Slice-Aware Elastic Resource Management

Sina Khatibi
Nomor Research GmbH
Munich, Germany
Email: khatibi@nomor.de

Irina Balan
Nokia Bell Labs
Munich, Germany
Email: irina.balan@nokia.com

Dimitris Tsolkas
University of Athens
Athens, Greece
Email:dtsolkas@mobics.gr

***Abstract*—Network softwarisation and network slicing are two important 5G technology enablers. Implementing mobile networks over the commercial data-centres has proven to have considerable benefits. Realising cloud-based mobile networks and serving multiple network slices with different requirements over the same virtualised physical infrastructure are challenging tasks. The IP tsunami and dramatic temporal and spatial variation of traffic demands made the situation worse. In this situation, slice-aware elastic resource management approaches are required to guaranteed the quality of offered services to the slices while minimising the capital and operational expenditure of networks. This paper provides a brief overview on the design guidelines for slice-aware elastic resource management.**

I. Introduction

The monthly mobile traffic demand is going to pass 41 PB in 2021 [1] due to the proliferation of traffic-hungry applications. The traffic demands growth is becoming even more problematic in the Internet of Things (IoT) era, where a considerable portion of "things" are smart [2] with diverse Quality of Service (QoS) requirements. This situation increases the competition among different service providers as well as Over-The-Top (OTT) providers. However, serving these diverse demands with scarce available resources is not only challenging but also significantly increases the CAPital EXpenditure (CAPEX) and OPerational EXpenditure (OPEX) and reduces the Return On Investment (ROI). Hence, in contrast to the "one-fit-all" architecture in LTE, the next generation of mobile networks (i.e., 5G networks) are going to offer service-tailored connectivity and top-notch QoS in a multi-tenancy environment [3] while providing on-demand, flexible, and re-configurable networks.

Based on the discussion above, *i)* Network Function Virtualisation (NFV), and *ii)* network slicing are two important technology enablers in 5G. The revolution in implementing Network Functions (NFs) on Commercial Off-The-Shelf (COTS) computers or data-centres instead of proprietary hardware has proven a great advantage to achieve the 5G requirements. Also, virtualisation of network resources and network functions enable sharing the physical infrastructure while offering isolation, network element abstraction, and ease-of-use [4]. The combination of NFV and network slicing offers multi-tenancy and on-demand service/resource provisioning, which can reduces CAPEX and OPEX.

Realising a virtualised environment hosting multiple network slices with different (and even contradictory) QoS requirements is a non-trivial task. Managing the physical resources is a critical issues in the 5G networks. First, the dramatic temporal and spatial traffic demands make the design of systems for only rush-hour (i.e., the time interval with the highest traffic demands) no longer acceptable since it leads to high CAPEX and OPEX. Therefore, the resource management should be even more elastic comparing to the former generation of networks. Meeting the different slices or tenants QoS requirements over the shared infrastructure in the cost-efficient manner is the second issue.

The resource elasticity of a communications system can be defined as: "the ability to gracefully adapt to load changes in an automatic manner such that at each point in time the available resources match the demand as tightly and efficiently as possible"[5]. The two critical resources in 5G systems are radio resources as well as computational resources (e.g., CPUs, RAM, and storage). While management of the former is very well known and studied, the elastic management of the latter is a new topic in communication systems.

This paper provides a brief overview on slice-ware elastic resource management in 5G systems in addition to high-level design concepts and guidelines.

II. Slice-Aware Elastic Resource Management

The goal in slice-aware elastic resource management is to develop algorithms, which consider the QoS requirements, Service Level Agreements (SLAs), and demands of network slices operating on the same physical infrastructure to optimally allocate/deallocate a portion of available resources to each of them. Since these parameters vary during time, an elastic management of resources, either computational or radio resources, is required to avoid the resources shortage, on one hand and on the other hand, unnecessary increase of network's cost. For example, as the demands in one network slices increases more computational nodes will be allocated to that network slice and as demands decreases the extra computational nodes can be removed. The approaches for slice-aware resource management can be applied in the following phases:

1) **On Admission:** Flexible slice admission control and blueprint analysis may be applied during the slice instantiation and configuration phase, to reduce the probability of resource outage as a result of the activation of a new slice.
2) **On preparation:** Slice behaviour analysis can be a critical asset for slice-aware provisioning, since statistics can be exploited in the slice preparation phase to efficiently decide the basic configurations and set the network environment.
3) **On runtime:** Advanced sharing of resources among Virtual Network Functions (VNFs) of multiple slices may

provide resource elasticity to running slices (in operation phase) by exploiting multiplexing capabilities.

One example of such a slice-aware elastic management use-case looks at how the Radio Access Network (RAN) radio resource pool can be optimally utilised for achieving the slice specific requirements of all the ongoing services in a cell or network region. This use-case is looking at an outdoor touristic city scenario comprising of several cells and two active slices, a browsing slice (background traffic) and a tourist slice (high throughput, low latency Augmented Reality (AR) or Virtual Reality (VR) application.

