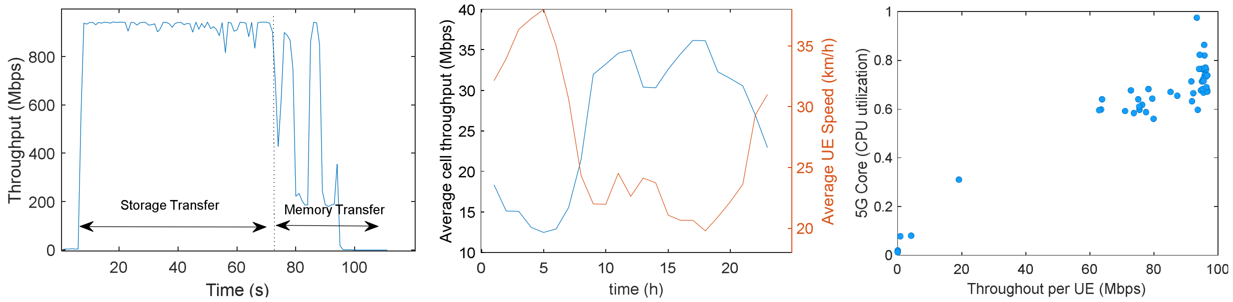
III. Mobility Consideration in 5G Networks supported by optical Transport

To minimize the deployment costs of 5G systems, 5G RAN and 5G Core elements are hosted at the same compute nodes with the end-user applications. All these elements are implemented in software and are hosted in Virtual Machines (VMs) (or Containers) running on compute nodes placed at the network edge or deeper in the core network. However, the integration of MEC with 5G systems brings new issues and challenges that need to be resolved. On the one hand, edge nodes usually have limited capabilities and are responsible to provide services targeting small geographical areas. On the other hand, mobile users such as smartphones and intelligent vehicles, tend to frequently move in between those small covered areas. Therefore, a main issue that needs to be resolved is how network and compute resources are allocated when a user leaves the area of coverage of a MEC node and enters the area served by another MEC node [6].

Another challenging aspect is associated with the reservation of sufficient resources across all elements of the 5G system (RAN and CORE and transport network providing connectivity between these) to support mobility. As users move from one gNB to another, PDU sessions with the required QoS Flow Identifier should be established. This requires reservation of specific resources to set up the appropriate Data Radio Bearer (DRB) tunnels between the UE and the gNBs and N3 GTP-U tunnels between the gNB and the UPFs. In addition to the PDU sessions, for services requiring access to a specific data network (i.e. MEC server) N6 tunnels should be established between the UPFs and the MEC and maintained for the whole duration of the connection of the mobile user. Therefore, a critical decision that needs to be taken by the Session Management Function (SMF) is when and over which elements these sessions should be established to ensure service continuity.

To address this challenge, the present study considers the adoption of joint user handover and VM migration to ensure service continuity in MEC-assisted 5G environments supporting advanced transport network connectivity. As an example of the supported functionality, consider the scenario shown in Fig.2 where a mobile UE moves from a source gNB to a target gNB. This relocation triggers a handover-related signaling procedure that is implemented in 5G systems using the N2 interface. In the simplest scenario where the UE moves from gNB1 to gNB2, the handover process will trigger SMF to establish a new N3 tunnel from UPF1 to gNB2. As the UE moves from gNB2 to gNB3 a new intermediate UPF (UPF2) is inserted by the SMF. This new UPF is hosted in MEC2 and is used to provide the necessary connectivity between the gNB3 and the APP server through UPF1. As before, the SMF will establish an N4 session with the UPF3 in order apply the necessary rules to UPF3 and create an N9 tunnel between UPF1 and UPF3 and an N3 tunnel between gNB3 and UPF3.

In the above cases the application server is hosted in MEC1 and therefore, the connection through the N6 tunnel interconnecting the MEC with UPF1 remains unaltered. However, as the user moves to gNB4 the distance between the UE and the VM where the APP server is hosted increases leading to an increase in the end-to-end delay. In this case, the application will be transferred to a server that is closer to the location of the mobile user. To realize this, a path interconnecting the source (MEC1) with the target (MEC3) server should be established to enable migration of the user context from MEC1 to MEC3. This process, also known as live Service Migration can be used to move active VMs (or containers) along with their applications to appropriate servers. When considering the concept of VM migration in 5G environments it is clear that this decision should be taken jointly with the placement of the UPF. In Fig. 2 it is shown that once the migration has been completed a tunnel interconnecting gNB4 with MEC3 should be established through UPF3. It is obvious that to ensure service continuity for MEC-assisted 5G services a complex chain of several processes needs to be performed ensuring efficient allocation of connectivity between the UEs and the MEC nodes [4]. To successfully complete all these processes in a timely manner reducing service disruption, several issues need to be considered during the service provisioning phase including



a) b) c)

