# Mobile Network Architecture: End-to-End Network Slicing for 5G and Beyond

## The path from concepts to practice: The 5G PPP Phase 2 project 5G-MoNArch

Lars Christoph Schmelz[1], Albert Banchs[2], Mauro Boldi[3], Ömer Bulakci[4], Emilio Calvanese Strinati[5], David Gutiérrez-Estévez[6], Diomidis S. Michalopoulos[1], Jose Enrique G. Blazquez[7], Heinz Droste[8]

[1] Nokia Bell Labs, Munich, Germany; [2]Universidad Carlos III de Madrid, Spain; [3]Telecom Italia, Torino, Italia; [4]Huawei Technologies GRC, Munich, Germany; [5]CEA-LETI, Grenoble France; [6]Samsung R&D, Staines-upon-Thames, UK; [7]Atos, Madrid, Spain; [8]Deutsche Telekom, Darmstadt, Germany

***Abstract*—5th generation mobile networks (5G) aim at enabling a large variety of applications, services and use cases for vertical industries and markets. In order to fulfil the verticals' requirements, the 5G network architecture has to be adaptive, flexible and programmable. Network slicing, where multiple virtual networks share a common infrastructure spanning over technical domains end-to-end (E2E), is being developed as the corresponding solution. Considerable effort has already been spent by industry, academia, and standardization bodies at conceptual level. Now, with the 5G networks roll-out getting closer, these concepts need to be brought to practice. The 5G-MoNArch project will consolidate the developments and standardization progress in the area of 5G network architecture and network slicing, and will implement and deploy two testbeds along the requirements of two vertical use cases: Extreme mobile broadband in a touristic city, and secure and reliable industrial communication in a sea port environment. With the hands-on experience from these testbeds and the corresponding experimental results, the capabilities of the concept of E2E network slicing will be proven.**

***Keywords—5G Mobile Networks, 5G-MoNArch, 5G Network Architecture, Network Slicing, Cloud-enabled RAN, Verticals***

## I. INTRODUCTION

The 5th generation (5G) of mobile communication networks will clearly change the access, perception and experience of wireless communication technology to all kinds of tenants, such as, consumers, vertical markets, and industries. Current technology follows a one-fits-all approach, i.e., the provided network services are similar for all tenants and their requested services independent of the requested quality. With 5G, a large variety of customized use cases shall be supported, each having a dedicated composition of requirements regarding service availability, flexibility, reliability and security, but also with respect to the service characteristics, location, the number of terminals, and lifetime. The reduction of the service creation time furthermore plays an important role. Vertical markets and industries like automotive, smart factory, smart city, media & entertainment, or healthcare [1] are the main drivers behind these use cases. Examples are autonomous vehicles (such as vehicle-to-vehicle and vehicle-to-infrastructure communication), predictive maintenance for vehicles and machines, highly flexible production environments with a transition from production chains to production cells, adaptive traffic control and steering in cities, augmented and virtual reality applications or remote surgery. As it is clear that a separate physical infrastructure for each of these use cases is simply impossible due to high cost, inflexibility to changing requirements or lack of available radio spectrum, a flexible and adaptive approach for network sharing is required.

End-to-end (E2E) network slicing represents the corresponding architectural solution that has been developed for 5G. It provides the ability to deploy and operate a multitude of vertical-, service- or application-specific logical network instances sharing a common physical infrastructure [12]. This represents the step from providing simply connectivity towards providing a network of services. To enable this approach, it is necessary to shift from the current network of nodes and entities towards a network of capabilities [7] with a completely new network architecture, where multiple services are supported by adapting the network operation to the services' requirements and at the same time providing multi-tenancy support. While two facilitating techniques build the baseline for enabling network slicing, namely, Software Defined Networking (SDN) and Network Function Virtualization (NFV) [2], new business and stakeholder models as well as new trust models including privacy and security mechanisms are required. Finally, communication and information processing resources need to be tightly integrated and orchestrated across multiple technical domains such that the sharing by multiple tenants is possible.

