Deliverable 0.1

Reasoning for a highly flexible and modular control plane

Editor:	Dirk Trossen, InterDigital Europe Ltd
Deliverable nature:	(R) Document, Report
Dissemination level: (Confidentiality)	Consortium (CO)
Actual delivery date:	March 2017
Suggested readers:	3GPP and other standard representatives
Version:	V3.1
Total number of pages:	43
Keywords:	5G drivers, requirements, use cases, verticals, network resources, access agnostic, network functions, context awareness, modularisation, slicing, roaming, 5G network architecture

Abstract

This document outlines the reasoning for the architecture of the CONFIG modular control plane, capturing the main principles and concepts that underlie the control plane, capturing the rationale that has led to the specific choices being made by CONFIG.

Disclaimer

Impressum

Full project title: COntrol Networks in FIve G

Short project title: CONFIG

Number and title of work-package: Principle framework document

Number and title of task: not applicable

Document title: Reasoning for a highly flexible and modular control plane

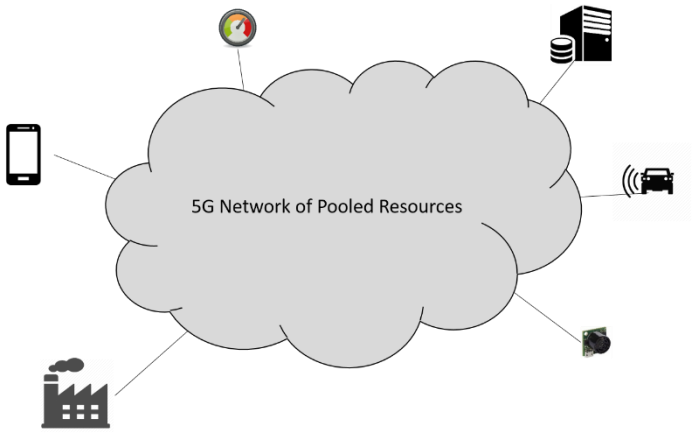
Editor: Dirk Trossen, InterDigital Europe Ltd. and co-editor: Hans J. Einsiedler, Deutsche Telekom AG

Work-package leader: not applicable

Copyright notice

Executive Summary

The transition between LTE/EPC system and 5G is driven currently by the need to create a future proof telecommunication system that is able to fulfil diverse and conflicting requirements, as they will arise, with the intention to cost efficiently integrate vertical industries into a common 5G system [1]. In such common 5G system, the perception of fully customized, yet cost-efficient communication networks that are independently operated and instantiated, is created through inherent capabilities of the underlying network architecture, laying the foundation for vertical industry integration through such customized solutions.



Such perception counters in many ways the view of a network architecture in today's Internet with a common dominator, namely that of best effort IP packet forwarding. Instead, higher flexibility across such large set of requirements lead to the need to re-think the definition of such **core network architecture**, where the main principles of resource pooling and isolation are provided by virtue of a modularized architecture that can be assembled in a plug-and-play manner as a dedicated customer solution. *With this, we can formulate our vision of a 5G system as that of a network of pooled resources across many last hops towards things, devices and users alike.*

The **objective** of this document is to provide **a) rationale**, **b) guidelines** and **3) evaluation methodology** for the future **5G network architecture**. For this, we provide a **common scope** for a 5G network architecture (Section 1.1) by outlining the main drivers for the need for such new 5G network architecture (Section 2), followed by a **common basis** for the main principles and concepts that play a key role in 5G (Section 3 and 4, respectively). This will lead to an initial yet **appropriate set of modular building blocks** based on a well-formulated rationale (Section 5). This set of building blocks will particularly redefine the functional scope of the core network, i.e., the split between access dependent and access independent building blocks, including addressing the question if looking at the notion of a CORE network in isolation might not suffice for providing vertical solutions[1]. We will also provide insights, in Section 6, into the evaluation of the **efficacy** of said modularisation against an evolving set of concerns defined by technology and socio-economics alike.

This document shall serve as a foundation for the wider 5G network architecture discussions within suitable SDOs and forums concerns with 5G as well as a specific validation for the current Modular Architecture proposed in CONFIG D1.3 "Overall 5G Convergent Control Plane Design".

[1] It is this foreseen re-thinking that lets us position the 5G network as the **network of the many last hops**, diminishing the role of specific access architecture and subsuming them into the wider scope of the resource pooling that the network architecture provides, as illustrated by our figure above.

