H2020-ICT-2016-2
5G-MoNArch
Project No. 761445

5G Mobile Network Architecture

for diverse services, use cases, and applications in 5G and beyond

Deliverable D3.1

Initial resilience and security analysis

Contractual Date of Delivery	2018-05-31
Actual Date of Delivery	2018-06-08
Work Package	WP3 – Resilience and Security
Editor(s)	Diomidis Michalopoulos (NOK-DE), Beatriz Gallego Nicasio Crespo (ATOS), Gopalasingham Aravinthan (NOK-FR)
Reviewers	Peter Schneider (NOK-DE), Haithem El Abed (NOK-FR), Rakash Sivasiva Ganesan (NOK-DE)
Dissemination Level	Public
Type	Report
Version	1.0
Total number of pages	77

Abstract: This document contains the initial view of the 5G-MoNArch project on a resilient and secure 5G architecture design, which corresponds to the work conducted within the framework of project’s work package 3 (WP3).

The features associated with resilience span both the radio access (RAN) and the telco cloud part of the network. For the former, RAN reliability approaches such as macro diversity with data duplication and network coding are considered. For the latter, carefully designed network fault management techniques are put forward, considering the corresponding infrastructure redundancy as well as scalability of the network controllers. As far as security is concerned, this document highlights the fundamental features of a secure network design, with emphasis on the concepts of security trust zone profiling and characterisation, as well as that of security monitoring and active learning. In addition, this report contains an initial joint study on resilience and security, analysing the fundamental steps of a trade-off process that balances the resources between resilience and security purposes. The developments on resilience and security modules carried out in the WP3 framework of 5G-MoNArch are incorporated into the overall architectural structure developed in WP2.

Keywords: RAN reliability, Resilience-Security trade-off, Fault management, Security threat detection, security threat prevention, telco cloud resilience

Executive Summary

Together with resource elasticity, *resilience and security* constitute the major pillars on which the functional innovations of the 5G-MoNArch project are constructed. This report summarises the initial conceptual developments of the project on the topics of resilience and security, as this is realised through the framework of work package 3 (WP3).

The role of WP3 within the 5G-MoNArch project is the design and development of the security and resilience innovations. Such innovations are necessary for *instantiating network slices with a secure and resilient functionality*, on the basis of the overall architecture defined in WP2. WP3 represents the "resilient and secure" counterpart of WP4, where resource elasticity is treated. Both WPs have the goal to leverage the baseline architecture of WP2, for providing network slices with customised functionality. The motivation for secure and resilient slices stems from the demanding requirements on 5G networks, towards a robust infrastructure and reliable services and applications that customers and end-users, in particular from industry, can safely trust. Part of the innovations developed in WP3 are planned to be validated in the context of the Smart Sea Port Testbed, as described in WP6.

This document presents the work conducted within the WP3 framework across three different research topics, namely i) *RAN reliability*; ii) *resilience in telco cloud,* and iii) *security*.

- With respect to ***RAN reliability***, this report focuses on the fundamental approaches followed to increase the reliability at the radio access part, which are *macro diversity* with data duplication and *network coding*. Although such approaches are available in the literature since long time, their application to address RAN reliability requirements is new, and so is the corresponding design of the respective network functions.
- As regards ***resilience in telco cloud***, the main topics captured in this report are related to *fault management* approaches including infrastructure redundancy and *scalability of the controllers*. An overview of the adopted fault management approach is provided, and a novel technique on supporting the resilience on telco cloud via *context-aware network function virtualisation* is highlighted.
- As far as ***security*** is concerned, this report summarises the processes of *security monitoring and active learning* used for counteracting security incidents and mitigating their effects. In addition, the concept of *security zones* is addressed, and the major elements that constitute the characterisation of security zones as well as their classification into profiles are listed. Moreover, this report includes an initial analysis on a joint study between resilience and security, yielding a respective *resilience-security trade-off process* that is used to balance the available resources based on the existing resilience/security level and the anticipated performance.

Within the WP3 framework, the most relevant techniques associated with each of the aforementioned topics are assessed, describing appropriate alternatives to be used in the project. In this regard, this document presents a proposal on how to integrate such techniques into the architectural elements developed in WP2. Such integration involves depicting the modules developed in WP3 into the architecture elements developed in WP2. The reasoning for such integration is to ensure an efficient and seamless interaction with the existing building blocks, carried out in four different architectural layers: i) *service*; ii) *management & orchestration*; iii) *controller* and iv) *network*. The contributions to all four layers are described at the level of each specific innovation function, along with potential interactions between security and resilience components.

In summary, this deliverable captures the main work conducted in the 5G-MoNArch framework towards instantiating a specialised network slice functionality that addresses requirements on resilience and security. The fundamental theoretical concepts are introduced, and directions towards customising such concepts in the 5G-MoNArch architecture and the respective use case study are put forward.

List of Authors

Partner	Name	E-mail
NOK-DE	Diomidis Michalopoulos Borislava Gajic Christian Mannweiler	diomidis.michalopoulos@nokia-bell-labs.com borislava.gajic@nokia-bell-labs.com christian.mannweiler@nokia-bell-labs.com
DT	Jakob Belschner	jakob.belschner@telekom.de
NOK-FR	Gopalasingham Aravinthan Bessem Sayadi	gopalasingham.aravinthan@nokia-bell-labs.com bessem.sayadi@nokia-bell-labs.com
HWDU	Onurcan Iscan Yunyan Chang Oemer Bulakci	onurcan.iscan@huawei.com yunyan.chang@huawei.com oemer.bulakci@huawei.com
ATOS	Beatriz Gallego Nicasio-Crespo Susana Gonzalez Zarzosa Ruben Trapero	beatriz.gallego-nicasio@atos.net susana.gzarzosa@atos.net ruben.trapero@atos.net
CERTH	Stavros Papadopoulos Anastasios Drosou	spap@iti.gr drosou@iti.gr
UNIKL	Bin Han	binhan@eit.uni-kl.edu

Revision History

Revision	Date	Issued by	Description
1.0	08.06.2018	5G-MoNArch WP3	Final submitted version 1.0

List of Acronyms and Abbreviations

2G	2nd Generation mobile wireless communication system (GSM, GPRS, EDGE)
3G	3rd Generation mobile wireless communication system (UMTS, HSPA)
4G	4th Generation mobile wireless communication system (LTE, LTE-A)
5G	5th Generation mobile wireless communication system
3GPP	3rd Generation Partnership Project
5G-PPP	5G Public Private Partnership
AAA	Authentication, Authorisation and Accounting
APT	Advanced Persistence Threat
ARQ	Automatic Repeat Request
BSS	Business Support System
CAPEX	CAPital Expenditure
CNM	Cognitive Network Management
CoMP	Coordinated Multipoint (Transmission and Reception)
COTS	Commercial off-the-shelf
CP	Consumer Plane
CU	Central Unit
C-RAN	Cloud RAN
D2D	Device to Device
DDoS	Distributed Denial of Service
DU	Distributed Unit
E2E	End-to-End
eMBB	extreme Mobile BroadBand
FEC	Forward Error Correction
FM CF	Fault Management Cognitive Function
HPA	Hamburg Port Authority
HSRP	Hot Standby Router Protocol
HetNets	Heterogeneous Networks
ICIC	Inter Cell Interference Coordination
ICT	Information and Communication Technologies
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
ISC	Intra-slice Controller
KPI	Key Performance Indicator
LTE	Long Term Evolution
MANO	Management and Orchestration
mMTC	massive Machine Type Communication
MIMO	Multiple Input Multiple Output
NIDS	Network Intrusion Detection System
NC	Network Coding
NE	Network Element
NFV	Network Function Virtualisation
NFVO	Network Function Virtualisation Orchestrator
NR	New Radio
NS	Network Slice
NGMN	Next Generation Mobile Networks
ODL	Open DayLight project
ONOS	Open Network Operating System
OPEX	Operational Expenditure
PAN	Personal Area Network
PDCP	Packet Data Convergence Protocol

PDU	Protocol Data Unit
PNF	Physical Network Function
QoE	Quality of Experience
QoS	Quality of Service
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RRC	Radio Resource Control
SDN	Software-defined Networking
SIEM	Security and Information Event Management
SLA	Service Level Agreement
SMAL	Security Monitoring and Active Learning
SMS	Short Message
SMm	Security Monitoring Manager
SON	Self Organising Network
STZ	Security Trust Zone
STZm	Security Trust Zones Manager
SthD	Security threat detection
SthP	Security threat prevention
SthR	Security threat reaction
ThIntEx	Threat Intelligence Exchange
UE	User Equipment
UP	User Plane
URLLC	Ultra-Reliable Low Latency Communications
VNF	Virtual Network Function
XSThIntEx	Cross Slice Threat Intelligence Exchange
XSC	Inter-Slice Controller

Table of Contents

List of Figures

List of Tables

1 Introduction

5G mobile networks will support diverse services with different requirements that were described and analysed by 5G-PPP Phase 1 projects. Work package 3 (WP3) of 5G-MoNArch deals with resilience and security aspects of future mobile networks. Special attention is put on these two aspects since both resilience and security have an important impact on the overall network operation and in the quality of service offered to the operator's customers. Moreover, failing in achieving the agreed security requirements may turn into major resilience issues, and vice versa. For instance, a distributed denial of service (DDoS) attack that is not quickly detected and promptly isolated may spread and cause downtime in the entire operator's infrastructure, as well as major disruption in the service offered to all tenants making use of it. From another viewpoint, a robust infrastructure where the necessary and adequate means to guarantee resilience at each level (i.e., RAN reliability and resilience of virtual network functions in telco cloud) have been put in place, would represent a more challenging barrier for attackers.

In the framework of the WP3 of 5G-MoNArch, security and resilience are studied on a common ground. The main reasoning for such common study stems from the requirements of the services that they are usually associated with, and particularly from their common impact on the network performance and operational costs. With reference to the network slicing aspect of the 5G networks, such common impact of resilience and security leads to a common design into a common network slice. This also highlights the *importance of network slicing for achieving resilient and secure services*: Without network slicing it would be practically impossible to design a reliable architecture without incurring an unacceptable cost. In fact, by exploiting the concept of network slicing the level of resilience and security can be provided on-demand based on the use cases, thereby considerably increasing the efficiency of network deployment. As a result, with this novel approach based on network slicing, the role of resilience and security within the 5G ecosystem is enhanced.

It is important to note that the human factor influences the network operation to a large extent. On the one hand, the humans are the main creators of security threats that can severely jeopardise the network operation; on the other hand, humans can be a source of other network problems, which are caused even unintentionally. As statistics presented in [AD13] show, people directly cause approximately 12% of all service outages by incorrect maintenance, operation and planning. In addition, people can cause many other problems in an indirect way: Examples are usage of bad programming methods susceptible to bugs, and software patching, insufficient testing of network equipment, as well as other incorrect or erroneous actions. Having this in mind, and including also the intentional security threats caused by humans, it is interesting to note that the human factor in the correct network operation contributes to network outages approximately with 60% [AD13].

In view of the above, and in order to address the common impact on cost and performance, two major causes of network degradation are distinguished, which are related to resilience and security aspects of 5G networks:

- The "unintentional" human cause of network outage, which is mitigated via RAN reliability and telco cloud resilience mechanisms. This is treated in Chapters 2 and 3.
- "Human-intentional" threats, which are in principle addressed by specially-designed security mechanisms. This is treated in Chapter 4.

It is also worth mentioning that, in addition to their service-related binding, network resilience and security are dependent to each other also from deployment as well as performance perspective. In this respect, as will be elaborated in Chapters 3 and 4, the mechanisms for ensuring network resilience are usually associated with infrastructure redundancy, which affects the security performance. Specifically, redundancy in this context entails involving additional infrastructure elements in order to provide alternative mechanisms that guarantee service continuity in case of failure. This additional infrastructure results in an increase in the number of potential targets of cyberattacks, since a) hackers are in fact offered alternative paths for finding a weakness, and b) hackers are offered easy ways to escape once the damage is done. Consequently, this yields an interesting *trade-off between security and resilience*, which is treated in Chapter 5.

1.1 Objectives and Major Tasks

The primal aim of WP3 is to leverage a network slicing-based architectural approach for providing a customised network slice functionality that focuses on resilience and security. To this end, WP3 develops network functions (NFs) which provide secure and reliable communication, even in situations where meeting the high requirements on reliability and security is challenging due to poor radio conditions. Moreover, WP3 provides the theoretical groundwork of the network features developed for the Hamburg Sea Port testbed. Nonetheless, its scope is not limited to the given testbed but is rather versatile enough to be utilised in industrial applications with given requirements in terms of resilience and security.

To achieve this main objective, the study in the WP3 framework is carried out across three main tasks, namely *Reliable RAN operation; Resilience in telco clouds; Security*. A detailed view of the work conducted in these tasks is as follows.

- ***Reliable RAN operation**:* Reliability in RAN refers to the success probability of transmitting a certain packet to its destination within a given delay requirement. In fact, ensuring high reliability is a challenging task, especially for certain use cases such as high-mobility scenarios. To improve RAN reliability, two approaches are studied within this work package. The first approach is multi-connectivity, and particularly its special case that is associated with data *duplication*. With data duplication, the same packets are transmitted multiple times to minimise the probability of erroneous reception. The second approach investigated in this task relates to the well-known method of *network coding*, which is based on performing operations at the intermediate nodes of a network to improve throughput and increase the reliability. The expected outcome of this task is the introduction and customisation of the considered techniques to the overall project architecture, including the derivation of the required characteristics for enabling multi-connectivity and network coding.
- ***Resilience in telco clouds**:* Resilience describes the ability to provide and maintain an acceptable level of services in case of faults. In this task, different approaches to maintain resilience are considered. The main approaches adopted in WP3 are i) *fault management* for troubleshooting, as well as for identifying and isolating network faults; ii) *resource redundancy* towards higher resilience levels; iii) dimensioning and configuring edge cloud resources to allow *autonomous operation* of basic network services at the edge, without requiring continuous connectivity to the central cloud; iv) *supplementary hardening* of the resilience of critical NFs. The expected outcome of this task is two-fold: It involves an investigation of the trade-off between availability and required level of resilience, along with a specification of edge cloud dimensioning and study and deployment of fault management techniques.
- ***Security**:* This task addresses the security aspects of the considered architectures within this project. The considered activities within this task include implementation of *security trust zones* for preventing propagation of security gaps as well as the implementation of security *monitoring and active learning* techniques for identifying, preventing and reacting to security threats. Moreover, this task involves a *case study analysis* of the Hamburg Smart Sea Port testbed, including a risk analysis which is not limited to the specific testbed but rather extends to industrial scenarios where similar security threats are encountered. The expected outcome of this task is an assessment of security solutions for cloud applications, and a trust zone security analysis based on the use case requirements.

This document contains a separate chapter for each of these tasks, including detailed analysis of each sub-task, followed by a chapter that jointly addresses resilience and security aspects. Before proceeding to elaborate the corresponding tasks, some important definitions are in order, followed by a state-of-the-art overview.

1.2 Requirements and Key Performance Indicators (KPIs)

This section contains an overview of the terms used throughout this document. For the reader's convenience, such terms are grouped into two major lists, namely those pertaining to *Requirements* and those pertaining to *Key Performance Indicators* (KPIs). Such requirements and KPIs have been originally defined in [5GM-D6.1], and they are highlighted here for emphasising their role in the subsequent resilience and security analysis conducted in the framework of WP3.

1.2.1 Requirements

As far as requirements are concerned, we distinguish two major categories, corresponding to i) general security requirements and ii) dedicated resilience and security requirements. A detailed view of such requirements follows.

1.2.1.1 General Security Requirements

As general security requirements we classify all requirements which are related to designing 5G systems in a way that they *secure the network, its users and their traffic effectively against cyber-attacks*. In particular, the following requirements are distinguished:

- the consumer plane shall be protected against denial of service attacks from user equipment (UEs),
- the UEs shall be protected against network denial of service attacks as a result of a security attack,
- UEs and the 5G network should be protected against denial of service attack from external networks, e.g. the internet, and from other UEs.

In view of the above, the security mechanisms should a) allow for the verification of the integrity of radio messages; b) provide *confidentiality* to protect voice, data and signalling, as well as subscriber's privacy; c) provide *authorisation and authentication services* for users, devices and networks both at a bearer level and at a services level; d) provide authorisation, integrity protection and confidentiality between network elements and between networks, as well as provide authorisation, integrity protection and confidentiality for new 5G services.

On the other hand, it should be emphasised that the security mechanisms design should be *flexible and configurable***.** This is to ensure that such mechanisms are able to adapt to the needs of the different use cases in terms of 5G technology evolution (e.g. new 5G services). The security mechanisms should also adapt to changing performance requirements (e.g. high-speed communications), applicable regulations and laws, and should also be *extensible* to enable new algorithms and procedures to be incorporated where appropriate.

1.2.1.2 Resilience and Security Requirements

For assessing whether the operation of the considered architecture design meets the required standards in terms of resilience and security, there are three dedicated groups of requirements as described below.

- *Protection requirements* refer to requirements used to define how efficiently the network can protect itself from encountering any type of malfunctions.
- *Detection requirements* are associated to requirements related to the so-called "problem space". That is, the requirements for identifying a problem to the network of any kind.
- *Reaction requirements***,** which correspond to the performance measure used to assess the ability of the system to recover after a problematic functionality has occurred. This is also referred to as the "solution space".

In summary, the *protection requirements* refer to a targeted network design towards minimising network faults, the *detection requirements* reflect the ability of the system to detect and diagnose a malfunction or anomaly, while the *reaction requirements* focus on the ability of the network to withstand such undesirable situations.

1.2.2 Key Performance Indicators (KPIs)

As far as the KPIs for the work in WP3 are concerned, focus is put on the following list, which tabulates the specific KPIs that permits evaluating the 5G-MoNArch system with regards to the security and resilience aspects [5GM-D6.1]. These KPIs are planned to be evaluated in the future work of WP3, in conjunction with the WP6 framework where the validation of the proposed functionalities will be carried out on a larger scale.

Mean time to repair (by ETSI): The MTTR is the statistic mean downtime before the system/component is in operations again.

Reliability (based on 3GPP/ITU-R/5G PPP/NGMN): Refers to the percentage (%) of the network layer packets successfully delivered to a given system node (incl. the UE) within the time constraint required by the targeted service.

Resilience (based on ITU-R): Resilience is the ability of the network to continue operating correctly during and after a natural or man-made disturbance, such as the loss of mains power.

End-to-end reliability: This KPI equals the probability that all network components, including the virtualised and non-virtualised part of the network, are capable to support a required function (taken from the set of computation; networking; storage) for a given time interval.

Reliability of the telco cloud: Probability that a telco cloud component can perform a required function (taken from the set of computation; networking; storage) under stated conditions for a given time interval. This KPI reflects the ability of a telco cloud to withstand any network faults or malfunctions which might have negative impact on system or service performance.

Service restoration time: Time span required between a point in time when a service related malfunction has started (independently of whether this has been diagnosed or not, c.f. network fault detection requirement), until the service has been completely recovered. With this KPIs the 5G system will be assessed in terms of its ability to restore an affected service within a given, usually strongly limited time.

Security threat identification: Percentage (%) of security threats (where any type of security intrusion attempt is regarded as security threat) that are identified by threat identification algorithms, evaluates the effectiveness of security threat algorithms for anomaly detection.

Security failure isolation: Complementary percentage (%) of propagated security failures, i.e., of security failures that pass the security zone (i.e., the zone where certain security measures to be implemented). This metric evaluates he ability of the 5G system to isolate artificially security failures.

1.3 State of the Art

Before proceeding in providing the 5G-MoNArch analysis associated with resilience and security, a state-of-the-art overview is put forward. This overview is structured in three major parts, namely the state-of-the-art pertaining to RAN reliability; resilience; security. In addition, a special reference to the state-of-the-art developments from previous 5G-PPP projects is provided at the end of this section.

1.3.1 RAN Reliability State of the Art

Reliable communication through a wireless link is a challenge due to the time-varying nature of the radio channel. At the time of transmitting a signal, the influence of the channel can be estimated, yet there is always uncertainty involved. This, in addition to interference, renders the quality of the wireless signal reception unpredictable. There is a set of different approaches that can be used and combined to enable a reliable communication over a wireless link, as described below.

- *Link adaptation*: The transmitter of a wireless link adaptively selects a modulation scheme depending on the quality of link (e.g. the power at the receiver or the amount of interference and noise). By using a robust modulation scheme even under good channel conditions, unexpected short-term fading and interference can be compensated to a certain extend.
- *Power adaptation*: A similar approach applies to the transmit power. That is, by using a higher transmit power than the power required, potential problems can be compensated.
- *Forward error correction (FEC)*: This technique adds different levels of redundancy (depending on the targeted error rate) to the payload data. As a result, the receiver can decode the data even in the presence of errors.

- *Interference coordination*: Techniques pertaining to interference coordination schemes, such as inter cell interference coordination (ICIC) [BPV09], can be used to reduce interference at the receiver, which can be a source for errors in the reception.
- *Repetition*: Retransmissions can correct errors by sending data again upon negative feedback. However, they increase the latency of the communication, which is in contrast to the 5G target of an ultra-low latency.

In addition, a single wireless link between two antennas can be improved my means of diversity, which can be grouped into microscopic and macroscopic schemes [PSL+15]. Diversity aims at reducing the probability of a negative impact of fading in the wireless channel, as multiple (in the best case independent) links can be used. In particular,

- *Microscopic diversity* schemes rely on deploying two or more antennas at the receiver and the transmitter, i.e. the usage of multiple input multiple output (MIMO) communication.
- *Macroscopic diversity* relies on sending or receiving the same data from geographically independent locations, which can be implemented in the form of coordinated multi point (CoMP) [LSC+S12].

Within the context of 5G, multi-connectivity is an important means to provide a high reliability. With multi-connectivity, the UE establishes two independent wireless connections as a form of macroscopic diversity. To facilitate this, 5G new radio (NR) introduces the so call packet data convergence protocol (PDCP) split in which a central unit (CU) can distribute data flows over multiple distributed units (DUs) [3GPP38.801]. Another possibility to enhance reliability is to utilise network coding (NC) [ACL+00]. NC is a technique that makes use of the topology of the network, where the nodes in the network do not just forward their packets to the destinations, but perform certain operations on the incoming packets that allow better utilisation of the resources and improve the reliability. Nevertheless, besides the recent advances of multi-connectivity and network coding, *there was no particular focus in the literature on utilising these approaches towards higher RAN reliability*. The work in WP3 targets to cover this gap, and provide the necessary network function design that provides a reliability-customised operation of the RAN. In addition, WP3 provides an analysis of such techniques in the framework of the 5G-MoNArch architecture, in the sense that the RAN reliability techniques developed in WP3 are studied in terms of their role within the architecture. More details on such concepts and how to use it for RAN reliability are provided in Chapter 2.

1.3.2 Telco Cloud Resilience State of the Art

The relevant literature that relates to the work conducted in WP3 is classified into two major groups, namely the work focusing on *network fault management* in telco clouds and *software-defined controller frameworks* with focus on resilience. These two groups are treated separately below.

1.3.2.1 Network Fault Management in Telco Clouds

In the telco cloud domain, there exist different approaches for increasing the overall resilience. Some of the common techniques for mitigating the network faults in traditional network are self-healing SON solutions [HSS12]. Self-healing SON aims at automating the mitigation of outages on the level of individual network cells, including outage detection and root cause analysis. Within such framework different improvements of detection and diagnosis processes can be applied as presented in [N13, NS12].

The introduction of network function virtualisation (NFV) in network design and deployment brought new challenges in handling the network faults. As the faults can occur on different deployment layers, e.g. physical, virtual, application, the fault management needs to be enhanced in order to master the increased complexity in fault localisation and isolation. The work targeting the fault management issues in virtualised environment has been presented in [MHS15] where distributed fault management approach has been chosen. However, despite the considerable progress in this field, *the majority of 5G network architecture proposals did not explicitly (or to a large extent) target addressing the resilience levels of URLLC.*

The requirements on resilience have mainly been implicitly addressed by the management and control entities and mechanisms that are designed in a way to promptly react to unexpected events. For instance, as reported in [Y17], after a violation of quality of service (QoS) requirements is detected on centralised controllers, the problem mitigation is attempted through network reconfigurations. This might involve reconfigurations of network functions parameters, as well as link reconfigurations. In the case that this was not sufficient to overcome the problem, the centralised controllers send a trigger to management and orchestration (MANO) blocks, such as the orchestration entity, in order to perform the action needed for problem mitigation. This might include scale out actions if the resources of network functions are scarce, as well as relocation of existing functions and deployment of new functions.

Although such architecture is capable of reacting to unexpected traffic/network events and mitigate their negative influence to a certain extent, the architecture and mitigation mechanisms are not built under the concept of resilience. In other words, *there is no detailed resilience consideration* intrinsic to the network design, in the sense that there are no specialised network functions for empowering the resilience or service-specific resilience requirements built-in to the network design. Therefore, the *aforementioned mitigation actions and processes are suboptimal and cannot meet different reliability requirements in an efficient way*. In this regard, an approach that highlights the potential of network fault management to provide a customised functionality of the telco cloud with emphasis on resilience is conducted within the WP3 framework, and discussed in Chapter 3.

Typical Fault Scenarios in Telco Clouds

Fault in telco clouds can be caused by various factors and it is very frequently caused by a very complex interaction of different contributing elements. This additionally makes the root cause analysis and mitigation a complex task. The most prominent telco cloud vulnerabilities are hardware, software, network, human and environment. Those factors can contribute to the service fault to a different extent and based on statistical data they appear in telco cloud with different proportion [AD13]. Table 1-1 gives an overview of main causes of telco cloud service outage along with their distribution.

Table 1-1: Telco cloud fault causes

Outage Cause	**Proportion of Outages**
Hardware	15%
Software	19%
Network	21%
People	12%
Environment	21%
Miscellaneous	12%

Each cause of outage listed in Table 1-1 can be further derived into smaller cause groups. Table 1-2 summarises the derived causes.

Table 1-2: Telco Cloud fault causes (detailed view)

Hardware Fault	**Proportion of Outage**
• Servers	65%
• Storage	35%
• Power Supplies	5%
Software Fault	
• Software Bugs	56%
• Upgrades	30%
• Failover Faults	14%
Environmental Fault	
• Power	72%
• Storms	11%

• Cooling, fire, other	17%
Miscellaneous	
• Cyber Attacks	45%
• Capacity	41%
• Other	15%

Statistics illustrated in Table 1-1 and Table 1-2 [AD13] show that apart from hardware and software problems, human factor has significant impact on telco cloud faults. The humans can unintentionally cause the errors/problems in telco cloud operation by, e.g., failover faults (lack of testing), improper fall-back planning after problems in software updates, improper programming (bugs creation) and backup planning, etc. However, also malicious human actions, e.g., cyber-attacks represent a considerable threat for telco cloud operation.

1.3.2.2 Resilient Software-Defined Controller framework

Advantages of software-defined networking (SDN) are growing rapidly in telecommunications due to its capability efficiently manage end-to-end networks and provide the necessary scalability and flexibility. Such scalability and flexibility can bring benefits to network management and maintenance. In general, SDN brings several advantages to mobile network architecture such as high flexibility, programmability, complete control of the network from centralised vantage point, and enables operators to easily deploy new applications, services and tune network policies.

SDN and NFV are two closely related technologies that are often used together in cloud paradigm to complement and benefit from each other. The integration of SDN framework in cloud RAN (C-RAN) can provide several advantages such as dynamic control over fronthaul transport network to allocate available capacity while maintaining overall QoS requirements, realisation of centralised SON (e.g., coordinated scheduling) and configuration and load balancing between virtual base band units (vBBUs) [GRT+16]. Although SDN is a quite matured technology, most of the SDN frameworks have been designed and developed with the major focus on supporting several use cases in fixed and transport networks. However, SDN is an important aspect that can enable dynamic control of radio and networking resources in telco cloud by re-programming/re-configuring VNFs in real-time. Due to the stringent QoS requirements of 5G mobile networks, the SDN framework should introduce low latency, as well as high resilience and scalability in order to be adapted as a controller framework.

With the introduction open network operating system (ONOS) and open daylight project (ODL), the controller framework can be deployed in distributed mode avoiding single point of failure and also improving performance, scalability and resilience [S15]. The distributed architecture is a key feature of ONOS to support both scaling and fault-tolerance by instantiating and linking multiple instances in the cluster. In such approach, each instance can be an exclusive master for set of switches and failure of any instance leads to the selection of new master for those set of switches by the other instances. Raft consensus [OO14] algorithm is used for data synchronisation and state management between distributed instances in ONOS. ODL has a similar clustering model build with Infinispan NoSQL data-store. Although the distributed design is intended to improve the controller layer resilience, it introduces challenges related to timing, consistency, synchronisation and coordination for its adaptability in low latency and time constraint mobile network infrastructure such as telco cloud.

1.3.3 Security State of the Art

1.3.3.1 Anomaly Detection in Mobile Networks

Denial of service (DoS) attacks against the core network can be launched utilising either the control (signalling) or the data (billing) planes. The signalling plane contains all the signals that are necessary for the operation of the different network services (e.g. call handover, enabling/disabling call forwarding). For example, Traynor et al. [TLO+09] propose an attack against the Home Location Register (HLR) that overloads it with Call Forwarding enable/disable signals. On the other hand, the billing plane contains the actual information exchanged between the mobile devices, such as Call Detail

Records (CDR), voice/text messages, Internet traffic etc. For instance, by intensively increasing the volume of the Short Message Service (SMSs) sent through a cellular mobile network, one can degrade its availability [KMP13] [ETM+05] and deny voice service in large cities [ETM+05].

While intrusion detection systems (IDS) can be installed on mobile devices to detect malwares [BHS+08] [LYZ+09], these solutions are not feasible for mobile network operators which have only access to the signalling and billing-related information. Therefore, many researchers focus on anomaly detection techniques using only network-level information.