The concepts for the 5G network architecture and of network slicing have been discussed in industry fora, such as, Next Generation Mobile Networks (NGMN, [6]) and developed in various research projects, such as, 5G-NORMA [9] and Selfnet [10] that led to an overall vision and concept [7]. Standardization has already started within several working groups of the 3rd Generation Partnership Project (3GPP), for example, on services [3], management and orchestration [4] or on the radio network architecture [5]. These efforts led to the definition of an E2E architecture with the key features:

- Use case dependent flexible instantiation and location of network functions (NF)
- Flexible configuration and control of NFs using software-defined control and orchestration
- Joint optimization of access and core NFs through intelligent interaction design and co-location

While the challenges, requirements, key features and generic concept of the 5G network architecture and network slicing have been strongly addressed during the past years, with the 5G roll-out getting closer it is inevitable to bring these concepts to practice. For this reason, the 5G-MoNArch [11] project has been defined and recently started.

The remainder of the paper is structured as follows: Section II provides an overview of the general concept of 5G-MoNArch together with the key enabling and functional innovations. Section III shortly introduces the use cases and testbeds, and Section IV concludes the paper.

## II. 5G-MoNArch Concept

The project's main goal is to enhance and complete the existing concepts for the 5G mobile network architecture and to develop, implement and deploy two vertical use cases in real-world testbeds, namely, an (i) industrial use case for a sea port environment focusing on highly resilient and secure communication, and a (ii) resource elastic extreme Mobile Broadband (eMBB) use case in a touristic city. For both use cases, dedicated functional innovations will be developed and the required network will be designed and implemented. Furthermore, in order to fill gaps not yet addressed, the existing 5G architecture concepts will be enhanced and completed by three enabling innovations: (i) Inter-slice control and cross-domain management that aim at the coordination across multiple network slices and domains; (ii) the experimentally-driven modelling and optimization of a virtualized environment in order to enable the design and proof of high-performance network functions and algorithms; and (iii) a cloud-enabled protocol stack to reduce interdependencies between network functions and allowing for a more flexible function placement in the network.

Fig. 1 shows the approach of 5G-MoNArch. As it can be seen, this approach is to merge the common architecture building blocks of the existing 5G architecture concepts with the common architecture building blocks of the 5G-MoNArch enabling innovations into the overall 5G-MoNArch architecture. The instantiation of this architecture, together with the specific functions of the project's functional innovations, will be implemented into the two use case-specific network slices, namely, the *secure and resilient network slice* and the *resource-elastic network slice*. Each of these two network slices will then be deployed within one of the two testbeds.

In the following paragraphs, the cornerstones of the 5G-MoNArch architecture are described in more detail.

### A. Flexible and adaptive architecture

5G mobile and wireless communications networks need to address the unprecedented diversity of use cases and the associated requirements. At the same time, the future mobile and wireless networks shall also overcome the challenge on enabling new business use cases in a cost-efficient way by using a common infrastructure. Network slicing provides the necessary framework to support novel businesses, i.e., vertical industries.

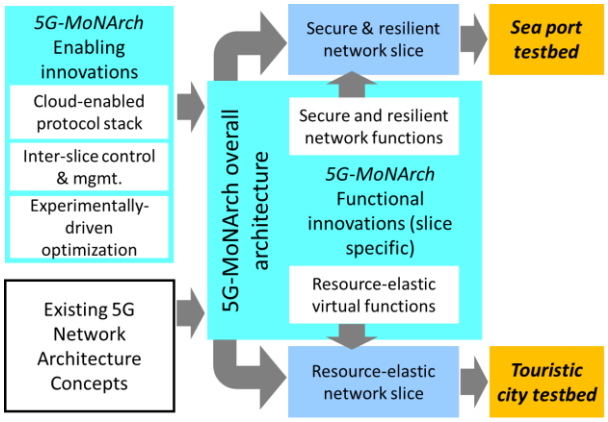


Fig. 1: 5G-MoNArch approach: Building blocks and innovations

On this basis, a flexible and adaptive architecture design is of utmost importance to reach the aforementioned goals. 5G-MoNArch will identify gaps in the state-of-the-art, in order to extend and complement the baseline architecture with key enabling innovations. In doing so, E2E network slicing is placed into the core of the design and an experimentally-driven optimization will follow. The specification of the overall architecture will then include the technical domains of core and radio access networks along with functional control and user plane architecture as well as the management and orchestration planes. Particular focus is put on the following three architecture enhancements relative to the state-of-the-art.