List of authors

Company	Author
DTAG	Hans J. Einsiedler (co-editor), Kay Hänsge, Dirk von Hugo
Huawei	Xueli An, Riccardo Trivisonno
IDC	**Dirk Trossen (main author)**
ITAV	Daniel Corujo, Augusto Neto, Rui Aguiar
NEC	Marco Liebsch, Filipe Leitão
Orange	Xiaofeng Huang
Telenor	Kashif Mahmood
Thales	Damien Lavaux

Table of Contents

Abbreviations & Definitions

Abbreviation	Definition
BB	Building Block
CAPEX	Capital Expenditure
CPE	Customer Premises Equipment
HSS	Home Subscriber Server
ITS	Intelligent Transport System
OPEX	Operational Expenditure
QoE	Quality of Experience
QoS	Quality of Service
SDN	Software Defined Network
SDO	Standards Developing Organisation
SEA	Service endpoint Agent
UE	User Equipment
VM	Virtual Machine
WAMCS	Wide Area Monitoring and Control Systems

1 Framing the Scope of Work

When thinking about defining a network architecture, any kind as well as the particular one for 5G, a number of issues need addressing. These issues frame the scope of the work at hand.

Firstly, we need to understand the **main drivers** for 5G in general and those affecting the 5G network architecture in particularly. Section 2 aims at outlining these main drivers.

Secondly, we need to understand the **main principles** as well as enabling **concepts** that surround the resource pool view onto the 5G network architecture, as outlined in the executive summary. Particularly important is the right understanding of the resources that comprise the 5G system overall. This drives our view on resources as well as the main principles and concepts that underlie the objective to efficiently control and manage those resources through a 5G network architecture.

Specifically, we need to understand the main drivers for *resource isolation and sharing*, formulating a rationale for both that goes beyond a narrow slicing discussion that aims purely achieving resource isolation for the sake of enabling vertical businesses with the possibly widely varying requirements. Instead, we believe that only a joint discussion on isolation within a shared pool of resources will lead us to the right approaches for a 5G network architecture that accommodates the many vertical use cases within an ultimately constrained pool of physical resources that realizes these use cases.

Thirdly, we need to understand the **main concerns** that impose upon a 5G network architecture. We argue later in this document that these concerns drive a particular modularisation of the architectural functions, while also blurring the notion of traditionally separate core and access architectures from a control perspective[2].

Fourthly, we need to understand what enables the **delivery of an end-to-end service experience** from a network architecture perspective. While we recognize that many aspects of service delivery are beyond the scope of a network architecture, we do believe that flexible function chaining and placement will be crucial in order to accommodate varying service requirements and cope with access heterogeneity, i.e., any 5G network architecture approach must provide answers to how to support E2E service delivery at the level of the common networked resource pool.

In order to address the aforementioned issues, we have structured the documents into four main parts. We will first outline the main drivers and assumptions for the 5G network architecture in Section 2. We then follow with the main principles and concepts utilized for the design of a network architecture in general and for 5G specifically, presented in Section 3 and 4 respectively. On the backdrop of these principles and concepts, we will then outline our choice for a 5G network architecture in Section 5, while outlining an approach for rationalizing as well as evaluating any proposed modularization of a network architecture in Section 6. We conclude our work with outlining a roadmap approach, including a tracking proposal, for the development of a 5G network architecture in Section 7.

[2] While we believe that a traditional access function will prevail, the overall architecture is seen as managing a network of pooled resources, where differences might exist in the 'last hop' to include the interface to the user experiences, such as devices, non-5G environments, cars, …

2 Main Drivers & Assumptions

Currently, core networks of operators are designed and implemented based on the history of the cellular industry, in which the fixed Internet was realized largely through fixed communication technologies, whereas the mobile Internet was following mobile communication architecture principles. There is now the possibility to merge the two worlds. The following subsections provide some insights into the main identified drivers and assumptions that underlie our work. We refer appropriately to those forums and activities that have identified those drivers and assumptions.

2.1 Customers' Needs

Previous generations of the telecommunication access and core technologies address a single category of customer needs into silo-based architecture. Given that, each vertical silo runs a specific access technology and specific control-plane and user-plane functions for signalling and forwarding.

In 5G, a proper gathering of customer requirements is required, because new use cases from various business sectors are emerging. As example, next to the typical broadband Use Case, the NGMN addresses already eight use case families, which are in the scope of 5G [24]: Broadband Access in Dense Areas, Broadband Access Everywhere, Higher User Mobility, Massive Internet of Things (IoT), Extreme Real-Time Communications, Lifeline Communication, Ultra-reliable Communications, and Broadcast-like Services.

Each of them has different requirements on a technical and functional view of the network and the users' perception. In the context of 5G, we address verticals and use cases. To find a common understanding, the customers' needs are the essential element inside each vertical and its use cases. From an operators' point of view, the various number use cases and their dynamic requirements are reflected inside these needs.

2.2 Multi-Tenancy

Multi-tenancy originally refers to an architecture in which a single instance of a software application serves multiple customers (the tenants). Mainly driven by the various 5G use cases especially from verticals, each with their widely varying requirements and, multi-tenancy becomes a must in order to provide the customization of 5G networks and services with the adequate and differentiated technical characteristics (e.g. QoS, security, robustness and performance).