Billing Information

SMS-related anomalies are an emerging network problem [DBG12] [KMP13] [ETM+05]. Most methods used in this area use the content of SMSs to extract relevant features for anomaly detection/classification (e.g., [AHY11] [YKG+11] [JBW10]). These methods, however, require the monitoring of the SMS content, which sacrifices user privacy in addition to high cost. To this end, researchers have proposed methods that use only high-level SMS information (e.g. time, source, destination) to detect anomalies. Kim et al. [KMP13] propose multiple statistical metrics based on the SMS reply rate to identify the mobile devices involved in SMS-flooding attacks. Their performances are evaluated through simulations. Murynets et al. [MJ13] utilise a combination of two algorithms for the detection of anomalous SMS activities using different levels of abstraction (i.e. aggregate, cluster, and individual device). The first method detects large activity changes with respect to the frequent SMS contact list of a mobile device, while the second method detects changes in the volume of the SMSs sent. Xu et al. [XXY+12] propose multiple features that are extracted from CDR, in order to detect spam SMSs using Support Vector Machine (SVM) classifiers.

DoS attacks against the core network and the mobile subscribers can also be initiated without utilising SMSs. For example, Gorbil et al. [GAP+15] and Abdelrahman et al. [AG14] propose an attack method against the Radio Resource Control (RRC), in which malware-infected mobile devices send periodically Internet packets, so as to force the network to provide them with highly energy consumptive bandwidths.

Since the billing-related data contain high-level relational information, i.e. the source and destination of a communication event, they can be naturally represented using graphs. This fact suggests using graph-based anomaly detection techniques to identify billing related anomalies in a mobile network. For example, Xu et al. [XXY+12] utilise graph-based features in addition to other non-graph-based features for the improvement of the detection rate of spam SMSs. Papadopoulos et al. [PDT16] proposed multiple graph-based features for the detection of billing related anomalies in mobile networks. The effectiveness of proposed features is demonstrated in multiple simulated scenarios, including SMS flood, spam SMS, and RRC-Based Attacks [GAP+15] [AG14]. The authors also show that the proposed features capture information related to the propagation of malwares through the network.

Signalling Information

There exist several methods proposed in the literature for the detection of signalling-related anomalies in mobile networks. Most of them detect anomalies by identifying the difference in activity with respect to a normal/baseline activity. To identify variations from the normal activity, there are two prominent categories of methods in the literature: 1) *statistical methods*, and 2) *machine learning methods*.

With respect to the statistical methods, Gurbani et al. [GKM+17] propose two detectors for differences between the traffic distributions in an abnormal and a normal period. The first one is a non-parametric approach based on the Chi-Square test, and the second is a parametric approach based on Gaussian Mixture Models. Similarly, Bodrog et al. [BKK+16] use a simple metric to capture the difference of a set of KPI values from their average value. Falk et al. [FCS+17] proposes to use Histogram-Based Outlier Scores for the comparison of histograms from normal and abnormal periods and the subsequent detection of anomalies.

As far as machine learning methods are concerned, Gupta et al. [GJJ17] uses hidden Markov models to model normal network traffic and detect differences from that state as anomalous. Gogoi et al. [GBB+14] uses a combination of supervised and unsupervised outlier detection methods for efficient detection of attacks in the network. Similarly, Papadopoulos et al. [PDD+15] use Bayesian Robust Principal Component Analysis in order to model aggregate network data traffic and detect abnormal network behaviour. In contrast to previous work, this method takes account of the periodic characteristics of the mobile network traffic, and thus, has the potential to reduce the false positive detection rate.

1.3.3.2 Decentralised Network Collection for Threat Detection

One of the essential features required in 5G networks is to provide security monitoring, and the first step to achieve this goal is the collection of network traffic. Such network traffic collection is necessary, not only for detecting potential attacks and threats, but also for investigating them and provide the measures to prevent from future incidents. This necessity, in conjunction with the dynamic and flexible nature of 5G networks where the concept of network slicing is present, calls for a decentralised network collection that ensures a more reliable and efficient data collection.

There are different proposals of distributed network forensics frameworks in the literature such as the ones that are analysed by G.S. Chhabra and P. Singh [CS15]. One of the most known frameworks is ForNet [SMS+03], a distributed network logging mechanism for wide area networks. The aim of this framework is to achieve a compromise where it is stored as much information as possible about the network traffic but at the same time reducing the storage requirements. They use for this purpose what they call different *synopsis engines*. It is worth noting that the main issues of this mechanism are i) how to identify the useful network events to store; ii) how to integrate information distributed across multiple networks such as the ones in 5G networks; iii) how to provide security for the own framework components. A more recent proposal to provide collection of network data in a secure and autonomous way is the forensics edge management system (FEMS) (Oriwoh, 2013). It should be noted, however, that this solution is specified for internet of things (IoT) or smart homes context.

1.3.3.3 Threat Intelligence Data Interchange

Nowadays, one of the challenges to face when we talk about security monitoring is the fast and huge growths in the amount, novelty and complexity of cyber-attacks. Consequently, it is crucial for protecting any infrastructure to have timely and updated threat intelligence information. These pieces of threat intelligence data are called IoCs (Indicators of Compromise). This means that any organisation would need to have some procedure to enable receiving these IoCs. This can be even more relevant in the environments such as 5G networks, where the security procedures need to be applied throughout different network slices. Consequently, the most up-to-date threat intelligence shall be shared across network slices and an automated mechanism to ensure this requirement must be put in place.

During the last years many standards related to threat intelligence data interchange formats have appeared. Currently, the most used and also the most promising format for describing cyber threat information is structured threat information expression (STIX), with trusted automated exchange of indicator information (TAXII) as its counterpart for sharing this information in an automated and secure way [D-D4.1]. Both STIX and TAXII have recently being recognised by the European Union for their use in public procurement[1], and although it does not mean supporting these standards become mandatory, it becomes a decisive feature/functionality in cybersecurity solutions.

There are also in the market some threat intelligence platform solutions that allow collecting, aggregating, correlating and analysing threat data from multiple sources in real time to support defensive actions, and sharing the refined intelligence with trusted partners. Some of the most known are: MISP[2], CRITs[3] or Soltra Edge[4].

1.3.3.4 Security advances in WP3

The aforementioned state of the art pertains to related work on anomaly detection and network data collection for threat detection and threat intelligence exchange. In the context of WP3, the above works are leveraged and studied from a network slicing viewpoint. The objective of WP3 in this regard is to i) to extend such advances on security and study their *applicability in the 5G-MoNArch architecture*; ii)

[1] Commission Implementing Decision (EU) 2017/2288 of 11 December 2017: http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515520575463&uri=CELEX:32017D2288

[2] MISP, “MISP Threat Sharing”, http://www.misp-project.org/

[3] CRITS, “Collaborative Research Into Threats”, https://crits.github.io/

[4] Soltra Edge, https://www.soltra.com/en/products/soltra-edge/

to provide a secure functionality which is targeted to *meet the requirements of a specific use case scenario*, namely the Hamburg Smart Sea Port testbed. Details of this security approach are provided in Chapter 4.

1.3.4 State of the art from 5GPPP Phase I and SDOs

This section captures the major developments of 5GPPP Phase I projects, along with the recent developments in standardisation organisations (SDOs). The 5G RAN (a.k.a. NG-RAN in 3GPP) design enables faster operation of certain network functions than in legacy systems. Such example is the operation of traffic steering on a faster time scale on lower protocol stack layers, i.e., packet data convergence protocol (PDCP) level, as opposed to "hard switching" on radio resource control (RRC) level between different RATs (e.g., between 3G and 4G) or access. Similar to the approach adopted in this document, this is achieved by exploiting *multi-connectivity* where traffic flow adaptation can be performed dynamically considering the radio link conditions at different frequency bands.

On this basis, by employing packet duplication, the signal-to-interference-plus-noise ratio (SINR) can be increased especially in the low regime, which can improve the cumulative link reliability (see, Figure 1-1) [MII-D52] [MII-D24] [MBQ+18] [5GARCH17-WP]. In particular, duplicated packets are sent over co-operating links, which is seen as a single transmission from the UE perspective (similar to single frequency network concept). SINR improvements at 50-th percentile cumulative distribution function (CDF) read as around 1.5 dB and 3 dB for two and three links cooperating, respectively, as compared to long-term evolution (LTE) baseline. Such a construct can reduce the need for retransmissions and, thus, delay within the RAN. It should be noted that packet duplication is also considered in the 3GPP new radio (NR) specification, where the completion is aimed for the end of Release 15 [3GPP-38.323].

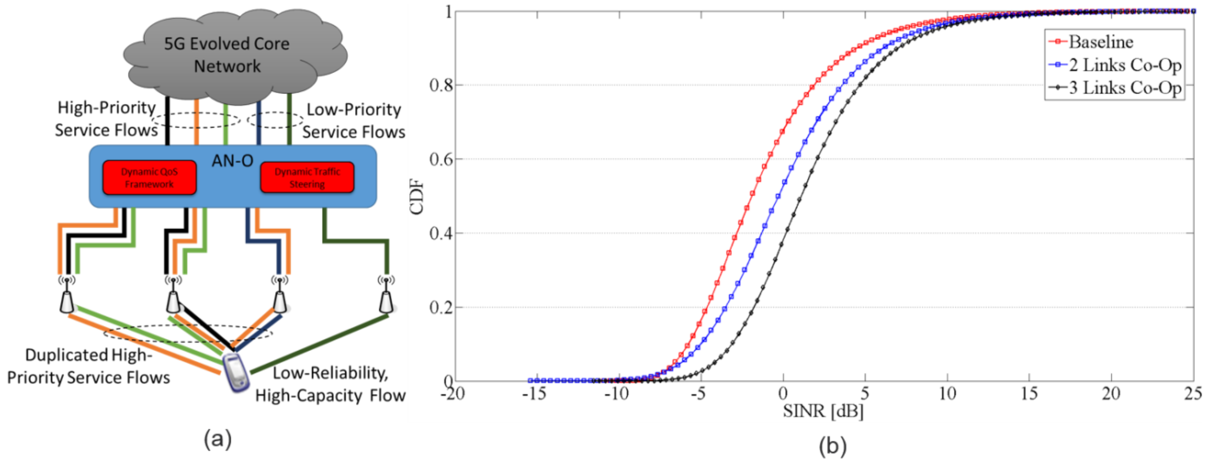


Figure 1-1: Dynamic traffic steering can exploit multi-connectivity to increase the cumulative link reliability [MII-D52]

Reliability improvements via SINR gain can also be achieved by utilising interference management schemes along with the additional degree of freedom provided by the unlicensed bands [MII-D52]. This becomes particularly important in case of using unlicensed band (for instance, as is the case in licensed assisted access (LAA)) in dynamic radio topologies (such as vehicular nomadic nodes (NNs)). In such cases, the radio link conditions, uncontrolled interference sources (e.g., WiFi access points), and network topology can change frequently, and the resource allocation and interference management schemes required to adapt changes in an agile way. It should be also noted that different service types associated with different KPIs should be taken into account.

Another key challenge in LAA is the effect of listen before talk (LBT) to the actual performance, since the dense activation of LAA-enabled nodes can lead to severe interference, if there is no centralised LAA coordination. In one example case study [MII-D52], when some users cannot achieve the target reliability in licensed bands (according to the service requirements), LAA is utilised in parallel. That is, together with the NNs in licensed carriers, a set of LAA NNs provide multi-connectivity by parallel redundant links to the user at a reserved (but low) number of resource blocks. In Figure 1-2, the satisfaction ratio is provided (the proportional number of users which achieve 99.999% reliability). Satisfaction in terms of reliability can be measured at the 10-5 percentile of CDF of SINR where the threshold of 0dB is assumed as exemplary target reliability. It is observed that, when coordinating the LAA operation in a reliability-driven fashion to improve the worst case SINR (by allowing multi-connectivity with limited number of reserved resource blocks), the reliability requirement can be met.

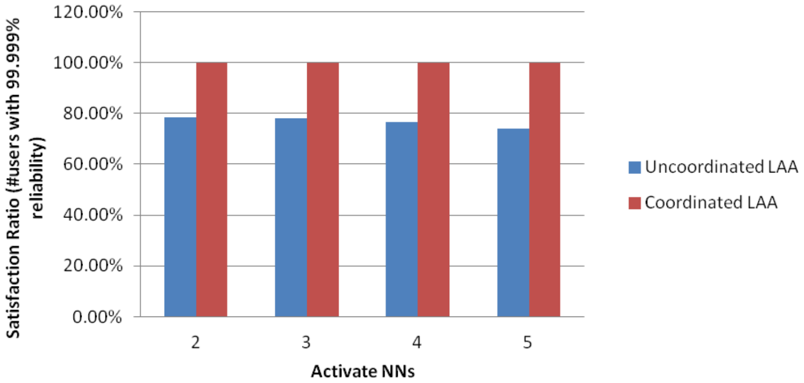


Figure 1-2: Satisfaction ratio for reliability enhancement by utilising multi-connectivity in unlicensed bands and coordinated resource allocation [MII-D52]

Coordinated group transmissions can also contribute to the reliability increase especially on the uplink (UL) and when the cell size is large. Here, device-to-device (D2D) groups can be formed, where group members transmit the UL data to the base station (BS) in a coordinated fashion simultaneously [MII-D52]. In particular, D2D communication within a group of UEs is used to distribute the UL user data to the group members. Such an enhancement can be critical for the UL due to the low Tx power of the user equipment (UE) compared to that of the BS. Such a scenario can be envisioned for, e.g., massive machine-type communications (mMTC), where devices can form groups and aid each other, deep indoor deployments, and emergency services in remote areas or where part of the infrastructure has been damaged. Figure 1-3 shows the gain for maximum achievable bit rate when groups of five users randomly dropped in hotspots transmit as a group (within 50m hotspot) compared to the maximum achievable bit rate if they would transmit on their own. The median relative bit rate gain is in this case between 1.25 for hotspots close to the base station and up to 5 times for hotspot far away from the base station, i.e. for groups in bad coverage [MII-D52].

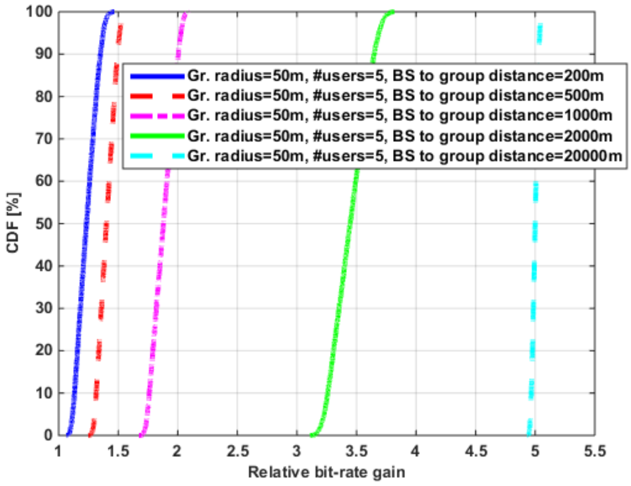


Figure 1-3: UL improvement by utilising coordinated group transmissions [MII-D52]

5G envisions to support a diverse set of use cases enriched by new business sectors aka vertical industries. This implies new architectural concepts and capabilities to enable such business models and to provide enhanced applications and services. To this end, *security architecture* shall be natively integrated into the overall 5G architecture [5GPPP17]. The security architecture in 5GPPP Phase 1 has been developed in the 5G-ENSURE project[5] and can be seen as an evolution based on the existing security architectures for 3G [3GPP-23.101] and 4G [3GPP-33.401]. That is, the basic concepts, e.g. domains and strata, remain but have been adapted and extended to fit and cover the 5G environment [5GPPP17] [5GPPPSEC17].

1.4 Structure of the Document

The remainder of this document contains the developments conducted within the framework of WP3 of 5G-MoNArch, and is organised as follows.

Chapter 2 describes two approaches, *data duplication* and *network coding*, which aim at contributing to achieve the strict levels of RAN reliability that 5G services require. Performance aspects and deployment options are presented, with focus on implementation on the overall 5G-MoNArch architecture.

Chapter 3 focuses on telco cloud resilience. The chapter presents four different approaches towards higher resilience in the telco cloud, namely *redundancy*, *fault management*, *resilient and scalable controller* and *autonomous failsafe operation*. Chapter 3 discusses on how these can be improved, adapted or evolved in order to deal with the new challenges brought by the so-called network cloudification.

Chapter 4 is devoted to describing how to protect 5G infrastructures, as well as the services built upon them, against the most likely threats, with the ultimate objective of minimising the risks and their consequences for the overall service operation. The chapter starts by presenting the Security Monitoring and Active Learning (SMAL) process which methodologically addresses security requirements. This is followed by a characterisation of the concept of Security Trust Zones, which offers the necessary means to implement the stages of the SMAL process, and proposes how security trust zones can be realised in the 5G-MoNArch baseline architecture. The chapter then concludes with an analysis of the Hamburg Sea Port Testbed case study from the security point of view, which serves to illustrate how security trust zones could be used to protect against the most likely threats and risks identified.

It is emphasised that *Chapters 2, 3 and 4 contain a section devoted to the new modules and functions added to the baseline 5G-MoNArch architecture*, as this was described in D2.1. This structure is used to highlight the role of WP3 into the overall architecture design.

Chapter 5 introduces the concept of jointly addressing the resilience and security requirements in 5G systems. In particular, a resilience-security trade-off is put forward, followed by a detailed view of the respective process as well as the trade-off criteria involved in both the resilience and the security domain.

Finally, *Chapter 6* summarises all contributions that the resilience and security innovations make to the baseline 5G-MoNArch architecture. The new modules and functions developed in Chapters 2-4 with respect to the baseline 5G-MoNArch architecture are presented in an aggregated form. Conclusions are drawn, and future work is outlined.

[5] 5G-ENSURE, http://www.5gensure.eu/

2 Towards Higher Reliability at the RAN

The most common 5G services are extreme mobile broadband (eMBB), massive machine type communication (mMTC) and ultra-reliable low latency communications (URLLC) [5GM-D6.1]. Such services differ in their nature and requirements. With respect to reliability, URLLC services are most challenging since the conventional methods used for increasing the probability of a successful reception in legacy systems, as described above, become insufficient. In particular, the extremely strict requirements on reliability reach the level of 99,999% probability of uninterrupted operation [5GM-D6.1]. As a result, in order to attain this extremely high level of reliability, alternative approaches are put forward, such as *data duplication* and *network coding*. These techniques are described in more detail in the following Sections 2.1 and 2.2. Section 2.3 then discusses how these approaches can be integrated into the 5G-MoNArch architecture.

2.1 Macro Diversity with Data Duplication

Data duplication refers to the case where redundant data transmissions are used as a means to protect the correct reception of the data. This case typically applies to scenarios where the UE is connected to two or more access points, which are sufficiently far apart from one another so as to ensure independency of the wireless links involved. From another viewpoint, data duplication represents a specific implementation of the concept of *macro diversity* [ODG+05] where multiple transmissions across uncorrelated links are combined at the receiver, thereby increasing the probability of correct reception.

In 5G-MoNArch, data duplication is studied as a special case of dual connectivity [3GPP-36.808], [3GPP-36.842]. Specifically, dual connectivity is a technique introduced in the LTE standards, as an attempt to create dually aggregated links where data throughput is considerably increased. However, in the context of 5G-MoNArch and particularly of WP3, the *dual connectivity* concept is *modified* such that data is not split but rather duplicated, so as to contribute towards higher reliability rather than higher data rate.

In the remainder of this section, we highlight the technical features of implementing data duplication techniques in 5G. We first discuss the deployment characteristics of data duplication, seen as an extension of LTE's dual connectivity; then, we elaborate on the particular features of its implementation, as this affects the RAN protocol stack. We distinguish between two major implementation options of data duplication, corresponding to duplication in Heterogeneous Networks (HetNets) and in cell boundaries. These two implementation options are treated in the ensuing Sections 2.1.1 and 2.1.2, respectively.

2.1.1 Data Duplication in Heterogeneous Networks (HetNets)

A typical deployment scenario used for the dual-connectivity approach in the LTE standards to increase the throughput is the HetNet approach [3GPP36.842], which is also anticipated to provide coverage for data duplication. With this approach, the UE connects simultaneously to both a macro cell and a small cell, which usually operate in different frequency bands. It is also assumed that 5G networks will adopt a *centralised architecture, where networks functions are split between two RAN units, namely the central unit (CU) and the distributed unit (DU)* [3GPP-38.801]. It is also highlighted that with this approach no mobility is assumed, in the sense that the UE is assumed to remain static during its communication session within the HetNet.

An illustration of the HetNet deployment in the centralised architecture is provided in Figure 2-1. As can be seen from Figure 2-1, it is assumed that the coverage area of the small cell falls within that of the macro cell. This allows that the UE is simultaneously connected to a macro and a small cell.

On the basis of the centralised architecture [3GPP-38.801], the lower layers of the protocol stack of both the macro- and the small cell take place at the corresponding DUs. Then, the integration of the signal flow to both links involved is carried out at the CU, which contains the higher RAN layers. It is noted that, in the context of NFV, the CU can run in a virtualised implementation, such that it represents part of the telco cloud itself. In such case, the orchestration of the CU resources represents part of the telco cloud management, as described in Chapter 3.

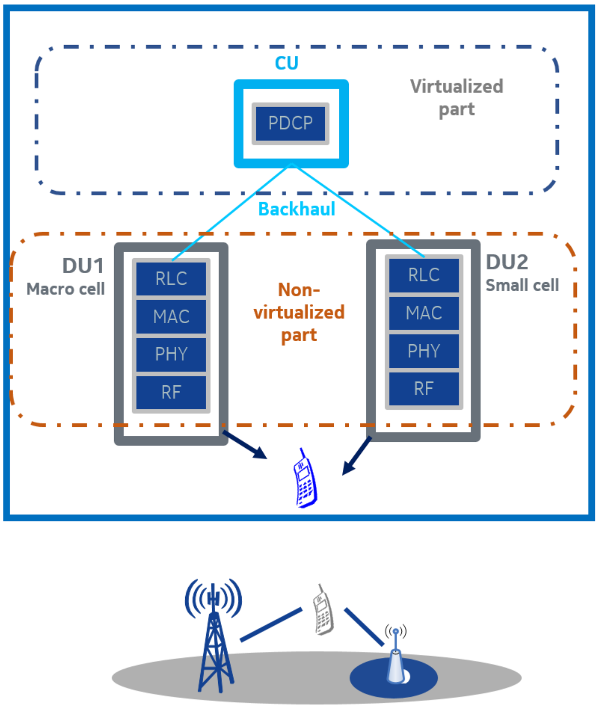


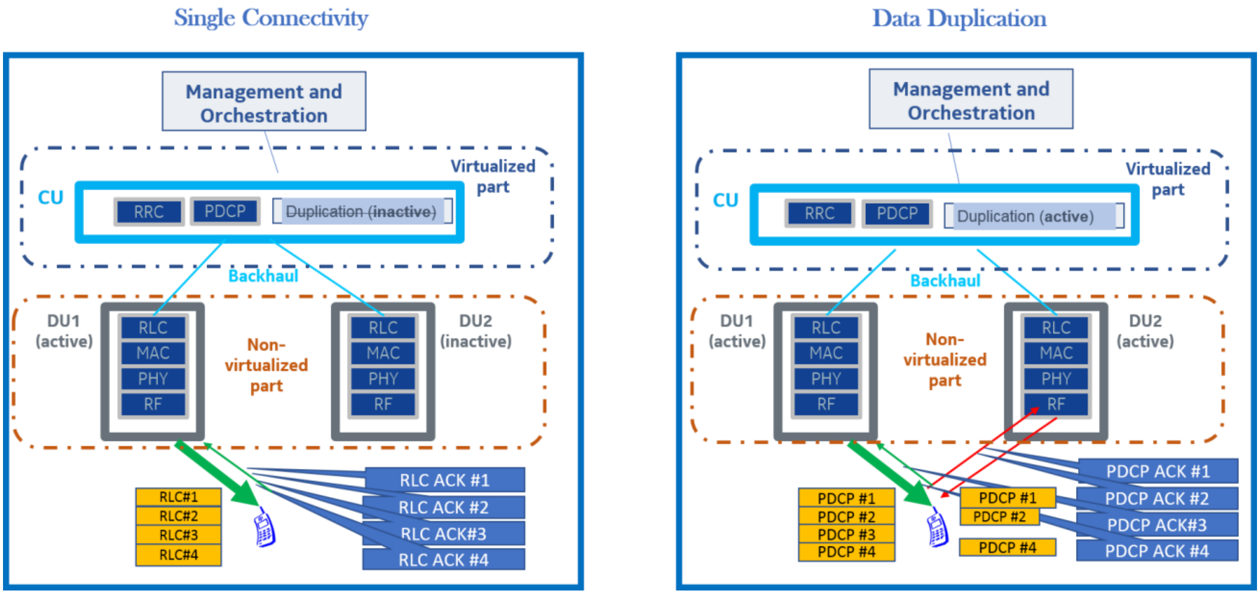
Figure 2-1: HetNet deployment under the centralised architecture, where network functions are split between the CU and the DU

2.1.1.1 Effect of Unbalanced Links

The implementation of data duplication requires special coordination of the signal flow at the CU. Specifically, the CU needs to take care that duplicate packets are delivered correctly to the UE, and that the overhead of the extra resources needed is minimised. Particular attention should be put in scenarios where the two links involved are *unbalanced*, in the sense that their corresponding data rates are considerably different. In such case, the coordination of the duplicated packets is not a straightforward process, since the rate at which the two links deliver data is not identical hence any lost packet from any of the links would correspond to a different packet sequence number at the other link.

To this end, it becomes evident that novel coordination techniques should be introduced in the data duplication process, which take into account the special features of non-balanced links. In this regard, two major points where data duplication differs from existing approaches, from the perspective of the underlying technology.

- *Introduction of Packet Data Convergence Protocol (PDCP) acknowledgments*. In LTE standards, the packet acknowledgment feedback (ACK) sent from the receiver to the transmitter in order to indicate whether the transmission was correctly received is carried out in two layers: At the medium access control (MAC) layer by means of hybrid automatic repeat request (HARQ), and at the radio link control (RLC) layer by means of outer ARQ. On the contrary, given that the RLC layers of the two involved links do not process the exact same packet sequence (i.e., \an RLC packet number \#2 for DU#1 is not necessarily identical to RLC \#2 for DU#2), in the data duplication case feedback should be sent to the PDCP packet numbering instead. This process is expected to be introduced to 5G systems, and is illustrated in Figure 2-2.



Single Connectivity

Data Duplication

Figure 2-2: Single connectivity versus data duplication, as seen via the prism of the new signalling involved to and from the RAN and the management and orchestration layer

- *Management and Orchestration*: The activation of the data duplication process is followed by utilisation of additional resources which require special administration and control. In this respect, the data duplication function that resides in the CU (c.f. Figure 2-2) is orchestrated by a higher-level entity which resides in the MANO layer. Besides the overall orchestration of the virtualised resources, this entity is responsible for deciding whether the data duplication function should be activated, and if so, what is the amount of virtualised resources allocated to it. More details on the architectural aspects are part of Section 2.3. The RAN radio resource control (RRC) entity is then assigned the task of allocating radio resources between the two involved links, based on the needs of the underlying service.

2.1.2 Duplicated Transmissions in Cell Boundaries

The second deployment option of data duplication aims at increasing the mobility-related reliability of 5G networks. This is in line with the high reliability requirements of 5G systems, which dictate that the interruption caused during handovers is minimised, since otherwise such interruption would affect the delivered service. As a result, the mobility-related duplicated transmissions are used in *cell boundaries*, usually of macro cells, where mobility events such as handovers take place. It is emphasised that, in the context of critical industrial applications, any interruption could potentially harm the quality of the delivered service, hence conventional handover approaches are in principle not sufficient for meeting the requirements of mission critical industrial applications [VML+18].

It is noted that duplicated transmissions in cell boundaries are complementing the concept of data duplication in HetNet deployments, in the sense that they contribute towards increased reliability not only in “hot spots”, which HetNet deployments aim to cover, but also in cell boundaries where a large portion of access interruptions take place.

Control Plane Duplication

The main challenge associated with mobility events in cell boundaries pertains to the minimisation of the connection interruptions during handovers. Such interruptions are usually associated with a loss of the control plane signal at the UE, leading to time-consuming connection re-establishment processes.

The solution anticipated to tackle the problem of interruptions during handovers is the duplication of the control plane signal at cell boundaries. An overview of such technique is illustrated in Figure 2-3. As can be seen, the main difference of the control plane duplication as opposed to conventional handovers is the fact that the control plane (i.e., the red connection line in Figure 2-3) flows from the CU to the UE via both DU1 and DU2. This allows for higher control plane connection robustness, aiming at minimising the cases where the control plane connection to the network is lost. That is, to minimise the cases leading to handover failures.

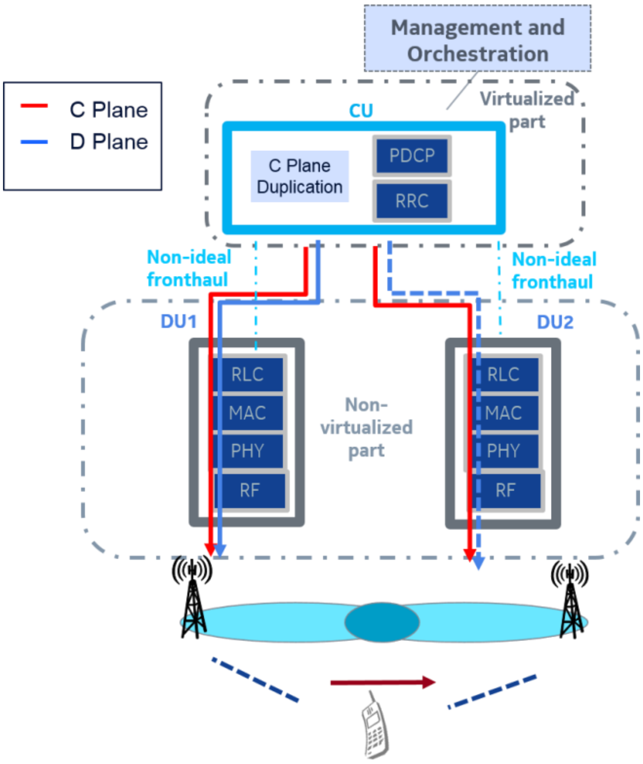


Figure 2-3: Control plane duplication in cell boundaries

From a technical point of view, control plane duplication is different from Coordinated Multipoint Transmissions (CoMP), since the link that connects the CU with the DUs is non-ideal. As a result, the latency introduced by this link will be prohibitive for deploying synchronised coordination methods such as CoMP. Moreover, control plane duplication is technically different from the data plane duplication described in Section 2.1.1, as explained below.