*Inter-slice control and cross-domain management:* Network slicing has a strong impact on the Radio Access Network (RAN) design, both in the user plane and (particularly) in the control plane. Indeed, a new design paradigm is needed to allow the simultaneous operation of multiple network slices, each with tailored access functions and functional placements to meet their target key performance indicators (KPIs). While each network slice is considered as a fully operational (logical) network on its own, multiple network slices are operated on a common physical / virtualized infrastructure, which requires specific inter-slice control and management functions. To fulfil the objective of supporting multiple network slices in a common infrastructure, the inter-slice control and cross-domain management enhancement enables the interworking of different network domains operated and used by different tenants while providing cross-domain service guarantees. Each network slice can have tailored network functions and functional placements to meet their target KPIs coupled with business-driven service-level agreements (SLAs). This architecture enhancement therefore includes the development of solutions for network slice lifecycle management, multi-slice resource management, the design of functions and mechanisms for network slice prioritization and context sharing, and appropriate interfaces to allow network slice control across infrastructure domains.

*Experimentally-driven modelling and optimization of a virtualized environment:* The optimization of the operation of 5G-

MoNArch's cloud environment needs to consider the computational behavior of nodes and functions. Since some of the nodes in the telco cloud, i.e., the entirety of edge and core nodes in different locations, may be equipped with limited resources, the placement of NFs in nodes needs to consider the availability of computational resources in addition to other criteria such as service requirements. Where traditional approaches assume that NFs and nodes consume / offer a fixed amount of resources, a widely flexible and adaptive network architecture needs an accurate model of the computational behavior to determine the corresponding performance impact. This enhancement leverages on the implementation of 5G-MoNArch architecture to gain experimental insights into the architecture and use these insights to create realistic models for resource utilization and accordingly enable the design of optimized algorithms and functions for resource optimization.

*Native cloud-enabled protocol stack:* One of the key concepts the 5G-MoNArch architecture builds on is the flexible decomposition and allocation of NFs, i.e., decoupling mobile NFs from the underlying physical infrastructure and enabling their flexible placement within the different network nodes. The cloud-enabled protocol stack enhancement reduces dependencies within the network functions in the protocol stack to enable a more flexible placement of such functions within the network. This concept builds on state-of-the-art 5G mobile network architecture which, relying on orchestration and virtualization technologies, decouples network functions from the underlying hardware infrastructure and enables their flexible placement within the different nodes that conform the physical network.

### *B. Resilience and security*

The need for functional innovations in general stems from the fact that application specific functions are necessary for deploying network slices with particular requirements such as dedicated KPI targets. In this regard, the structure of the flexible architecture described above is leveraged such that specific network functionalities are instantiated, which are designed to meet these given KPI targets.

Resilience and security comprise one of the two major functional innovations of 5G-MoNArch. The main objectives of this functional innovation are two-fold: (i) To maintain a *failsafe mobile network operation*, even under extraordinary situations, such as, outages in parts of the infrastructure or substantial changes in radio propagation, and (ii) to provide an *enhanced level of security*, considering the challenges posed by the use of shared resources within a common infrastructure. To meet the above objectives, the work associated with this functional innovation is split into three major domains, namely, *RAN reliability; resilience in telco clouds*; and *security and data integrity*. These three domains contain the functional elements for an E2E secure and reliable network operation. The work planned for each domain is elaborated below in detail.

