Bulakci, Ö., et al. Overall 5G-MoNArch architecture and implications for resource elasticity, in 2018 *European Conference on Networks and Communications (EuCNC), June 18-21, Ljubljana, Slovenia*

Overall 5G-MoNArch Architecture and Implications for Resource Elasticity

Ö. Bulakci[1], Q. Wei[1], C. Mannweiler[2], M. Gramaglia[3]

[1]Huawei GRC, Germany; [2]Nokia Bell Labs, Germany;
[3]UC3M, Spain

D. M. Gutierrez-Estevez[4], M. Shariat[4], P. Arnold[5],
N. di Pietro[6], G. Dandachi[6], D. Tsolkas[7]

[4]Samsung R&D Institute, UK; [5]Deutsche Telekom, Germany; [6]CEA-LETI, France; [7]Mobics, Greece

***Abstract*— The fifth generation (5G) of mobile and wireless communications networks aims at addressing a diverse set of use cases, services, and applications with a particular focus on enabling new business cases via network slicing. The development of 5G has thus advanced quickly with research projects and standardization efforts resulting in the 5G baseline architecture. Nevertheless, for the realization of native end-to-end (E2E) network slicing, further features and optimizations shall still be introduced. In this paper, essential building blocks and design principles of the 5G architecture will be discussed capitalizing on the innovations that are being developed in the 5G-MoNArch project. Furthermore, building on the concept of resource elasticity introduced by 5G-MoNArch and briefly re-summarized in this paper, an elasticity functional architecture is presented where the architectural implications required for each of the three dimensions of elasticity are described, namely computational, orchestration-driven, and slice-aware elasticity.**

I. Introduction

Since the early research phase of the fifth generation (5G) starting in 2012 [1], the development of concepts for the 5G system (5GS) has progressed at a rapid pace. Within the 5GS, end-to-end (E2E) network slicing spanning over network domains (e.g., core network, CN, and radio access network, RAN) where multiple logical networks corresponding to different business operations, aka verticals, are sharing a common infrastructure, is seen as the fundamental pillar. Diverse and continuously emerging new communication services driven by the verticals require the mobile communication industry to support multiple telecommunications services with heterogeneous key performance indicators (KPIs) in a cost efficient way. 5G, powered by network virtualization and network slicing, shall give mobile network operators unique opportunities to offer new business models to consumers, enterprises, verticals, and third-party tenants and address such various requirements. To this end, both research projects [2][3][4] and standardization efforts [5] have described the main elements of the 5G architecture. Third generation partnership project (3GPP) has already completed the early-drop “non-standalone” release of 5G by the end of 2017.

Although all these aforementioned efforts have provided a solid baseline architecture, in our view there is still room for 5G system (5GS) enhancements to better fulfil the 5G vision of supporting diverse service requirements while enabling new business sectors often referred to as vertical industries. To this aim, we have performed a thorough 5GS gap analysis, in order to identify the features and optimizations that can be included in the future refinements of the 5G architecture. Such enhancements can be considered, for example, in the ongoing and future study items and work items of standardization developing organizations (SDOs), e.g., Release 16 and Release 17 of 3GPP for the 5GS. In particular, we found that current baseline architectural work, although understanding the importance of proper slice-coordination schemes, is still not addressing them at full steam. On this basis, herein, we present a functional architecture laying the foundation of such possible enhancements. Along these lines, we have also identified the concept of resource elasticity as a key innovation for 5G network architecture. Despite having been widely studied in the cloud computing domain [6], and having been traditionally exploited 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 architectural implications of the computational aspects of resource elasticity for 5G networks, as we identify the management of computational resources in networks a key challenge of future virtualized and cloudified systems.

This paper is a dissemination result of the 5G-MoNArch project, which started in July 2017 with the commitment of a consortium which comprises key global vendors, leading mobile network operators and key research groups as well as small enterprises. Further details regarding the overview of the project, including objectives, structure and expected impact can be found in [7]. The rest of the paper is organized as follows. Section II provides an overview of overall functional architecture being developed in the 5G-MoNArch project, and Section IV provides a conclusion.