The control plane duplication process during handovers, sketched in Figure 2-3, is as follows: As the UE approaches DU2 from DU1, the reception power from DU2 increases until it reaches a threshold value. In case of a conventional handover, such a threshold value is usually set higher than the reception power from DU1, in order to allow for some level of certainty towards a handover decision. In contrast to conventional approaches, in the control plane duplication process DU2 is considered as a secondary connection to the UE, which is triggered *earlier* than the conventional handover time.

Technically speaking, the early activation of the secondary cell implies that the addition of DU2 as a secondary node is triggered by a threshold which is lower than that of the conventional handover. Adding DU2 as secondary node implies that the control plane information is exchanged with the UE via both DU1 and DU2. It is noted that the main connection via DU1 is still maintained. However, for as long as DU2 remains the secondary connection, the control plane signal from/to DU2 is used as a fall-back option in case the connection to the master node (that is, DU1 in this case) fails. Then, as the UE moves deeply into the coverage area of DU2, DU2 becomes the main connection while DU1 becomes the secondary option.

Eventually, and assuming that the UE indeed moves towards the coverage area of DU2 and does not return to its original position, DU1 will be removed as secondary connection and the UE will be single connected to DU2. This implies that, in the long run, the UE will connect exclusively to DU2, hence the described process will converge to that of the conventional handover. However, duplicating the control plane includes an additional step that reduces handover failures and, as a consequence, connection interruptions as well.

2.2 Network Coding Approaches, Tailored for Higher Reliability

Besides data duplication, network coding (NC) is a technique with high potential in improving the reliability and throughput of networks. In this direction, the NC concept, tailored for high reliability, comprises two relevant changes compared to legacy systems such as LTE:

- In traditional networks, signals from different nodes are treated separately, and the intermediate nodes within the network are only allowed to perform routing operations, i.e., the intermediate nodes forward their received signals to their destinations without performing any kind of processing. In the pioneering work [ACL+00], however, it was shown that for certain networks the performance can be improved if the intermediate nodes are allowed to perform operations, where they combine their received packets and forward these combinations to their destinations, where they are decoded. Depending on the application as well as the requirements involved, this improvement in the *performance can be converted into gains in terms of transmission rate, reliability, as well as transmission power*.
- Network Coding can be an alternative to the used Automatic Repeat Request (ARQ) approach: In legacy communication systems, the sender of a packet waits for a negative acknowledgement by the receiver to initiate potentially required retransmissions that compensate errors. Instead, the NC encoder continuously generates linear combinations of subsets of the source data packet. As soon as the NC decoder received a sufficient number of linear combinations, it is able to decode the source data packet. It then sends a single acknowledgment, such that the NC encoder can stop sending or proceed with new source data. Note that, although NC can be seen as an alternative to ARQ, it does not stand for a direct replacement of it in the sense that *such techniques can be employed together to benefit from the advantages of both of them.*

In the framework of 5G-MoNArch, network coding is mainly studied to improve the RAN reliability. In the following, we first describe in Section 2.2.1 an approach for implementing NC in the CU/DU architecture as proposed for 5G NR (Figure 2-1). We then describe two advanced implementation scenarios in Section 2.2.2, for downlink (Section 2.2.2.1) and uplink (Section 2.2.2.2).

2.2.1 Application to the CU/DU Architecture

The second property of NC described above is well suited for increasing RAN reliability: It enables that source information can be sent out with a configurable level of redundancy, by adapting the rate at which the NC encoder generates linear combinations. Figure 2-4 shows an example for a highly reliable service, where the NC encoder can generate linear combinations out of an incoming IP packet at a high rate (upper part of the Figure). The NC decoder will then be able to decode the original data within a short period of time, even if single transmissions of linear combinations fail. The time-consuming ARQ process (shown in the lower part of the Figure) can be avoided.

There are different options to integrate Network Coding within the 3GPP protocol stack: [KVT12] proposes an implementation of Network Coding at the MAC Layer of LTE. The Network Coding implementation (“MAC-RNC”) augments the traditionally used HARQ and RLC retransmissions. A Protocol Data Unit (PDU) received from the RLC layer is divided into a set of source-symbols, from which a stream of network-coded symbols is produced. The UE receives this stream and sends a single acknowledgement after it was able to decode the RLC PDU.

In addition, two solutions are proposed for the case where the UEs are connected to multiple DUs:

- In the first solution, the RLC PDU is forwarded to two transmission points who execute independent implementations of MAC-RNC. The UE can use the network-coded data from both transmission points to recover the RLC PDU.
- In a second solution, a secondary transmission point does not have access to the PLC PDUs, but to the output of the MAC-RNC implementation running in the first transmission point. Therefore, the secondary transmission point operates based on network coded packets. Out of the network-coded packets the secondary transmission point received, it generates additional new linear combinations and transmits them towards the UE.

It is noted that, although the MAC layer approaches have been proposed for LTE, they can be applied to 5G NR as well.

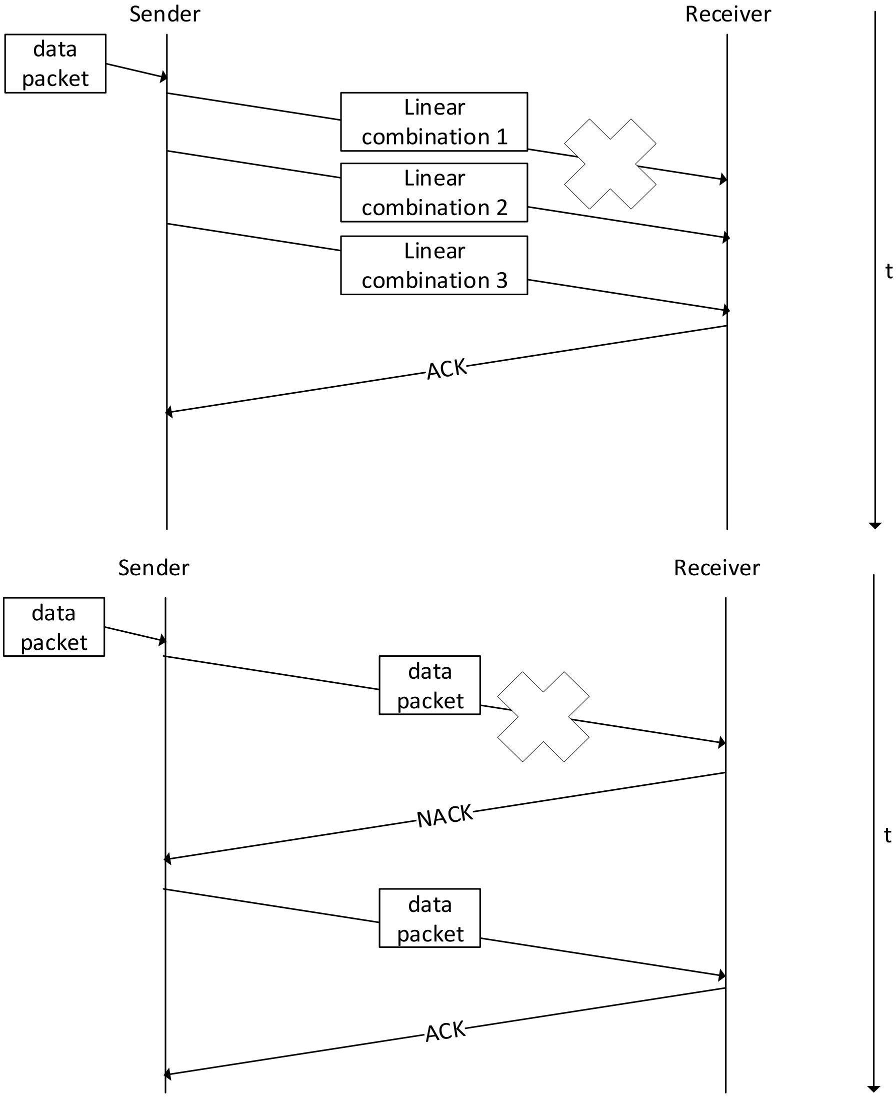


Figure 2-4: Network coding example for a highly reliable service

A network coding implementation at the PDCP layer, named PDCP-RLNC, is proposed in [VTD+18]. Such implementation receives PDCP PDUs and generates out of one or multiple PDUs one or multiple streams of linear combinations, which are to be transmitted by lower layers. Retransmissions at RLC and HARQ are not required any more. For multi-connectivity, the 5G NR architecture supports a CU which is connected to multiple Distributed Units (DUs), as depicted in Figure 2-1. PDCP-RLNC is well suited for the 5G NR CU / DU architecture: If PDCP-RLNC is implemented in a CU, it can send different linear combinations to the UE through two or more DUs, thus exploiting the full benefit of multi-connectivity.

With respect to performing operations at intermediate nodes of the network, the topology of the network plays an important role. For instance, network coding cannot be applied on unicast communications with a single transmitter and a single receiver. In fact, the smallest network that can benefit from processing at intermediate nodes consists of three nodes, which are able to hear the messages intended to nodes other than themselves. In 5G-MoNArch we consider the *CU/DU architecture as a baseline and develop our techniques according to this structure*. In CU/DU architecture, UEs can have a direct physical connection to (multiple) DUs, but they are not connected to CUs directly. Instead, the DUs can communicate to CUs, where the connection between CUs and DUs are more stable and assumed to have high capacity.

2.2.2 Considered Scenarios

Network coding is a broad concept and any technique where the nodes in the network perform coding operations on multiple packets can be considered as a network coding method. In the framework of 5G-MoNArch, we are interested in network coding techniques that may be beneficial with the given system architecture. As a result, we focus on two different scenarios, namely the downlink and uplink scenario, and study them in detail.

In the following, we discuss uplink and downlink communication with network coding. For the uplink scenario, we consider that the UEs are transmitting their messages to DUs, which are then combined (network coded) and forwarded to CUs. For the downlink scenario we consider a simpler setup, where we only consider DUs that transmit network coded packets to multiple UEs, depending on the feedback they receive from the UEs.

2.2.2.1 Downlink Scenario

In traditional cellular communication systems, reliability in the physical layer is ensured by forward error correction codes and retransmissions based on the ACK/NACK feedbacks. If downlink communication is considered, the retransmissions are managed by the base-station, and are performed for each user separately. Moreover, due to the nature of the wireless medium the signals transmitted to one user can also be received by other users. However, users in general do not process signals which are not intended for them. By proper utilisation of network coding approaches, the fact that users can overhear messages can be converted into a gain, e.g., in terms of the number of retransmissions.

We consider a downlink communication from a DU to at least two UEs. The considered scenario can be several unicast communication links, where each UE demands different packets, or a multicast communication where the same packets are transmitted to each user. We further assume that there exists a feedback channel, where the UEs can inform the DU about the reception of the packets. This form of downlink communication is already addressed by the current communication standards such as LTE, however the utilisation of network coding is not considered in the existing systems.

In the following, we explain a network coding method for such a downlink communication. The main assumption pertaining to our network coding approach are as follows.

- *The UEs* that are going to take part in the network coded downlink communication *are grouped by the DU*. The grouping is made according to the relative positions of the UEs, i.e., UEs in proximity are grouped together, and each UE knows to which group it belongs to.
- The UEs have a buffer, where they can store a certain number of received packets (that may be intended also for other UEs).
- The transmitted packets contain a unique ID number allowing them to be distinguished within the buffer.

Network Coding for Downlink

The main idea of the proposed technique is to prepare smart retransmissions according to the ACK/NACK feedback from the UEs. For a better comprehension of the proposed technique, let us consider the following example with one DU and two UEs.

The DU wants to transmit the packet A to UE_1 and packet B to UE_2. Due to the nature of the wireless medium, both UEs receive signals containing the information in A and B. Let us assume, that UE_1 could not decode the packet A, but was able to decode the packet B, and UE_2 could not decode the packet B, but decodes the packet A. In a conventional system, the DU would initiate a retransmission for UE_1 and another retransmission for UE_2. However, as the DU already knows which packets are decoded by which user, it can initiate a network coded retransmission, e.g., the DU can generate a linear combination of the Packets A and B (C=A+B) and transmit this packet to both users. As UE_1 already buffers packet B, it can extract A by simply reversing the network coding operation C+B=A. Similarly, UE_2 can recover its missing packet. This example is depicted in Figure 2-5.

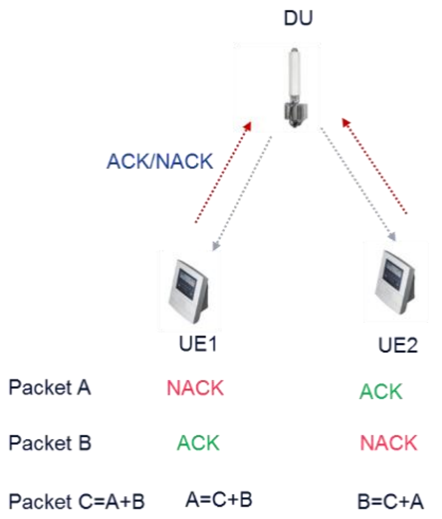


Figure 2-5: Network coding for downlink

Note that for this example the required number of retransmissions is reduced by 50%, i.e., the same reliability can be obtained by using fewer resources, or the remaining resources can be used to improve the reliability further. This approach can be further extended to more users.

Rate Performance

The described network coding strategy can be evaluated analytically, as the links between the DU and the UEs can be described as packet erasure channels (PEC). The maximum achievable (normalised) transmission rate of a PEC is defined by its capacity, given by 1-*e*, where *e* is the packet loss probability. Assuming two users with packet loss probabilities e_1 and e_2, one can describe the normalised achievable rate region $\hat{R}$ as

$$\hat{R} = \left\{(R_1, R_2) \geq 0 : \frac{R_1}{1-e_1} + \frac{R_2}{1-e_2} \leq 1\right\}.$$

An example with e_1=e_2 is depicted in Figure 2-6, where the light grey area describes the $\hat{R}$. Any points on the border can be achieved by using time-sharing, i.e., allowing different amount of resources to each user. As can be seen from the Figure, the sum of the rates however stays constant. It was shown in [GT09], that by utilising the feedback for the network coded retransmissions, one can increase the normalised rates and obtain the following rate region $\tilde{R}$.

$$\tilde{R} = \tilde{R}_1 \cap \tilde{R}_2$$

with

$$\tilde{R}_1 = \left\{(R_1, R_2) \geq 0 : \frac{R_1}{1-e_1} + \frac{R_2}{1-e_{12}} \leq 1\right\}$$

$$\tilde{R}_2 = \left\{(R_1, R_2) \geq 0 : \frac{R_1}{1-e_{12}} + \frac{R_2}{1-e_2} \leq 1\right\}$$

Figure 2-6 also depicts $\tilde{R}$, and one can clearly see that the rate region where the feedback is utilised with network coding is larger.

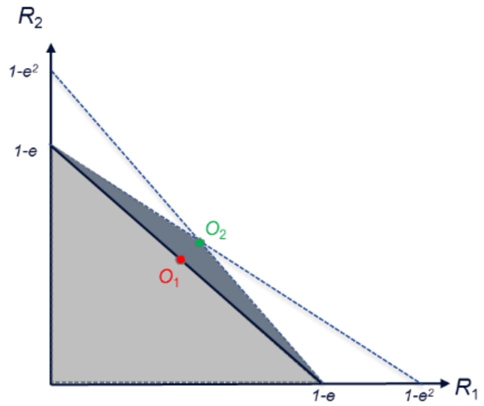


Figure 2-6: Achievable rate regions with and without network coding

Note that the current mobile communication systems are designed with the target error probability of 10% for the first transmission. Taking this as a baseline and assuming e_1=e_2=0.1, the reference approach without network coding would operate the point denoted by O_1 in the Figure, if equal time sharing (with symmetrical rates) is assumed. On the other hand, by utilising network coding the operating point shifts to O_2. For example, in order to successfully transmit 45 packets to each UE with the conventional approach, the DU needs on average 100 total transmissions. By utilising network coding, 100 total transmissions can lead to 47 successfully transmitted packets to each UE, saving 4 of 10 missing packets.

As shown in this example, *an increase in the rate region leads to transmitting packets in a more efficient way*, such that less retransmissions are needed. This means that the same reliability can be obtained by using less resources. The resources saved by network coding can either be used to start with the transmission of the new packets, or those resources can also be used to further protect the packets for transmission errors by using more redundancy (e.g. by lowering the channel coding rate) to decrease the target error probability. It is further noted that the gain obtained by network coding depends on the packet loss rate, which can be different depending on the layer and application, where network coding is utilised. Moreover, implementing network coding in different layers can pose different challenges (e.g. see [SI12] for an implementation in the physical layer). Therefore, the layer where network coding is implemented also affects the network coding performance.

Note that the presented scheme is a simple but powerful network coding technique that was analysed for different setups and conditions before (e.g. in [GT09] and [SI12]) from an academic point of view, without considering the integration to an existing system. In the framework of 5G-MoNArch we investigate this method further from a practical point of view and check how it can be integrated to the existing architecture.

2.2.2.2 Uplink Scenario

In general, the performance in terms of reliability of an uplink transmission is significantly deteriorated due to channel collisions when multiple devices concurrently access a shared wireless channel, which leads to large interference for the radio links. This occurs with relatively high probability especially when the number of connected devices in the network grows sufficient large.

Two widely-considered approaches to this problem are interference avoidance where the objective is to avoid channel collisions by scheduling the devices on orthogonal radio resources (e.g., in frequency or time domain) and interference cancellation in which case the interference is cancelled out at the receiver-side. However, these approaches either do not scale properly with the network size or are notoriously difficult to implement in practical systems. As a result, a paradigm shift in the RAN operation is essential to enable more efficient and reliable transmissions.

In this section, instead of making efforts to avoid channel collisions, we propose a potential solution by doing exactly the opposite; that is, harnessing channel collisions. The basic idea is to *exploit channel collisions at multiple DUs to reliably decode linear combinations of the transmitted messages*. The linear equations are then forwarded to a CU that solves a system of linear equations to reconstruct the original messages. The proposed concept is described in more detail in the following sections.

In this work, we consider a two-hop wireless network, which consists of the following main network elements:

- Multiple devices deployed over some geographical area.
- One CU.
- A dense network of DUs that have high-capacity links to the CU.

Furthermore, we make the following assumptions:

- Communication and coordination between the DUs are not provided.
- The devices have no channel state information (CSI).
- The channels between devices and DUs are frequency flat and constant during the transmission period, and the channel coefficients are known to the DUs a priori.
- The average transmit power (per node) is constrained to some real-valued P>0.
- Each device transmits at equal rates R.
- The links between the DUs and CU are of high capacity C (i.e., C>>R)

Network Coding for Uplink

As a network coding approach for uplink, we consider compute-and-forward strategy [NG11]. The key idea is to *exploit the additive nature of the wireless channel at DUs to transform channel collisions into decodable linear equations*. In fact, this can be seen as an instance of the idea of network coding on the physical layer, and hence it has become known as physical-layer network coding, and referred to as compute-and-forward. The proposed approach mainly consists of the following four steps:

1. By using a physical-layer network coding strategy, the devices in the network transmit concurrently on the same wireless resources.
2. Instead of individual messages, DUs decode linear equations of the messages from (i).
3. Reliably decided linear equations are forwarded to CU, together with the equation coefficients.
4. The CU collects equations from DUs until it is able to solve for the original messages.

Therefore, in contrast to a conventional approach such as interference avoidance and interference cancellation, compute-and-forward do not avoid the interference caused by concurrent transmissions but converts the channel output into a set of reliable linear equations.

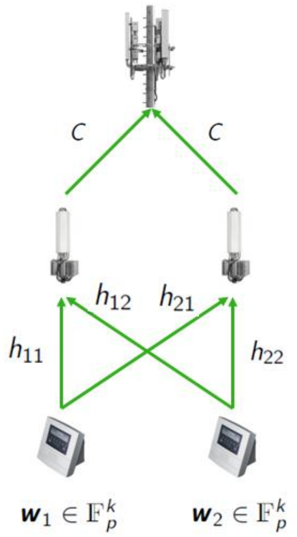

C
C
h_{12}
h_{21}
h_{11}
h_{22}
$\mathbf{w}_1 \in \mathbb{F}_p^k$
$\mathbf{w}_2 \in \mathbb{F}_p^k$

Figure 2-7: Uplink Compute-and-Forward toy example with 2 devices, 2 DUs and 1 CU

For the sake of simplicity, we use an instructive toy example as shown in Figure 2-7 to illustrate the proposed approach. Let both devices have a length-k message w_i to be transmitted, which is uniformly drawn from a finite field F_p (i.e., $\mathrm{w_i} \in \mathrm{F_p^k}$). Using the same encoding function $\Im\colon F_p^k \rightarrow C_n \subset R^n$, each device maps its message to a length-n codeword x_i subject to an average power constraint $\frac{1}{n}\left|\left|x_i\right|\right|^2 \leq P$. The corresponding message rate in bits per channel use is then

$$R = \frac{k}{n}\log_2 p. \qquad (1)$$

For every fixed block-length n, the codebook C_n is a subset of an n-dimensional lattice. Thus, each encoder simply maps its finite field messages to lattice points and transmits them over the channel.

Decoding linear equations at DUs

For ease of exposition, let the channels between the devices and the DUs be real-valued (the extension to complex valued channels is straightforward). Then, the signal observed by a DU i can be modelled as the two-user Gaussian multiple-access channel

$$y_i = h_{i1}x_1 + h_{i2}x_2 + z_i, \quad i = 1,2 \qquad (2)$$

where $h_{i1,}h_{i2} \in R$ are the channel coefficients from devices to DU i, and $z_i{\sim}N(0, I_n)$ is i.i.d Gaussian noise with zero-mean and unit variance. Now, instead of the individual messages w_1, w_2, the DU wishes to decode with high reliability a linear combination thereof. That is, for some $\beta_{i1}, \beta_{i2} \in F_p$,

$$u_i := \beta_{i1}w_1 \oplus \beta_{i2}w_2, \qquad (3)$$

where $\oplus$ denotes addition modulo p. Therefore, each DU applies a decoding function $D := R^n \rightarrow F_p^k$ that maps its channel output y_i to an estimate $\hat{u}_i = D(y_i)$ of (3). It is shown in the seminal work of Nazer and Gastpar [NG11] that the equations in (3) can be reliably decoded from the received signals (2) if n and p are sufficiently large and the message rate fulfils

$$R < \mathrm{R}^* := \min\{R\ \ (h_1, a_1), R\ \ (h_2, a_2)\} \qquad (4)$$

with

$$R(h_i, a_i) = max_{a_i}\, \frac{1}{2}\log_2\left(\left|\left|a_i\right|\right|^2 - \frac{P\left|h_i^T a_i\right|^2}{1+P\left|\left|h_i\right|\right|^2}\right)^{-1}, \qquad (5)$$

where h_i is the vector of channel coefficients from devices to DU i and a_i is a vector of equation coefficients in the way $\beta_i = [a_i]mod\ p$. Therefore, $R(h, a)$ is referred to as the computation rate, and it has been shown in [NG11] that $R(h, a)$ is achievable for any given h and a.

Based on the knowledge of h_i, the DU i is able to choose integer vector a_i, and therefore the equation coefficients $\{\beta_{i1}, \beta_{i2}\}$, in order to carry out the maximisation operation in (5). Low complexity algorithms for doing that can be found in [WC16].

Decoding Messages at the CU

Once the DUs have successfully decoded the linear equations, they forward their estimates $\hat{u}_1$ and $\hat{u}_2$ along with the respective coefficients β_1 and β_2 to the CU over backhaul links of capacity $C \gg R$. If the equation coefficients have been chosen such that the matrix $B := (\beta_1, \beta_2)$ is invertible over F_p, the CU obtains estimates of the original messages by solving the linear system

$$(\widehat{w}_1, \widehat{w}_2)^T = B^{-1}(\hat{u}_1, \hat{u}_2)^T \qquad (6)$$

The estimates are sufficiently accurate if the coding block length n is chosen sufficiently large.

Selection of the Coefficients

In order to optimise the end-to-end rate performance of the proposed strategy, each DU tries to decode the equations that maximise its computation rate in (5). Therefore, the equation coefficient $\mathrm{a_i}$ has to be appropriately adapted to the channel coefficient $\mathrm{h_i}$ where the maximisation in (5) is achieved when $\mathrm{a_i} = h_i$. However, since the equation coefficients are taken over some finite field, it is possible that the decoded equations at different DUs are linearly dependent. As a result, the CU is unable to retrieve the original messages although all equations are successfully decoded at each DU.

It can be shown that the probability of solving the system of equations over F_2 at the CU decreases significantly with higher SNR values. This is mainly due to the fact that the equations decoded by the DUs are linearly dependent with high probability, if the coefficients are selected not carefully. Hence, in order for the CU to solve the system of equations over some finite field, a smart choice that guarantees the linear independency of the equations decoded by the DUs is crucial for reliable end-to-end performance. In the following, we propose two approaches to solve this problem, namely *DU configuration* and *DU cooperation*, respectively.

a) DU Configuration:

The key idea behind this approach is to guarantee the linear independency of the equations by *allowing the CU to configure each DU* on which equation to decode a priori. The mechanism is summarised as follows:

- First, the DUs collect CSI and forward it to the CU. CSI may be obtained via long term observation or based on geographical location information.
- Based on the forwarded CSI the CU determines a set of equation coefficients for each DU. The equation coefficients are selected such that the equations, decoded by different DUs, are linearly independent and the end-to-end rate is maximised.
- Then the CU configures the DUs with a set of determined equation coefficients.

It is emphasised that, in this case, no coordination between either the UEs or DUs is needed. Hence, the control and maintenance overhead is significantly reduced. However, the decision of the set of equation coefficients by the CU is highly dependent on the accuracy of the reported CSI by the DUs. Therefore, it is not possible to guarantee that the pre-configured equations at each DU can be decoded if the CSI is not accurate.

b) DU Cooperation:

Another possibility to obtain linear independent equations at the CU is to enable cooperation between the DUs. Contrarily to the DU configuration case where the CU configures the DUs, in this scenario the *DUs cooperate* towards the goal that all original messages can be reliably reconstructed at the CU, given that each DU is able to decode its equation successfully and the end-to-end rate is maximised. The cooperation procedure is elaborated as follows:

- DUs decode linear equations sequentially following a pre-defined order.
- The coefficients of already decoded equations by other DUs are communicated via the network.
- Based on the knowledge of the received equation coefficients, the DU selects an equation coefficient that is linear independent from the set and at the same time maximising the computation rate.
- The CU decodes the original messages until collecting enough equations from DUs.

The information of the equation coefficients of already decoded equations can be exchanged between the DUs either via the help of CU or via inter-DU cooperation. In this case, each DU is guaranteed to decode its own combination of message, and the maximised end-to-end rate can be achieved in the meantime. However, it also comes at the cost of excessive coordination between the DUs, thus resulting in large communication overhead.

Rate Performance

As it is assumed that the links between the DUs and the CU are of high capacity, the performance is mainly determined by the rate achievable on the first hop. As mentioned previously, it is assumed that each DU is able to reliably decode its linear equation as long as the message rate commonly used by the devices fulfils (see Eq. (4)). As a result, the individual messages can be reliably reconstructed at the CU if the message rate satisfies

$$R_{CF} < \mathrm{C}_{\mathrm{CF}} \coloneqq \frac{1}{2} R^*, \tag{7}$$

where R* is defined on the right-hand side of (4). In the following, we compare the rate performance of the proposed strategy with decode-and-forward, where each DU tries to reliably decode the individual messages w_1 and w_2 and forwards them to the CU. In this case, the achievable rate is strictly limited to the intersection of the capacity regions of the two multiple access channels (MAC) on the first hop. It is widely known that the capacity region of a Gaussian MAC is given by

$$R_{MAC}(h_i) \leq \log(1 + \frac{P}{N}\left||h_i|\right|^2). \tag{8}$$

Therefore, the best rate that decode-and-forward can achieve is obtained by

$$R_{DF} < C_{DF} := \frac{1}{2}\min\{R_{MAC}(h_1), R_{MAC}(h_2)\}. \tag{9}$$

In Figure 2-8 we compare the rate performance of compute-and-forward as in (7) and that of decode-and-forward as in (9) by averaging over Gaussian distribution of the channel coefficients. It is shown that the proposed strategy has the potential to achieve much higher rates than classical approaches where each individual message is decoded at DUs. The main reason for this improvement is that the interference between the nodes is not avoided, but used to improve the performance. Compared to the classical approach with decode-and-forward strategy, each user practically uses twice as much resources, although the total amount of used resources remains the same. In fact, this represents an alternative explanation for the increase in the rates.

We emphasise that, besides improving the rate performance, this increase can help to support more reliable RAN operations for the 5G network. That is, less physical resources are required to transmit the same amount of information with the same reliability, the remaining resources can be used to further improve the reliability or the transmission rate.

As a side note, the presented scheme is studied in different works (e.g. in [NG11]) from an academic point of view, and its benefits are shown for different scenarios. Nevertheless, an investigation from an integration point-of-view is missing. In the framework of 5G-MoNArch, our aim is to investigate the suitability of such a technique to the considered architectures, highlight its advantages and challenges in regarding the considered architecture and propose (if possible) modifications.

