Department of Computer Engineering

Dr. D. Y. Patil Institute of Technology, Pimpri, Pune

(2021-2022)

A Seminar Report

On

"The Path Towards Resource Elasticity for 5G Network Architecture"

Submitted to the

Savitribai Phule Pune University

In partial fulfilment for the award of the Degree of

Bachelor of Engineering

In

Computer Engineering

Bharat Korade

Under the guidance of

Dr.S.V.Chobe

Dr. D. Y. Patil Institute of Engineering , Technology, Pimpri, Pune

CERTIFICATE

This is to certify that Bharat Sudam Korade from **Third Year Computer Engineering** has successfully completed his / her seminar work titled **"**The Path Towards Resource Elasticity for 5G Network Architecture**"** at Dr. D. Y. Patil Institute of Technology, Pimpri, Pune in the partial fulfillment of the Bachelor's Degree in Engineering.

Dr.S.V.Chobe	**Dr.S.V.Chobe**	**Dr. Pramod Patil**
(Guide)	(Head of the Department)	(Principal)

ABSTRACT

Vertical markets and industries are addressing a large diversity of heterogeneous services, use cases, and applications in 5G. It is currently common understanding that for networks to be able able to satisfy those needs, a flexible, adaptable, and programmable architecture based on network slicing is required. Moreover, a softwarization and cloudification of the communications networks is already happening, where network functions (NFs) are transformed from monolithic pieces of equipment to programs running over a shared pool of computational and communication resources. However, this novel architecture paradigm requires new solutions to exploit its inherent flexibility. In this paper, we introduce the concept of resource elasticity as a key means to make an efficient use of the computational resources in 5G systems. Besides establishing a definition as well as a set of requirements and key performance indicators (KPIs), we propose mechanisms for the exploitation of elasticity in three different dimensions, namely computational elasticity in the design and scaling of NFs, orchestration-driven elasticity by flexible placement of NFs, and slice-aware elasticity via cross-slice resource provisioning mechanisms. Finally, we provide a succinct analysis of the architectural components that need to be enhanced to incorporate elasticity principles

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who helped in making of this seminar.

I would like to express my heartfelt gratitude to our seminar coordinator Dr.S.V.Chobe for their valuable guidance, constant encouragement and creative suggestions on making this seminar.

I am grateful to Dr. PRAMOD PATIL, Principal and Dr. S.V.Chobe, Head of Department, Department of Computer Engineering, for their necessary help in the fulfilment of this seminar.

I am also grateful to all my teachers for helping me to make this seminar.

Table of Contents

1] Introduction

The 5th generation (5G) of cellular systems will change the access to technology for users, vertical markets and industries. Thanks to the 5G-enabled technical capabilities, they will experience a drastic transformation that will trigger the development of cost-effective new products and services. A large number of uses cases and corresponding requirements for representative vertical markets such as automotive, health, factories of the future, energy, and media and entertainment will need agile access to network support functionalities [1]. This will require a fundamental rethinking of the mobile network architecture and interfaces. The expected diversity of services, use cases, and applications in 5G requires a flexible, adaptable, and programmable architecture. To this end, network architecture must shift from the current network of entities to a network of capabilities.

In the context of 5G network architecture, a few key concepts have been introduced in the last years by Standards Development Organizations (SDOs) and research efforts. The first one is the concept of network slicing , which allows the network to run multiple network instances in parallel. It was introduced as an effective way to meet all of the heterogeneous requirements from supported use cases and services by means of a cost-effective multi-tenant shared network infrastructure. Another fundamental enabler that emerged as an initiative from the industry to increase the deployment flexibility and the agility with which a new service is integrated within the network is network function virtualization (NFV) and its management and orchestration (MANO) architecture . NFV is a framework where network functions (NFs) that traditionally used dedicated hardware are now implemented in software that runs on top of general purpose hardware, effectively enabling a hardware-software separation that reduces both capital and operational expenditures.