II. Overall 5G-MoNArch Architecture

This 5G-MoNArch functional architecture considers the requirements from the project’s use cases and the ones initially defined in [8]. The baseline architecture of the overall architecture has its roots in the results of 5G PPP Phase 1 projects (especially 5G-NORMA and METIS-II), that are summarized in the White Paper [4] of the 5GPPP Architecture WG.

A. 5G-MoNArch overall functional architecture

This work has been performed in the framework of the H2020-ICT-2016-2 project 5G-MoNArch. The authors would like to acknowledge the contributions of their colleagues. This information reflects the consortiums' view, but the consortium is not liable for any use that may be made of any of the information contained therein.
.

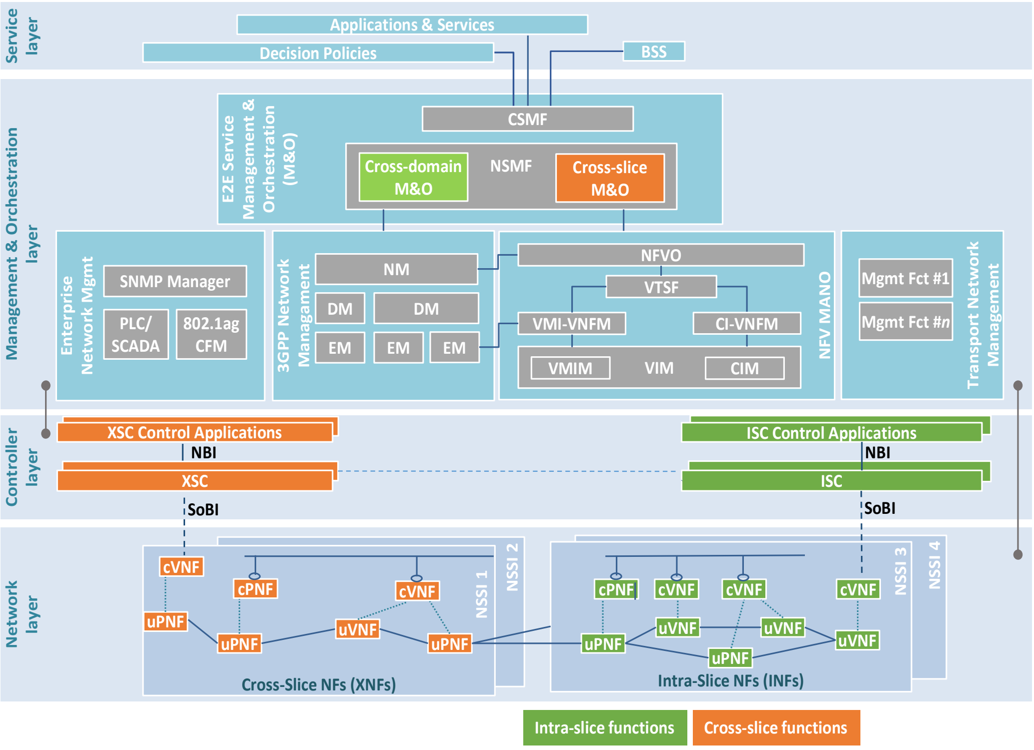

Figure 1. Initial 5G-MoNArch overall functional architecture

Fig. 1 depicts the four fundamental layers of the architecture. For each of these layers, a set of architectural elements that deliver the system's functionality, including the key network functions, their responsibilities, the interfaces exposed, and the interactions between them, are defined. The Service layer comprises Business Support Systems (BSS), business-level Policy and Decision functions, and further applications and services operated by a tenant or other external entities. The M&O layer is composed of the M&O functions from different network and technology domains, including, but not limited to, 3GPP network management, ETSI NFV MANO, management functions of transport networks (TNs) and private networks. The Network layer comprises the virtualized and physical NFs of both control and user plane, e.g., 5G RAN and CN network functions defined in 3GPP Rel. 15. The (optional) Controller layer accommodates two controller types: (1) the Cross-slice Controller (XSC) for cross-slice NFs and (2) Intra-slice Controller (ISC) for Intra-slice NFs. In the following subsections, we describe each layer in details.