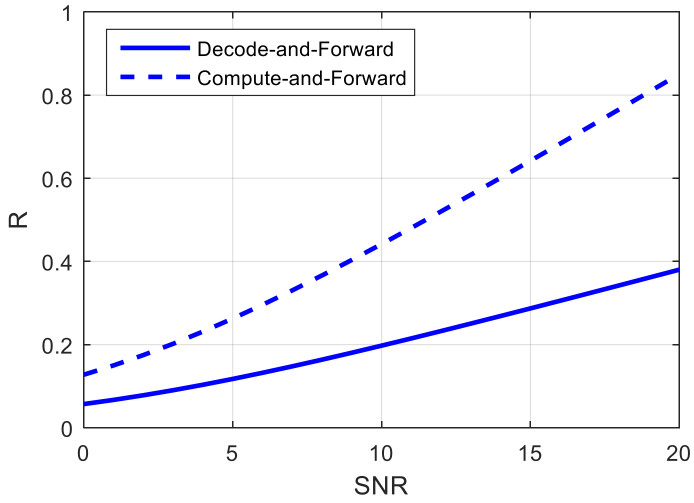


Figure 2-8: Rate performance of Decode-and-Forward and Compute-and-Forward

Both methods mentioned in the previous subsections for downlink and uplink promise an increased reliability in the RAN operations, by making use of the network concepts. However, as network coding is not included in any of the legacy wireless communication standards, *certain protocol stack modifications are required* in order to benefit from the gains of network coding. In the following, we briefly summarise the required modifications for the downlink and uplink scenario.

- ***Downlink scenario:*** The downlink network coding approach requires the UEs to be grouped such that each group (for instance, a group with two UEs) can be served by network coded packets. Moreover, the transmitted packets need to be identified (by using packet IDs) and each UE needs to store a certain number of packets for decoding the network coded packets. The UEs also need to send an ACK/NACK feedback according to the reception of the packets. In this approach, the lower layer operations such as detection and channel decoding are unchanged.

- ***Uplink Scenario:*** In this approach, the channel access for both UEs needs to be coordinated carefully, such that they use the same time/frequency slots simultaneously. Moreover, the channel coefficients and the modulation scheme need to be communicated to CU, such that the CU can obtain the required coefficients and inform DU about these coefficients. It is further noted that, depending on the parameters, the DUs may need to modify some of their PHY-layer operations, such as de-mapping.

Although these modifications can be included in the current architectures, they may make the overall implementation more challenging, particularly if cross layer operations are required. Especially the presented uplink scenario requires a certain level of cross-layer operations that require more detailed feasibility studies. On the other hand, the downlink approach does not involve any cross-layer operations, and the required modifications are more straightforward to be included in the considered architecture of 5G-MoNArch.

In light of the above, we leave a more detailed study of the uplink approach as a future work, and rather *focus on the downlink approach*. To this end, further protocol aspects are planned to be investigated, such as the layer at which the network coding operations should be performed.

2.3 Architectural Considerations

A network slice that is supposed to support services with strict requirements on reliability and latency needs specialised network functions. With respect to the RAN domain, this can be, as discussed above, network functions for data duplication or network coding. In view of the above, in this section we consider the *architectural issues related to the implementation of the RAN reliability* techniques, as described above. In particular, we highlight the role of the developed RAN reliability modules in the 5G-MoNArch architecture.

The role of the RAN reliability-related developed modules is illustrated in Figure 2-9. It is emphasised that an aggregated view of the modules developed in WP3, including not only the modules related to RAN reliability (as shown in Figure 2-9), but also the modules related to telco cloud resilience and security as well, is given in Figure 6-1, provided at the last chapter of this document.

For a better understanding of Figure 2-9, we first give a brief description of the 5G-MoNArch architecture, which is described in detail in [5GM-D2.1] and [5GM-D2.2]. The 5G-MoNArch architecture consists of the following layers:

- The Service layer hosts business applications and services as well as Business Support Systems (BSSs).
- The MANO layer contains a set of functionalities required for the operation of virtual network functions (VNFs).
- The controller layer consists of two software defined controllers which are amongst others responsible for fulfilling QoE/QoS constraints.
- The network layer accommodates VNFs and physical network functions (PNFs) required to transport and process user data.

The *new modules pertaining to higher RAN reliability*, introduced by WP3 into the 5G-MoNArch architecture, *are shown in Figure 2-9 as separate boxes* within dashed black frames. Their functionality is summarised as follows.

WP3 extends the 5G-MoNArch architecture by a reliability sub-plane in the network layer and a reliability control application in the controller layer. The role of these functionalities in the corresponding layers is as follows.

- *Network layer*: Throughout the instantiation of a network slice with URLLC services, the end-to-end (E2E) service MANO functionality implements and activates a specialised set of functions. In this regard, a RAN reliability function in the network layer is instantiated. Such a

RAN reliability function can be either a *data duplication* or a *network coding* function, which serves as a user plane functionality that processes data according to the principles described in Sections 2.1 and 2.2. In Figure 2-9, this function is depicted in the network layer, in the "*reliability sub-plane*" frame.

- *Controller layer*: The controller layer functionality introduced in [5GM-D2.1] involves the main controllers, namely the intra-slice controller (ISC) and the inter-slice controller (XSC). Those controllers are enhanced with a control application for RAN reliability, labelled "*reliability control*". This control application ensures that URLLC traffic undergoes the required processing by the corresponding network layer functions. In addition, this application is responsible for guaranteeing the envisioned reliability of the URLLC data streams by controlling the corresponding network layer function. In case it is impossible to achieve the required QoS, a re-orchestration is requested by the control application. Such re-orchestration in the controller layer takes place in the intra-slice level, thereby affecting the ISC controller.

Finally, it is noted that the network layer functionalities are integrated into the service chain of ultra-reliable services via the network function virtualisation orchestrator (NFVO). Further details on the role of the architecture modules in Figure 2-9 will be soon available in [5GM-D2.2].

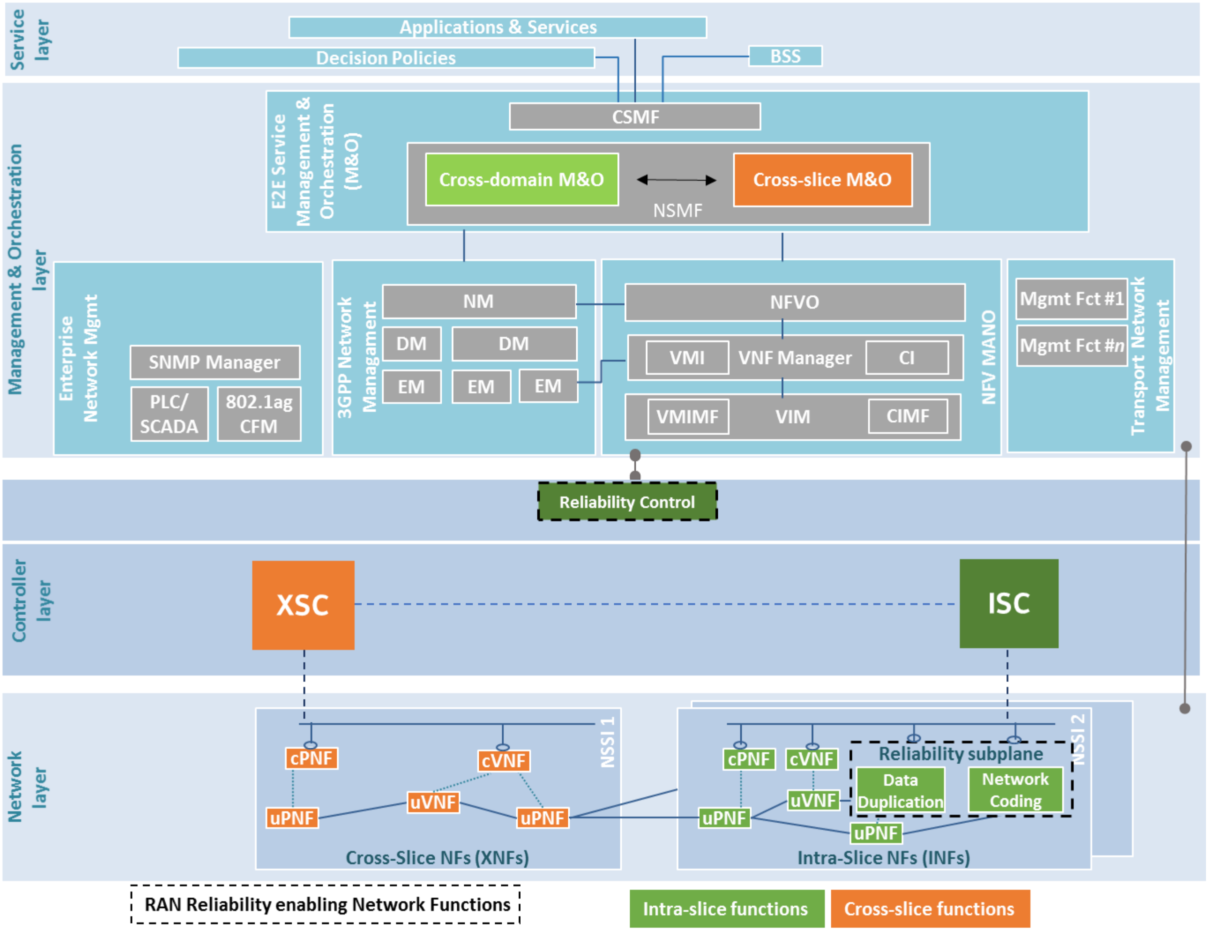


Figure 2-9: The 5G-MoNArch Architecture, enhanced by network functions for ultra-reliable services (Reliability Control and Reliability sub plane)

3 Resilience in the Telco Cloud

The deployment of 5G network can include the network functions running on virtualised infrastructure (e.g. centralised or core network functions) as well as on the specialised physical hardware instances (e.g. parts of RAN). In fact, RAN reliability and resilience in the telco cloud are two study areas which are equally important for achieving a sufficient level of E2E resilience. It is worth noting that RAN reliability and resilience in the telco cloud represent a common feature in the overall architecture design, i.e., that of E2E resilience. Nonetheless, their specific implementation is independent, since their design is driven by different resilience issues as well as different mitigation mechanisms.

The network functions that run on virtual infrastructure is denoted within 5G-MoNArch project as *telco cloud*. The 5G URLLC services are challenging the entire network setup and operation due to their high reliability requirements. As mentioned above, such strict reliability requirements can even reach 99,999% probability of uninterrupted operation [5GM-D6.1]. As a telco cloud represents a part of the E2E network architecture of 5G URLLC services, it also needs to comply with its strict reliability and resilience requirements. However, the network cloudification, that brought significant changes in the network setup and operation poses new challenges and requires new measures in achieving the so-called "five nines" reliability.

In the context of telco cloud resilience 5G-MoNArch considers different approaches in order to fulfil required resilience of a telco cloud such as redundant hardware and network function setup, improved fault management, in-built resilience of network functions etc. Chapter 3, along with Sections 3.1. 3.2, 3.3 and 3.4, aims at describing in more detail the different telco cloud resilience approaches and their realisation options in the 5G-MoNArch architecture.

3.1 Redundancy for Higher Resilience

The usage of cloud technology presents many advantages: The minimisation of infrastructure resources costs and the elasticity property, which allows services to be scaled up or down according to the current demand. To meet all the requirements defined at service level agreements (SLA), there are many challenges to be overcome. In this regard, high availability is the biggest challenge. A set of techniques have been designed to implement such high availability, such as the checkpointing (copying the state of a system), load balancing, and redundancy.

Redundancy can offer different levels of availability depending on the redundancy model and the redundancy strategy (active, passive). The redundancy model can combine active and standby replicas of hosted VNFs. Four models have been proposed in the literature: 2N, N+M, Nway, and Nway active [AMF16]. N represents the number of instance able to handle active assignments. M represents those with standby assignments. Due to its simplicity, the 2N model is preferred in terms of implementation.

The redundancy strategy is divided in two classes: active and passive redundancy [AD13]. In active strategy, there are no standby replicas and all the replicas work in parallel. When one node fails, tasks executing at the failed node can be resumed in any remaining node. This is similar to the behaviour of Hot Standby Router Protocol (HSRP) introduced in RFC 2281 [IETF-RFC2881]. In passive redundancy, there is one working replica whereas remaining replicas are standby.

Although redundancy enables high availability, it imposes the higher cost for realisation of the network service. In order to keep such cost at acceptable level the addition of redundant instances needs to be driven by the actual resilience level that needs to be achieved keeping in mind the cost limitations with regard to network service implementation. In other words, a careful *analysis of a trade-off between gained resilience by increasing the redundancy level and imposed costs due inclusion of additional resources needs to be performed*. Furthermore, applying redundancy might impose additional complexity of managing the redundant instances and updating their states in order to keep them prompt for taking over the functionality of faulty instances.

Redundancy can be used in many ways for improving the overall resilience of the network service, e.g. it provides more options for network function healing done by fault management or reaction options on security threats. Furthermore, as described in Section 3.4 the redundancy can be used as a tool for autonomous failsafe operations, e.g., by replicating the network functions from the central cloud to the edge cloud, thus providing the minimal network operation in the edge cloud in the case of a problems in the central cloud etc.

In general, one may conclude that redundancy needs to be supported by proper, use-case driven deployment decisions and management procedures in order to bring the cost- and complexity-conscious benefits to resilience.

3.2 Fault Management Approaches

The main goal of network fault management is to enable the resilience to network failures by monitoring the network state and provide solution to the problems that cause the network performance degradation or failure. As a first step, based on input from monitoring tools the detection of changes, potential problems and anomalies in network behaviour needs to be done. Furthermore, the actual cause of the problem needs to be determined in order to perform the suitable recovery actions. The root-cause analysis enables the localisation of the actual problem and consequently its isolation such that the propagation of the fault effects and impact to the rest of the network can be minimised [HSS12].

Figure 3-1 illustrates the main processes and actions involved in fault management. Such fault management techniques need to be adapted and extended towards the 5G network slicing context. The fault management characteristics and parameters need to be adjusted to the actual service that is supported. This might include e.g., the service-aware design of triggers and thresholds for alarms creation, start of recovery actions, etc.

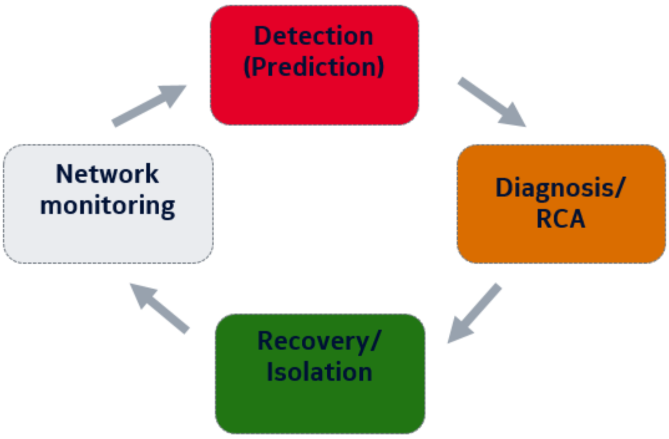


Figure 3-1: Fault management: Processes and actions involved

The virtualisation of traditional network elements broke up the tight coupling between hardware and software and introduced additional complexity in handling the faults of network functions. In virtualised networks three layers of deployment can be identified: network function/service logic, virtual infrastructure (e.g. virtual machines, containers, etc) and physical infrastructure (e.g., commercial off-the-shelf (COTS) servers, compute and storage components) as illustrated in Figure 3-2. In such an environment there might be different implementation and deployment options for network functions, i.e. there can be many to many relationships between layers of network functions logic, virtual infrastructure and physical infrastructure where the network function resides.

Such layered implementation of network function requires enhanced fault management logic which considers the actual deployment and interrelations between the layers. In general, the network fault should be handled at the layer where it occurs, ideally discovered before the major effects take place and/or propagate among different layers. As the faults can be related to different layers of network function deployment the correlation between fault occurring at different layers is essential for *root cause analysis in virtualised networks*. Furthermore, the correlation between the resource failures and the impact on the service performance and ultimately on the user satisfaction can create a baseline for better resource provisioning, prioritisation and maintenance. However, the correlation is complex task due to many-to-many relations between infrastructure and network functions, service providers, deployments in multi-site and multi-domain data centres, etc.

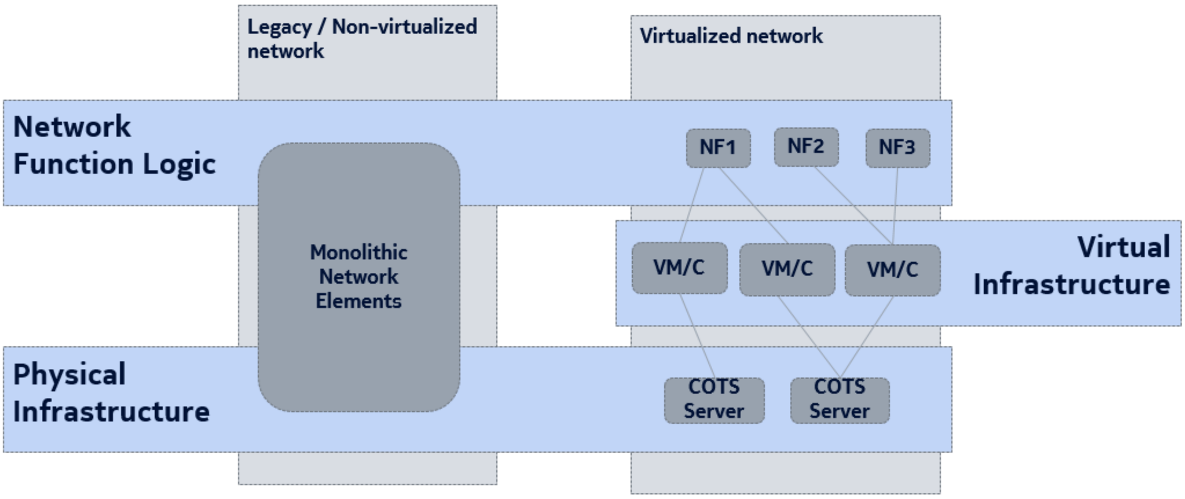


Figure 3-2: Fault management deployment in virtualised networks

Despite the fact that fault management might be more complex in virtualised networks, the virtualisation can be seen as a facilitator for network resilience through much easier and cost-effective redundancy implementation. As the network functions can be implemented on the commodity hardware, the network functions can be more easily multiplied and moved across the network. Furthermore, adding redundancy in virtualised environment is more cost-effective as the infrastructure resources of redundant network functions can be more easily re-used. Adding redundancy is especially important for critical network functions or network functions with higher importance/priority. For example, the SDN controllers which have central role in network control might be designed with more redundancy than other network functions, as the outage in network controller might have severe impact on overall network operation. Nevertheless, careful consideration on trade-offs in applying redundancy, e.g. in terms of overprovisioning and resource reservation, needs to be done in order to design efficient and resilient network.

3.2.1 Cognitive Network Management and Fault Management Cognitive Functions (FM CFs)

The network can encounter faults originated in different parts or deployment layers of the architecture which need to be handled by the fault management. Furthermore, the network slicing imposes a clear need to adapt the network operation and management to the slice requirements and determined Service Level Agreements (SLAs) with the tenant. The slice-tailored network management framework implies slice-specific fault management and consequently slice-specific self-healing. The main goal of the slice-specific fault management is to provide the resilience to the network faults according to the actual slice requirements as requested by the tenant, avoiding the overprovisioning and thus cost increase while fulfilling the targeted quality of service. But due to foreseen dynamicity of network slices the according network and fault management procedures need to be automated as much as possible.

The first step in terms of automating the OAM of mobile networks, and in particular, mobile radio networks, was the introduction of Self Organising Network (SON) solution, which allows in particular the *autonomic optimisation of certain network configuration parameters based on measurements from the individual network elements*. To enable a joint operation and configuration of a different SON function instances at the same time, concepts for SON coordination and SON management have been introduced. However, the SON concept has certain issues that appear as suboptimal in dynamic context of 5G network slicing, namely a rather static nature of the logic of deployed SON function seems as unsuitable. While SON management allows a modification of some parameters of a SON function such that the behaviour of the SON algorithm can be slightly modified (and thereby its effects on the network configuration), the SON algorithm as such (including the algorithm inherent state machine and state transitions) remains unchanged. More sophisticated adaptations of the SON algorithms therefore need to be done manually through the SON manufacturer. Such manual intervention might not be acceptable for highly dynamic nature of 5G networks, thus the new solution that enable more automation in SON adaptation need to be developed.

The aim of Cognitive Network Management (CNM) [MDM2016] is to make the *automation of OAM processes in mobile networks more flexible and adaptable to current network context*. The main idea of CNM is to better extract the characteristics of the network environment so that it can decide on the most suitable configurations of network functions having in addition the information about current network states. The CNM introduces the so-called cognitive functions (CF), which represent more intelligent SON functions that learn from historical data on network operation in different contexts. Furthermore, the CFs can be designed in a way that their logic can be adapted automatically by extending the knowledge space for a certain network context. Such a knowledge extension can be gained by applying the network setups that were not used before and learn from the corresponding network performance in a given context. Thus, the CFs of the CNM go beyond traditional SON solutions where each SON function merely matches combinations of KPIs to pre-configured network constellations.

Within the 5G-MoNArch project, we focus on the design of CFs that implement the Fault management (including self-healing) operations for slicing enabled 5G networks. We further refer to such functions as Fault Management Cognitive Functions (FM CFs). Depending on the slice/tenant requirements and priorities, criticality of individual network functions etc., the FM CFs need to be adapted. Furthermore, the interaction between different FM CFs at different deployment layers, subnets and network slices need to be carefully designed. In order to meet the stringent latency requirements, where applicable the troubleshooting should be done more locally/distributed, avoiding the case of transmission of all relevant data to hierarchically higher management entities and performing centralised data processing and troubleshooting.

3.2.2 Network Slice Fault Management in 3GPP

Fault Management of a network slice is addressed in 3GPP TR 28.801 [3GPP2018] Release 15 on the NSMF level i.e. on the network management level of an entire slice instance and on the NSSMF level, i.e. on the network management level of slice subnet instance. The FM information is collected from all network slice subnet instances NSSIs which take part in network slice instance NSI. For NSSIs shared across different NSIs the NSMF will allow or suppress the FM information based on the FM requirements of the specific slice managed by NSMF. Similarly, the NSSMF will receive the alarm notifications from the subnets it manages as well as its constituents. In the case of shared constituents, the alarm notifications might be applicable only to a specific subnet.

Moreover, 3GPP TR 28.801 assumes that automated healing can happen either on NSI or NSSI level and it is driven by the pre-configured self-healing policies as illustrated in Figure 3-3. Further adaptation of self-healing process in terms of algorithms used and operational granularity is not considered.

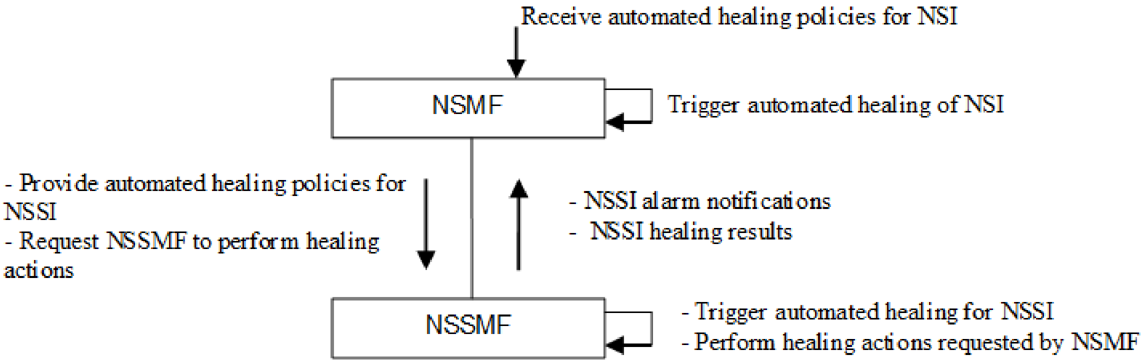

Receive automated healing policies for NSI
NSMF
Trigger automated healing of NSI
- Provide automated healing policies for NSSI
- Request NSSMF to perform healing actions
- NSSI alarm notifications
- NSSI healing results
NSSMF
- Trigger automated healing for NSSI
- Perform healing actions requested by NSMF

Figure 3-3: Controlling of automated healing on an NSI [3GPP2018]

3.2.3 FM CF design in 5G-MoNArch

In order to enable adaptable, slice-aware Fault Management, the 5G-MoNArch opts at designing the Fault Management Cognitive Functions (FM CFs) that can be mapped to the different network entities, functions or parts of the infrastructure. For instance, a single FM CF can be responsible for the fault management of entire NSI or building blocks of the network slice, that is network slice subnet instances NSSIs, network function chains, individual network functions as well as individual deployment layers of network functions.

The exact scope of operation as well as the logic/algorithm of FM CF are determined by the *slice characteristics/requirements along with SLAs agreed with the tenant*, as well as the requirements of individual network functions. The parameters that drive the design of FM CF are:

- Slice requirements, particularly in terms of resilience
- Existing service level agreements with the tenant
- Type/criticality of the network function (e.g. user or control plane) along with its resilience requirements (e.g. in terms of required level of redundancy or required time for restoration in the case of fault)
- Affinity among network functions e.g. if NFs are usually appearing together in the network function chain or if the output of one function is the direct input of the other function
- Deployment characteristics of the network functions in terms of the mapping between physical, virtual and functional deployment layers

As an example, in the case of URLLC slice which has critical control function as a constituent, the most suitable option for FM CF design would be to dedicate one FM CF only to that single control NF. Furthermore, the FM CF algorithm needs to be designed in a way to be extremely reactive to any anomaly in the NF operation. Optionally, in order to minimise the effect of other NFs to that critical NF, the critical NF might be implemented on a single VM, container or physical server (so that fault localisation and isolation/self-healing might be facilitated more swiftly).

On the other hand, in eMBB slice which has more relaxed resilience requirements, one FM CF might be responsible even for entire slice or subnet(s). The constituents of such slices/subnets can span across multiple VMs/containers and physical servers with many inter-dependencies among infrastructure layers where fault localisation and isolation might be more complicated and thus take longer time. Furthermore, the algorithm of such FM CFs can be adapted to more relaxed resiliency and fault recovery requirements, hence the FM CF might be designed to react only on a specific set or number of alarms/events. For instance, they can react only to those events that can seriously endanger the network operation and thus the E2E service fulfilment.

As a network slice (and consequently subnet and NF) realisation in virtualised environment can include the existence of different deployment layer (that is physical, virtual, functional layers), the *responsible FM CF needs to perform a consolidated fault management* on all layers by considering the inputs from all the layers jointly. Alternatively, the dedicated layer-specific FM CFs can be mapped to different layers. Furthermore, based on handled FM events and ability to localise, isolate and resolve the occurred issue the FM CF can give a feedback to the NSMF (in particular to the orchestration process) in order to improve the NF placement/deployment. For example, in case of very critical control network function which constituents were initially deployed across multiple physical servers, yielding to difficulties in fault detection and isolation, the FM CF of that network function might recommend to the NSMF/Orchestrator to deploy the NF differently, e.g. on the same physical server. Similarly, in the case that very critical control function of URLLC slice shares the infrastructure, e.g. the physical server with another NF that can be prone to failure or security threads, the critical NF can be affected by the problems caused by other NF. The corresponding FM CF (e.g. responsible for both NFs) will provide feedback to the NSMF that such deployments should be avoided.

3.2.4 FM CF deployment options

The network slice and its constituents can have different realisations in virtualised environment, and the FM CFs can have different mappings to network slices, network functions and deployment layers. Figure 3-4 illustrates such different options for FM CF deployment. Figure 3-4a) shows the identified deployment layers (physical, virtual, functional) of a network slice (NS) (and constituent NFs), as well as one possible deployment option for FM CF (i.e. at functional deployment layer). Figure 3-4b) shows one example implementation of FM CFs responsible for managing three network functions as well as corresponding mapping to the virtual and physical infrastructure on which NFs are deployed. In this example, it is noticeable that *multiple overlaps* in the layers mapping a single FM CF exist (for instance both NF2 and NF3 components are implemented on the same physical server). As a result, in order to perform the fault localisation and isolation, one needs to exclude the input that might be related to the NF3 which is not under the responsibility of a given FM CF (for instance, FM CF 1 in this example). It is noted that this comes in addition to correlating the performance indicators and alarms related to different layers of the infrastructure.

Alternatively, the FM CFs can be deployed as *layer-specific FM CFs* as shown in Figure 3-4c). In such case their scope of operation is limited to a specific deployment layer (e.g. physical or virtual). Such layer-specific FM CFs can to a certain extent autonomously act within a dedicated layer. As an example, in the case of physical Network Interface Card (NIC) outage the physical layer FM CF can automatically trigger the switching of traffic to another NIC. However, the layer-specific FM CFs need to exchange the info between each other and/or with the FM CF of the network function to which the layer-specific FM CFs are related to.

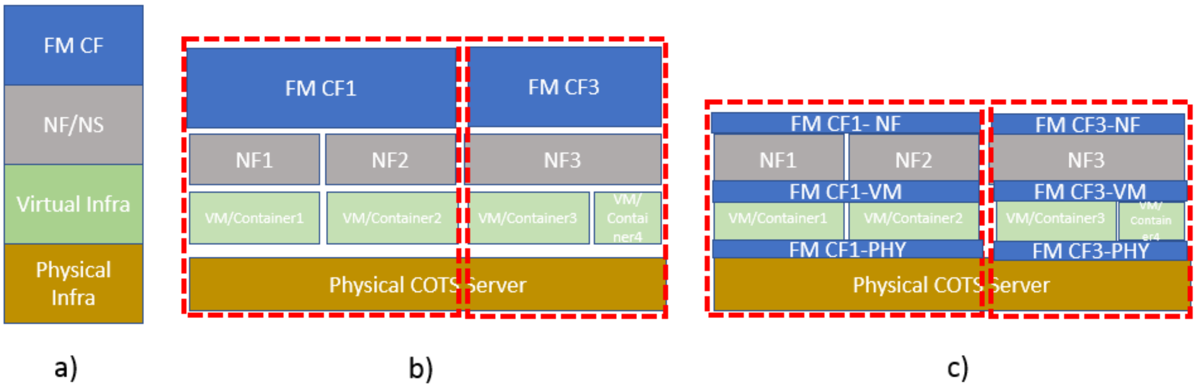


Figure 3-4: a) Deployment layers of network function/slice implementation in virtualised environment and one example deployment of FM CF at functional deployment layer. b) possible mapping of FM CFs to corresponding network function and deployment layers, c) an ex example of layer-specific FM CFs deployed on dedicated deployment layers

Figure 3-5 illustrates the case where FM CFs can implement the fault management algorithms tailored to the NFs (a single NF or a group of NFs with a similar resiliency requirements) they are managing where self-healing can be done on virtual infrastructure (e.g. VM) level. Such self-healing approach is possible as VMs/containers on which one NF is deployed can be easily isolated without affecting other NFs. For instance, if the fault happens on the virtualisation deployment level of NF1 (i.e., VM/Container1 or VM/Container2) the self-healing will be done on this level without affecting the NF2 and NF3.