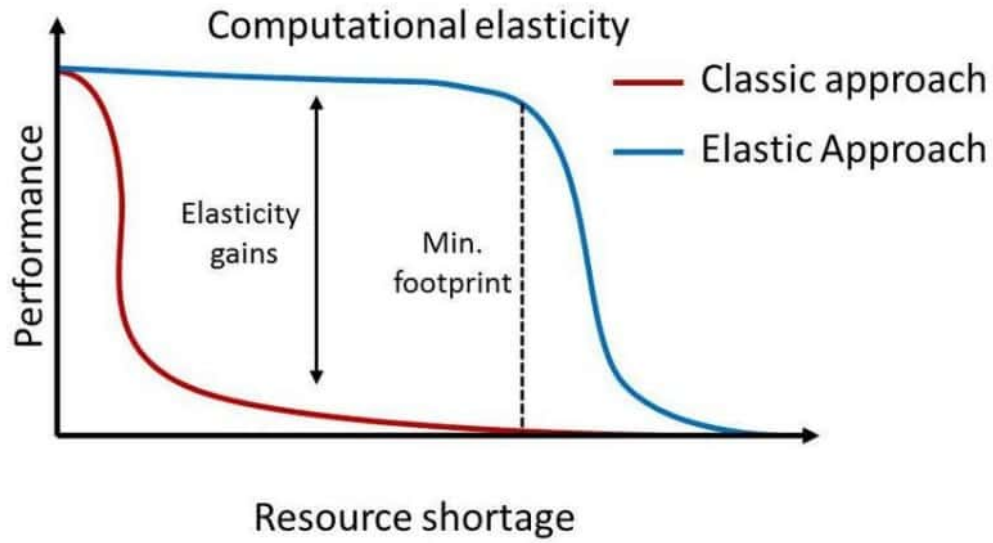
In this paper, we focus on an architectural concept for 5G network architecture that we believe will be key given the above well-established innovations. We refer to this concept as resource elasticity. Elasticity is a well-studied concept in cloud computing systems defined as the degree to which a system is able to adapt to workload changes by provisioning and deprovisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible . In networks, temporal and spatial traffic fluctuations require that the network efficiently scales resources such that, in case of peak demands, the network adapts its operation and re-distributes available resources as needed, gracefully scaling the network operation. We refer to this flexibility, which could be applied both to computational and communications resources, as resource elasticity. Although elasticity in networks

has already been exploited traditionally in the context of communications resources (e.g., where the network gracefully downgrades the quality for all users if communications resources such as spectrum are insufficient), in this paper we focus on the computational aspects of resource elasticity, as we identify the management of computational resources in networks a key challenge of future virtualized and cloudified 5G systems.

The dedicated to describe in depth the concept of resource elasticity, is organized as follows. Section II presents a definition of elasticity, along with the main associated requirements and key performance indicators (KPIs). In Section III we cover the main challenges and envisioned mechanisms for provisioning resource elasticity. Section IV shows the architectural components involved in resource elasticity

2)Resource Elasticity

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 closely and efficiently as possible. Hence, elasticity is intimately related to the system response when changes occur in the amount of available resources. The term gracefully in the definition of elasticity to imply that, for a relatively small variation in the amount of resources available, the operation of the service should not be disrupted. If the service produces a quantifiable output, and the resource(s) consumed are also quantifiable, then the gracefulness of a service can be defined as the continuity of the function mapping the resources to the output; sufficiently small changes in the input should result in arbitrarily small changes in the output (in a given domain) until a resource shortage threshold is met where the performance cannot keep up. We refer to this resource shortage threshold as minimum footprint. Follwing figure shows a conceptual example of the operation of an elastic system compared to a non-elastic one, where the elastic performance is capable of achieving graceful degradation with resource shortages until the minimum footprint is met. An elastic VNF should thus be able to cope with variations in the availability of resources without causing an abrupt degradation in the outputs provided by the function.



3) Elastic Operation Requirements

Resource elasticity can be exploited from different perspectives, each of them being a fundamental piece required to bring overall elasticity to the network operation.

The first requirement for an elastic network operation is the need for elasticity at the VNF level. In general, the concept of Fig. 1. Illustration of gains achieved by elastic computation. elasticity for a NF has not been directly applicable to legacy physical network functions (PNFs). Especially for the case of distributed NFs, the functionality is provided by a physical box that is the result of a thorough joint hardware/software design. Therefore, they have traditionally been designed without any major constraint on the available execution resources as they were expected to be always available by design. In addition, in networks with centralized VNFs, the joint hardware/software design is not possible anymore: VNFs are pieces of software that run on virtual containers on heterogeneous cloud platforms with standard interfaces.

A second requirement for elastic network operation can be characterized as elasticity at intra-slice level. The elastic design of a VNF has an impact on the elasticity of a network slice, defined as the chain of VNFs that provide a telecommunication service. Indeed, chaining and orchestrating a sequence of VNFs with different elastic KPIs (as described in Section II-C) will result in an overall elasticity associated to a tenant running a service using a single network slice.