B. *Management and Orchestration Layer*

This layer is substantially the gateway between the different services that could be instantiated by different tenants such as the verticals, and the real network operation. This layer shall support different ways of operation (i.e., Infrastructure as a Service, Platform as a Service or Network Slice as a Service), offering two main functionalities: *i)* the translation of high-level service requirements into real VNF deployments by means of network slice *blueprinting* and *instantiations*, and *ii)* the efficient lifecycle management and resource orchestration of the network slices running on the shared infrastructure. The latter operation shall directly deal with novel technologies, such as resilience and resource elasticity, which will be discussed in Section III [9].

C. *Control Layer*

This layer bridges the high-level directives mandated by the M&O layer with the specific VNF configuration that finally compose a network slice. The control functionality follows the principles of network programmability: rapid reconfiguration of VNF is achieved by getting information from both the orchestration layer and the network layer. Within 5G-MoNArch, we envision that this role is played by flexible network controllers that are especially useful for the coordination of different tenant on scarce resources such as the radio ones.

D. *Network Layer*

The flexibility brought by pillars enablers, such as NFV and SDN, allows the fast instantiation, management and configuration of the, formerly monolithic, now modular network functions that compose a slice. This layer includes both control VNFs (cVNFs) and user plane VNFs (uVNFs). Control VNFs may be further arranged as in a Service Based Architecture, if needed, while uVNFs may implement the concepts of resilience and elasticity. Besides the enforcement of the quality of experience (QoE)/QoS policies mandated by the M&O layer through the control layer, the Network layers also proactively reports monitoring data that can be used by, e.g., Big Data -based algorithm for the overall network optimization.

E. Design Principles

This high level architecture follows three fundamental design principles: (1) Split of control and user planes, (2) support for E2E network slicing, and (3) network programmability. 5G-MoNArch applies a consistent split of control plane and user plane throughout different network domains, including RAN, CN, and TN. This allows for flexible network architecture deployment according to the different characteristics and requirements of control plane/user plane functions. For instance, 5GC control plane functions can follow service based architecture as defined in 3GPP, and user-plane functions can be distributed over the network either deployed as virtual network function of physical network function. The architecture allows for different levels of slicing support across different layers. The network functions in the network layer are either slice specific (INFs) or slice common (XNFs), and can be controlled by ISC and XSC respectively. In the management and orchestration layer, the cross domain M&O function takes care of per slice management, and cross slice M&O function performs the management cross different slices. A network slice in the network layer can be flexibly configured/ programmable based on the decision from management plane/control application. This enables the network to flexibly adjust its behavior according to the requirements of various use cases in high dynamic environment.

III. Elasticity Functional Architecture

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. As introduced in [9], resource elasticity can be exploited from different perspectives, referred to as elasticity dimensions, each of them being a fundamental piece required to bring overall elasticity to the network operation. The first dimension is computational elasticity, is to improve the utilization efficiency of computational resources by adapting the NF behavior to the available resources without impacting performance significantly. Secondly, orchestration-driven elasticity 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. Finally, slice-aware elasticity 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.

In this section, we provide a description of the elasticity functional architecture developed within 5G-MoNArch. The next subsections deal with the logical relation between elasticity dimensions and their mapping to the 5G-MoNArch reference architecture described above.

A. Logical Interactions among Elasticity Dimensions

Implementing elasticity in a network is a challenging task that involves several elements in the architecture. Still, their high level interaction can be summarized as depicted in the figure below.

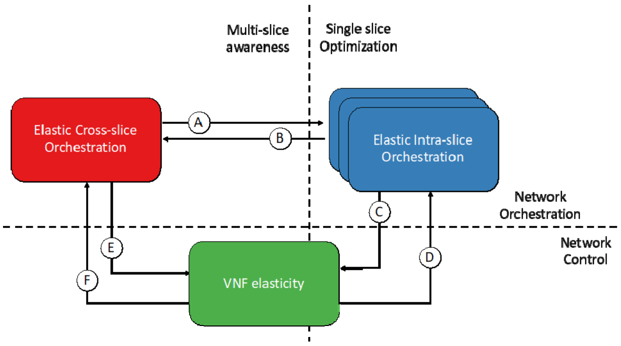
Multi-slice awareness
Single slice Optimization
Elastic Cross-slice Orchestration
A
B
Elastic Intra-slice Orchestration
C
Network Orchestration
Network Control
D
E
F
VNF elasticity