The network virtualization gives the chance to the operators (and/or new actors, for example coming from IT world) to enter the market by leasing resources and setting up virtual networks, without owning large and expensive infrastructures. Furthermore, multi-tenancy allows for multiple users and providers (e.g. operators, services providers) to share a common infrastructure by virtualizing hardware and sharing resources without exposing the private data and traffic outside of their virtual boundaries. This provides the benefits on the cost-saving, adapted Quality of Service (QoS) and better choices for the end-user. In particular, the virtualization will create new network trends based on Multi-Tenancy, characterized by:

- Isolation (separation of services provided to each tenant)
- Scaling conveniently with the number and size of tenants
- Meet SLAs for each tenant
- Support for per-tenant service customization
- Support for backup, upgrade, …
- Secure data processing and storage
- Support for regulatory law (per legislator, per tenant)

There are a number of challenges that are introduced by the multi-tenant nature of 5G. Examples for those are the possible heterogeneity of control policies (e.g. access control) that might exist in individual administrative domains (and therefore the possibility of conflicts when needing to interconnect), the dynamic integration when connecting a new element (not belonging to the same owner) in the network and so on.

2.3 Context awareness is a Must

One of the assumptions going towards 5G is that context information will play a crucial role for optimizing control and data plane actions in Next Generation systems. For e.g. efficient user plane and efficient content delivery are highlighted as some of the many key features of smarter network operation in 5G [25]. Context information pertaining to the network, the device and the user can play a crucial role for many of such features. Furthermore, a QoS framework has been proposed in TR23.799 (Sec 6.2.3) [26], which explicitly requires context awareness function both in the core and the RAN. Likewise, the concept of lightweight context cookie is introduced in TR23.799 for stateless context management for data network sessions. These are the some of the many use-cases, which require context awareness. However in order to address these use cases efficiently, a richer context information is required than what is available today. For e.g. a D-plane anchor point reselection is typically triggered on the mobility of UE or load/ failure of gateway. However, a richer context which describes the best match according to location of UE, corresponding service, underlying available best network, and available anchor points can result in an even optimised D-plane anchor point reselection. Furthermore, context information can be used for taking proactive actions even before the actual violation occurs.

2.4 Unified and Access Independent Control Plane

Today's Evolved Packet Core (EPC) network was designed for cellular access and supports a variety of valuable features in addition to device registration and handover, e.g. support of bearers and QoS differentiation, discontinuous reception (DRX), idle mode and paging. Basic connectivity and handover are considered for non-cellular wireless access through associated gateways, allowing mobile devices to attach and handover between cellular access and trusted, or non-trusted, non-cellular access. This is enabled by sharing the Packet Data Network Gateway (PGW), which serves as mobility anchor while the mobile device performs handover, irrespective of the access technology being used. However, most valuable features, which are available for cellular access, are not supported when using non-cellular access. For cellular access, features such as bearers and paging require support on the involved components in the core network, access network as well as the mobile device. Though non-cellular technologies may include support for power save mode and even for QoS differentiation, these mechanisms are typically not aligned with the functional support and associated protocol operation in the core network. Nowadays, when handover between cellular and non-cellular access is required by a mobile device, different protocols are used in between mobile cellular specific functions and the core network's gateway, which is shared irrespective of the access technology being used by the mobile device (Fig. 2.4.1).

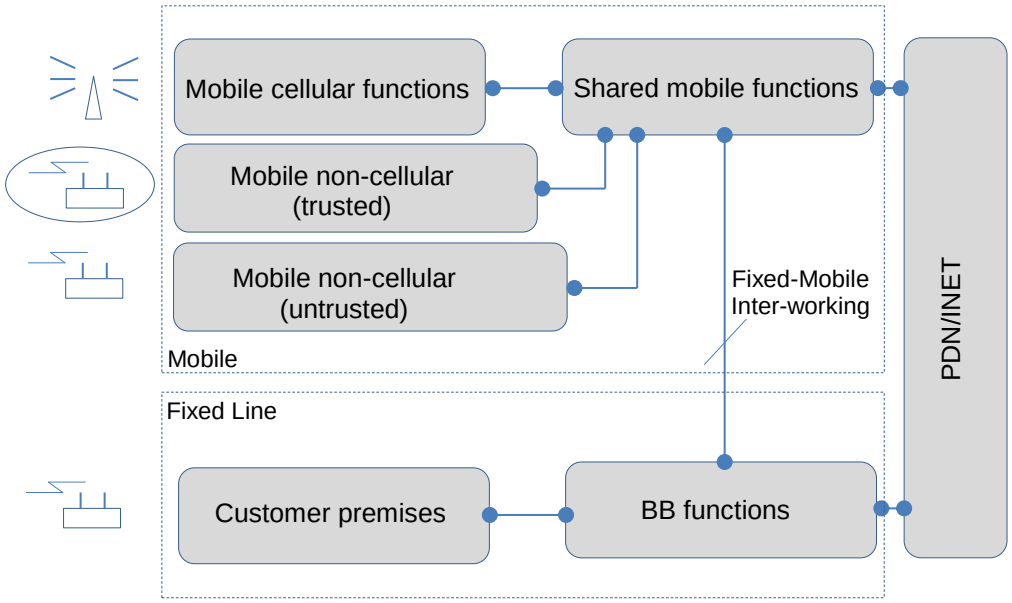
Some standardization efforts have been made to enrich the existing architecture with features, protocols and procedures, which are used for non-cellular access, such as QoS [13] or paging [14]. However, associated solutions, which tried to make support for such features independently of a particular access technology, have not been adopted by 3GPP technical specifications so far, mostly because of lack of consensus. Further efforts have been made to bring together the networks of mobile operators with fixed line operators for Broadband (BB) access. Related activity has been commonly treated as Fixed-Mobile-Convergence (FMC), though the current status of such effort is more about inter-working between relevant components, such as policy controllers and AAA infrastructure components. An overview of enhancements done in 3GPP and the Broadband Forum (BBF) standards towards inter-working can be found, for example, in [15].