*RAN reliability:* This domain comprises the work towards a reliable operation of the RAN. This is measured by the time a connection can be established and kept which fulfils the minimum requirements of the respective service or application. Alternatively, reliability is determined according to the probability that messages or data are exchanged between user equipment (UEs) and base stations. The two major technical components that contribute to achieving reliability are *multi-connectivity* and *network coding*. Multi-connectivity provides simultaneous connections to multiple radio access points and therefore uses a wider set of resources. Related techniques are, e.g., carrier aggregation and dual connectivity as already provided with the 4$^{th}$ generation (4G) of mobile wireless networks (Long Term Evolution, LTE). Network coding refers to network nodes transmitting composite versions consisting of two or more messages, which are then inferred at the destination rather than being directly decoded. However, 5G-MoNArch foresees to use these techniques for increasing reliability instead throughput as this has been the intention with LTE. In particular, data duplication approaches will be considered, involving transmission of radio packets with additional redundancy, in conjunction with network coding techniques tailored for reducing the radio outage probability. Furthermore, for both multi-connectivity and network coding, the existing concepts will be examined in terms of their applicability to virtualized network environments, also considering implementations with limited computational capabilities. In all above approaches, assumptions pertaining to machine-type-communications (MTC) will be considered, involving specific characteristics in terms of packet size and arrival.

*Resilience in telco clouds*: The work in this domain is directed towards a failsafe operation of the telco cloud, i.e., failure management solutions that aim at keeping the operability of the network. The major activities in this regard involve *fault isolation*, i.e., to ensure that a fault occurring in one part of the network does not affect other network parts as well. Furthermore, the *prioritization and scaling of network functions* based on their role and importance in the network will be applied. To this end, machine learning techniques are planned to be devised, thereby automating the implementation and optimizing the operation of the corresponding methods and algorithms. This may include a graceful degradation of the operation of the resilient and secure slice in areas with insufficient network coverage. Finally, the concept of semi-autonomous *5G islands* will be developed. In case complete parts of the network fail or are not reachable by the tenant or application, such a 5G island, consisting of a semi-autonomous edge cloud, can provide all relevant network- and application-level functionalities such that service continuity at least at a certain quality level can be ensured.

*Security and data integrity*: Within the framework of this domain, the notions of *security trust zones* and *security fault isolation* are used for instantiating a resilient and secure network slice. Specially designed techniques concerning the prevention and detection of security threats are planned, as well as proper techniques for reaction against such threats, aiming at their mitigation. This tackles both, inter-slice deployments where different slices shall be sealed from each other, and geographic deployments where security gap propagation from one to neighbor areas shall be prevented. Furthermore, security and data integrity across operational domains, for example, corporate enterprise deployments operated together with a mobile operator domain, will be considered as well. The planned work comprises an investigation of the case of concurrent security and network operation faults, resulting in a potential trade-off between security and resilience. Machine learning techniques will be devised

for auditing the data integrity level of the slice associated with resilience and security, including anomaly detection in both hardware and software components.

### C. Resource elasticity

Resource elasticity comprises the second main functional innovation of 5G-MoNArch. It addresses the need for assigning and scaling computational, storage and communication resources where and when they are needed. For instance, in the case of a typical urban downtown scenario, the required services and applications may range from audio and video streaming via augmented reality to video chats and instant messaging, each imposing different service and quality requirements over a certain period of time at a specific location. To holistically address the problem of spatial and temporal traffic fluctuations in a cost-efficient manner, the mobile network must be able to assign, scale and cluster resources to those areas and parts of the network where and when they are needed, and the network functions need to be elastic enough to adapt to the available resources without impacting performance significantly. The ability to efficiently scale those resources according to the demand and to gracefully scale the network operation when insufficient resources are available is what is denominated as *resource elasticity* within 5G-MoNArch.

The concept of elasticity is well known in the area of cloud computing. An efficient cloud platform needs to automatically provision and release computing resources on demand as workloads change. However, solutions from cloud frameworks will need to be enhanced as (i) timescales involved in RAN functions are usually much shorter than those considered in cloud solutions, hence leading to possible outages, and (ii) cloud resources are typically limited at the edge, sometimes preventing centralized solutions to exploit multiplexing gains. 5G-MoNArch will introduce the concept of elasticity at both edge and central clouds, considering the associated constraints of the cloud infrastructure and the mobile network. Furthermore, the elasticity framework must consider the fact that cloud resources may be shared between different slices and their availability may change according to the dynamic request of tenants.