On the one hand, in contrast to the computational resources, RAN radio resources are limited in nature (i.e. spectrum, Physical Radio Blocks (PRBs), number of beams per cell). On the other hand,simultaneously active slices in the network have significantly different requirements and thus would need different RAN configurations (e.g. dedicated beam sets/ widths/ number per slice, adaptive scheduling scheme, adaptive cell coverage, etc.). Furthermore, the optimal RAN configuration may vary in spatial domain due to different radio link conditions at different geo-location and in temporal domain depending on dynamic network conditions: traffic load over the time of day, User Equipments (UEs) using a service /slice appearing in / disappearing from a certain area. Thus, the goal of the slice-aware elastic function is to adapt the network configuration in a proactive and automated fashion to match the dynamic changes in the service specific needed capacity. One possible solution is adapting the beam patterns to achieve locally and momentarily an optimal capacity/coverage trade-off [6].

The slice-aware elastic resource management approaches include the following main steps:

1) **Forming the available resource pool:** In the first step, the algorithm has to identify the available physical resource (e.g., available CPU cores, the CPUs' operating frequency, and available PRBs) and form the shared resource pool for the serving slices. Based on the slices requirements and their SLAs, the algorithm allocates the available resources to each slice.
2) **Estimating the total network throughput and computational resource impact on the network performance:** in this step, the algorithm estimates the total throughput based on the available radio resources in addtion to estimating the expected network performance (e.g., network function processing time as the function of the input variables such as allocated PRBs in the RAN) using the available computational resources. The algorithm uses the output of this step to decide whether to admit any new slice in addition to determining the service level each slice can receive.
3) **Allocation of available resources to different slices**: The algorithm allocates a portion of available resources pool to each network slice. Regarding computational resources, it has to allocate required computational resource to the NFs of each slice (horizontal or vertical scaling [7] ensuring an acceptable total processing time. The allocation procedure should consider SLAs type and slice priorities. Based on [8], there are three main SLA types, as follows:
   - **Guaranteed slice**, where the offered service quality is guaranteed to be kept in between the minimum and maximum level.
   - **Best-effort with a minimum guaranteed slice:** the SLAs of these type of slices guarantees a minimum quality of service. However, the higher quality of service will be provided in a best-effort manner.
   - **Best-effort Slice**: the slice will be served in best effort manner without any guarantee on the offeredQoS.

   While the SLAs define the constraints for resource allocation optimisation problems, defining serving weights for each NF of each of the slices is necessary to enable prioritisation among them. The NFs with higher serving weight will have higher priority in the resource allocation process. Consequently, the violation weights define the order of importance to violate the NFs requirements in the undesirable case where the shortage of resources happens.
4) **Observe the network performance and re-allocate the resources**: The resource management algorithm observes the changes in the network performance as the results of the changes in the resource demands or resource availability and updates the resource allocation accordingly.

ACKNOWLEDGEMENT

Part of this work has been performed within the 5GMoNArch project, part of the Phase II of the 5th Generation Public Private Partnership (5G-PPP) program partially funded by the European Commission within the Horizon 2020 Framework Program.

REFERENCES

[1] Cisco Systems, "Global Mobile Data Traffic Forecast Update, 2016 - 2021," Cisco Systems, California, USA, Tech. Rep., 2017. [Online]. Available: http://www.cisco.com

[2] F.Z. Yousaf, "Network function virtualisation (nfv) enabling technology for 5g," in *Tutorial Session of IEEE/IFP Netowrk Operation and Management Symposium*, Taipei, Taiwan, Apr. 2018.

[3] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing & softwarization: A survey on principles, enabling technologies & solutions," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2018.

[4] S. Khatibi and L. M. Correia, "Modelling of virtual radio resource management for cellular heterogeneous access networks," in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, Washington, DC, USA, Sept 2014, pp. 1152–1156.

[5] D. Gutierrez-Estevez, M. Gramaglia, N. P. A. Domenico, S. Khatibi, K. Shah, D. Tsolkas, P. Arnold, and P. Serrano, "The path towards resource elasticity for 5g network architecture," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW): Workshop on Flexible and Agile Networks (FlexNets)*, Barcelona, Spain, Apr. 2018.

[6] Nokia Networks, "Liquid Radio Making radio networks active, adaptive and aware - Nokia white paper," Nokia, Espoo, Finland, Tech. Rep., May 2014.

[7] B. Wilder, *Cloud Architecture Patterns*, ser. Develop cloud-native applications. O'Reilly, 2012.

[8] S. Khatibi and L. M. Correia, "A Model for Virtual Radio Resource Management in Virtual RANs," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 68, 2015. [Online]. Available: http://jwcn.eurasipjournals.com/content/2015/1/68