Similarly, Figure 3-6 illustrates the case where the FM CFs can implement NF-tailored FM algorithms and isolation can be done on physical infrastructure level as well, e.g. self-healing or isolation of NF2 either on virtual of physical level will not affect NF1 and NF3.

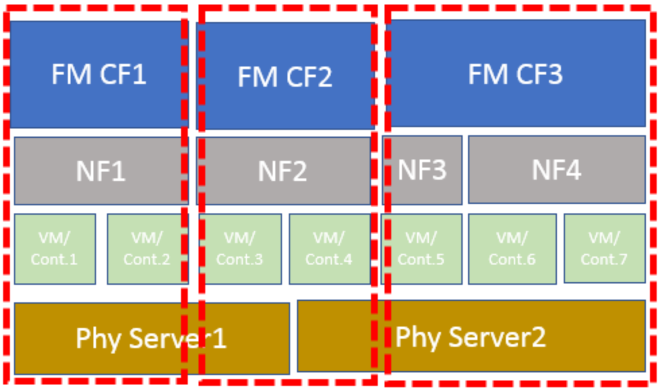


Figure 3-5: Mapping of FM CFs to NFs and infrastructure components and deployment layers where more isolation is possible on functional and virtual infrastructure layer

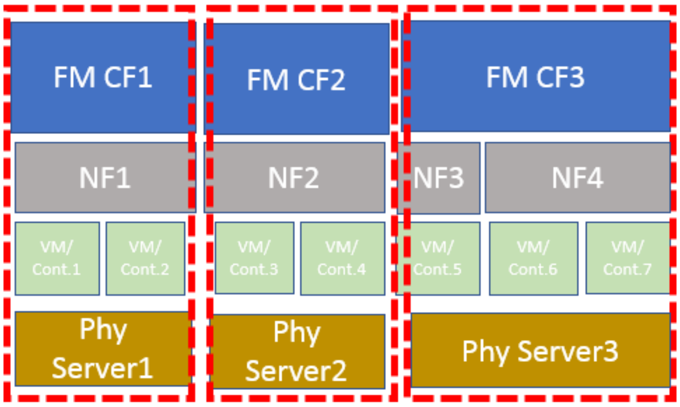


Figure 3-6: Mapping of FM CFs to NFs and infrastructure components and deployment layers where fault isolation is possible on functional, virtual and physical infrastructure layer

The different implementation approaches can be used for realisation of FM CFs namely, a distributed, centralised or hybrid implementations can be suitable. However, certain advantages and disadvantages are bound to every implementation option, especially in terms of ability to detect and isolate faults locally using NF-specialised algorithms and the need to coordinate the operation with other instances of FM CFs.

For example, the advantage of distributed (NF-specific) implementation of FM CFs is more efficient handling of FM events both in terms of applying the algorithms suitable for specific type of network function (or a group of NFs with same/similar resiliency requirements) and its implementation/deployment, as well as in terms or more local processing of fault event notifications and thus faster reaction to faulty events. However, the distributed, highly NF-specific FM CFs cannot work independently from each other. This is due to many dependencies among network functions and overlapping deployments in the underlying infrastructure, which might lead into contradicting decisions at the FM CFs. Therefore, a highly distributed implementation of FM CFs would require either very close interaction between FM CFs or the existence of coordination entity for consolidation of FM CFs operation and decisions. This means that the corresponding interfaces between FM CFs and coordination point as well as among FM CFs need to be defined in order to allow for exchange of data.

By a more centralised implementation of FM CF concept (e.g. merging the different FM CFs into a single one, or assigning a single FM CF to entire network slice), such interaction becomes less critical as decisions are taken centrally at one single entity, yet the responsiveness of such implementation might become lower. Furthermore, implementing the NF-specific healing solutions might be more challenging.

Due to trade-offs between distributed and centralised approaches neither a fully distributed nor a fully centralised implementation for FM CFs are optimal. The actual level of centralisation/distribution of FM CFs depends on the use case, e.g. slice characteristics in terms of required reliability, responsiveness to faults etc.

3.3 Resilience and Scalability of the Controller

Many telco operators are deploying SDN in tandem with NFV. The SDN architectural framework abstracts the data plane switching infrastructure allowing the decoupling of the control plane and data plane functionality. This enables the control plane to be implemented as a centralised resource. In practice, a centralised controller layer running on COTS servers can control multiple data planes, providing both a global network view and automation. In addition to this, an SDN architecture enables applications to request and manipulate services provided by the network and detect the network state, enabling thus agility, efficiency, rapid response and innovation. In a nutshell the approach of SDN can brings re-programmability to the network infrastructure. It is worth noting that, in a SDN implementation, *the centralised controller system proactively monitors and controls the whole network* using predefined policies. If it detects any abnormalities, such as network congestion, it can take corrective steps and reduce the impact on customers by defining and executing policies from M&O layer.

In telco cloud, the VNFs corresponding to RAN and core of particular network slice can be deployed across distributed cloud segments such as front-end unit, edge and central cloud located in physically separated locations. Moreover, each slice can have different QoS requirements: For example, the URLLC slice requires low latency and high throughput throughout its life cycle management starting from deployment to resource allocation. In such scenarios, the controller framework needs to have different level of performance and behaviour corresponding to different deployment scenarios and use cases. The current implementation of both ONOS and ODL [ONOS][ODL] has its drawbacks by not considering the load in the selection of the master controller instance for set of devices along with higher data synchronisation time i.e., in milliseconds.

Considering the stateless and distributed master/slaves controller design is one of the approach to solve the problem of scalability and resiliency. In such design, the number controller node instances can be automatically added up with increase in load with distributed decision management residing in both master and slaves. Scalable control plane design can further improve resiliency in telco cloud control. Adding also load aware and load predictive features in the master selection in the control selection framework can further improve and/or satisfy the latency requirement of the VNF chain correspond to different use cases.

In summary, as shown in Figure 3-7 the successful realisation of SDN for telco cloud requires a controller framework that is able to provide scalability and resilience, while satisfying the stringent performance requirement of each use cases. Such framework needs to be load aware in selecting master controller instance for each set of devices.

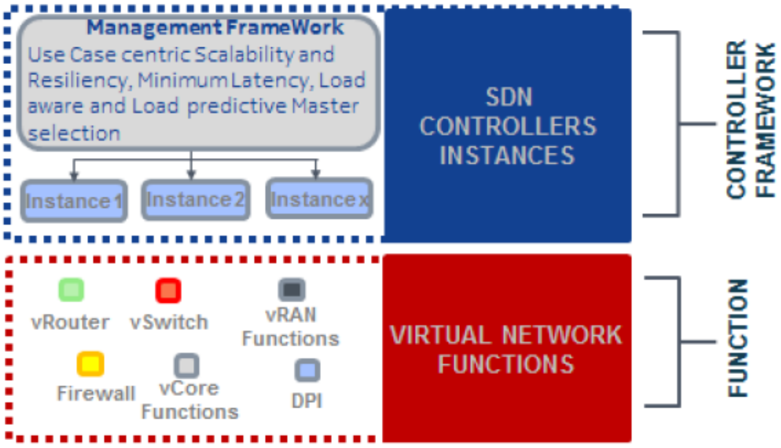

Management FrameWork
Use Case centric Scalability and Resiliency, Minimum Latency, Load aware and Load predictive Master selection
Instance 1
Instance 2
Instance x
SDN CONTROLLERS INSTANCES
CONTROLLER FRAMEWORK
vRouter
vSwitch
vRAN Functions
Firewall
vCore Functions
DPI
VIRTUAL NETWORK FUNCTIONS
FUNCTION

Figure 3-7: Load-aware scalable and resilient controller framework

3.4 Autonomous Failsafe Operation through Context-Aware NFV with Redundancy

In the context of cloudified networks resilience, one concern of 5G-MoNArch is that, during the operation, one or more virtualised network functions in central cloud become unreliable, i.e. unable to provide the desired KPI. We refer to such a case as a VNF outage, where the edge clouds should be able to provide a temporary backup solution autonomously, in order to maintain a minimal network availability. Classified with respect to the error location, such unavailability of telco cloud can be caused by two main reasons, namely:

- a disconnection between the edge and central clouds (backhaul outage);
- failure or an error at the central cloud server.

Alternatively, according to the error source, VNF outages can be triggered by:

- malicious attacks, such as distributed denial of service (DDoS) attacks;
- planned events, e.g. a ship departing from the port to an ocean voyage;
- unintentional disasters, such as intensive core network congestions or infrastructure damages.

To address this issue, we propose the use of a "*5G Island*", which enables autonomous failsafe operations in the above scenarios. The main concept is to create distributed redundant VNF deployment in edge clouds, instead of relying on the centralised servers in the core network (central cloud), so that the edge clouds can provide local or regional services in the aforementioned emergency cases to enable necessary failsafe operations.

As the cost of infrastructure, operation and privacy protection can be quite high for a fully distributed redundancy of all VNFs, our approach proposes to provide only a minimised VNF set in every edge cloud, which requires only limited static cost to remain always standby. As extension, advanced network functions, especially the VNFs customised for tenant slices, can be dynamically implemented in different edge clouds upon the online-estimated real-time service requirements.

This dynamic implementation involves preparations of VNF data, subscriber profiles, agreements with service providers, temporary authentication and authorisation (as the authentication, authorisation and accounting (AAA) server in central cloud can also become unavailable in these use cases, temporary local AAA processes can be required to enable network services). This preparation phase consumes time in addition to data traffic, and therefore shall be executed in an intelligent and predictive way, as described below.

According to UE context information (e.g. position and mobility) and edge cloud context information (e.g. geolocation, user mobility statistics, backhaul status), the network is expected to estimate the outage probability of each central cloud VNF in a specific edge cloud for a specific user through real-time network status monitoring and prediction, as illustrated in Figure 3-8.

First, given a certain edge cloud, for every user *u*, the arrival probability in the next time period of *t* can be estimated as $f_{\mathrm{arrival},u}(t)$. Given its current position, a user has more chance to arrive in a given edge cloud with certain coverage when its mobility increases. For a user with given mobility level, i.e. speed, the nearer to the edge cloud it is currently located, the more likely it will arrive at the edge cloud in the next certain period of time.

Next, the probability density function of time that the user *u* stays in an edge cloud, $f_{\mathrm{stay},u}(\tau)$, can be estimated. Given certain mobility level, the expected time of a user to stay in the coverage of the edge cloud depends on the mobility model and the coverage area of the edge cloud. For instance, a user is expected to stay longer in edge clouds with larger coverage and more traffic jams, such as an urban sector in the central of Paris City, than in the edge clouds with smaller coverage and less obstructed public traffic, such as a small village near a high-speed railway.

Meanwhile, with assistance of network monitoring, the error probability of the VNF on central cloud server and the backhaul connection in the next time period of *t* can be estimated as $\mathrm{f}_{\mathrm{error}}(t)$. For instance, edge clouds suffering from dense data congestions or continuously increasing network delay have generally higher risk of VNF outage.

As a result, the expected time that a user u suffers from outage of a given central cloud VNF in a certain edge cloud in the next time period of T can be finally estimated as:

$$\mathrm{E}\{t_{\mathrm{o},u}\} = p_{\mathrm{o}}\eta_u \int_0^T f_{\mathrm{arrival},u}(t) \int_0^{T-t} f_{\mathrm{stay},u}(\tau)\tau \,\mathrm{d}\tau\mathrm{d}t \,.$$

It is worth to note that outages of different VNFs, depending on their roles in network services, can lead to different losses. Therefore, an individual opportunity cost of service outage shall be computed independently for every different VNF in every different edge cloud. Summing up the opportunity costs for all users that may arrive / stay in an edge cloud during a coming period, the edge cloud can obtain an overall opportunity cost of outage of a certain VNF is given by

$$c_{\mathrm{o}} = \sum_{u\in U} \mathrm{E}\{t_{\mathrm{o},u}\}l,$$

where each U is the set of all UEs that can possibly be served by the edge cloud in the next period T, and l is the loss per unit time caused by each user suffering from the central cloud VNF outage.

Correspondingly, we can estimate the online implementation cost c_{i} of this VNF in this edge cloud, taking account of the data traffic cost (e.g. for downloading the VNF from central cloud), operation cost, etc. By comparing this opportunity cost c_{o} to the cost of locally and temporarily implementing the VNF c_{i}, the edge cloud is able to decide if to trigger such an implementation, and thereby to prepare essential data and execute the necessary processes.

The most important technical challenge here is to precisely estimate p_{o}, η_u, $f_{\mathrm{arrival},u}$ and $f_{\mathrm{stay},u}$. While p_{o} can be obtained in a straight-forward way from historical statistics and the real-time network state monitoring, the estimation of the rest three can be accomplished in two possible ways.

For stateful VNFs that contains critical data related to the UEs they serve, e.g. a virtualised HSS, the UE-relevant data must be synchronised to the local edge cloud, so the edge cloud should know the identifications of the UEs that are likely to arrive. In this case, the network estimates for every individual UE u, depending on its mobility and current position, at which edge clouds it could arrive in the next period. Then for every such edge cloud, $f_{\mathrm{arrival},u}$ and $f_{\mathrm{stay},u}$ will be individually estimated with help of the mobility model. Meanwhile, η_u can be obtained from the historical log of u.

For stateless VNFs that have no dependency on the UEs they serve, e.g. gateways, the edge cloud does not need to know the exact identifications of the served UEs, but only the overall statistics of their arrivals, stays and VNF utilisations. In this case, c_{o} can be briefly estimated as

$$\hat{c}_{\mathrm{o}} = \bar{N}\bar{\eta}\bar{\tau}_{\mathrm{stay}}p_{\mathrm{o}},$$

where $\bar{N}$ is the average number of UEs simultaneously served by the local edge cloud and relying on the target central cloud VNF, $\bar{\eta}$ is the average duty factor of the target central cloud VNF in the local edge cloud, $\bar{\tau}_{\mathrm{stay}}$ is the average UE stay time in the local edge cloud. Based on historical data and current network status, $\bar{N}$ and $\bar{\tau}$ can be estimated by the AMF, and $\bar{\eta}$ can be estimated by the VNFM.

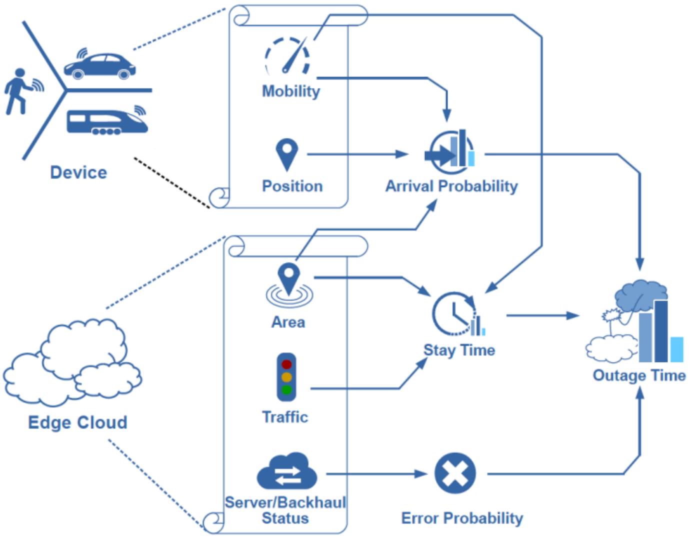


Figure 3-8: Estimating the service outage probability with the context information of device & edge cloud

3.5 Architectural Considerations

In addition to RAN-related NFs that can increase the reliability, the specialised network functions for increasing the resilience and reliability can be developed and implemented in the telco cloud, as described in Chapter 3. In fact, this is a pre-requisite for supporting the strict reliability requirements of URLLC services on the level of entire service, i.e. in an end-to-end level.

Similarly as with Figure 2-9, where the role of the RAN reliability modules is depicted into the overall architecture picture, Figure 3-9 illustrates the role of telco cloud-related network functions to the 5G-MoNArch architecture, designed for enabling high resilience. The network modules related to resilience in telco cloud are marked with a black dashed frame in Figure 3-9; they are furthermore underlined in this section for the sake of clarity. It is further noted that, as indicated in Chapter 2, the aggregated view of the resilience and security modules developed in WP3 of 5G-MoNArch and their role in the overall architecture is provided at the last chapter of this document, by means of Figure 6-1.

The telco cloud-related network functions are instantiated and configured based on the actual resilience demand of the service, as well as the SLAs agreed with the tenant. The service and tenant's requirements are translated into the network slice templates to be deployed on the available infrastructure. Such template includes the network functions that need to be deployed in order to support the required resilience. The Management and Orchestration layer is responsible for actual deployment, configuration, management and control of such functions.

The Management and Orchestration layer comprises the M&O functions from different network and technology domains (3GPP network management, ETSI NFV MANO, management functions of transport networks (TNs) and private networks) [5GM-D2.2]. In 3GPP network management domain, this can comprise element management (EM), domain management (DM) and network management (NM) functions. The 3GPP network management domain functions would also implement ETSI NFV MANO reference points to the VNFM and the NFVO, which are essential for fault management in virtualised environment and correlation of event related to physical and virtual domains. In legacy networks, the fault management is a part of network management FCAPS (Fault, Configuration, Accounting, Performance, Security) functionality. The 5G-MoNArch evolves the fault management in the direction of slicing awareness and cognition. The fault management takes into account slicing requirements as well as inputs from virtualised environment in order to perform and learn the optimal decisions. Thus, the concept of *5G fault management* spans different blocks of 5G-MoNArch Management and Orchestration layer as illustrated in Figure 3-9. The adaptability and dynamicity of 5G-MoNArch fault management is achieved through the concept of Fault Management Cognitive Functions (FM CFs). As described in Chapter 3, the Fault Management Cognitive Functions can be deployed at different layers of 5G-MoNArch architecture, (e.g. network slice, individual NFs, etc), depending on the desired level of centralisation/distribution of fault management implementation.

The *Cross-slice M&O* function responsible for inter-slice management will incorporate *x-slice Security & Resilience Management* function specialised for addressing jointly the security and resilience considerations (further details on security are provided in Chapter 5). The *Cross-domain M&O* function is responsible for the coordination/negotiation between different management domains (RAN, Core, edge cloud) within a single slice and it can incorporate the functionality for joint dealing with security and resilience issues, i.e. *x-domain Security & Resilience Management* (further details provided in Chapter 5). Additionally, as the concept of "*5G islands*" advocates the realisation of selected NFs or in extreme case entire network on the edge cloud, both the 3GPP Network Management and NFV MANO need to be involved in operation of "5G island", as depicted in Figure 3-9.

Finally, a crucial building block for improving the telco cloud resilience is load-aware, leading to a *scalable and resilient controller framework* (c.f. Figure 3-9) which can be applied to XSC and ISC controllers of 5G-MoNArch. For supporting high resilience as well as scalability, the controller framework (ISCs and XSCs) operates in a cluster mode. In this regard, in order to maintain the overall topology state of the network, the data *Store* modules are introduced. Such modules are internal and distributed data base modules available in both the ISC and XSC controller nodes. As a result, data Store has a dedicated interface for data synchronisation and state management, which is depicted in Figure 3-9 in addition to the northbound interface (NBI) and southbound interface (SoBI) of the controller layer. In this respect, a *context synchronisation* operation is anticipated between the store nodes associated with each of the controllers.

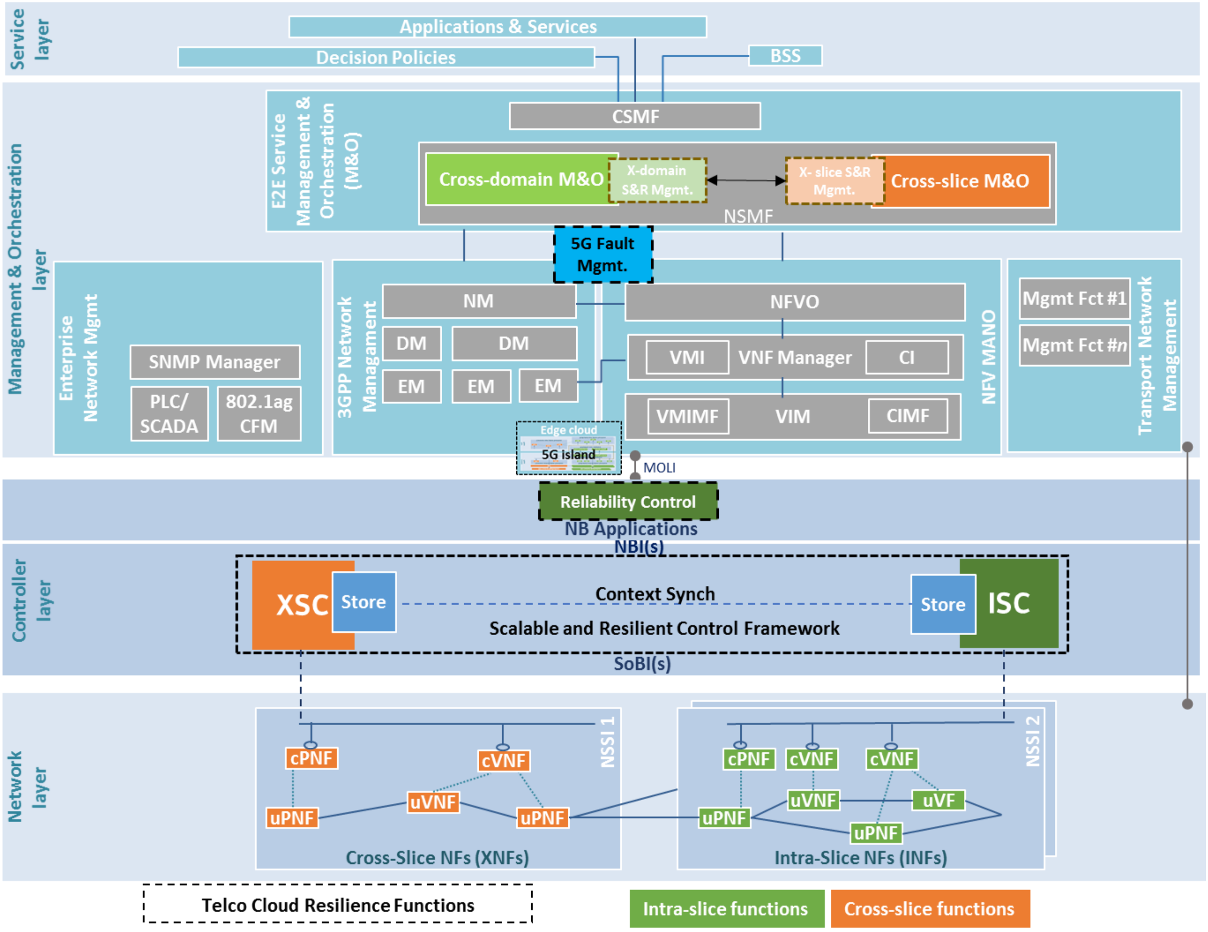


Figure 3-9: Mapping of the 5G-MoNArch architecture to telco-cloud resilience functions

4 Security Issues

Traditionally, security assurance has been a transversal aspect independent from the rest of properties that characterise an ICT infrastructure. While performance has been seen as the main factor affected by the impact of mechanisms required to protect an infrastructure (i.e., normally more security entails less performance), there are, as it will be shown, other aspects affected. This is especially relevant for 5G networks, where the aim of covering many different (and specific) domains while optimising resources (specially costs), prevails over other aspects. In this regard, resource optimisation can be done in many different ways. Chapter 2 described techniques for the optimisation of the RAN, while Chapter 3 described techniques to make the infrastructure resilient to internal failures. However, it is not just about reacting to internal eventualities such as a faulty base station. External factors, such as security threats exploited by attackers, might also impact on the infrastructure, reducing performance, availability, exposing critical data, and, as a result, increasing costs and usage of computational resources. This chapter describes techniques for the security protection of 5G infrastructures, adapted to the dynamicity of these networks while optimising the resources used.

It is known that any system exposed to the environment and the human interaction is subject to be a target of attacks. Depending on the degree of exposure and the nature of the elements that compose the system, some threats are more likely to occur than others. In the case of IT services based on 5G network infrastructures, there is a wide range of threats that can affect both network tenants and telco operators [V17]. Since different technologies are involved, intertwined by multiple software and hardware infrastructure layers, the number of critical assets to protect increases. As a result, the vulnerabilities and weaknesses that can be exploited increase as well.

Security incidents have a direct impact on the overall service operation at different levels. For instance, by progressively (and silently) exhausting network resources through a “man-in-the-middle” attack on a mobile connection. Another example is a targeted DoS attack which may cause disastrous business downtime, loss of data and application service, turning into a huge economic impact and damage to the brand image. Section 4.3.1 outlines the most usual threats and risks in industrial scenarios.

One of the most common strategies to holistically address security threats is security monitoring. This is a conservative mechanism that relies on well-proven security directives that permit detecting an (attempt of an) incident with high accuracy. However, advanced cyber threats such as *advanced persistence threats (APTs)*, due to their increasing frequency, sophistication, importance and difficulty in countering in recent years [EETLR16], make any organisation efforts to fight them with traditional approaches almost useless. In fact, hackers can easily circumvent any new patch or security obstacle deployed in the system, and render the recently updated detection rules outdated shortly after these are rolled-out. Therefore, it is crucial to implement a proper strategy which a) methodologically identifies the threats that may affect the (changing) system under analysis, assessing their risk and impact in the system and b) define the most appropriate mechanisms to apply in order to proportionally address them.

Section 4.1 elaborates exactly on this strategy and presents a high-level architecture design to support each of the stages of the associated process taking into account the characteristics of the 5G architecture. In particular, the network slicing concept calls for security architectures that are able to work autonomously within a slice, even in a disconnected way (e.g. at a cloud edge) [MDS+17]. Security trust zones (STZs) is a concept introduced in [HWM+S17] to describe an architectural security solution for 5G networks that enhances the so-called AAA security functions at edge clouds.

Section 4.3 describes how the operation of STZs can be additionally equipped with the necessary mechanisms to detect security incidents, take decisions and apply custom countermeasures locally and fast. This so-called prescriptive security is based on automating simple and specific threat analysis tasks with sophisticated machine learning and artificial intelligence [GBT17]. In addition, STZs shall have the capabilities to share certain threat intelligence with other zones to avoid propagation and remain self-defending. STZs may cover multi-geographic areas and spread across different network slices, therefore, an inter-slice security management function would be required to govern and orchestrate the overall security response.

Finally, Section 4.3 takes a case study, the *Hamburg Smart Sea Port Testbed*, to illustrate how all these concepts are applied to a real scenario.

4.1 Security Monitoring and Active Learning (SMAL)

Figure 4-1 depicts the SMAL process, which was originally introduced in [MGG+17]. SMAL is a strategy that comprises a combination of continuously monitoring of the landscape and active learning, in order to counteract security incidents and mitigate their effects, avoid their propagation, prevent them from happening again and consequently, as well as limit their impact to operative business.

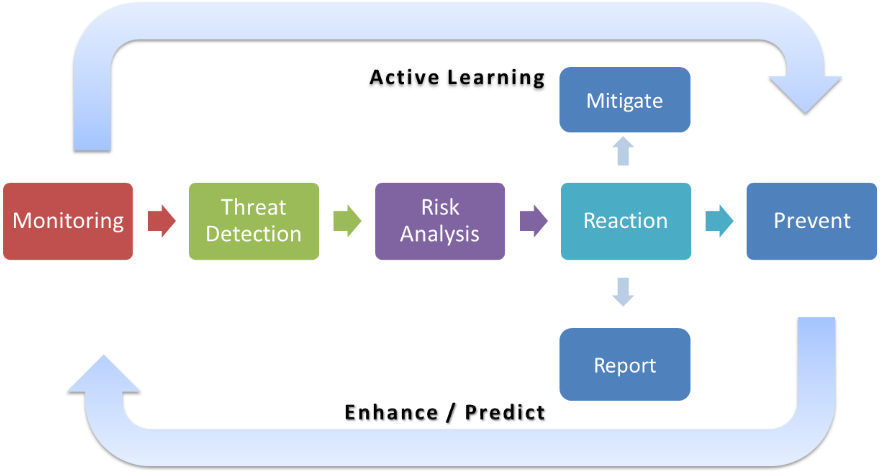


Figure 4-1: Security monitoring and active learning (SMAL) process

Security monitoring is the activity of observing the infrastructure to be protected (by collecting events from multiple and diverse data sources in the managed infrastructure) looking for known patterns of attacks or security incidents which allows for their detection (*Threat Detection* in Figure 4-1) as soon as they occur. The reliability of these detections is very high in most cases since the patterns are well proven in industrial settings and must come together with evidence (e.g. system events, logs, etc.) that proves the accuracy of the detections, which are also subject to audits conducted afterwards.

Security information and event management (SIEM) systems have the capacity to collect, store and correlate events generated by a managed infrastructure. These events represent the actual behaviour of a system in a certain point in time. Most importantly, SIEMs have processing capabilities that correlate collected and stored events through rules or security directives. These directives permit looking for certain patterns of activity in the collected events (either malicious or legitimate), and trigger an alarm to raise awareness of such pattern detection with high levels of accuracy.

Recent versions of leading SIEM systems [D-D2.1] already include some features for doing *Risk Analysis*, taking into account the criticality of the assets of the managed infrastructure. This feature evaluates the most appropriate reaction to be taken on the event of a security alarm raised, as it is described later in Section 4.1.1.

This high reliability allows taking the appropriate countermeasures safely with a minimum delay, minimising the impact and avoiding propagation. However, this is proven not enough when dealing with changing context conditions, and especially with hackers learning from experience and releasing new attack vectors, malware, ransomware growing at tremendous rates per day [GBC+17]. In fact, there is no security and contingency plan in the world able to deal with such numbers, unless opting for an active learning process and prescriptive security approach. This approach is described later in Section 4.1.2.