Figure 2: High level interactions across elastic modules

Elasticity happens at different levels: at the VNF level, at intra-slice level and at infrastructure level across slices. Their interactions form hence a triangle, in which each vertex is represented by one of the dimensions of the elasticity. The edges are the needed functional interfaces that shall be included in the overall architecture, to finally enable an elastic operation.

Elasticity mainly addresses two domains: network orchestration and network control. The former shall incorporate the elements needed to (i) flexibly assign resources to different slices and (ii) find the best location of a VNF belonging to a certain network slice within the infrastructure. The latter, instead, shall provide an inner loop control of network functions, to enforce elasticity at faster time scales, such as the ones needed in the RAN. Also, the dimensions of elasticity span across different domains with respect to the network sharing policies. Some elasticity aspects have necessarily to take into account different services at the same time while performing resource assignment or when to decide “elasticity level” of a network slice, as it is composed by a set of elastic VNFs. On the other hand, modules that are devoted to intra-slice elasticity targets one slice only, at least from the logical perspective. Then, depending on the implemented service provisioning model (e.g., Network Slice as a Service, or Infrastructure as a Service), these logical entities may be operated by different stakeholders.

Fig. 2 also shows the required interfaces among modules. We briefly describe them in the following, noting that their full definition will come by the end of the project lifetime.

- *Interfaces A/B*: these interfaces regulate the information exchange between the cross-slice elastic orchestration modules and the (logically) different intra-slice orchestration modules that take care of the intra-slice orchestration. Interface A transfers information related to the instantiation of an elastic NS on a shared infrastructure, i.e., the selected VNFs needed to provide a given service, and the associated resources. This information is needed by the intra-slice orchestration module to perform then an optimization within the resources granted to the slice. Also, this interface is used to exchange the information related to the lifecycle management of the network slice. The cross-slice orchestrator has an E2E view of the available resources, so it can command to a given slice to acquire or release new resources, as needed. In turn, interface B has mostly the role of reporting the used resources by a given network slice, that is eventually used by the cross-slice orchestrator to keep the resource map up-to-date. Also, this interface is used to deliver re-orchestration triggers coming from inside a slice (e.g., when the computational resources originally granted to a slice are not enough anymore to provide the required service).

- *Interfaces C/D*: these interfaces cross the Orchestration / Control domain shall be used to exchange information regarding the behaviour of specific elastic VNFs that build a network slice. For instance, interface C is used to transmit to the VNF the information needed for the elastic behaviour such as the total amount of available computational resources or memory. Also, interface C logically covers the one used to instantiate a VNF in a specific location of the system (e.g., in the edge). On the other hand, interface D is used mainly for reporting by the VNFs: information about the current resource utilization, together with other network related metrics (e.g., in case of a RAN function, the number of used physical resource blocks (PRBs) or the average signal to noise ratio (SNR) of each user). In general, this information is the used by artificial intelligence (AI) and big data analytics modules.
- *Interfaces E/F*: these logical interfaces are not mapped to a specific interface in the overall 5G-MoNArch architecture, but rather specify a functional behaviour. Interface E represents the abstraction of a network slice blueprint that is composed by a certain number of elastic VNFs and, finally, achieve an elasticity level defined as a function of them. Similarly, the information about the elastic behaviour of each VNF shall be available at the cross-slice orchestration. That is, the information about the used resources and the graceful degradation trends of a VNFs need to be taken into account by the cross-slice orchestrator when performing the slice admission control and onboarding operations.

B. *Architectural Enablers for Elasticity*

This section describes the architectural enablers for each of the elasticity dimension.