Real convergence by sharing common components, such as policy controller, AAA backend, and Data-plane anchors is desired to ease management and operations, e.g. using SIM authentication for both, wireless and wired access, and having a single control point to manage subscriber profiles and determine policies and charging rules. The Broadband Forum (BBF), in cooperation with the 3GPP, developed architectural enhancements and solutions for the converged policy and charging management for wireline and wireless networks that resulted in [16]. An overview of the standardization efforts made to enable policy and charging convergence can also be found, for example, in [17].

On a converged Data-Plane, routing, and traffic steering can be flexible and optimized. This includes the use of a common anchor point during mobility between mobile and fixed access, offloading traffic at a gateway in the fixed line network, or in the view of better handling hybrid access from customer premises. An ultimate goal for 5G should be the convergence of core network functions to support various access technologies and access networks. This includes the support of valuable features, such as QoS, location tracking and paging, for different cellular and non-cellular technologies. Common to all access technologies should be functions and utilized protocols for mobility management and policy control, handover, charging, and service lifecycle management. Convergence is also desirable in the view of identity federation.

Access technology specific operations may be abstracted towards core network functions at an adapter function or layer in between the common core network functions and each specific access network. Whereas the core network functions can use the same protocol suites and operation for all access technologies, the adapter component should perform mapping to access technology specific protocol operations.

Though the core network operates access independent, it may not be entirely agnostic to access specific information associated with an attached device, mainly to enable and maintain differences in, e.g., policing and charging, according to the used access technology. Also, some decisions maybe be taken in the core network dependent on the type and status of an access technology, e.g. network-initiated handover between accesses, which benefit from the identification of the access technology provided or type semantically.



(PDN: Packet Data Network, NET)

Fig. 2.4.1: Abstract view of heterogeneous access support and fixed-mobile interworking in EPC

2.5 Technologies are ready – Virtualisation and Software Defined Networking

According to Open Network Foundation (ONF) [17], Software-Defined Networking (SDN) is an emerging architecture that is dynamic, manageable, cost-effective, and adaptable. SDN defines an approach to network design and management that separates the control from the forwarding plane of the network and thus enables their independent evolution. The OpenFlow protocol is one of the foundational element to enable SDN solutions [18]. SDN is considered as one of the major enablers of the 5G system, especially core networks [19]. The idea behind SDN is to abstract everything as a flow and to move the complexity of flow treatment to a single logical entity, i.e. SDN Controller. Such approach reduces all network elements to dumb flow treatment devices, which are only responsible for flow processing. SDN approach provides the hint on how to split control and user plane in conventional mobile networks that control and user planes are intertwined in the network entity like Serving GW (SGW) and PDN GW (PGW).

Virtualization, especially ETSI's Network Function Virtualization (NFV) concept is considered in the same ecosystem together with SDN. The NFV defines a method to provide network functions. Compared to the conventional approach, which depends on the closed and proprietary appliances based deployment, NFV approach addresses the operational challenges and reduce both operational (OPEX) and capital costs (CAPEX) [17]. NFV leverages the advantage from virtualization technology from IT domain to virtualize the entire network functions that are traditionally implemented in dedicated hardware.

NFV and SDN are conceptually independent. NFV seeks solution to achieve network function provisioning agility while reducing both CAPEX and OPEX. In comparison, SDN seeks solution to optimize the underlying infrastructure that supports the operation of network functions. Therefore, these two concepts are complementarity on the road map towards 5G system. The agility introduced by NFV and SDN, can bring 5G architecture design in a new era, for instance, multiple control planes may coexist upon the same infrastructure (due to virtualization introduced in Section 3.3 and slicing concept introduced in Section 4.1), meanwhile, such control planes may also potentially sharing common resources, e.g. common network function, common database, etc. The driver of variant architecture design should come from the requirements and constraints of the use cases, e.g. vertical domain.