4.1.1 Reaction to Security Incidents

Up to now, we have described mechanisms that are used to monitor, detect and possibly prevent attacks, which are mainly derived from the complex and dynamic nature of 5G infrastructures. Nonetheless, the major challenge remains to apply automated responses to cybersecurity incidents in a timely and automated manner.

There are several options to react when a security incident is detected:

- *Reporting actions*: In many cases, the first and most natural reaction to a security incident is to report about it to the entity in charge of dealing with it. This is a passive attitude that may delay the countermeasure and contribute positively to threats to spread and succeed in their ultimate malicious objective. On the other hand, however, it is a safe mechanism in cases where the security infrastructure does not have the required permissions to further intervene in the managed infrastructure, or when there is not enough evidence to take a decision that might hinder or damage the overall business service operation.
- *Mitigation actions*: Such actions aim at controlling the negative effect of the security incident and avoid its propagation, e.g., changing the configuration of firewalls or network intrusion detection system (NIDS). These are invasive actions which require, in many cases, administrative permissions over the managed infrastructure. Therefore, the application of mitigation countermeasures is not possible in all cases.
- *Preventive actions*: It is in principle possible to extract some lessons learnt from any detection (positive or negative), and use these to enhance the overall security operation in many different ways by using advanced analytics such as machine learning or artificial intelligence techniques. Besides enhancing existing security monitoring rules (or directives) to adapt to context changes, it is possible to learn new potential malicious behaviours and thus, predict the attack vectors of tomorrow.

In any of the above listed cases, it is necessary that the reaction taken is proportionate and appropriate to the situation. In order words, *several aspects of the infrastructure, the service operation and the business agreement in place* should be considered. The ensuing two sections elaborate on these aspects.

Risk-based Reactions

Given the implications that certain reactive approaches may have in the overall service operation and in turn, in the fulfilment of the SLA between the 5G service provider and its tenants, it seems reasonable to have a mechanism in place that evaluates certain factors before deciding on which reaction to take.

There are different ways in which existing security monitoring and management tools implement the risk analysis. A basic and static approach is the prioritisation of detections based on their categories, or using a formula to calculate over certain characteristics of the detection (e.g. asset relevance, priority, reliability, etc. and which can be customised to certain point to align better to customer needs). Some more advanced systems perform a real-time risk assessment based on a scoring model of the infrastructure assets, which is configurable by the security administrator, and will determine if a detected security incident should be notified to the administrator or not. Other approaches cross-correlate security events with vulnerability scans on the affected assets, to determine if the detected event targets to exploit a particular existing vulnerability and thus, have a higher probability of success. Finally, there are dynamic approaches that update the risk score based on pattern matching on the observed activity of assets (i.e. identify actions that raise the risk profile of assets), and thus adapting to context changes.

4.1.2 Active Learning towards Threat Prevention

A thorough analysis of more recent versions of leading SIEMs conducted in the context of H2020 project DiSIEM (Diversity-enhancements for SIEMs) [D-D2.1] shows that nowadays the trend is to integrate with application- and user-based anomaly detectors. This particularly includes user and entity behaviour analysis (UEBA) systems, which comprise the analysis of the behaviour of employees, third-party contractors and other collaborators of the organisation, as well as the use of machine learning techniques for detecting misbehaviour, i.e. by using outlier detectors or classifiers. In contrast to the legitimate or normal behaviour, by actively learning from the analysis of anomalous behaviour allows coming up with new patterns or evolutions of known ones. Overall, this active learning process is a way to autonomously enhance the threat intelligence knowledge database, adapt to dynamically changing attack vectors and prevent from future security incidents and from their propagation.

Prevention mechanisms permit learning from experience and enhance detection rules and reconfigure the security monitoring infrastructure to adapt to new scenarios. However, the main drawback of prevention mechanisms based on machine learning algorithms is the high rate of false positives. The reliability of the alarms raised by such tools is usually not high (especially when the training data is not extensive, rich or varied enough) and thus, the triggered countermeasures must be simply preventive rather than reactive. As such, these signals should be used to *prepare the system for the worst scenario*, which can last for a predefined period of time or until the preventive system identifies that the threat is no longer probable to materialise.

4.2 Security Trust Zones (STZs)

4.2.1 Characterisation

An STZ defines a logical area of infrastructure and services where a certain level of security and trust is required. Security refers to the quality of being protected against threats and the measures put in place to guard against these. Trust relates to the assurance or confidence in that certain expectations will be indeed meet, throughout a defined period of time.

Many factors determine the most appropriate security controls to apply in order to protect a system and there exist methodologies, standards and control frameworks [ISO/IEC 27001:2013, NIST SP 800-53, USDoD Instruction 8500.2, ITU-T X.800 Recommendation, ISO 7498-2, etc.] to guide security practitioners in this task. The majority of these frameworks require a great deal of manual work. Even for the tasks where some degree of automation has been achieved, any adaptation to the changing context conditions might entail a complete re-design of the security strategy.

One way towards a fully automated deployment of the right security controls in the target infrastructure that we aim to protect, is to define *STZ templates*. These templates, as a sort of blueprints, describe the list of security services and infrastructure elements that need to be provisioned, as well as the default configurations to ensure that a particular level of security and trust is achieved. In addition to automation, templates help in easily restoring in case of failure or misconfiguration, facilitates updates and patch management and in general, overall management in distributed deployments. STZ templates include also the necessary means to equip the security infrastructure elements with some degree of autonomy and self-healing capabilities.

Table 4-1 lists all elements included in the descriptor of a STZ template. Six groups of properties have been identified to describe a STZ. The General group characterises the STZ by defining the security and trust level we aim to achieve. The security level has been decomposed into two properties: privacy and integrity. This could be extended to other security properties such as availability. The STZ level would be a simplification to identify the type of STZ template considered in each case. In the Table there are also three groups of properties (detection, prevention and reaction) that describe the capabilities of the STZ in order to achieve the security and trust level promised. The group listing the self-healing capabilities of the STZ aim at describing the workflow to adapt to context changes and recover from failure to provide business continuity. The Threat Intelligence group focus on the mechanisms that enable the exchange of threat intelligence between STZs to avoid propagation of threats.

Table 4-1: STZ profile template

Group	Property	Description
General	STZ Level	e.g. L (low), M (medium), H (high) e.g. [1 … 5]
	Privacy level	Determines the privacy-preserving mechanisms put in place, e.g. when sharing threat intelligence between zones
	Integrity level	Determines the resulting integrity level to achieve, which is the objective of the security measures deployed
	Resilience/Security Trade-off factor	This is a placeholder that needs to be further defined and clarify how this factor influences the STZ functioning
Detection capabilities	Threats	According to the Threat Taxonomy, the list of threats able to detect

	Rules Deployed	Set of available detection directives (not all might be active all the time)
	Rules Active	The actual set being monitored by default (may change at runtime)
	Sensors Deployed	Set of available sensors deployed (not all might be active all the time)
	Sensors Active	The actual set activated by default (may change at runtime)
	Events	The events understood by the infrastructure (type, XSD schema)
	Alarms triggered	The alarms output (to trigger actions)
Prevention capabilities	Threats	According to the Threat Taxonomy, the of threats able to prevent
	Rules Deployed	Set of available prevention directives (not all might be active all the time)
	Rules Active	The actual set being monitored by default (may change at runtime)
	Sensors Deployed	Set of available sensors deployed (not all might be active all the time)
	Sensors Active	The actual set activated by default (may change at runtime)
	Events	The events understood by the infrastructure (type, XSD schema)
	Alarms triggered	The alarms output (to trigger actions)
Reaction capabilities	Countermeasures	According to a Countermeasure Taxonomy, the list of countermeasures able to trigger
	Rules Deployed	Set of available reaction rules (not all might be active all the time)
	Rules Active	The actual reaction rules applicable by default (may change at runtime)
	Actuators Deployed	Set of available reaction mechanisms deployed (not all might be active all the time)
	Actuators Active	The actual set of reaction mechanisms that could be invoked by default (may change at runtime)
	Alarms	The alarms understood by the infrastructure (type, XSD schema)
Self-healing capabilities	Reconfiguration rules	Under certain conditions, the actual configuration of the STZ may be changed to adapt to the context condition
	Autonomy rules	Enables the STZ infrastructure to work in isolation (disconnected) totally or partially (e.g. by logging events/alarms produced and countermeasures triggered, so these can be send back to the central node once the connectivity is restored)
Threat intelligence exchange	Conversion Plugins	Convert from/to different events/alarms formats/schemas
	Normalisation Plugins	E.g. when data ranges are in different scales (e.g. H,L,M scale vs 0..5 scale), or IP v4 vs IP v6
	Privacy-preserving Plugins	Applies privacy measures on the information contained in exchanged events/alarms (e.g. obfuscation, anonymisation, pseudo-anonymisation)

4.2.2 Profiling Methodology

Traditionally, the way to conduct security profiling was by analysing the target system to protect and the information flows, performing a risk assessment to identify critical assets and prioritise them, and based on that, drawing the so-called security perimeters. Within the boundaries of each perimeter, a suite of security controls was deployed, adequate to the level of protection required in each case. A similar approach is possible with STZs, yet several other factors need to be taken into account in order to exploit the new characteristics introduced by STZ templates. Such factors are prevention mechanisms, self-healing capabilities and threat intelligence exchange.

Table 4-2 lists the criteria that need to be analysed in order to determine the different STZ profiles that live together in the target 5G infrastructure and services. Such criteria have been grouped into three dimensions. The *security/risks dimension* encompasses the traditional risk assessment and security framework guidelines, which will give a first approach agnostic to the particularities of the specific business and 5G context. The *business dimension* refers to the different requirements that are driven by the tenants, which are actually sharing the same 5G infrastructure. The *services/infrastructure dimension* takes into account the actual set of assets to protect and the technical resources available.

The business dimension has an impact on the application of some security controls rather than others, which in principle may be judged more appropriate or efficient, only because corporate policy or applicable national regulation impose them. Highly regulated environments such as eHealth or financial services are some examples where the business dimension criteria will weigh more than other dimensions. On the other hand, the terms agreed between tenant and service provider may influence the selection of controls with lower costs or footprint for example, in favour of guaranteeing a compromised service performance and value for money.

The services/infrastructure dimension would determine the most appropriate set of security components to deploy (and their configuration) in order to implement certain security controls. The overall capabilities of the STZ to protect against threats, in terms of detection, prevention and reaction, would be influenced by these criteria. In cases of limited resources, some less critical capabilities will not be activated or not even deployed in the first place.

Table 4-2: Criteria to assess the most appropriate STZ template to apply

Dimension	Criterion	Impact
Security	Risk assessment results	Determine the critically of the assets to protect and prioritise some security aspects over others
	Security Control Framework	Best practices on how to better secure the infrastructure
Business	Compliance to applicable Regulation	Strict regulations applicable may force to implement privacy measures despite the apparent lack of threats likely to occur
	Corporate/Organisation Security Policy	Some organisations oblige implementing security measures which are not apparently proportionate
	SLA (e.g. performance, resilience level, multitenancy/isolation)	This will force to relax the security measures in favour of maintaining certain level of performance or availability agreed between client (tenant) and the CSP / Network provider
5G Services / Infrastructure	Geographical dispersion/distribution (NSs)	The actual Network Slice configuration applicable, with the hardware, software and virtual elements involved will influence the most appropriate STZ configuration (threats,

		type of sensors, possible countermeasures, etc.)
	Connectivity (Domains)	The more isolated (disconnected) the less prone to threats (in principle). This will also imply a higher degree of self-healing capabilities
	Resources available	This determines the number and type of sensors that can be deployed, or e.g. the correlation processes that can be running in parallel

4.2.3 Architectural Considerations

As also mentioned above, one of the key concepts in 5G-MoNArch architecture is the use of the network slicing concept. In the context of security, this means that the deployment of the different STZs defined for a specific target environment needs to be slice-aware. This slice-aware deployment strategy is required not only to prevent and isolate each network slice in case of detecting some security issue to avoid propagation and ensure the normal operation in the rest of the system, but also to consider the specific network slice constraints (the available software and hardware infrastructure in each slice) and assets to protect in the slice for the deployment of security functions.

In light of the above, we can distinguish several levels in this deployment strategy based on slices, namely inter and intra slice components. Moreover, we distinguish two main security enablers, namely security threats monitoring and security trust zone manager, as described below. These enablers, along with the levels of the slice-aware security deployment strategy, are illustrated in Figure 4-3.

Security threats monitoring (SM)

On the lowest level, in each STZ defined in a network slice, three types of sub-components of the security threats monitoring can be deployed:

- *Security threat detection (SthD):* this component will integrate a set of sensors (such as Intrusion Detection Systems) to perform the detection of security incidents and suspicious behaviours in the 5G network traffic of a STZ. The events and alarms generated by these sensors will be collected, normalised to a common format and sent to the security monitoring manager for its processing.
- *Security threat prevention (SthP):* this component will integrate the prevention mechanisms or sensors that allows to learn about potential incidents and enhance the detection in a STZ. The events generated by this component (including not only the incidents detected but also forecast occurrence of a threat and its likelihood) will be also normalised and sent to the security monitoring manager for its processing.
- *Security threat reaction (SthR):* this component will apply the required countermeasures or mitigation actions when some incident is detected in a STZ. This component will be triggered by the security monitoring manager.

The selection and deployment of these components will depend largely on the STZ profile where they are located as well as on the available infrastructure for a specific network slice. If the resources are available, these components will support self-healing capabilities in the sense of self-adapting and autonomy to deal with loss of connectivity.

On the network slice level, a *security monitoring manager (SMm)* will be deployed. This component will receive the data provided by the different SthD, SthP and SthR components deployed through the STZs of a same Network Slice. The SMm will consolidate the information received to be presented to the user to know the overall status of the different STZ and network slice. With this purpose, this component will include a set of correlation and aggregation rules that will trigger high level alarms.

At the same time, the SMm will monitor the status of those security components deployed in its managed network slice and will interact with them to active/deactivate them when required.

The SthD and SthP components deployed in each STZ will have a predefined configuration and set of detection/prevention rules that will be initially activated or not depending on the specific STZ profile. This configuration and rules can be updated or activated/deactivated in real-time from the SMm when necessary on those components supporting these reconfiguration capabilities.

A *threat intelligence exchange (ThIntEx)* component will be also deployed at network slice level to share threat intelligence data between the different network slices. The ThIntEx will interact closely with the SMm in a bidirectional way: That is, to share incidents detected in a network slice as well as the information related to those incidents (e.g. STZ profile where an incident has been detected, the SthD components involved in the detection with the rules triggered and the reactions applied), and to receive notifications from another network slice to be considered by the SMMs to act in consequence if necessary.

A *cross slice threat intelligence exchange (XSThIntEx)* will also be deployed at the inter-slice level. The XSThIntEx acts as broker between the ThIntEx components that are deployed in different network slices, managing the subscription to and exchange of information among ThIntEx components (i.e., incidents detected, new rules to detect incidents, etc.). Figure 4-2 represents the interactions between the XSThIntEx and the ThIntEx of different slices. The ThIntEx will be subscribed to certain types of information exchanged by the rest of the network slices (i.e., information about new rules). The XSThIntEx will publish the security information retrieved from the ThIntEx of the different slices, while every ThIntEx component will receive the security information to which it is subscribed to. At the inter-slice level the ThIntEx components will communicate to their respective SMms about security information detected at other slices. Decoupling the SMms from the ThIntEx component allows for flexibility and robustness of the deployment, as it separates the technology used for the exchange of security information between ThIntExs and XSThIntExs from the normal operation of the SMms. For example, it would be possible to change from a RabbitMQ message broker to a Kafka based one or even to migrate to a synchronous solution without the need to interrupt the operation of the SMm.

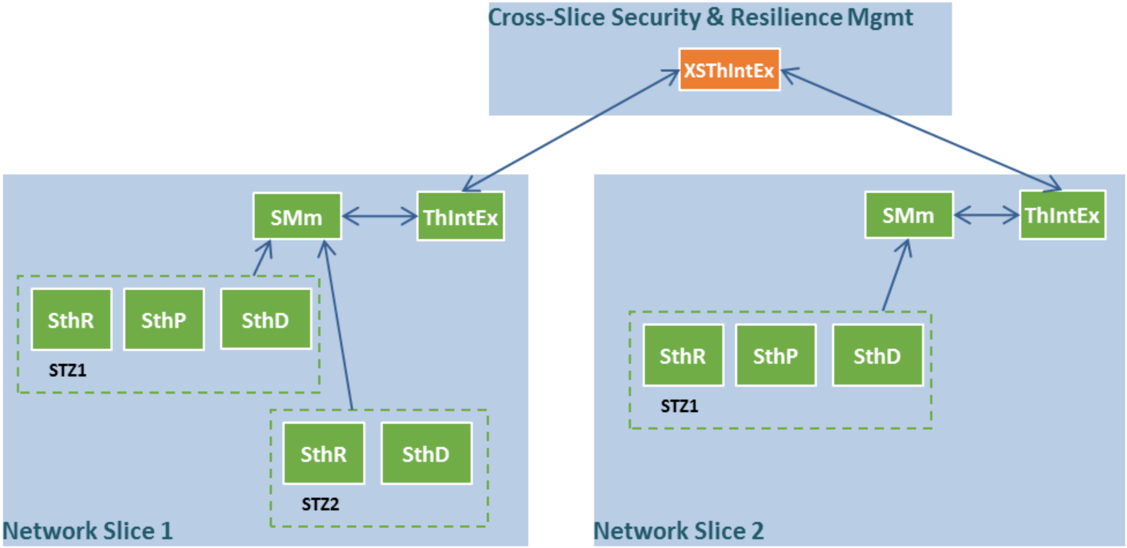


Figure 4-2: XSThIntEx and ThIntEx interactions

Security trust zones manager (STZm)

The main function of the STZm is to centralise and control what is happening in the STZs deployed across the different network slices. Consequently, this component will consult the security and resilience trade-off component from the service layer to determine the most appropriate security level to be applied throughout the different architecture layers. The STZm interacts with the different SMms deployed to ensure the security level provisioned is achieved. This component also holds a knowledge base of threat intelligence to keep up to date any STZ deployed cross-slices, with the latest cybersecurity intelligence events and security directives.

X-slice and X-domain security and reaction management

These two elements provide additional processing capabilities based on information retrieved from the rest of the security infrastructure. For example, these components will carry out the security and resilience trade-off evaluation (details are given in Chapter 5). The X-domain security and reaction management component operates at the network slice level based on information obtained from the SMm, and interacting with the components of the STZ (for instance to trigger a reaction at the SthR according to the result of the Security and Resilience trade-off evaluation). The X-slice Security and Resilience Management will operate at the MANO layer, coordinating the specific aspects that are relevant for the activities of the X-domain element (for instance, retrieving specific security and resilience requirements from the service layer).

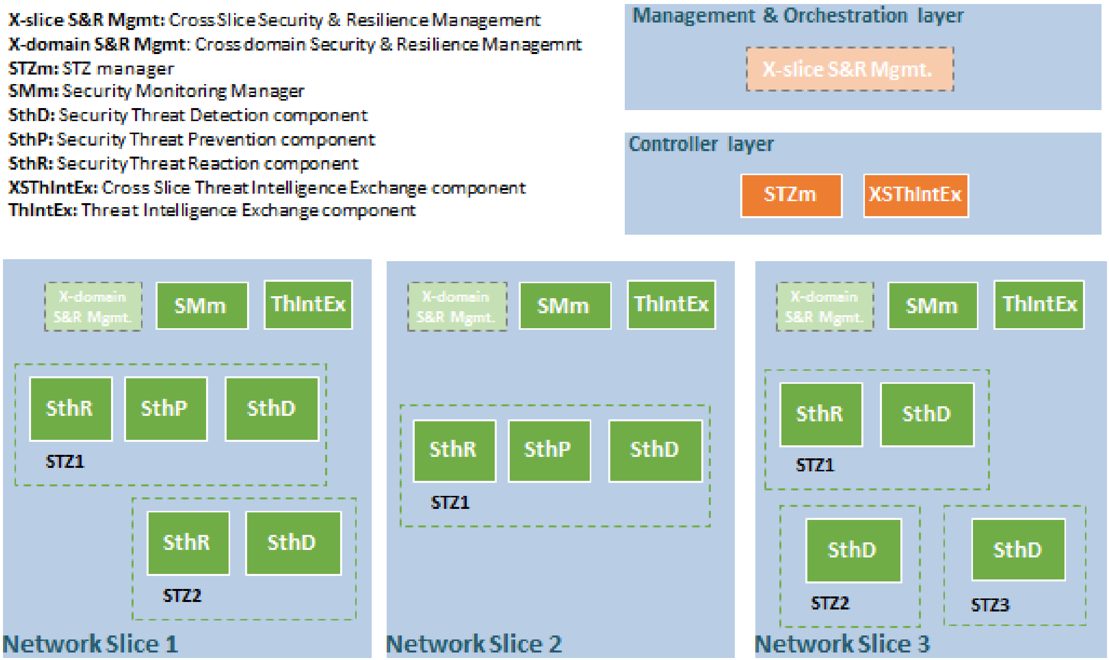

X-slice S&R Mgmt: Cross Slice Security & Resilience Management
X-domain S&R Mgmt: Cross domain Security & Resilience Managemnt
STZm: STZ manager
SMm: Security Monitoring Manager
SthD: Security Threat Detection component
SthP: Security Threat Prevention component
SthR: Security Threat Reaction component
XSThIntEx: Cross Slice Threat Intelligence Exchange component
ThIntEx: Threat Intelligence Exchange component
Management & Orchestration layer
X-slice S&R Mgmt.
Controller layer
STZm
XSThIntEx
X-domain S&R Mgmt.
SMm
ThIntEx
SthR
SthP
SthD
STZ1
SthR
SthD
STZ2
Network Slice 1
X-domain S&R Mgmt.
SMm
ThIntEx
SthR
SthP
SthD
STZ1
Network Slice 2
X-domain S&R Mgmt.
SMm
ThIntEx
SthR
SthD
STZ1
SthD
STZ2
SthD
STZ3
Network Slice 3

Figure 4-3: Slice-aware STZ deployment strategy

Translating this strategy into the 5G-MoNArch architectural view, Figure 4-4 illustrates the role of the security-specific modules in the architecture. In particular, Figure 4-4 reflects how the *STZm* component is located at the *controller layer* in order to reach for different network slices and coordinate the information exchanged by the *ThIntEx* from different network slices by means of the *XSThIntEx* network function. . The *SMm* component controls the security monitoring and active learning inter-slice network function elements of each active STZ, within the same network slice.

Figure 4-4 also depicts the slice-specific security modules, namely *SthR*, *SthP* and *SthD*, together with their interaction with the ThIntEx component in the *network* layer. Their joint operation consists the *STZ template*, which refers to a set of security network functions that interfaces with the controller layer and the respective controller functions considered there. The composition of each STZ instance, i.e. the actual available and active network functions, will depend on the template deployed by the MANO layer as well as their specific configuration, coordinated by the *X-slice and X-domain Security & Resilience Mgmt* component.

Finally, in the service layer, a *Cyber Security* dashboard service will permit the interaction with the security and systems administrators. Such interaction will allow to, on the one hand, correctly configure the entire security architecture elements according to the actual 5G Infrastructure and Tenants/Users requirements, and on the other hand, report on the actual security status of the system at any point in time.

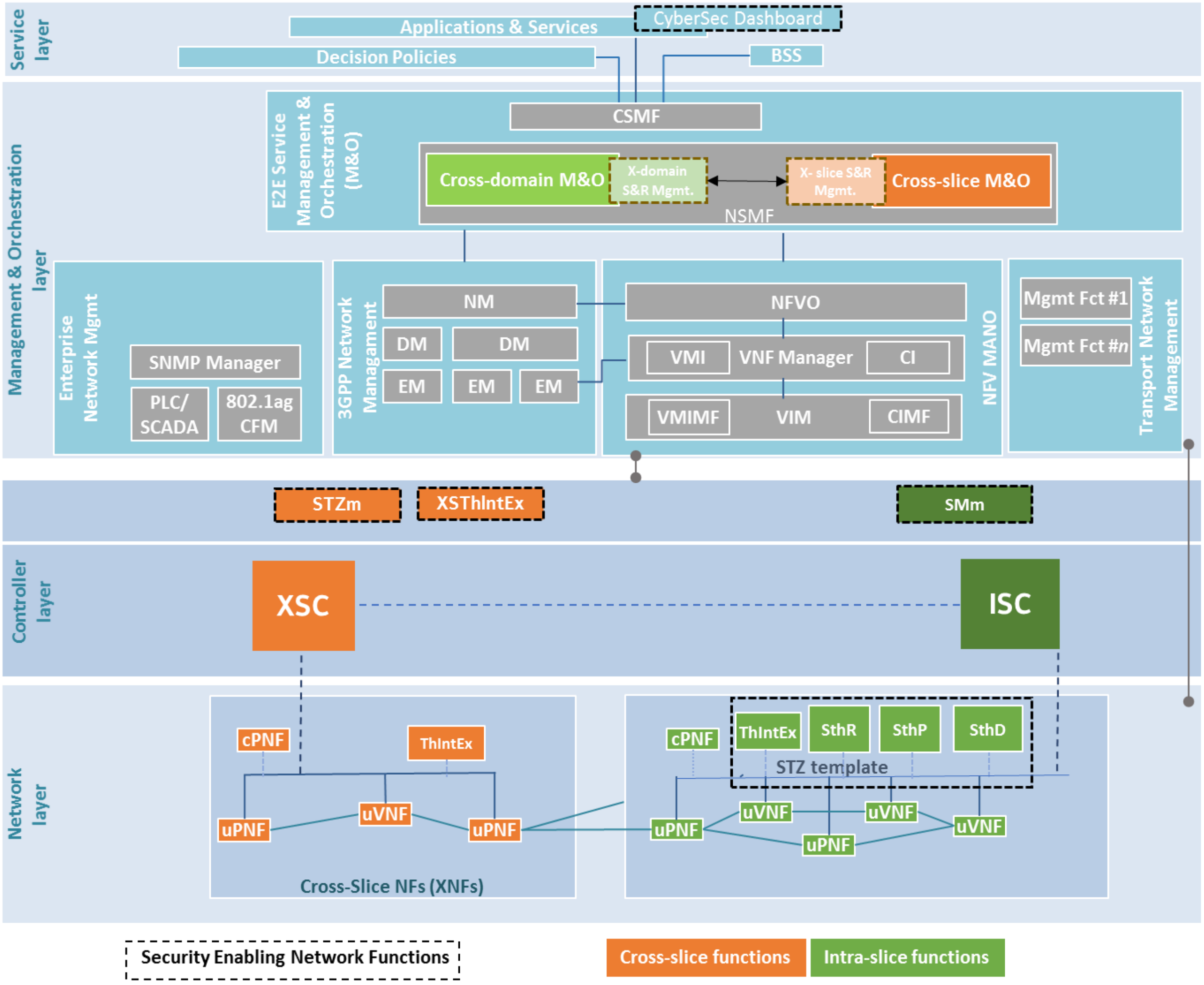


Figure 4-4: View of the 5G-MoNArch architecture with security-enabling architectural components

4.3 *Case study: Hamburg Smart Sea Port Testbed*

This section analyses the characteristics of the Hamburg Smart Sea Port case study, in terms of business services and infrastructure, in order to identify the risks and threats that may affect the normal business operation and compromise the security of the information stored and data flow. The work presented in this section *complements the work conducted in WP6 of 5G-MoNArch* (see, e.g., [5GM-D6.1]), where a business case analysis and an evaluation of the relevant performance indicators is carried out on a larger scale.

The section also proposes a way to protect the assets (that is, information, devices, infrastructure, services, etc.) by using the concept of security trust zones. It is noted that this is just a first proposal which aims at illustrating the use of the STZs in a real case, and thus, it should not be taken as the final solution, which will require further details on the infrastructure and business services, and a more thorough security analysis.

4.3.1 Threat and Risk Analysis for Industrial Scenarios

This section presents a taxonomy of threats and risks that can appear in industrial scenarios. Later, in Section 4.3.1, we identify the threats and possible solutions related to the Smart Sea Port scenario.

In addition to cyber-threats, the presented threat taxonomy contains also physical-threats, related to causing damage to information technology assets. Most of the threats/risks presented in this section are based on the ENISA Threat Taxonomy document [E15].

Next, a comprehensive list of threats and risks that can appear in a wide range of possible industrial scenarios is presented.

