Zanzi, L., Salvat, J. X., Sciancalepore, V., García-Saavedra, A. y Costa-Pérez, X. (2019). Latency-driven Network Slices Orchestration. In *Proceedings of the IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*.

DOI: https://doi.org/10.1109/INFCOMW.2019.8845216

Latency-driven Network Slices Orchestration

Lanfranco Zanzi, Josep Xavier Salvat, Vincenzo Sciancalepore, Andres Garcia-Saavedra, Xavier Costa-Pérez
NEC Laboratories Europe, Heidelberg, Germany
emails: {name.surname}@neclab.eu

***Abstract*—The novel concept of network slicing is envisioned to allow service providers to open their infrastructure to vertical industries traditionally alien to mobile networks, such as automotive, health or factories. In this way multiple vertical services can be delivered over the same physical facilities by means of advanced network virtualization techniques. However, the vertical service requirements heterogeneity (e.g., high throughput, low latency, high reliability) calls for novel orchestration solutions able to manage end-to-end network slice resources across different domains while satisfying stringent service level agreements.**

In this demonstration we will show a novel orchestration solution able to handle one of the most stringent requirements: *end-to-end latency*. Our testbed—evolution of the work presented in [1]—implements all the resource brokerage schemes and allocation operations necessary to complete the life-cycle management of network slices. In addition, the novel *overbooking* concept is applied to pursue the overall revenue maximization when admitting network slices. Finally, an advanced network slicing monitoring system will be provided as a user-friendly dashboard allowing users to interact with the proposed solution.

I. Introduction

The envisioned highly-demanding services by industry verticals (e.g. automotive, health, etc.) is fostering mobile network operators to redefine their business models for 5G networks. In this context, the well-known concepts of software-defined networking (SDN) and network function virtualization (NFV) provide the technological means to radically change the way mobile services are delivered, i.e., evolving from relatively complex and inflexible mobile network architectures to a *cloudification* of networking, computing and radio resources, enabling traditional network functions to run in virtualized environments. Building on top of these trends, network slicing appears as a key technology in the mobile network landscape, able to provide network operators the capability to offer, via proper abstractions and isolation, underlying pools of physical resources to vertical industries or Over-The-Top (OTT) service providers, traditionally alien to the telco domain. This opens a set of new technical challenges that must be addressed such as cross-domain latency, throughput performance guarantees and logical isolation among slices to avoid resource wastage and service degradation. Operators must deal with a dynamic environment and be able to assure an adequate resource provisioning to meet the service level agreements (SLAs) of each instantiated network slice [2]. Therefore, an automated network orchestration solution is needed to manage the deployment of vertical services within a network slice considering different ranges of resources.

While most of the literature works address the network slicing problem focusing on domain specific issues (e.g. [3]), in this demonstration we deal with a multi-domain solution, i.e., accounting for spectrum at radio sites, network resources in the backhaul and transport domain, and computing resources for virtual eNBs and virtualized services as well as storage resources at geographically distributed data centers. See Fig. 1 for an illustration. We extend our previous work [1] to accommodate end-to-end latency constraints of network services and orchestration of eNBs as virtualized network functions.

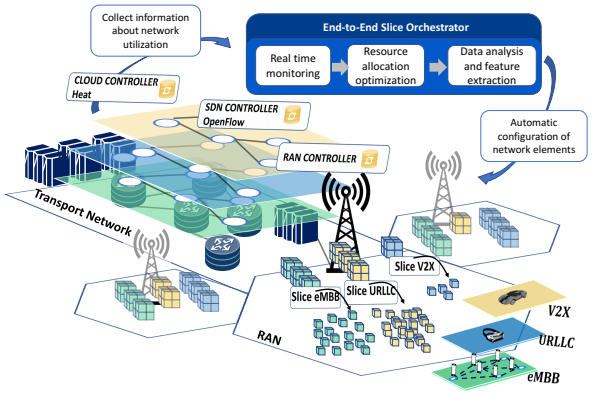

Fig. 1: End-to-end Network Slicing Orchestration

Our solution relies on the concept of *slice overbooking*, well-known in the field of yield management, prioritizing high-rewarding slices and leveraging on past resource usage patterns. We achieve this by balancing the risk of potential resource deficit (when overbooking slices) and the revenue attainable when accepting new slice requests. To summarize, our demonstration showcases a network slicing orchestrating solution that i) collects multi-domain resource utilization statistics, ii) applies machine-learning to study and predict service traffic patterns, iii) implements admission control policies that pursue overall revenue maximization, and iv) allocates physical and virtual resources on multiple domains to satisfy specific service requirements including latency.

II. Testbed Architecture

We designed and implemented a hierarchical control plane architecture as depicted in Fig. 2. The slice manager oversees the setup of an incoming slice request translating the incoming raw information into a detailed template describing the service requirements, which is then forwarded to the End-to-End (E2E) Orchestrator through a REST interface. The E2E orchestrator is the main entity of our system and oversees all the admission control and resource reservation tasks. The decision process is assisted by controllers, logically one per domain, with a two-fold contribution: on the one side, they abstract the underlying physical topology exposing only the most significant details, e.g., real-time monitoring information on resource availability and utilization, numbers of nodes, etc., and, on the other side, they enforce resource allocation policies in the corresponding domain.