The last requirement for elastic operation is elasticity at the infrastructure level, i.e., a requirement that involves the infrastructure on which elastic VNFs run. The choice of how many network slices are hosted in the same infrastructure depends on the infrastructure provider who run e.g., admission control algorithms to guarantee that the service level agreement (SLA) with the various tenants are always fulfilled.

4)Computational Elasticity

The goal of exploiting computational elasticity is to improve the utilization efficiency of computational resources by adapting the NF behavior to the available resources without impacting performance significantly. Furthermore, this dimension of elasticity addresses the notion of computational outage, which implies that NFs may not have sufficient resources to perform their tasks within a given time. In order to overcome computational outages, one potential solution is to design NFs that can gracefully adjust the amount of computational resources consumed while keeping the highest possible level of performance. RAN functions in particular have been typically designed to be robust only against shortages on communication resources; hence, the target should be directed at making RAN functions also robust to computational shortages by adapting their operation to the available computational resources. An example could be a function that chooses to execute a less resource-demanding decoding algorithm in case of resource outages, admitting a certain performance loss. In addition, the scaling mechanisms, i.e., the modification of the amount of computational resources allocated to such computationally elastic NFs may help in exploiting the elasticity of the system if they are properly designed.
There are two significant ways to scale a NF:
(i)horizontal scaling, where the system is scaled up or down by adding or removing new identical nodes (or virtual instances) to execute a NF, and
(ii) vertical scaling, where the system is scaled out or in by increasing or decreasing the allocated resources to the existing node (or virtual environment) . As an example in the RAN domain, supporting higher system throughput by adding additional access points is referred as horizontal scaling, whereas an increase in operating bandwidth is referred as vertical scaling.

5)Orchestration-driven Elasticity

This innovation focuses on the ability to re-allocate NFs within the heterogeneous cloud resources located both at the central and edge clouds, taking into account service requirements, the current network state, and implementing preventive measures to avoid bottlenecks. The algorithms that implement orchestration-driven elasticity need to cope with the local shortage of computational resources by moving some of the NFs to other cloud servers which are momentarily lightly loaded. This is particularly relevant for the edge cloud, where computational resources are typically more limited than in the central cloud. Similarly, NFs with tight latency requirements should be moved towards the edge by offloading other elastic NFs without such tight timescale constraints to the central cloud servers.

To efficiently implement such functionalities, special attention needs to be paid to

(i)the trade-off between central and edge clouds and the impact of choosing one location for a given function, and

(ii) the coexistence of Mobile Edge Computing (MEC) and RAN functions in the edge cloud. This may imply scaling the edge cloud based on the available resources, clustering and joining resources from different locations, shifting the operating point of the network depending on the requirements, and/or adding or removing edge nodes

6) Slice-aware Elasticity

this section addresses the ability to serve multiple slices over the same physical resources while optimizing the allocation of computational resources to each slice based on its requirements and demands, a challenge earlier referred to as E2E cross-slice optimization. Offering slice-aware elastic resource management facilitates the reduction of CAPEX and OPEX by exploiting statistical multiplexing gains. Indeed, due to load fluctuations that characterize each slice, the same set of physical resources can be used to simultaneously serve multiple slices .