2.6 Verticals – Essential Drivers for 5G

The integration of verticals industries is one of the key differentiators between 4G and 5G systems to allow truly global markets for innovative digital business models. Use-cases originating from verticals have to be considered as drivers of 5G requirements from the onset with high priority and covered in the early phases of research activities and standardisation process. Currently, digital use cases from most important vertical sectors in Europe are analysed, namely: Factories of The Future, Automotive, Health, Energy and Media & Entertainment. How their requirements impact 5G design is being investigated. The vision of 5G is driving the standards developments needed to address the entire network, including new and evolved Radio Access Technologies (RAT), new Radio Access Networks (RAN), and core network architectures based on fundamental changes to business models and eco-system. 5G architecture is expected to accommodate a wide range of use cases with advanced requirements, especially in terms of latency, resilience, coverage, and bandwidth. Thus, one of the major challenges identified is to provide end-to-end network and cloud infrastructure slices over physical infrastructures (see Section 4.1) in order to fulfil vertical-specific requirements as well as (mobile) broadband services and Fixed-Mobile Convergence as direct operator customer in parallel.

In the long run, it will not be sufficient to explore the requirements of the vertical industries but also conduct a proper analysis of market trends in order to sense new, upcoming technology especially through companies outside the industrial mainstream. Potentially disruptive technologies typically

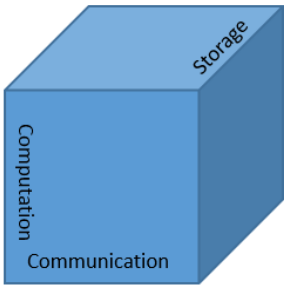
grow widely undetected by the established industry but certainly have a large potential to become drivers for significant technical change and innovation. Unanticipated 5G features are likely to emerge from future technological, legal, societal, and socio-economic considerations.

3 Main Principles

A **principle** is defined as "a fundamental truth or proposition that serves as the foundation for a system of belief or behaviour or for a chain of reasoning". Since we assume that the design of a network architecture shall be based on a chain of reasoning, albeit often not explicitly expressed in terms of requirements, we believe that outlining the main principles for a 5G network architecture is paramount for a proper argumentation of any choice being made, both at the level of individual design decisions as well as at the overall level. In the following, we outline seven basic principles in the context of viewing a 5G network architecture as a *networked pool of resources*.

3.1 Resource Pooling

Resource pooling is a principle that recognizes the necessity to holistically utilize resources along all three dimensions of computing, storage and communication in order to provide solutions to problems through distributed systems.



'Problems' here relate to anything a distributed system could possibly be utilized for, such as the use cases outlined by the 5G community. Following the definition of Russel and Norviq, a problem involves "a collection of information that an implementation can use to decide what to do"[3]. Within the context of a distributed system, such information can be stored, transformed (i.e., computed over) as well as communicated to other computing and storage resource. Hence, it is almost natural to consider the solution to any problem as a point in space within all three resource dimensions, as illustrated in the figure on the right.

Resource pooling is not limited to the Internet but defines a design principle for many other (digital but also analog) systems. Modern operating systems provide abstractions to access resources along all resource dimensions, enabling programmers to provide optimal (according to some problem-specific policy) utilization of the resources towards the solution to the problem. For this, resources (both physical as well as virtual ones – see also the resource virtualization discussion in Section 3.3) are abstracted and represented through a unified layer, called the **device abstraction layer** as shown in Figure 3.1.1. Resource usage within the joint pool can often be coordinated through a dedicated API, while applications are realized on top of the abstraction layer. A key problem in such frameworks (and their solutions) is the discovery and integration of resources at runtime. Modern device OSes provide plug-and-play capabilities where resources can be added and removed during runtime; therefore dynamically expanding or shrinking the joint resource pool.

Infrastructure solutions, specifically those emerging from more recent Software-defined Networking (SDN) concepts aim for a similar exposure of resources towards a common resource pool with the ultimate goal of providing a network operating system (NetOS), with ONOS[4] being a representative of such initiative. The added complexity compared to device-centric architecture often lies in the multi-domain nature of the underlying physical resources, often resulting in fragmented ownership structures. Nonetheless, a similar plug-and-play nature is desirable to that of device-centric solutions, while still preserving the ability to add resources stemming from different ownerships (similar to the addition of, e.g., line or graphic cards in computers without requiring to use the original manufacturer's solution). With that in mind, proposals for the unified control over a common resource pool resemble in many ways the frameworks being used for device-centric platforms, as

[3] S. J. Russell, P. Norvig, "Artificial Intelligence: A Modern Approach", 2nd Edition, Pearson Educ., 1998

[4] http://onosproject.org/

illustrated in Figure 3.1.2, while focussing the innovation on the aspects that are particularly introduced by the aforementioned multi-domain nature.