1. **Physical attack**
    a. *Fraud:* Fraud committed by employees or others that are in relation with entities, who have access to entities' information and IT assets.
    b. *Sabotage:* Intentional actions aimed to cause disruption or damage to IT assets.
    c. *Vandalism:* Act of physically damaging IT assets.
    d. *Theft:* Stealing information or IT assets, e.g. theft of mobile devices, fixed hardware, documents, and backups.
    e. *Information leak /sharing:* Sharing information with unauthorised entities. Loss of information confidentiality due to intentional human actions
    f. *Unauthorised physical access / Unauthorised entry to premises:* Unapproved access to facility.
    g. *Coercion, extortion or corruption:* Actions following acts of coercion, extortion or corruption.
    h. *Damage from the warfare:* Threats of direct impact of warfare activities.
    i. *Terrorist attack:* Threats from terrorists.
2. **Unintentional damage / loss of information or IT assets**
    a. *Information leak /sharing due to human error:* Information leak / sharing caused by humans, due to their mistakes, e.g. sharing via verbal communication, mobile applications, web applications, and network eavesdropping.
    b. *Erroneous use or administration of devices and systems:* Information leak / sharing / damage caused by misuse of IT assets (lack of awareness of application features) or wrong / improper IT assets configuration or management. For example: maintenance errors / operators' errors, configuration/ installation error, and user errors.
    c. *Using information from an unreliable source:* Bad decisions based on unreliable sources of information or unchecked information.
    d. *Unintentional change of data in an information system:* Loss of information integrity due to human error (information system user mistake).
    e. *Inadequate design and planning or improper adaptation:* Threats caused by improper IT assets or business processes design (inadequate specifications of IT products, inadequate usability, insecure interfaces, policy/procedure flows, design errors).
    f. *Damage caused by a third party:* Threats of damage to IT assets caused by breach of security regulations by third party.
    g. *Damages resulting from penetration testing:* Threats to information systems caused by conducting IT penetration tests inappropriately.
    h. *Loss of information in the cloud:* Threats of losing information or data stored in the cloud.
    i. *Loss of (integrity of) sensitive information:* Threats of losing information or data, or changing information classified as sensitive.
    j. *Loss of devices, storage media and documents:* Threats of unavailability (losing) of IT assets and documents, e.g. devices/ mobile devices, storage media, and documentation of IT Infrastructure.
    k. *Destruction of records:* Threats of unavailability (destruction) of data and records (information) stored in devices and storage media, e.g. through malware infection or abuse of storage.
3. **Disaster**
    a. *Natural/environmental disaster:* Large scale natural disaster, e.g. natural earthquakes, floods, landslides, tsunamis, heavy rains, heavy snowfalls, heavy winds.
    b. *Fire:* Threat of fire.

   c. *Pollution, dust, corrosion:* Threat of disruption of work of IT systems (hardware) due to pollution, dust or corrosion (arising from the air).
   d. *Thunder strike:* Threat of damage to IT hardware caused by thunder strike.
   e. *Water:* Threat of damage to IT hardware caused by water.
   f. *Explosion:* Threat of damage to IT hardware caused by explosion.
   g. *Dangerous radiation leak:* Threat of damage to IT hardware caused by radiation leak.
   h. *Unfavourable climatic conditions: C*limatic conditions that have a negative effect on hardware, e.g. high humidity or high or low temperature.
   i. *Threats from space / Electromagnetic storm:* Threats of the negative impact of solar radiation to satellites and radio wave communication systems - electromagnetic storm.
   j. *Wildlife:* Threat of destruction of IT assets caused by animals, e.g. mice, rats, or birds.

4. **Failure/Malfunction**
   a. *Failure of devices or systems*: Failure of IT hardware and/or software assets or its parts, e.g. data media, hardware, applications and services, and parts of devices (connectors, plug-ins).
   b. *Failure or disruption of communication links:* Threat of failure or malfunction of communications links, e.g. due to problem in cable/wireless/mobile networks.
   c. *Failure or disruption of main supply:* Threat of failure or malfunction of power supply.
   d. *Failure or disruption of service providers:* Failure or disruption of third party services required for proper operation of information systems
   e. *Malfunction of equipment:* Threat of malfunction of IT hardware and/or software assets or its parts, e.g. improper working parameters, jamming, and rebooting.
5. **Outage**
   a. *Absence of personnel:* Unavailability of key personnel and their competences.
   b. *Strike:* Unavailability of staff due to a strike.
   c. *Loss of support services:* Unavailability of support services required for proper operation of the information system.
   d. *Internet outage:* Unavailability of the Internet connection.
   e. *Network outage:* Unavailability of communication links.
6. **Eavesdropping/Interception/ Hijacking**
   a. *War driving:* Threat of locating and possibly exploiting connection to the wireless network.
   b. *Intercepting compromising emissions:* Threat of disclosure of transmitted information using interception and analysis of compromising emission.
   c. *Interception of information:* Interception of information which is improperly secured in transmission or by improper actions of staff, e.g. corporate/national espionage or unsecured Wi-Fi.
   d. *Interfering radiation:* Failure of IT hardware or transmission connection due to electromagnetic induction or electromagnetic radiation emitted by an outside source.
   e. *Replay of messages:* A valid data transmission is maliciously or fraudulently repeated or delayed.
   f. *Network Reconnaissance, Network traffic manipulation and Information gathering:* Threat of identifying information about a network to find security weaknesses.
   g. *Man-in-the-middle/Session hijacking:* Relay/alter communication between two parties.
7. **Nefarious Activity/Abuse**
   a. *Identity theft:* Threat of identity theft action, e.g. using malware.
   b. *Receiving unsolicited E-mail:* Threat of receiving unsolicited email which affects information security and efficiency, e.g. spam, malware infected e-mails.

c. *Denial of service:* Threat of service unavailability due to massive requests for services, or protocol attacks, e.g. requests for access to network services from malicious clients, use of multiplication/amplification methods.

d. *Malicious code/ software/ activity:* Threat of malicious code or software execution, e.g. worms, Trojans, rootkits, virus, web exploits etc.

e. *Social Engineering:* Threat of social engineering type attacks that target to manipulate the behaviour of personnel, e.g. phishing attacks or spear phishing attacks.

f. *Abuse of Information Leakage:* Threat of leaking important information, e.g. through malware, infected web applications, and network traffic.

g. *Generation and use of rogue certificates:* Threat of use of rogue certificates, e.g. through exploitation of the web session control mechanism or fake OS updates.

h. *Manipulation of hardware and software:* Threat of unauthorised manipulation of hardware and software, e.g. use of anonymous proxies, cloud to launch attacks, 0-day vulnerabilities, and alternation of software.

i. *Manipulation of information:* Threat of intentional data manipulation to mislead information systems or somebody or to cover other nefarious activities (loss of integrity of information, e.g. DNS poisoning, falsification of record, autonomous System hijacking, address space hijacking etc.

j. *Misuse of audit tools:* Threat of nefarious actions performed using audit tools.

k. *Misuse of information/ information systems:* Threat of nefarious action due to misuse of information / information systems.

l. *Unauthorised activities:* Threat of nefarious action due to unauthorised activities, e.g. devices, systems, software, networks, data records.

m. *Unauthorised installation of software:* Installation of unwanted malware software

n. *Compromising confidential information:* Threat of data breach.

o. *Hoax:* Threat of loss of IT assets security due to cheating, e.g. false rumour and/or fake warning.

p. *Remote activity:* Threat of nefarious action by attacker remote activity, e.g. Remote Command Execution, Remote Access Tool (RAT), botnets.

q. *Targeted attacks:* Sophisticated, targeted attack which combine many attack techniques, e.g. malware, phishing, watering hole attacks.

r. *Failed business process:* Damage or loss of IT assets due to improperly executed business process.

s. *Brute force:* Unauthorised access via systematically checking all possible keys or passwords until the correct one is found.

t. *Abuse of authorisations:* Using authorised access to perform illegitimate actions.

8. Legal threats

a. *Violation of rules and regulations / Breach of legislation:* Threat of financial or legal penalty or loss of trust of customers and collaborators due to violation of law or regulations.

b. *Failure to meet contractual requirements:* Threat of financial penalty or loss of trust of customers and collaborators due to failure to meet contractual requirements.

c. *Unauthorised use of IPR protected resources:* Threat of financial or legal penalty or loss of trust of customers and collaborators due to improper/illegal use of IPR protected material.

d. *Abuse of personal data:* Threat of illegal use of personal data.

e. *Judiciary decisions/court order:* Threat of financial or legal penalty or loss of trust of customers and collaborators due to judiciary decisions/court order.

4.3.2 Assessing Security Threats for the Smart Sea Port

In the Smart Sea Port considered in 5G-MoNArch, there are three main services considered, as described in Table 4-3.

Table 4-3: The services of the Smart Sea Port scenario

Service	Description
Traffic Light Control (URLLC)	Traffic lights (static as well as mobile, e.g., in case of construction site along streets) connected through wireless links in a reliable and resilient way under consideration of data integrity
Video Surveillance (eMBB)	Video control entrance to sea port area or parts of it, up-to-date status information related to those areas; also, data integrity and security as important aspects
Sensor Measurements (mMTC)	Measurements about, e.g., environmental pollution on mobile barges connected through wireless terminals or at stationary locations

The analysis of threats and their possible countermeasures is of high importance in the considered scenario, since cyber-attacks or software bugs can have a large impart in operation of a critical infrastructure such as a sea port. According to Cerrudo [C15], simple bugs can cause big problems in critical infrastructures:

- May 2012 California: Placer County Courthouse system accidentally summoned 1,200 people to jury duty on the same morning causing traffic jam
- November 2013 Bay Area Rapid Transit (BART): major software glitch, service was shut down by a technical problem involving track switching, it affected 19 trains with about 500 to 1,000 passengers on board
- August 2003 Northeast: blackout, primary cause was a software bug in the alarm system at a control room of the FirstEnergy Corporation, 55 million people affected

Other related attacks described by Cerrudo [C15] include:

- Manipulate city management systems to send workers to dig a hole to wrong place (gas, water pipes).
- Manipulate sensors to fake seismic detection or flood detection.
- DDoS attacks can take services off line.
- Public transportation attack by influencing people behaviour, e.g. by displaying wrong information etc.
- Street lighting attack, i.e. black out big city area.

Another team, led by University of Michigan computer scientist J. Alex Halderman [GBH+14], found three major weaknesses in the traffic light system: Unencrypted wireless connections, the use of default usernames and passwords that could be found online, and a debugging port that is easy to attack. Using these vulnerabilities and a computer that can communicate at the same frequency as the intersection radios—in this case, 5.8 Gigahertz—the researchers could access the entire unencrypted network. It is noted that it took just one point of access to get into the whole system.

This section presents the relevant threats presented in Section 4.3.1 related to the Smart Sea Port and the services available in this testbed. Specifically, Table 4-4 presents a comprehensive list of the possible threats that can occur in the Smart Sea Port testbed, and assesses the possible impact in resilience, as well as in economy/business. Three levels of impact assessment are used for simplicity: low, medium, high. Additionally, Table 4-4 also presents possible countermeasures for each considered threat.

Table 4-4: The most likely threats in the Smart Sea Port scenario

Threat category	Threat	Impact in Resilience	Economic/Business Impact	Possible countermeasure
Physical attack	Sabotage / Vandalism / Theft	High: Damage/Lack of IT infrastructure can cause the services to be not available	High: Unavailability of URLLC/eMBB services can cause physical damages. Additionally, the cost of restoring the infrastructure	Use of security personnel to guard IT infrastructure
Disaster	Natural / environmental disaster / pollution, dust, corrosion / water	High: Damage in IT infrastructure can cause the services to be not available	High: Unavailability of URLLC/eMBB services can cause physical damages. Additionally, the cost of restoring the infrastructure	Use of waterproof, corrosion-proof solutions
Failure / Malfunction	Failure of devices or systems / Malfunction of equipment	Medium: Failure of IT infrastructure can cause the degradation of the quality of services	Medium: Cost to restore the failed infrastructure	Use of malfunction monitoring system for fast restoration
Eaves-dropping / Interception / Hijacking	War driving	Low: Detection of sensor/camera locations	Low: Detection of sensor/camera locations	-
	Interception of information / Man in the middle	Medium: Changing the content of the communication messages can cause degradation of the quality of services	Medium: Loss of sensitive information	Intrusion Detection Systems (IDS), Cryptography
	Replay of messages	Low: Does not affect resilience	Low: Can cause unwanted behaviour in the services (e.g. make all traffic lights in an intersection green)	-
Nefarious Activity / Abuse	Denial of service / botnets	High: Can cause a failure of a certain network function or multiple network functions running on affected machine	High: Unavailability of URLLC / eMBB services can cause physical damages.	Anomaly detection methods using data/control plane information
	Malicious code / software / activity (Privilege escalation / malware / rootkits / data tampering)	Medium: Privilege escalation can be used to deliberately cause the failure of network functions or hosts	High: Information leakage, unwanted behaviour in the services	Intrusion Detection Systems (IDS)

	Misuse of audit tools (port scanning)	Medium: hacker might use learned vulnerabilities to deliberately cause failure of network function or host/infrastructure	Low: Does not have Economic / Business Impact	Anomaly detection methods using data/control plane information
Outage	Network outage	High: Lack of network connectivity can cause the services to not be unavailable	High: Unavailability of URLLC / eMBB services can cause physical damages.	Use multiple means of network communication (e.g. cable and wireless)

The analysis presented in Table 4-4 gives also the rough indication of security threats relevance with respect to resilience issues. In general, the security threats might have as a target either the information (e.g. retrieving the confidential data) or the infrastructure (e.g. affecting the hardware/software operation). The latter category of security threats has very high impact to resilience as well, as the infrastructure on which the network is operating is directly endangered. The former category of security threats might potentially have an impact to the resilience in the case that retrieved data is used in order to prevent the normal functionality of the network infrastructure, e.g. by deliberately causing failure or malfunctioning of network functions or hosts.

4.3.3 Protection of the Hamburg Sea Port Testbed: Proposal of Security Trust Zones Deployment

In this section we illustrate how the slice-aware STZ deployment strategy could be applied to the particular case of the Hamburg Smart Sea Port testbed. Figure 4-5 depicts the logical setup of the testbed, showing key infrastructure elements such as i) mobile terminals (i.e. traffic lights, video cameras and barges on boats) at the top left corner; ii) the VPN that serves as entry point to the Hamburg (HH) tower; iii) geographically distributed servers in Hamburg, Munich or Nuremberg, right below the mobile terminals. Outside the VPN, there is a lifecycle management node in Munich, shown on the right bottom corner, and a switch connecting to application servers on the top right corner of the picture. In this setup, the infrastructure is distributed among various geographical locations and managed by different entities, that is Hamburg Port Authority (HPA), Nokia and Deutsche Telekom. Additionally, it is noticeable that a VPN is already put in place to access some of the infrastructure elements. All these factors would influence, among others, the selection of the appropriate STZ levels to deploy, as will be explained later in this section.

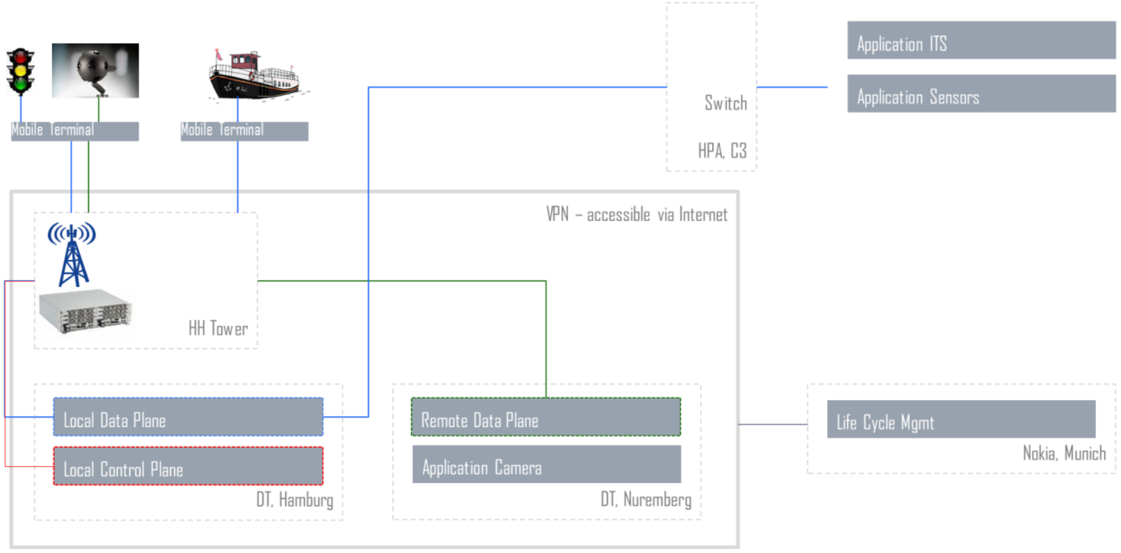


Figure 4-5: Hamburg Smart Sea Port testbed logical setup

As it has been already described in Section 4.2, there are different STZs defined in a target infrastructure. Each of them is associated with a STZ template, where the levels of security and trust to be provided in the specific logical area covered by a STZ are described. Three different dimension criteria (security, business and 5G infrastructure/services) have to be considered to identify which are these logical areas and the STZ templates that best match to the infrastructure where we are going to do the deployment.

Next, we analyse these criteria in the particular case of the Hamburg Smart Sea Port testbed by considering the logical setup available in Figure 4-5. The preliminary results of such analysis on the security, business and 5G infrastructure and services level, are as follows.

Security

In a first approach to identify the *criticality of the assets to be protected* and its connectivity (i.e., whether they are isolated from some parts of the infrastructure or exposed to the public internet), we could differentiate between the mobile terminals which are outside the sea port VPN, the HH Tower (which is the input to the mobile communications received), the servers inside the VPN and accessible via internet (Local Data and Control Planes as well as the Remote Data Plane and the Application Camera), and the servers outside the VPN (Life Cycle Management and the HPA Applications connected through a Switch). In addition, since the HH Tower is the input to the VPN from the mobile devices, this asset should have a higher security level than other servers (e.g. the Application Camera).

Business

Depending on the information stored and processed by the different servers (e.g. Local and Remote Data Planes), the *privacy and integrity levels can change* and the security procedures to be applied e.g. to maintain compliance with specific regulation or legislation application and consequently they could be included in different STZs. For example, the Remote Data Plane could be in a different STZ than the Application Camera. On the other hand, depending on the corporate security policies we could need to use different STZ on the servers located in DT than the servers located in Nokia to deploy specific or additional prevention or reaction mechanisms.

5G infrastructure and services

There are four different geographical locations in the testbed: Hamburg, Nuremberg and Munich. However, based on the available information, it is not clear if the network slices are defined geographically or there is a network slice through different locations to associate a STZ to a specific one. Nevertheless, if we consider that connectivity of the mobile terminals such as the pollution sensors on board a boat is more susceptible to be disconnected, the STZ including these mobile networks should have more *self-healing capabilities to work autonomously* during the time they are isolated.

The resources available is also something to be considered when defining the STZs, since the sensors deployed to detect and prevent security incidents as well as the reaction mechanisms would be different for each of them. For example, we could deploy a host intrusion detection system (HIDS), such as OSSEC, to detect attacks in the servers with Local Data/Control Planes at application level. However, the deployment of this type of sensor makes no sense to, for instance, monitor accesses to a Radio Access Network, where a sensor to detect anomalies in mobile networks is more suitable. In the same way, a NIDS such as Snort could be useful to detect intrusions analysing the network traffic in the connections to the DT VPN or to the Switch that gives access to Central NW premises.

In summary, it can be concluded that it is not a unique and easy way to identify and determine the STZs for a target infrastructure. Instead, a deep analysis of the different profiling criteria for the specific use case is required, in order to accomplish this task. Nonetheless, in Figure 4-6 we illustrate a possibility of distribution of STZs for the Hamburg Sea Port testbed, considering the criteria commented above. Each blue box represents a different STZ and the number identifies a type of STZ template. In this example, we could have:

- a STZ (STZ1) for the mobile terminals that would need specific anomaly detectors for mobile networks;
- two STZs located geographically separated and in different network slices but with the same template (STZ2) for the HH Tower and the Switch, assuming the criticality of these assets and

their risk assessment is the same since they are exposed and vulnerable incoming points to VPN infrastructure;

- two STZs with different templates for the Local Data/Control Plane (STZ3) and Remote Data Plane (STZ4), assuming that although the detection mechanisms applicable can be the same due to the similar nature of the assets to be protected, they can have different privacy levels and can require the application of different prevention/reaction mechanisms to be compliant with the regulations. Besides, we are assuming that it is higher than the privacy levels required for the Application Camera;
- two STZs with a same template (STZ5), one for the Application Camera in DT premises and another for the Application ITS and Application Sensors in HPA, assuming that the privacy and security levels for these applications is the same and the same detection/reaction/mechanisms can be deployed in both organisations and network slices;
- the infrastructure in Nokia located in Munich could be in a separate STZ with a different template (STZ6), assuming they have different corporate policies.

In each STZ, according to its template, it would be deployed one or the three types of sub-components of the security threats monitoring to perform security threat detection (SthD), prevention (SthP) and/or reaction (SthR). In this example, we are assuming there are two network slices defined in the Hamburg Sea Port testbed (one for the mobile terminals and the assets in the VPN and the other for the assets outside the VPN). As it is reflected in Figure 4-6, for each network slice (referred to as NS1 and NS2, respectively) we would have a different Security Monitoring Manager (SMm) instance deployed at the *Controller Layer* to interact with the security monitoring components deployed in the different STZs of a same network slice. An instance of the Threat Intelligence Exchange component (ThIntEx) would be also deployed at this level to share security information between the different STZs. Finally, a Security Trust Zones Manager (STZm) and a ThIntEx would be required at the XSC Controller Layer for orchestration and management of the different network slices.

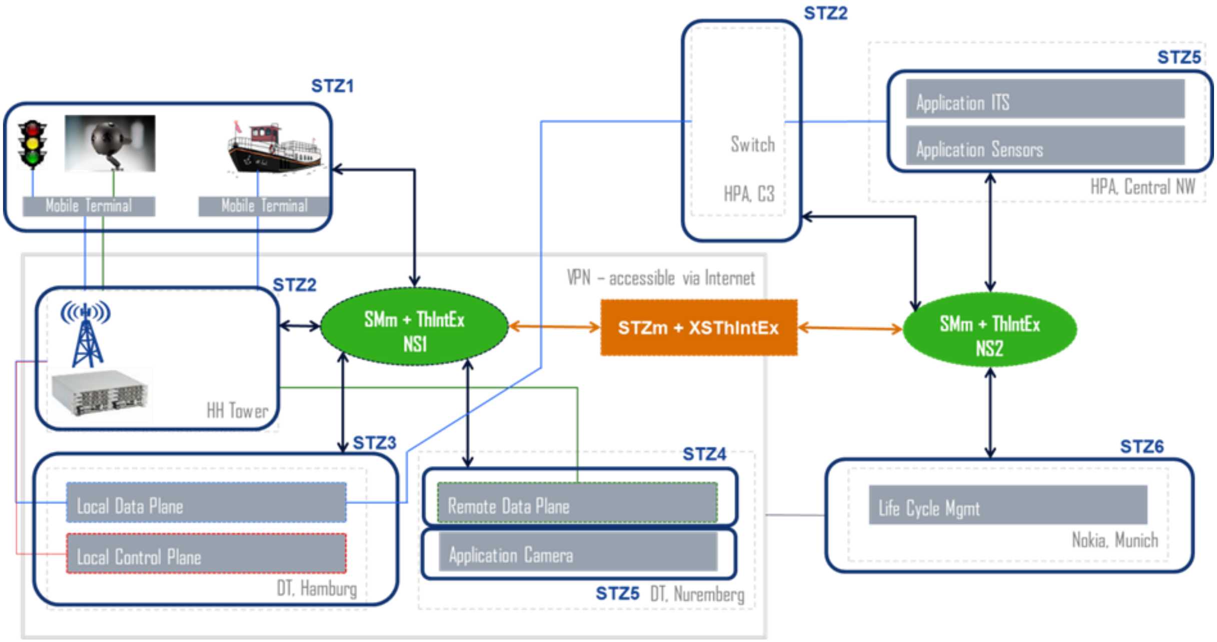


Figure 4-6: Example of slice-aware STZ deployment for Hamburg Smart Sea Port testbed

5 Joint Study of Resilience and Security

Security and resilience are two related concepts with mutual effect on one another. In particular, a large number of security-related threats can affect to different extent the resilience and functionality of the network fault management. For example, the DoS attack can result in an unavailability of machines and network functions running on top of affected machines. Such effect will be detected by the network fault management which will attempt to solve such issues using its restoration capabilities. If redundant machines, network functions and links are available, the security threat might be mitigated using the existing redundancy. However, this may lead to lowering the current redundancy and consequently the resilience level of the network.

Depending on the actual service and agreed SLAs with the network tenant as well as the actual severity of the security threat, this relaxed resilience level might or might not be acceptable. In certain cases, lowering the resilience level for handling the threat might be unacceptable. This is true, for example, in situations where the security threat is assessed to be minor and does not jeopardise the normal network operation, whereas redundancy needs to be kept at the certain level due to risks of software and hardware problems. In such a case security might be compromised for achieving the required resilience/redundancy. On the other hand, as certain security threats might result in severe problems in the functionality of individual network functions or the network as a whole, handling such threats might have highest priority, even at the cost of lowering the current redundancy/resilience level. In such a case, the resilience might be compromised for security.

In general, measures need to be put in place to guarantee that a *certain degree of resilience could not pose new threats* or attack paths that could be exploited with malicious purposes. Duplicating network resources to ensure availability of a service operation could give attackers another entry point to the system, if such resources are not properly secured. Nevertheless, the solution may not be as straightforward as simply duplicating the security as well, i.e., applying the same security mechanisms to the duplicated network branch. On the contrary, it requires reconsidering the resilience as well as the security strategy of the system as a whole, whereby including the duplicated network branches and any other plausible resilience mechanism.

5.1 Resilience-Security Trade-off Process

One of the paramount characteristics of 5G infrastructures is the dynamic allocation of resources depending on the application domain. Given the heterogeneity of the application domain where the 5G infrastructure can operate, the security and resilience requirements might also differ a lot between them. While for some critical applications the security requirements of the service provided are very high, for others it is more important to give priority to the resilience, regardless of the security threats. Additionally, if the cost of responding to a security incident is very high with respect to the impact over the infrastructure, it might be more convenient to deal with the consequence of the ongoing incident. 5G-MoNArch is capable of deciding whether resilience prevails against security or vice versa.

This decision is made upon a *trade-off evaluation process*, which uses information from the targeted infrastructure (received from Management and Orchestration layer and from the service layer), which is correlated with the security incidents detected and with the available counter measures designed to react to them. The *trade-off evaluation is carried out at slice level* with the SMm analysing the severity of the incidents detected. The resilience levels are configured at the SMs based on the importance of the assets deployed within the slice. This information is inferred based on information received from the Management and Orchestration layer.

Figure 5-1 represents the process for the trade-off evaluation process.

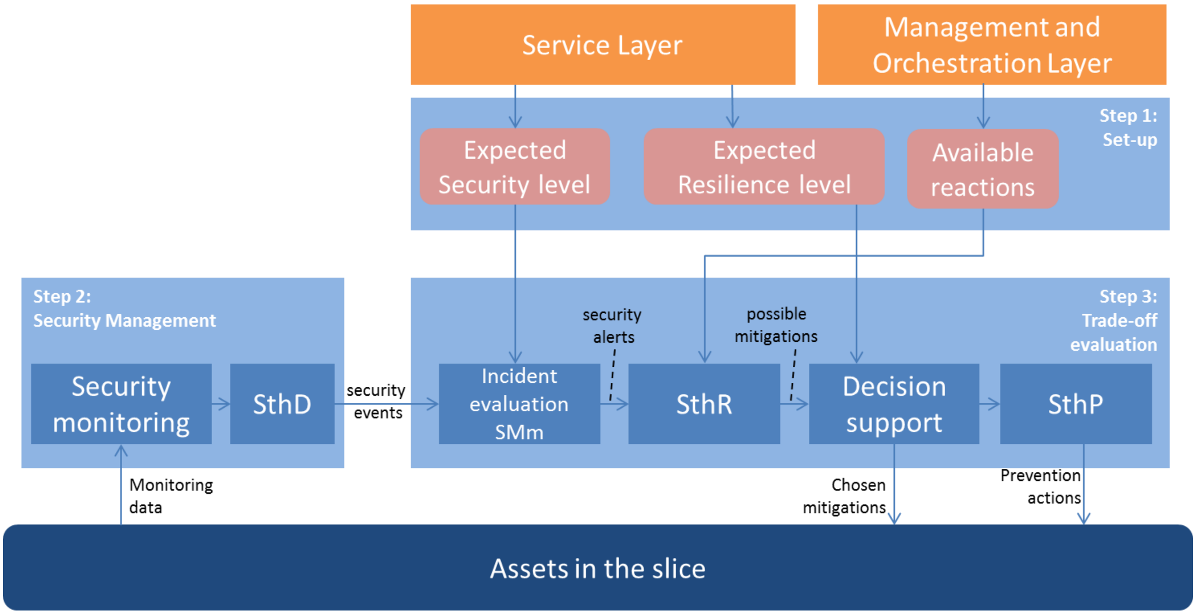


Figure 5-1: Trade-off between resilience and security

The complete process is completed in three steps:

- *Step 1: Set-up*. This step sets up the trade-off evaluation, and will establish the thresholds that will determine whether a mitigation action should be enforced or not, depending on the resilience and security requirements obtained from the Service Layer. Security and resilience levels are built upon several levels, which allows not only to align the predefined levels for resilience and security but also to makes the trade-off evaluation easier and more accurate. These requirements are retrieved from different parts of the infrastructure:
    - The *Expected Security level* is obtained by evaluating the importance of the assets deployed within a slice. In general, expert knowledge will determine the criticality of the individual assets running within the slice, which is received through the service layer. The aggregation of the individual criticality levels will determine the global security level required for the slice. A good approximation for the classification of security levels can be adapted from the IoT Security Compliance Framework as follows [IOT-SCF] (see Table 5-1):

Table 5-1: Security levels according to the impact of incidents against devices (adapted from IoT Security Compliance Framework)

Class (Security Level)	Description of the security level
0	In case of incident the infrastructure must be able to deal at least with incidents that involve data generated or controlled in the event of a security breach that have the potential to affect critical infrastructure or cause personal injury.
1	In addition to level 0, in case of incident the infrastructure must be able to protect sensitive data including sensitive personal data.
2	In addition to level 1, the infrastructure must be able to resist attacks on availability that would have significant impact on an individual or organisation, or impact many individuals.
3	In addition to level 2, the infrastructure must be able to deal with incidents that involved data generated or controlled by the device which results in no more than a limited impact on an individual or organisation.

4	In addition to level 3, the infrastructure must be able to deal with all type of incidents that involve data generated or controlled by devices, even if they result in little discernible impact on an individual or organisation.

- The *Expected Resilience level* is obtained from the requirements of the slice and requirements of individual network functions especially in terms of their criticality/importance for supporting E2E network slice, see Table 5-2.