Computational elasticity: To introduce computational elasticity, several requirements to the 5G-MoNArch architecture have been identified. In general, a VNF itself needs to have additional functionalities to react on dynamic shortages of computational resources, to reach graceful degradation in radio performance. Furthermore, it is necessary to have a centralized entity, which controls multiple VNFs to react on short-term dynamics. The central entity needs to react on lack of resources being aware of the impact on the radio performance. VNF individual decision might cause a lack of resources for other VNFs or unwanted decreased radio performance and thus possible SLA violations. A VNF will have a certain degree of resource isolation (e.g., CPU pinning). However, in a non-pinned setup, this may happen. Therefore, it is mandatory for the controller to have knowledge about the behavior of the VNFs. In other words, the containers and virtual machines (VMs) need to be associated with the functionalities/VNF running inside to react from a short-term perspective. To interact between the MANO and the Control Layer for re-/orchestration and life cycle management procedures, there is a need to define interfaces between those layers. Such interfaces need the ability to trigger re-orchestration, influencing and monitor VNF behavior as well as resource consumption. The necessary interaction among the considered layers to enable computational elasticity are described in the following.

- The MANO Layer especially the VIM needs to generally allocate the necessary computational resources among the VNFs during re- /orchestration phase (long term control loop).
- The Control Layer needs to support flow control to react on the radio performance, dynamically. This is done within the given maximum amount of resources allocated by the VIM in the MANO Layer for each VNF (short term control loop).
- The Control Layer can set parameters of the VNF how to handle the available computational capabilities (short term control loop).
- It is for further study if the Control Layer might dynamically reassign computational resources in tight arrangement with the VIM or if the VIM can react dynamically on a fast time scale.
- The VNF needs to be able to report on the virtual resource consumption to the Control Layer.
- The Control Layer needs to be able to request more computational resources from MANO or at least needs to trigger a re-orchestration of specific VNFs.
- The Control Layer needs to be able to influence the behavior of the VNF to react on short term.
- The Control Layer needs to know the influence on the radio performance of the computational resource shortages of specific VNFs.
- The Control Layer needs the ability to influence VNF behavior to react on computational resource shortages with respect to the radio performance.

Orchestration-driven elasticity: Fig. 3 shows a potential integration of the orchestration-driven elasticity within the 5G MonArch E2E Service Management & Orchestration architecture. The orchestration-driven elasticity functions are foreseen as part of the 3GPP Network Slice Management Function (NSMF), which is in charge of managing the overall slice life-cycle [10]. The decisions taken at this level are then transferred at the underlaying management function, the 3GPP Network Slice Subnet Management Function (NSSMF), which has the role to control a specific subset of NFs, and adjust the decision received from the NSMF locally. Finally, the management instructions are forwarded at the network function virtualization management and orchestration (NFV MANO) block and executed [11].

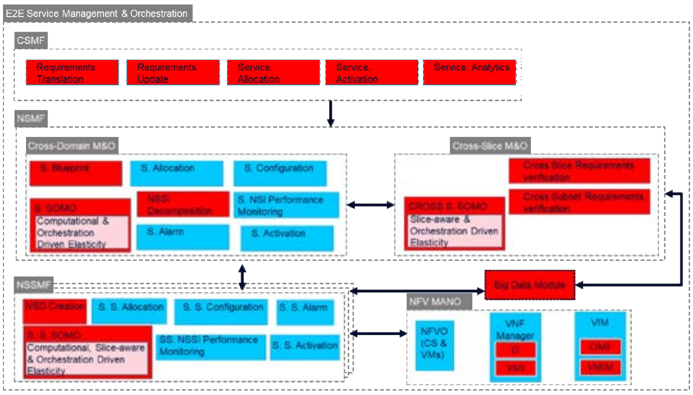

Figure 3. Orchestration-driven elasticity within the 5G-MoNArch E2E Service Management & Orchestration Architecture

Slice-aware elasticity: Slice-aware elasticity refers to the ability to serve multiple slices over the same physical resources while optimising the allocation of computational resources to each slice based on its requirements and demands. The algorithms/policies for slice aware elasticity reside at the E2E service M&O layer and more precisely at the Cross-Slice M&O entity (see Fig. 3), which triggers the above-mentioned resource optimisation via the 3GPP Network Management or the NFV MANO entities.