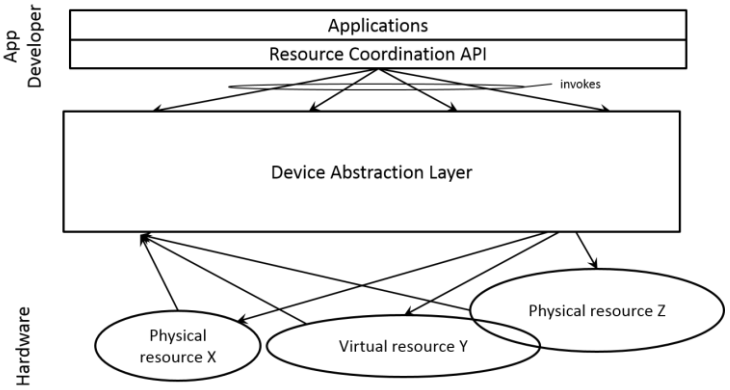


Figure 3.1.1: Resource Abstraction in a Joint Pool

Examples for such aspects include that of **resource naming** (see also Section 3.4) across different administrative domains as well as that of **control routing**. The latter refers to the problem of bootstrapping a control structure that will allow for the control of resources without assuming a well-established routing infrastructure to be in place. For instance, existing solutions for SDN-based resource management (i.e., the management of forwarding switches) requires an IP-based routing infrastructure to be in place for the TCP-based communication between controller and switch. Bootstrapping such communication fabric (for the control of the communication fabric itself) is a crucial aspect in enabling efficient and flexible resource pooling.

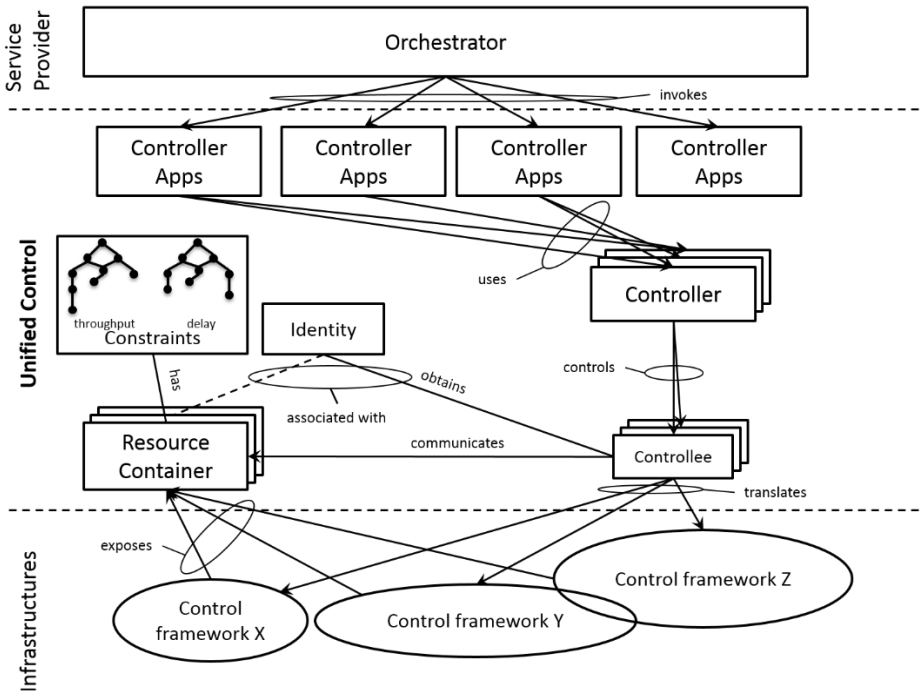
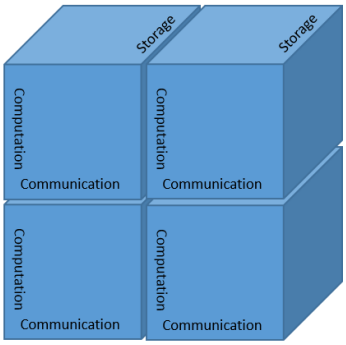


Figure 3.1.2: Unified Control over a Joint Pool of Networked Resources

As indicated in Figure 3.1.2, the **connection to existing control frameworks**, each coming with its own abstractions and methods, is a crucial aspect when realizing a resource pooling framework. Figure 3.1.2 indicates the translation of control command into framework-specific ones, e.g., through wrappers, and the exposure of framework-specific resource abstractions towards the unified resource control. The latter aspect also ties into the resource naming, not only for the resources itself but also for the constraints that govern the control over them.

3.2 Resource Isolation



Resource isolation is a principle that aims at providing resources dedicated to a (set of) task(s), aimed at addressing the objectives of isolating the needs of tenants in an otherwise shared infrastructure.

The provisioning of network slicing, as the means to allow the differentiated support of disparate verticals over the same set of networking resources, relies on the principles of flexibility and efficient resource allocation or sharing. According to [1], these principles become coupled to a set of requirements, which can not only extend to the radio spectrum, the infrastructure and the transport network, but also further incorporate the specific operational requirements placed by the different verticals themselves.