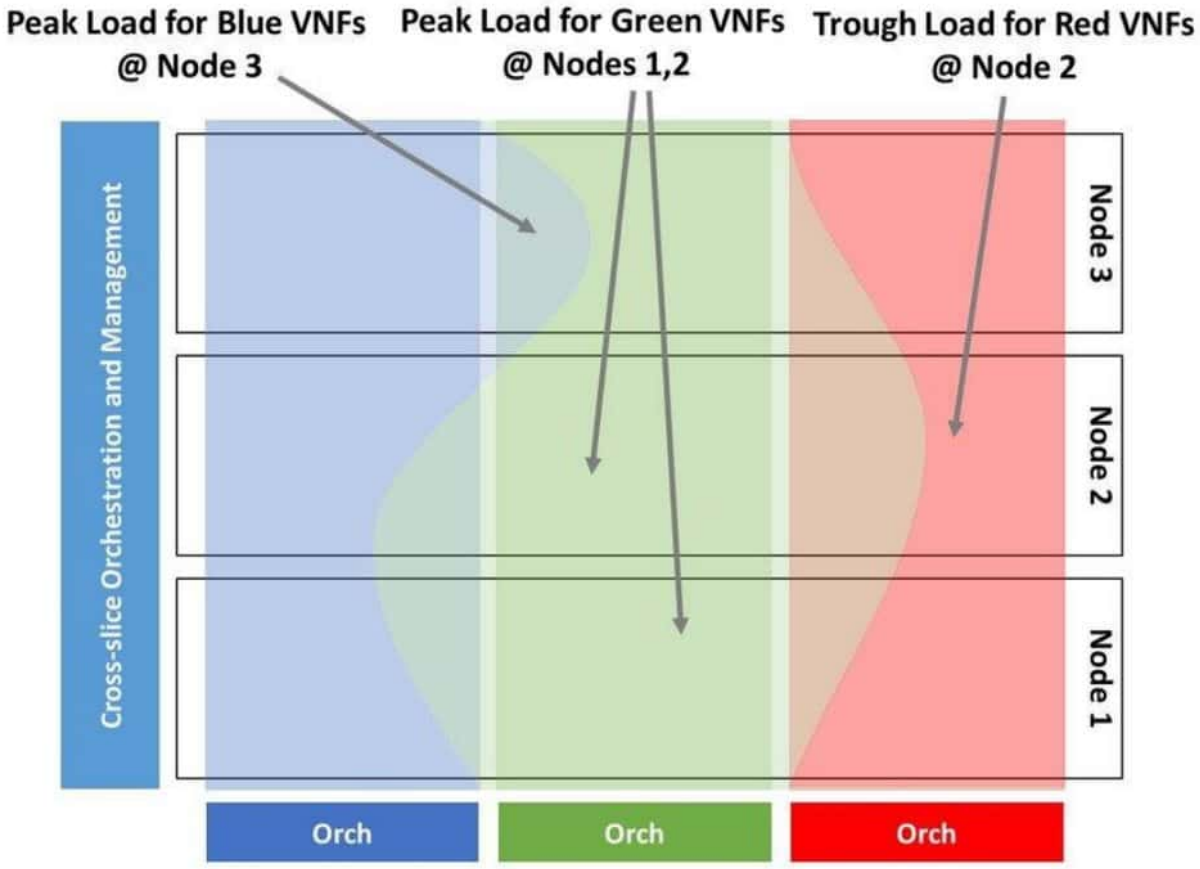


Illustration of slice-aware

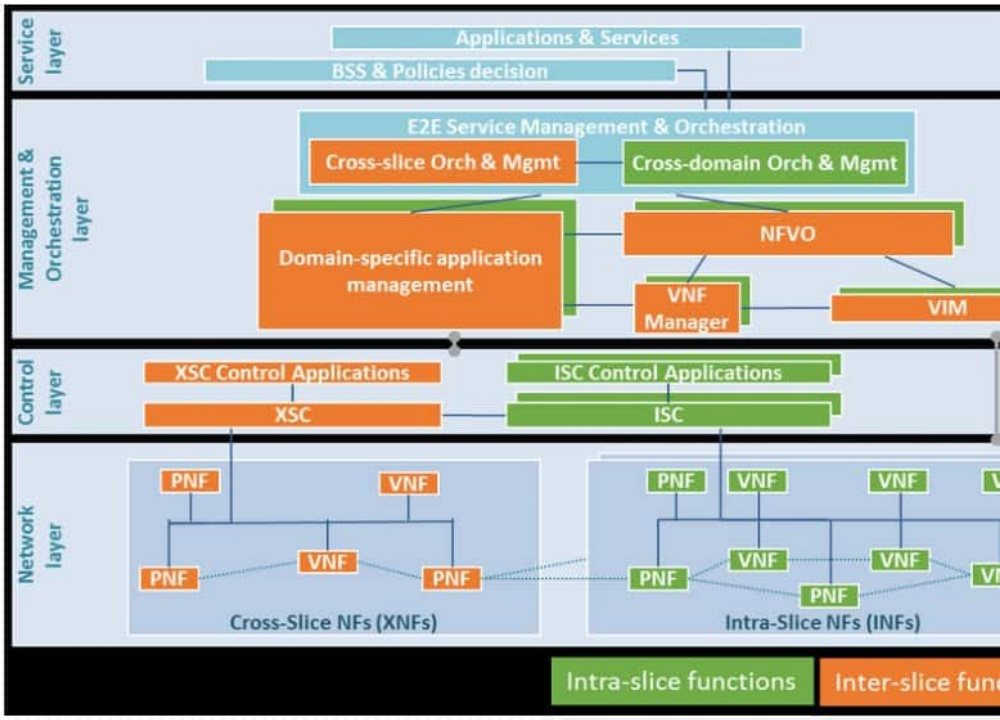
Adaptive mechanisms that exploit multiplexing across different slices must be designed, aiming at satisfying the slice resource demands while reducing the

amount of resources required. Hence, the solutions must necessarily dynamically share computational and communications resources among slices whenever needed. An elastic admission control system would be also required, as elastic slices need not have the same amount of available resources as e.g., a highly resilient slice where all resource demands must be fully satisfied at each point in time.