Fig 3: a) Time series showing the traffic generated during VM migration from a source to a target VM. b) Correlation between background mobile network traffic per gNB and speed per UE, c) Impact of average UE throughput on 5GC computational resources.

allocation of: i) sufficient network resources for the establishment of the necessary connections between the 5G RAN and the 5GCORE elements, ii) sufficient computational resources to host not only the virtualized 5G functions (CU, DU, UPF etc) but also end user applications and iii) availability of network resources for the interconnection of servers to perform live migration. In response to this, a multi-stage optimization framework is developed in which a decision related to the placement of each VM to the appropriate servers is taken at each process stage. The objective of the proposed framework is to minimize the network cost for the provisioning of the services to the end users with the required KPIs. This cost function takes into account the weighted average of the utilization of the network and compute elements and a penalty when service latency increases. The analysis is based on realistic statistics for network traffic and users' mobility patters as well actual measurements for the VM migration process overheads.

To solve the problem of joint VM migration and mobility management in 5G systems, a 5G testbed has been deployed over a virtualized cloud environment allowing the accurate estimation of network and compute resources consumed during the establishment of new UE sessions. These measurements are coupled with actual network traffic and user mobility statistics collected over an operational mobile network system. Fig 3a shows an example of the network traffic generated during the migration from a source to a target MEC server as measured in a lab environment. In this example, the VM hosts a 4K streaming video server. During this live service migration process, the memory and disk state of the VM is transferred from the source host to the destination host. Storage transfer is performed through a steady throughput, while memory transfer though multiple synchronization iterations. As mentioned above, a prerequisite for the success of the VM migration process is the availability of network and compute resources during the storage and memory copy phase. The availability of these resources depends on the area where UEs move and the background network traffic. Higher background network traffic is observed in densely populated areas (i.e. city centers) where the speed of the mobile UEs is also lower. The interrelation between the average mobile traffic per gNB and the average speed per UE within the area covered by the gNB is shown in Fig 3b. The relevant traces have been captured from an operational mobile environment, whereas average speed statistics have been collected from GPS trackers. The impact of the mobile network traffic on CPU utilization of the virtualized 5G platform is shown in Fig 3c) . As expected, the average traffic per UE increases the CPU utilization of the platform used to host the virtualized 5G system. It is concluded that possible migrations associated with a user moving from a gNB covering a sparsely populated region to a densely populated region, should be treated carefully as service disruptions may occur.

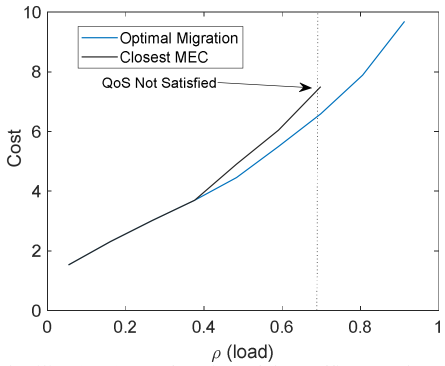


Fig 4: Total utility (cost) as a function of the traffic served per gNB

A comparison between the proposed VM migration scheme considering both the operation of the 5G system and the end-user services with a policy that assigns VMs to the MEC closest to the UE is shown in Fig 4. The total cost is the weighted average of the network and compute cost (increases with the increase of the network resources used) plus the end user service delay (increases with the increase of packet latency in the PDU sessions). We observe that under low loading conditions both schemes have similar performance. Therefore, VM migration may be applied in both cases providing similar results. Under high loading, for the closest MEC VM migration policy, MEC resources are not sufficient to handle both operational and user services (i.e. 5G CORE. 5G RAN and application server). In this case, a migration (if allowed) will overload the system resulting in degradation of the system performance. On the other hand, the model that considers all components of the 5G network, will optimally place VMs to appropriate servers ensuring service continuity for a wider range of inputs traffic loads.

IV. ACKNOWLEDGMENT

This work has been financially supported by the 5G PPP EU projects 5G-COMPLETE and 5G-VICTORI.

REFERENCES

[1] A. Tzanakaki et.al., "Optical networking interconnecting disaggregated compute resources: An enabler of the 5G vision," ONDM 2017,

[2] 5G System architecture for the 5G System (5GS) (3GPP TS 23.501 version 16.6.0 Release 16)

[3] O-RAN ALLIANCE, https://www.o-ran.org/

[4] I. Leyva-Pupo *et al.*,. Dynamic Scheduling and Optimal Reconfiguration of UPF Placement in 5G Networks. *MSWiM '20*, 2020

[5] H2020 Project 5G-COMPLETE, Deliverable D2.1

[6] A. Tzanakaki *et al.*, "Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services," *IEEE Comms. Mag*, vol. 55, no. 10, pp. 184-192,.2017