An example of such requirements is resource isolation. The operating principle of a network slicing relies on the capability of sharing the same infrastructure resources and allowing the elaboration of dedicated logical constructs over them. By design, under a logical perspective, slices are mutually isolated [3], in order to operate in an efficient way and without violating specific performance requirements. However, the underlying physical resources (i.e., bare metal) where such slices are deployed, are still shared, with concurrency being managed by a dedicated system or feature (i.e., virtualization technologies).

Naturally, with the possibility of slicing deployment at different aspects of the network as a whole (i.e., RAN, Core, and others), the term “resource isolation” can vary accordingly. On one hand, isolation could refer to spectrum resources in the access or, on the other, it could refer to the memory shared by Virtual Network Functions running concurrently at a datacentre in the network core. As such, considering the potential plethora of different requirements placed, it is paramount to establish the underlying generic aspects for the provisioning of resource isolation as a base feature for a 5G architecture.

In [4], slice isolation is referred to as the means to manage network and computing resources in a way that slice performance is not affected by other slices instantiated in the same set of resources. H2020 projects under the umbrella of 5G-PPP, such as METIS II [4] and NORMA [3] have undergone deep analysis of the impact of network slicing at the access network. In this respect, protective channel mechanisms need to be in place so that congestion in one slice does not have a negative impact on another. Such mechanisms already exist in today’s 3GPP-based network, including different barring services for access class or service specific, as well as admission control, and application congestion control for data communications. However, such mechanisms do not consider network slicing operations. METIS II, specifically, is actively researching service prioritization aspects in the access, combining Random Access Channel (RACH) preambles in order to differentiate different prioritized services. COHERENT [5] is also a project that is handling network slicing at the radio and cell levels.

At the core, one important resource isolation characteristic relates to security isolation, with [3] emphasizing tenant isolation and physical VNF separation aspects. In the first case, network slicing isolation requires that tenants are restricted to their assigned resources, and without the ability to interpose over the resources of other slices, despite infrastructure provider control procedures. This also accounts for privacy and legality aspects, with the potential for information leaking via side channels. An example of this, is withholding memory resources from previously instantiated slices, recovering still-visible or leaked data left from them. Another isolation aspect is the isolation from the infrastructure domain itself, in order to avoid having the tenant braking out of its own domain, or to mitigate issues regarding malicious infrastructure providers. In the second case, regarding physical VNF separation, it can be seen as a solution to prevent attacks between network functions running in the same hardware (which can happen sue to exploits in the virtualization software). However,

allocating different hardware resources in order to provide physical separation for slices (e.g., mission-critical communications) has to be realized without sacrificing the principle of flexible and efficient resource allocation.

The realization of key networking procedures via slices, unlike when functions are executed on bare-metal, which are characterized by predictable performance, conveys the potential for placing risks at resource isolation or security. For example, [6] indicates that "Consistency Availability and Partition tolerance (CAP) conjecture states that it is impossible for any distributed system to provide consistency, availability, and partition tolerance at the same time". It also recommends that a trade-off is necessary to be considered at design time "for different slice regions to control the partitioning and fulfil two out of the three guaranties". This places stringent requirements over the virtualization infrastructure dynamics, because "in order to maximize the infrastructure utilization, It is required to dynamically and freely relocate hardware resources depending on current and local needs, under the control of cloud operators". According to [7], requirements have to be relaxed in order to implement certain networking policies without sacrificing availability, which mandates further research regarding consistency models for network policies in network slicing environments.

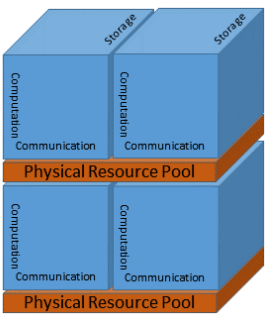
Finally, resource isolation needs to consider both inter and intra-slice management, or even more dynamic slice-management constructs, such as allowing for slice-specific network management (i.e., in [4], it is indicated that the RAN should allow offering slice-specific network management functions as a service) or allowing multiple slices to be managed simultaneously. These can relate as well to OAM support isolation, through the provision of usage and fault isolation at different network slicing utilization levels.

The perspective in affording strict performance isolation to assigned traffic at every network slice without interfering with traffic from other slices, imposes the need for deploying QoS provisioning capabilities in the 5G architecture. Efficient ways for providing QoS provisioning through automated C-Plane features are expected, with the perspective to assure strict performance at every network slice instance over time. Given the need to afford strict QoS at every network slice while fulfilling the corresponding traffic requirements, fine-grained QoS provisioning C-Plane is key. However, the task in provisioning per-slice strict QoS is more complex than in traditional network systems, by requiring to suite to a variety of aspects beyond those related to per-flow bandwidth, class-based traffic provisioning, and the like, which are regularly accomplished through highly over-provisioning network resources by a large factor to avoid QoS violation. In addition to the per-flow level view, the QoS provisioning C-Plane must cope with the per-slice level view, in order to commit to the multiple service-level agreements (SLA) end-to-end, and thus afford full QoS-guaranteed traffic transport and consequently minimum QoE to corresponding UEs.