Table 5-2: Resilience levels

Resilience levels	**Description of the Resilience level**
1	**Best effort resilience**: The resilience mechanisms need to be able to detect, isolate and mitigate the root cause of the network problem. The resilience mechanisms need to guarantee healing of affected network function under certain time constraints. This might affect the E2E service quality perception in an arbitrary manner. There is no guarantee on possible impact to the E2E service.
2	**Custom resilience**: The resilience mechanisms need to be able to detect, isolate and mitigate the root cause of the network problem such that the resulting effect on the E2E service is within agreed/guaranteed framework. E.g. the E2E service can be down x amount of time over certain period.
3	**High resilience**: The resilience mechanisms need to be able to detect, isolate and mitigate the root cause of the network problem in such a way that **no impact** on the E2E service quality perception is perceived

- *Step 2: Security Management*. Although this step is part of the *security monitoring* process, it is also required for the trade-off evaluation between resilience and security. Assets deployed at the STZs are monitored by the SthD of every STZ. The collected events are received by the SMm which are correlated looking for potential incidents. *Security Threat Detection (SthD)* is carried out at slice level although the severity of the incidents might be different depending on the importance of the assets deployed within single STZs. We consider the severity as a score given to an alarm that has been generated by correlating security events gathered from the STZ. The score depends on several factors, such as the importance of the assets deployed in the network, the rules configured from the service layer according to the type of events received or the frequency of the events received.
- Step 3: *Trade-off evaluation*. This step comprises four activities: *Incident Evaluation, Security Threat Reaction, Decision Support and, the Security Threat Prevention*.
  - The *Incident Evaluation* uses the expected security level set-up done during Step 1. It is worth noticing that the evaluation of the resilience-security trade-off is carried out at the slice level. However, as pointed out in Step 2, the severity of the incident is determined at STZ level, which increases the accuracy of the evaluation. For example, the same incident over the same type of asset might produce an alarm with different severity because the incident is targeting different STZs with different expected security levels. This is important as long as the final result of the trade-off evaluation will depend on the severity of the alarms generated in every STZ and the severity of the alarms will depend on the security level required in every STZ.
  - The *Security Threat Reaction* (SthR) uses the alarms (along with the level of severity for each alarm) to determine the possible reactions to mitigate the detected incidents. The SthR receives support from the Management and Orchestration Layer with details about the available reaction capabilities (i.e., traffic redirection with the deployment of virtual firewalls or instantiation of virtual honeynets). Additional information needs to be associated to the available reactions, such as the incident that the reaction is able to mitigate (along with its severity level) and the resources or cost associated to the enforcement of such reaction.

- The *Decision Support* is in charge of evaluating whether it is more convenient to mitigate the incident detected or to keep prevailing the resilience of the network. To this end, the Decision Support carries out a risk assessment which evaluate aspects such as the severity of the incident to mitigate, the cost of enforcing the mitigation and the resilience level expected of the slice. The result will determine whether the final recommendation is to deal with the incident (prevailing the resilience) or to enforce the mitigation actions to react to the incident. Figure 5-2 represents a possible schema for the decision on mitigating or not a security incident.
- The *Security Threat Prevention,* in charge of proposing prevention actions based on previous incidents, would receive such actions from the MANO, and would decide on the most suitable prevention action to be shared, for example, with other STZs or even other slices.

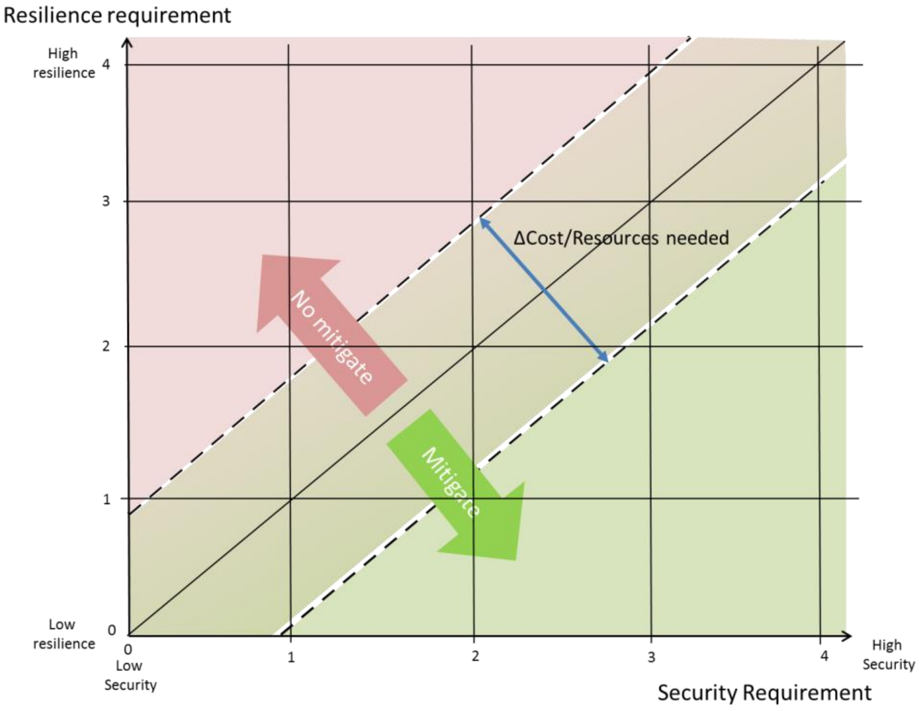


Figure 5-2: Decision criteria for resilience vs security trade-off

In case very high resilience requirements are present in conjunction with low security requirements, this will definitely drive the reaction into the no mitigation of the incident (pink area of Figure 5-2). In contrast, high security requirements with respect to low resilience requirement will result in mitigating the incident (green area in Figure 5-2). The area in between pink and green represents the not clear decision. In this case, *the cost or resources needed to mitigate the incident would be the criteria used for the decision*. The bisector from the origin of the coordinates axis represents the threshold between the mitigation or not mitigation of the incident. The concrete position of the threshold depends on the cost or resources used willing to be used for mitigating security incidents, which is configured at the service layer.

Additional elements such as the existence of SLAs can modify the requirements, both for security and resilience. The potential violation of the SLA might affect to some extent the trade-off analysis, either towards the mitigation or not of the security incident detected. Adding the SLA to the analysis entails additional implications that would require to be tackled. For example, if the trade-off analysis suggests ignoring the incident, it is possible that the SLA is violated, which might entail potential penalties to the network operator, and therefore additional costs, contracts termination, etc. Therefore, the implications of adding SLAs to the equation are higher and can entail implications that go beyond purely technical aspects.

6 Conclusions and Future Work

This document presented the preliminary work conducted in WP3 of the 5G-MoNArch project, which aims at designing the necessary network functions that provide an end-to-end failsafe and secure mobile network operation. The design of these functions is made in a way that such operation can be maintained even in situations where the radio link quality is not adequate to provide the desired performance.

We first analysed the state of the art, including 5G PPP Phase I working groups as well as the developments in SDOs. This facilitates assessing to what extent existing developments in RAN reliability, resilience of telco clouds and security would be able to deal with the specific challenges that 5G infrastructures and services bring. We provided a description of how 5G challenges are going to be addressed from three different viewpoints: i) RAN reliability, ii) telco cloud resilience, and iii) security monitoring activities.

We further proposed an extension to the baseline 5G-MoNArch architecture described in [5GM-D2.1] and [5GM-D2.2]. In particular, we highlighted the set of specific network functions, along with the enhancements to the controller, as well as to the management & orchestration layers. This permits the deployment and operation of network slices equipped with the corresponding resilience and security functional innovations. Such architectural extension is illustrated in Figure 6-1. Note that Figure 6-1 depicts a combined view of Figure 2-9, Figure 3-9, and Figure 4-4; that is, Figure 6-1 provides the aggregated picture that summarises the developments analysed in Chapters 2, 3 and 4.

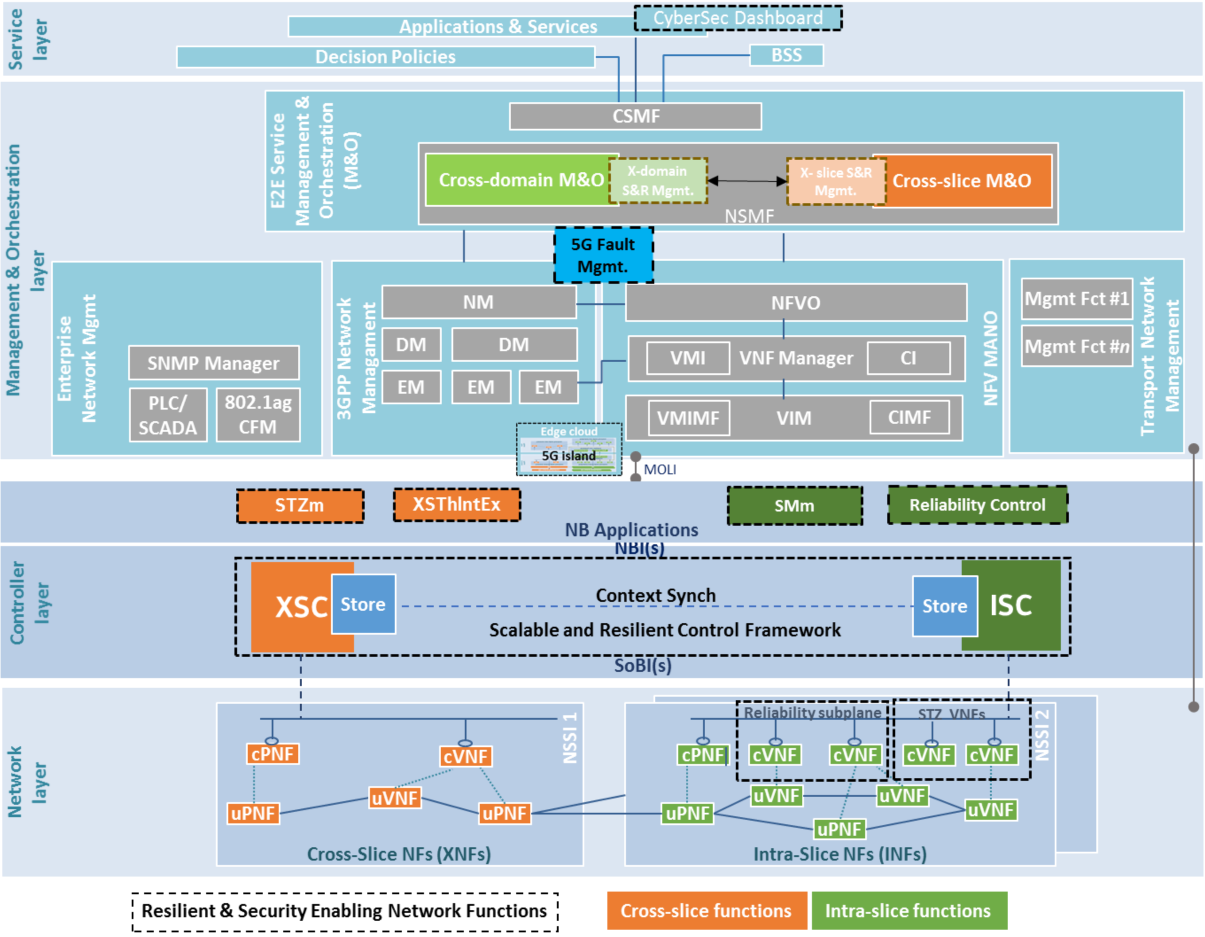


Figure 6-1: 5G-MoNARch architecture enriched with the necessary elements to enable resilience and security functional innovations

With reference to Figure 6-1, a summary of the corresponding contributions in each architectural layer, namely the network layer; controller layer; MANO layer; service layer, is as follows.

- **Network layer**:
    - o *Reliability sub plane*: The term "reliability sub plane" is used here to indicate that the additional functions considered within the framework of WP3 of 5G-MoNArch are taken from a pool of specially designed functions to meet the requirements on RAN reliability. In particular, a RAN reliability function is introduced, that can either be a data duplication or a network coding function. It mainly serves as a user plane functionality that processes data according to the principles described in Sections 2.1 and 2.2.
    - o *STZ VNFs*: This block represents the set of VNFs which is instantiated depending on the template, referred to as STZ template in Chapter 4, that is applied. Specifically, such VNFs include i) security threat detection and protection; ii) reaction to security threats; iii) threat intelligence exchange NF.
- **Controller layer**:
    - o The intra-slice Controller (ISC) [5GM-D2.1] is enhanced in this framework with two applications: i) a *Reliability Control* application responsible for RAN reliability functionalities, and ii) an application for controlling the security monitoring functions operating within a network slice (labelled "*SMm*" in Figure 6-1, which stands for security monitoring management).
    - o The inter-slice Controller (XSC) [5GM-D2.1] is enhanced with a control application (labelled "*STZm*" in Figure 6-1, denoting security trust zone management) that coordinates the activity of Security Trust Zones deployed across network slices within the same application domain. This element is in charge of avoiding propagation of threats across slices.
    - o Both the developments in ISC and XSC are incorporated within a block labelled *scalable and resilient controller framework*, which is used here to emphasise the conceptual relation of such blocks from the network resilience perspective.
    - o The controller framework (ISCs and XSCs) operates in the cluster mode for supporting both Scalability and Resiliency. Data store (labelled as *Store* in Figure 6-1) is an internal and distributed data base available in each controller node for maintaining the overall topology state of the network. Data Store has a dedicated interface for *context synchronisation* and state management.
    - o Security protection elements are included in the controller layer to i) support the deployment of security trust zones adapted to specific security levels demanded by network slices (executed by the *Security Threat Zone manager –STZm-*) and ii) to coordinate the exchange of security information among network slices (executed by the *Cross-Slice Threat Intelligence Exchange –XSThIntEx-*).
- **Management & Orchestration layer:**
    - o The *5G Fault Management function*, which takes into account slice requirements and performs joint handling of fault events originating from different deployment layers, e.g. functional, physical, virtual, is introduced in this framework.
    - o The Cross-slice M&O function responsible for inter-slice management incorporates the *x-slice Security & Resilience Management* function. This specialised function captures the modules developed in WP3 of 5G-MoNArch specialised for addressing jointly the security and resilience considerations across slices.
    - o In a similar manner, the Cross-domain M&O function incorporates the functionality for joint dealing with security and resilience issues within the same slice yet across different domains. This specialised block, developed in WP3 of 5G-MoNArch, is labelled *x-domain Security & Resilience Management* in Figure 6-1.

  - The concept of *5G islands*, which refers to the autonomous operation of parts of the network towards a higher resilience level is included in the Management & Orchestration layer, is also introduced in this framework.
- **Service layer**:
  - The *CyberSecurity dashboard* is a cybersecurity data analytics service to provide end-users with visualisation and awareness of the security status at different levels, tailored to the needs and interests of the different end-user roles involved in the security management. The aim is to provide a high-level view of the overall security status for the management board of the customer (e.g. HPA), a specific view for each customer's business service (i.e. corresponding to each network slice), and a view of the security status of the operated infrastructure.

Overall, this document presented the initial analysis and first results on resilience and security. Such analysis and first results were obtained as an outcome of the work conducted in the framework of WP3 of the 5G-MoNARch project, during the first eleven months of its execution time. Plans for the future extension include a refined view of each of the architectural elements, including evaluation results where applicable, along with a description of their interaction with the remaining architecture components.

7 References

[3GPP-23.101] 3GPP TS 23.101, “General Universal Mobile Telecommunications System (UMTS) architecture,” v14.0.0, March 2017.

[3GPP-33.401] 3GPP TS 33.401, “3GPP System Architecture Evolution (SAE); Security architecture,” v15.1.0, Sept 2017.

[3GPP-38.323] 3GPP TS 38.323, “NR; Packet Data Convergence Protocol (PDCP) specification (Release 15),” v2.0.0, Dec 2017.

[3GPP-38.801] 3GPP TR 38.801 “Study on new radio access technology: Radio access architecture and interfaces (Release 14),” March 2017

[3GPP-36.808] 3GPP TR 36.808, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Carrier Aggregation; Base Station (BS) radio transmission and reception (Release 10),” July 2013

[3GPP-36.842] 3GPP TR 36.842, “Study on Small Cell Enhancements for {E-UTRA} and {E-UTRAN} – Higher layer aspects (Release 12),” Sep 2014

[3GPP2018] 3GPP TR 28.801 Release 15 (2018-02), “Study on management and orchestration of network slicing for next generation network”

[5GM-D2.1] 5G-MoNArch Deliverable D2.1, “Baseline architecture based on 5G-PPP Phase 1 results and gap analysis,” October 2017

[5GM-D2.2] 5G-MoNArch Deliverable D2.2, “Initial overall architecture and concepts for enabling innovations,” June 2018

[5GM-D6.1] 5G-MoNArch Deliverable D6.1, “Documentation of Requirements and KPIs and Definition of Suitable Evaluation Criteria,” September 2017

[5GPPP17] 5G PPP Architecture White Paper, “View on 5G Architecture v2.0” December 2017

[5GPPPSEC17] 5GPPP White Paper, “5GPPP Phase1 Security Landscape,” June 2017

[ACL+00] Ahlswede, R., Cai, N., Li, S. Y., & Yeung, R. W., “Network information flow,” IEEE Transactions on information theory, No. 46(4), pp. 1204-121, 2000

[AD13] Alexandrov T, Dimov A, “Software availability in the cloud,” in Proc. 14th ACM International Conference on Computer Systems and Technologies, pp 193–200, 2013

[AD13] Report available at: http://www.availabilitydigest.com/public_articles/0803/outage_causes.pdf

[AG14] O. H. Abdelrahman and E. Gelenbe, “Signalling storms in 3G mobile networks,” in Proc. IEEE International Conference on Communications, pp. 1017–1022, 2014

[AHY11] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, “Contributions to the study of SMS spam filtering: New collection and results,” in Proc. 11th ACM Symp. Document Eng., pp. 259–262, 2011

[AMF16] “Availability Management Framework,” online available at http://devel.opensaf.org/SAI-AISAMF-B.04.01.AL.pdf. Accessed Oct 2016

[BHS+08] A. Bose, X. Hu, K. G. Shin, and T. Park, “Behavioural detection of malware on mobile handsets,” in Proc. 6th Int. Conf. Mobile Syst., Appl., Serv., pp. 225–238, 2008

[BKK+16] L. Bodrog, M. Kajo, S. Kocsis, B. Schultz, “A Robust Algorithm for Anomaly Detection in Mobile Networks,” in IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC): Workshop: 6th International Workshop on Self-Organising Networks (IWSON), 2016

[BPV09] Boudreau, G., Panicker, J., & Vrzic, S., “Interference coordination and cancellation for 4G networks,” IEEE Communications Magazine, No. 47(4), pp. 74–81, 2009

[C15] C. Cerrudo, “Hacking smart cities,” In Pro. RSA Conference, pp. 2-18, 2015

[CS15] G.S. Chhabra, P. Singh, “Distributed Network Forensics Framework: A Systematic Review,” International Journal of Computer Applications, 2015

[D-D2.1] Susana Gonzalez Zarzosa et. Al, “DISIEM- D2.1 In-depth analysis of SIEMs extensibility,” February 2017.

[D-D4.1] Pedro M. Ferreira et Al., “D4.1 Techniques and tools for OSINT-based threat analysis,” August 2017.

[DBG12] S. J. Delany, M. Buckley, and D. Greene, “SMS spam filtering: Methods and data,” Expert Syst. Appl., vol. 39, no. 10, pp. 9899– 9908, 2012.

[E15] ENISA, “ENISA Threat Taxonomy: A tool for structuring threat information”, January 2016

[ETM+05] W. Enck, P. Traynor, P. McDaniel, and T. La Porta, “Exploiting open functionality in SMS-capable cellular networks,” in Proc. 12th ACM Conf. Comput. Commun. Security, Alexandria, VA, USA, pp. 393–404, 2005.

[FCS+17] E. Falk, R. Camino, R. State, V. K. Gurbani, “On non-parametric models for detecting outages in the mobile network,” in IFIP/IEEE Symposium on Integrated Network and Service Management (IM), 2017

[GAP+15] G. Gorbil, O. H. Abdelrahman, M. Pavloski, and E. Gelenbe, “Modeling and analysis of Rrc-based signalling storms in 3g networks,” IEEE Trans. Emerging Topics Comput., vol. PP, no. 99, p. 1, 2015

[GBB+14] P. Gogoi, D.K. Bhattacharyya, B. Borah, J. K. Kalita, “MLH-IDS: A Multi-Level Hybrid Intrusion Detection Method,” The Computer Journal, vol. 57, issue 4, pp 602–623, 2014

[GBC+17] Brianna Gammons, “6 Must-Know Cybersecurity Statistics for 2017,” Barkly Blog, Jan 2017, online available at https://blog.barkly.com/cyber-security-statistics-2017

[GBH+14] B. Ghena, W. Beyer, A. Hillaker, J. Pevarnek, J.A. Halderman, “Green Lights Forever: Analyzing the Security of Traffic Infrastructure,” In Proc. 8th USENIX Workshop on Offensive Technologies (WOOT’ 14), pp.7-7, 2014

[GBT17] A. Grigory, P. Barnabé and G. Thomson, “Digital Vision for Cybersecurity,” Atos Whitepaper, September 2017, online available at: https://atos.net/content/dam/uk/white-paper/digital-vision-cyber-securityopinion-paper-new.pdf

[GJJ17] A. Gupta, R. K. Jha, S. Jain, “Attack modeling and intrusion detection system for 5G wireless communication network,” in International Journal of Communication Systems, vol. 30, issue 10, 2017

[GKM+17] V. K. Gurbani, D. Kushnir, V. Mendiratta, C. Phadke, E. Falk, R. State, “Detecting and predicting outages in mobile networks with log data,” In Proc. IEEE International Conference on Communications (ICC), 2017

[GRT+16] A.Gopalasingham, L. Roullet, N. Trabelsi, C. S. Chen, A. Hebbar, and E. Bizouarn, “Generalised Software Defined Network Platform for Radio Access Networks,” In Proc. IEEE Consumer Communications and Networking Conference (CCNC), Jan 2016, LasVegas, United States. 2016.

[HSS12] S. Hämäläinen, H. Sanneck, C. Sartori, “LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency,” John Wiley and Sons, 2012.

[HWM+S17] B. Han, S. Wong, C. Mannweiler, M. Dohler, H. D. Schotten, “Security Trust Zone in 5G Networks,” In Proc. 24th International Conference on Telecommunications (ICT), May 2017.

[IETF-RFC2881] https://www.ietf.org/rfc/rfc2281.txt

[JBW10] H. Jie, H. Bei, and P. Wenjing, “A Bayesian approach for text filter on 3G network,” in Proc. 6th Int. Conf. Wireless Commun. Netw. Mobile Comput. pp. 1–5, 2010.

[KMP13] E. K. Kim, P. McDaniel, and T. La Porta, “A detection mechanism for SMS flooding attacks in cellular networks,” in Proc. 8th Int. ICST Conf. Security Privacy Commun. Netw., pp. 76–93, 2013.

[KVT12] Khirallah, C., Vukobratovic, D., & Thompson, J., “Performance analysis and energy efficiency of random network coding in LTE-advanced,” IEEE Transactions on Wireless Communications, Vol. 11(12), pp. 4275–4285

[LSC+12] Lee, D., Seo, H., Clerckx, B., Hardouin, E., Mazzarese, D., Nagata, S., & Sayana, K., “Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges,” IEEE Communications Magazine, Vol. 50(2), pp. 148–155, 2012

[LYZ+09] L. Liu, G. Yan, X. Zhang, and S. Chen, “Virusmeter: Preventing your cellphone from spies,” in Proc. 12th Int. Symp. Recent Adv. Intrusion Detection, pp. 244–264, 2009.

[MBQ+18] P. Marsch, Ö. Bulakci, O. Queseth, M. Boldi (editors), “5G System Design: Architectural and Functional Considerations and Long Term Research,” John Wiley & Sons Ltd, 2018.

[MDS+17] D. S. Michalopoulos, M. Doll, V. Sciancalepore, D. Bega, P. Schneider, and P. Rost, “Network Slicing via Flexible Function Decomposition and Flexible Network Design,” In Proc. IEEE Personal, Indoor, and Mobile Radio Communications Conference (PIMRC), Workshop on New Radio Technologies, Oct 2017 .

[MGG+17] Diomidis Michalopoulos, Borislava Gajic, Beatriz Gallego-Nicasio Crespo, Aravinthan Gopalasingham, Jakob Belschner, “Network Resilience in Virtualised Architectures”, In Proc. 2017 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL), 2017.

[MDM2016] S. Mwanje, G. Decarreau, C. Mannweiler, M. Naseer-ul-Islam, C. Schmelz, “Network Management automation in 5G: Challenges and opportunities,” In Proc. PIMRC, September 2016 Spain

[MHS15] M. Miyazawa, M. Hayashi, R Standler, “vNMF: Distributed Fault Detection using Clustering Approach for Network Function Virtualisation,” In Proc. IFIP/IEEE International Symposium on Integrated Network Management (IM), Ottawa, Canada, 2015

[MII-D24] ICT-671680 METIS-II, Deliverable D2.4, “Final Overall 5G RAN Design,” June 2017.

[MII-D52] ICT-671680 METIS-II, Deliverable D5.2, “Final Considerations on Synchronous Control Functions and Agile Resource Management for 5G,” March 2017.

[MJ13] I. Murynets and R. P. Jover, “Anomaly detection in cellular machine-to-machine communications,” in Proc. IEEE Int. Conf. Commun., pp. 2138–2143, 2013.

[NS12] S. Novaczki, P. Szilagyi, “An Improved Anomaly Detection and Diagnosis Framework for Mobile Network Operators,” demo presented at the Second International Workshop on Self-Organising Networks, Paris, 2012.

[N13] S. Novaczki, “An Intelligent Anomaly Detection and Diagnosis Assistant for Mobile Network Operators,” In Proc. 9th International Conference on the Design of Reliable Communication Networks (DRCN), Budapest, Hungary, June 2013

[ODG+05] I. Ouachani, P. Duhamel, K. Gosse, S. Rouquette-Leveil and D. Bateman, "Macro-diversity versus micro-diversity system capacity with realistic receiver RFFE model,” In Proc. IEEE 6th Workshop on Signal Processing Advances in Wireless Communications, pp. 865-869, 2005

[ODL] OpenDayLight https://www.opendaylight.org

[ONOS] ONOS is building a better network, https://onosproject.org

[OO14] D. Ongaro and J. Ousterhout, “In search of an understandable consensus algorithm,” In Proc. of the 2014 USENIX conference on USENIX Annual Technical Conference (USENIX ATC'14), USENIX Association, Berkeley, CA, USA, 305-320.

[PDD+15] S. Papadopoulos, A. Drosou, N. Dimitriou, O.H. Abdelrahman, G. Gorbil, and D. Tzovaras, “A BRPCA Based Approach for Anomaly Detection in Mobile Networks,” In Proc. 30th International Symposium on Computer and Information Sciences (ISCIS), Sep, 2015.

[PDT16] S. Papadopoulos, A. Drosou, D. Tzovaras, “A Novel Graph-based Descriptor for the Detection of Billing-related Anomalies in Cellular Mobile Networks,” IEEE Transactions on Mobile Computing, Vol: 15, No: 11, pp: 2655-2668, Nov 2016.

[PSL+15] Pocovi, G., Soret, B., Lauridsen, M., Pedersen, K. I., & Mogensen, P., “Signal Quality Outage Analysis for Ultra-Reliable Communications in Cellular Networks,” In Proc. 2015 IEEE Globecom Workshops (GC Wkshps) (pp. 1–6), 2015

[S15] S. H. Sandra, “Design and deployment of secure, robust, and resilient SDN Controllers,” In Proc. IEEE 1st Conference on Network Softwarisation (NetSoft), 2015

[SMS+03] K. Shanmugasundaram, N. Memon, A. Savant and H. Bronnimann, “ForNet: A distributed forensics network,” In Proc. International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security, 2003.

[TLO+09] P. Traynor, M. Lin, M. Ongtang, V. Rao, T. Jaeger, P. McDaniel, and T. La Porta, "On cellular botnets: Measuring the impact of malicious devices on a cellular network core," in Proc. 16th ACM Conf. Comput. Commun. Security, pp. 223–234, 2009.

[V17] Verizon Threat Research Advisory Center, Data Breach Digest, "Perspective is Reality," Verizon Cybercrime Case Studies, online available at http://www.verizonenterprise.com/verizon-insights-lab/data-breach-digest/2017/

[VML+18] I. Viering, H. Martikainen, A. Lobinger, and B. Wegmann, "Zero-Zero Mobility: Intra-Frequency Handovers with Zero Interruption and Zero Failures," IEEE Communications Magazine, Mar-Apr 2018

[VTD+18] D. Vukobratovic, A. Tassi, S. Delic, and C. Khirallah, "Random Linear Network Coding for 5G Mobile Video Delivery," Information (submitted. Available: https://arxiv.org/pdf/1802.04873), 9(4), 72, 2018

[WC16] J. Wen and X. W. Chang, "A Linearithmic time algorithm for shortest vector problem in compute-and-forward design," 2016.

[XXY+12] Q. Xu, E. W. Xiang, Q. Yang, J. Du, and J. Zhong, "SMS spam detection using noncontent features," In Proc. IEEE Intell. Syst., vol. 27, no. 6, pp. 44–51, 2012.

[Y17] F. Z. Yousaf et al, "Network slicing with flexible mobility and QoS/QoE support for 5G Networks," In Proc. IEEE International Conference on Communications Workshops (ICC Workshops), Paris, France, 2017

[YKG+11] K. Yadav, P. Kumaraguru, A. Goyal, A. Gupta, and V. Naik, "Smsassassin: Crowdsourcing driven mobile-based system for SMS spam filtering," in Proc. 12th Workshop Mobile Comput. Syst. Appl (NetCod'09), pp. 1–6, 2011.

[GT09] Georgiadis, Leonidas, and Leandros Tassiulas, "Broadcast erasure channel with feedback-capacity and algorithms," In Proc. IEEE Workshop on Network Coding, Theory, and Applications, 2009

[SI12] Song, Xiaohang, and Onurcan İşcan, "Network coding for the broadcast rayleigh fading channel with feedback," In Proc. IEEE International Symposium on Information Theory Proceedings (ISIT), 2012

[NG11] Nazer, Bobak, and Michael Gastpar, "Compute-and-forward: Harnessing interference through structured codes," IEEE Transactions on Information Theory vol. 57.10, pp. 6463-6486, 2011