Due to the lack of *slice-ready* 5G equipment, we build our demo on top of a fully functional LTE mobile network com-

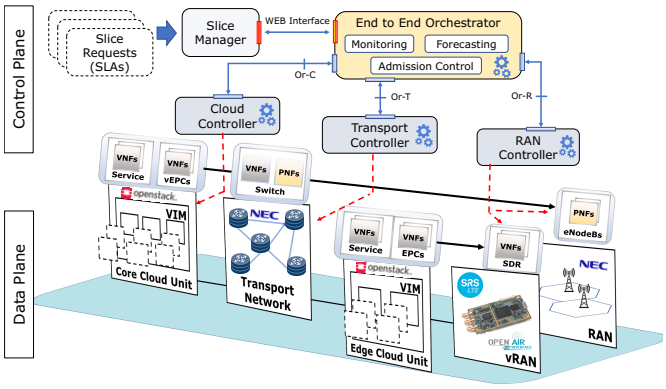

Fig. 2: End-to-end Network Slicing Orchestration architecture.

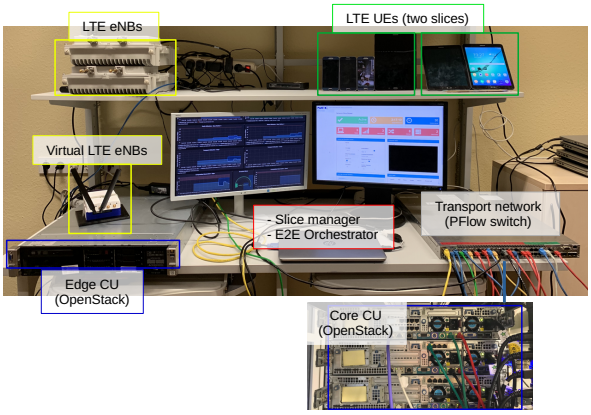

Fig. 3: Testbed deployment

prised of two commercial 3GPP rel.10 LTE eNodeBs (eNBs) and an additional radio access point shared by virtualized eNBs comprised of an USRP B210 software-defined radio (SDR) platform connected to our edge computing server where virtualized LTE stacks operate[1]. We exploit standard RAN sharing features of our eNBs and that of [4] to share the USRP radio frontend of virtualized eNBs. In this way slices of radio access points to the admitted network slices (identified each by a unique PLMN id) are provided. The SDN-based transport network is emulated through a programmable OpenFlow switch using 1 Gb/s Ethernet links building a simple mesh topology that connects the eNBs to the edge server (an off-the-shelf server with 16 CPUs) and to the core computing platform (with an aggregate of 64 CPUs across four compute nodes). We enforce 30-ms latency in our transport network to emulate certain geographical distance between vertical services and end-users. This encourage the E2E Orchestrator to deploy services with strict delay requirements (e.g., those expected by URLLC slices) in the Edge cloud premises. Each server provides a virtualized environment to host vertical services. We use OpenStack as compute infrastructure manager and its HEAT module to perform dynamic allocation of computing resources. To provide connectivity to end users, each service (and so each network slice) is associated to a customized virtual instance of OpenEPC [5] that includes all networking functionality of an Evolved Packet Core (EPC). A picture of our testbed is depicted in Fig. 3.

III. Operational Phases

A simplified example of our demonstration is available online[2]. In this video, we run across all the life-cycle management steps required by the network slicing orchestrator from the perspective of our monitoring dashboard. Network slice requests are iteratively introduced through a web-based interface (not shown in the video) with customized resource requirements (e.g., throughput, latency, time duration, etc.) conforming with the desired service provisioning (e.g., URLLC, mMTC, etc.). At first, a URLLC slice request with strict delay requirements is accepted into the system and the corresponding vertical service, including a virtualized EPC instance and the chain of all necessary network functions, is deployed into the edge premises. A second slice request comes, and it is also granted into the system due to resource availability. Being a slice supporting delay tolerant application (mMTC), the main service is enabled in the core cloud domain. The two slices share RAN and transport network, where proprietary interface in the eNBs allows to dynamically adapt the amount of Physical Resource Blocks (PRBs) allocated to each slice and SDN-based signalling guarantees capacity allocation in all the links involved. Simultaneously, the Cloud Controller triggers the computational resource reservation in the core (or mobile edge) premises where the tenants services are running. While traffic flows, we notice a gap between resource utilization and the resource allocation enforced in the different domains. As running slices may not fully exploit the set of resources they asked in the first place, an elastic management of the resource allocations has the potential to increase the overall system utilization (and so revenue) by admitting a higher number of tenants. Such *overbooking mechanism* must be carefully executed to avoid the revenue increase deriving from additional slices being penalized by growing the chances of SLA violations. Before each decision step, the service requirements coming from 3^{rd}-party requests are compared with real-time monitoring information provided by each domain controller. When overbooking, the resources allocated to previously accepted slices are adapted by balancing a prediction of their actual demand and the risk of making a mistake with said prediction. Low-latency slice requests are orchestrated in the edge server until its resource capacity is exhausted.

[1]We use srsLTE, a fully-fledged 3GPP rel. 8 LTE stack, running in docker containers for resource control.

[2]Available at https://youtu.be/Vr3X-2uBhjU.

Acknowledgements

This work was supported by the H2020 5G-Transformer Project under Grant 761536 and by the H2020-MSCA-ITN-2015 5G-AURA Project under Grant 675806.

References

[1] L. Zanzi et al., "Overbooking network slices end-to-end: Implementation and demonstration," in *Proceedings of the ACM SIGCOMM 2018 Conference on Posters and Demos*, pp. 144–146.

[2] GSMA, "Smart 5G Networks: enabled by network slicing and tailored to customers' needs," https://www.gsma.com/futurenetworks/wp-content/uploads/2017/09/5G-Network-Slicing-Report.pdf, September 2017.

[3] V. Sciancalepore et al., "Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization," in *IEEE INFOCOM*, Apr. 2017.

[4] Jose Mendes et al., "Cellular access multi-tenancy through small-cell virtualization and common RF front-end sharing," *Elsevier Computer Communications 2019*, vol. 133, pp. 59–66.

[5] CND. (2017) OpenEPC 7. http://www.openepc.com/.