7) Architecture

Many of the ongoing efforts to define a 5G architecture use a four-layer functional structure similar to the one depicted initially envisioned by the 5G-MoNArch project. The Service Layer, at the top of this structure, comprises business-level decision functions, applications, and services, operated by a tenant or other external entities. Such functions and services are applied to the network through operations in the Management & Orchestration Layer. This layer provides a multi-tenant, multi-service environment that enables E2E service and resource orchestration. Similarly to the ETSI NFV MANO architecture, the Management & Orchestration Layer incorporates components that deal with the life cycle management of the virtual resources (Virtual Infrastructure Manager (VIM)), the life cycle management of the VNFs (VNF manager) and the overall orchestration of the resources and the services on top of those managers (NFV-Orchestrator (NFV-O)). Additionally, it includes slice-aware and domain specific entities to manage the functional part of the VNFs.

The Management & Orchestration Layer further utilizes a Control Layer, which accommodates, using the intra- and cross-slice controllers based on SDN principles (ISC and XSC), the required translation of the northbound management and orchestration services into commands that are applied to the actual VNFs and PNFs. The VNFs and PNFs compose the lower layer in the reference architecture, referred to as Network Layer. In order to abide by the fundamental 5G direction for multi-tenancy support on top of a softwarized and slice-enabled network, the Network Layer incorporates separated control-plane and user-plane NFs, which are further divided into slice-specific NFs and shared NFs among different slices



5G baseline architecture for 5G-MoNArch

Elastic intra-slice orchestrator: Elasticity-aware algorithms are needed to orchestrate the different NFs that are part of the same slice. The tasks of such an elasticityaware orchestrator may include re-locating NFs (from central to edge cloud and vice versa, or from one server to another) depending on available resources, horizontally or vertically re-scaling the amount of resources allocated to one particular NF or a set thereof, clustering and joining resources from different locations, etc. Hence, this module would be responsible for implementing the dimension of elasticity

Elastic cross-slice orchestrator: The cross-slice orchestrator is in charge of performing the management and control of the multiple slices that share the architecture, i.e., enabling slice-aware elasticity as described in Section III-C. Some, or all of these slices may be elastic, i.e., slices that do not have totally stringent requirements but rather admit graceful degradation. For those cases, specific orchestration algorithms need to be designed.

8] CONCLUSION

In the quest to dramatically increase the flexibility of networks, in this paper we have introduced the concept of resource elasticity for 5G network architecture.

In addition to providing a definition, set of requirements and KPIs, we proposed the exploitation of elasticity along three different dimensions: computational elasticity, orchestration-driven elasticity, and slice-aware elasticity. Challenges and mechanisms for resource elasticity provisioning have been pointed out in each of the dimensions.

Finally, we provided a brief overview of the elasticity implications for the main architectural components of a 5G system.

9] REFERENCES

Journals / Periodicals:

- *5G-PPP, "5G Empowering Vertical Industries," white paper, Feb. 2016*
- *E. F. Coutinho et al., "Elasticity in cloud computing: a survey," Annals of Telecommunications, vol. 70, no. 7-8, pp. Aug. 2015.*
- *EU H2020 project 5G-MoNArch, Deliverable D2.1, "Baseline architecture based on 5G-PPP Phase 1 results and gap analysis*

Web Resources:

- https://www.google.com/searchq=resource+elasticity&oq=resource+elasticity&aqs=chrome..69i57j35i39l2j0i512j0i22i30l3.15777j1j15&sourceid=chrome&ie=UTF-8
- https://www.google.com/search?q=resource+elasticity+in+5g+network&sxsrf=AOaemvKVVOQD6OvL2PBdgCe39r2bAY2rpg%3A1638495116185&ei=jHOpYaXeCpCnrgSY05eQBQ&oq=resource+elasticity&gs_lcp=Cgdnd3Mtd2l6EAEYADIHCCMQsAMQJzIHCCMQsAMQJzIHCCMQsAMQJzIHCAAQRxCwAzIHCAAQRxCwAzIHCAAQRxCwAzIHCAAQRxCwAzIHCAAQRxCwAzIHCAAQRxCwA0oECEEYAFAAWABg3jFoAXACeACAAQCIAQCSAQCYAQDIAQrAAQE&sclient=gws-wiz
- 5g mano architecture - Google Search