Slice-aware elasticity is supported through three different approaches that can be mapped to three different phases of the lifecycle of a slice instance, namely during runtime, instantiation, and/or preparation phases. Referring to the 3GPP architectural components that undertake the major tasks for each one of the approaches, the functionality needed for the run-time approach is part of the NSMF and resides at the Cross-layer M&O. The input information needed is available at the NSMF via the NSSMF, while the output commands target the NFVO of the NFV MANO. Regarding the instantiation-time approach, input from the NSSMF or NFV MANO can be used, while info from CSMF, regarding the service requirements, might be exploited. Again, the output commands target the NFVO of the NFV MANO. Finally, additional functionality needed for the preparation-time approach can be provided by an AI and Big Data Analytics Engine, an architectural entity briefly described in the following.

In addition, all the above elasticity-related functionalities could be greatly enhanced with an AI-based engine similar to the one recently being proposed by the Experiential Networked Intelligence (ENI) group of ETSI [12]. Focused on optimizing the operators' experience, this engine would be equipped with big data analytics and machine learning capabilities that could enable a much more informed elastic management and orchestration of the network, often allowing proactive resource allocation decisions based on the history rather than utilizing reactive approaches due to changes in load. For example, reinforcement learning algorithms could be very suitable to determine optimal policies for horizontal or vertical scaling decisions of NFs, or better slice orchestration decisions could be made if real utilization data is gathered and processed from the underlying infrastructure. The detailed specifications of such a module including the particular algorithms it would apply as well as the description of its interfaces and data collection requirements are beyond the scope of this paper, but part of 5G-MoNArch future work.

IV. CONCLUSIONS

The first major milestone for the 5GS has been recently achieved, where the baseline architecture is already specified. Nevertheless, there is still room for new features and further enhancements to cover the envisioned full space of use cases and applications beyond enhanced mobile broadband (eMBB) and to realize a native support for E2E network slicing considering vertical requirements. In this paper, we present two contributions: we have first highlighted an initial functional architecture, where the main building blocks and design principles are presented. Secondly, we have described the implications in the previously mentioned architecture design of the concept of resource elasticity previously introduced in the context of the 5G-MoNArch project. These two innovations represent a step forward in the cost savings and resource efficiency achieved 5G networks.

ACKNOWLEDGMENT

This work has been performed in the framework of the H2020 project 5G-MoNArch co-funded by the EU. The views expressed are those of the authors and do not necessarily represent the project.

REFERENCES

[1] A. Osseiran, et al. "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS project," IEEE Communications Magazine, 52(5), 2014, pp. 26-35.

[2] P. Marsch, et al. "5G Radio Access Network Design: A Brief Overview on the 5G-PPP Project METIS-II," IEEE EuCNC, July 2015.

[3] 5G NORMA Deliverable D4.2, "RAN architecture components – final report," June 2017

[4] 5GPPP Working Group Architecture, "View on 5G Architecture," white paper v2.0, Dec 2017.

[5] 3GPP TS 23.501, "System Architecture for the 5G System," v15.0.0, Dec 2017.

[6] E. F. Coutinho et al., "Elasticity in cloud computing: a survey," Annals of Telecommunications, vol. 70, no. 7-8, pp. Aug. 2015.

[7] 5G-MoNArch project, see https://5g-monarch.eu/

[8] Next Generation Mobile Networks (NGMN) Alliance, "5G White Paper", Feb. 2015.

[9] D. Gutierrez-Estevez et al., "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), April 2018.

[10] 3GPP TSG Services and System Aspects, "TR28.801, Telecommunication management: Study on management and orchestration of network slicing for next generation networks," V15.0.0, Dec 2017.

[11] J. G. Herrera, and J. F. Botero. "Resource allocation in NFV: A comprehensive survey," IEEE Transactions on Network and Service Management, vol. 13, no. 3, pp. 518-532, Sept. 2016.

[12] ETSI ENI - Experiential Network Intelligence, Available Online: https://portal.etsi.org/Portals/0/TBpages/ENI/Docs/ETSI ISG ENI_Presentation.pdf