To meet the above described requirements and objectives, three main research areas in the context of resource elasticity are explored in 5G-MoNArch, namely *computational elasticity*, *orchestration-driven elasticity*, and *slice-aware resource elasticity*. These three areas account for an E2E elastic operation of the network, including the RAN. The specific research focus of each of them is elaborated in detail below.

*Computational elasticity*: This area addresses the notion of "computational outage", which implies that NFs may not have sufficient resources to perform their tasks within a given time. To overcome such computational shortages or outages, the project will design NFs that can gracefully adjust the amount of computational resources consumed while keeping the highest possible level of performance. An example therefore is an NF that chooses to execute a less resource-demanding decoding algorithm in case of resource outages, admitting a certain (as small as possible) performance loss.

*Orchestration-driven elasticity*: This area addresses the ability to re-allocate NFs within and between the edge cloud instances and the central cloud depending on the available resources. The corresponding methods and algorithms consider service requirements and the current network state, and implement preventive measures to avoid bottlenecks. This may imply scaling the edge cloud based on the available resources (e.g., releasing unneeded resources), clustering and joining of resources from different locations, shifting of the operating point of the network depending on the requirements, and/or adding or removing of edge nodes. Furthermore, edge cloud resources may also be required to provide Mobile Edge Computing (MEC) features, which may have higher priority than specific NFs depending on the service requirements.

*Slice-aware resource elasticity*: This area addresses the ability to accommodate multiple network slices within the same physical resources while optimizing the network size and resource consumption. This facilitates the reduction of capital and operational expenditures by exploiting statistical multiplexing gains across multiple network slices. Due to load fluctuations and variations that characterize each slice, the same set of physical resources can be used to simultaneously serve multiple slices, which yields large resource utilization efficiency and high gains in network deployment investments, as long as resource orchestration is optimally realized.

### D. Summary

With the envisaged innovations and solutions to be developed and implemented during the course of the project, the different cornerstones of 5G-MoNArch as introduced above will strongly contribute to closing the gaps in the existing 5G mobile network architecture. Fig. 2 shows these innovations allocated to the different logical planes of the mobile network architecture, namely, the user plane (the central part of the figure), and the control & management (bottom grey layer) and orchestration (top grey layer) planes (cf. Section II.A).

Fig. 2 furthermore shows how the different innovations will be associated to the use cases and the corresponding network slices the project will develop and implement in two testbeds, as it is explained in more detail in Section III.

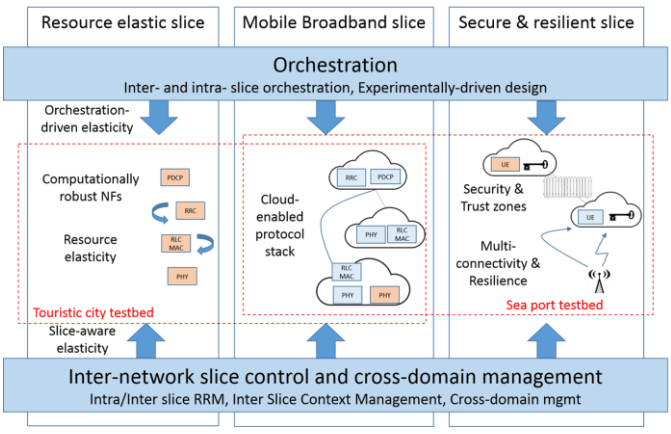


Fig. 2: 5G-MoNArch innovations for the different logical planes of the network architecture

## III. 5G-MoNArch Use cases and Testbeds

A core goal of the project is to validate the feasibility and performance of the developed overall architecture and use cases, through deploying two representative testbeds that can cover a large variety of requirements.

The first testbed implements a typical industrial enterprise scenario in a *sea port environment*, where highly reliable, resilient and secure communication is required together with supporting a large service diversity. Examples for the envisaged services are traffic light control, video surveillance, and sensor based measurements. The network functions aiming at a *resilient* network behavior, e.g., in case of an impairment of radio conditions, will incorporate techniques such as network coding and multi-connectivity. Security for the involved tenants and available services, also across multiple slices, plays an important role. The testbed will therefore implement the *5G island* concept. Finally, to support the service diversity, *multiple network slices* will be required, for example one supporting Ultra Reliable Low Latency Communication (URLLC) for traffic light control, and another one supporting mobile broadband for video surveillance. *Inter-slice control* functionality is hence inevitable to allow for a dynamic and efficient resource sharing.

The second testbed is implemented in a *touristic city environment*, where multimedia services such as interactive augmented and virtual reality (AR / VR) services will be provided. High-speed and low delay communication services are required therefore, and the *elasticity* of the computing and network resources to serve a frequently changing demand in terms of the number of users and service quality plays an important role. For this reason, *computationally elastic network functions* that can scale down their operation in case of a shortage of computational resources, without interrupting the overall service, will be implemented. Furthermore, a *cloud-enabled network protocol stack* will be provided, where the functions can be flexibly placed either at the edge or central cloud. Network orchestration functionality ensures that the function allocation is automatically performed depending on the service requirements, i.e., delay critical functions are placed close to the end-user such that low latency can be achieved. Also in this testbed, network slicing will be implemented such that an eMBB service can be run simultaneously with a low-latency interactive service and the AR / VR services.

## IV. Conclusions

E2E network slicing is a key technology for 5G and a key feature of the 5G network architecture. Network slicing allows a high degree of flexibility and efficiency to adapt the network to a large variety of vertical-driven requirements and use cases. Whereas the focus of research and standardization efforts up to now has been on the conceptual design, the approach of the 5G-MoNArch project introduced in this paper takes the next step towards bringing network slicing closer to reality. As it has been explained, several gaps in the existing concepts, which are critical for the applicability and practical deployment of network slicing, are being filled with a set of enabling innovations. As a key aspect of the project, two important vertical use cases and the required functional innovations are developed and implemented in real-world testbeds. In so far, 5G-MoNArch takes a big step towards validating and verifying network slicing as a core concept of the 5G mobile network architecture. With this practice-driven approach, and the strong footprint of the project through its consortium, the relevant standardization bodies shall be provided with valued and validated output such that the further definitions of the 5G standards can be driven forwards.

## V. Acknowledgement

Part of this work has been and will be performed within the 5G-MoNArch project, part of the Phase II of the 5th Generation Public Private Partnership (5G PPP) and partially funded by the European Commission within the Horizon 2020 Framework Programme. The authors would like to acknowledge the contributions of their colleagues.

## References

[1] 5G-PPP, "5G Empowering Vertical Industries," white paper, 2015.

[2] ETSI Specification GS NFV-MAN 001, "Network Functions Virtualisation (NFV); Management and Orchestration," version 1.1.1, December 2014

[3] 3GPP Technical Report 22.891, "Feasibility Study on New Services and Markets Technology Enablers; Stage 1 (Release 14)," version 14.2.0, Sep. 2016.

[4] 3GPP Technical Report 28.801, "Study on management and orchestration of network slicing for next generation network," version 1.2.0, June 2017

[5] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," version 14.3.0, August 2017

[6] Next Generation Mobile Networks (NGMN) Alliance, "5G White Paper", Feb. 2015, https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[7] Nokia white paper, "Dynamic end-to-end network slicing for 5G," 2016, [online] available at https://resources.ext.nokia.com/asset/200339

[8] 5G-PPP Architecture Working Group, "5G-Architecture White Paper: View on 5G Architecture," white paper, July 2016.

[9] EU H2020 project 5G-NORMA, website: http://www.5gnorma.eu/ [accessed August 28th, 2017]

[10] EU H2020 project Selfnet, website: https://selfnet-5g.eu/ [accessed August 28th, 2017]

[11] EU H2020 project 5G-MoNArch, website: https://www.5g-monarch.eu/ [accessed September 15th, 2017]

[12] Peter Rost, Albert Banchs, Ignacio Berberana, Markus Breitbach, Mark Doll, Heinz Droste, Christian Mannweiler, Miguel A. Puente, Konstantinos Samdanis, and Bessem Sayadi, "Mobile Network Architecture Evolution toward 5G" IEEE Communications Magazine, May 2016.