Under the perspective of a 5G architecture, the operational requirements of the supported verticals are the underlying providers of network slicing demand. Their heterogeneous nature, along with the upcoming realization of the 5G architecture control mechanisms, greatly impact how network slicing affects, for example, RAN design (on both the access network and user equipment sides, thus demanding further research.

3.3 Resource Virtualization

Virtualization has been existed in computing and communication domains for decades. It allows resource sharing among different hosts to increase usage efficiency. For instance, hardware virtualization creates a number of virtual machines, which behave as physical machines and share the same hardware resources. Virtualization technology provides a mean to provide an abstract view of the underlying resource, i.e. computing, storage and networking, which can be used by multi-tenant. This concept is also adopted by communication system as such, network functions can



be virtualized and running on Connection Oriented Transport Service (COTS) hardware instead of using dedicated hardware. Network entities like router/switches can be virtualized and even the network itself can be virtualized. Virtual network embedding (VNE) has been discussed enthusiastically within the academic field, which can be considered as an optimization problem with predefined constrains **Error! Reference source not found.**. Such constrains may come from service performance or network operational requirements.

Virtualization is one method to enable network slicing, but not the only one. On the other hand, network slicing does not equal virtualization. When network slicing was discussed under 5G context at very beginning, the concept novelty was questioned, because such concept has been used for instance in GENI **Error! Reference source not found.**, which is a running platform that enables "slicing" via network virtualization. "Slicing" is the concept used in GENI to guarantee that multiple experimenters are running multiple experiments at the same time without interfering each other. Therefore, "slices" for GENI means only resource isolation. IETF network working group **Error! Reference source not found.** also defines network slicing upon resource isolation. They refer that network slicing is a technology that can 'slice' a physical network into different pieces; each piece is logically independent from each other. However, slicing in CONFIG view is more than resource isolation that is brought by using virtualization. It is about composing customized virtualized network functions for dedicated use cases and such composed network functions can be run upon virtualized resources. Hence, being different from the conventional network function operation, which is upon dedicated hardware, e.g. ATCA, network functions in NextGen are expected to be decoupled from the underlying hardware and operate in software.

3.4 Resource Naming

The principle of naming resources aims at identifying resources across several usages and layers to meet the objective of avoiding ambiguity, providing assurance and possible accounting as well as aiding authentication and authorization objectives. With 5G a transition from hardware entities and fixed functionalities to flexible software instantiations and modular (virtual) functions which can be dynamically adapted to the requirements of slices and services is foreseen. These functions reside rather at logical than geographical locations so that traditional addressing principles will change to the need to identify uniquely abstract resources by names (thus extending the traditional understanding that "computer networks and distributed systems assign names to resources, such as computers, printers, websites, (remote) files, etc."[5]).

As such, resource naming is a principle that shall fulfil the challenging requirements related to security, multi-party communication models, slice and application complexity, efficient resource utilization, scalability etc. and in general aim at addressing the fundamental objectives of 5G. To cope with the trade-off between preciseness and a too high complexity and resulting processing effort in naming a hierarchical approach (local/global) as via e.g. (semi-)persistent naming across layers and a hybrid architecture with both central and distributed mapping instances has to be investigated. Novel architecture concepts have to be taken into account such as ICN (Information Centric Networking[6] where "everything is information") and approaches developed, for instance, in efforts like the Named Data Networking (NDN) project[7] or PURSUIT[8] developing a new Internet architecture based on naming data (content) instead of addressing their location. Such projects therefore are investigating issues as routing scalability, fast-forwarding functionality, trust models, network security, content protection and privacy, and fundamental communication theory.

[5] https://en.wikipedia.org/wiki/Namespace

[6] http://irtf.org/icnrg

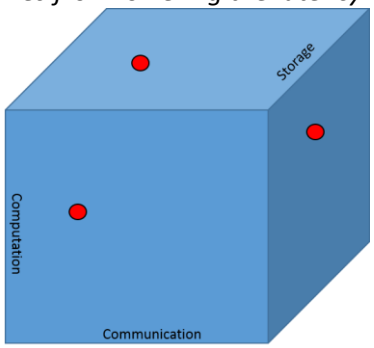
[7] https://named-data.net/project/

[8] http://www.fp7-pursuit.eu/PursuitWeb/

Naming of all kind of resources in a future 5G architecture is strongly related to identity management (IdM) (see e.g., [21][22][23]) - not only for users and their personal and/or business relationships but also for network entities and their authorized access and control - as well as security and authentication management (SAM) [20].

3.5 Flexible Function Placement

Flexible Function placement is a principle enabled by advances in the cloud computing and in virtualization technologies, aimed at addressing wide number of objectives from lowering the latency as well as the communication overhead to reducing energy consumption. With the development of cloud computing, and virtualization technologies, there has been an increasing interest in cloudification of mobile access and core network, mainly for reducing the OPEX and CAPEX by operators, but also to allow for a more flexible deployment of the mobile network. One such flexibility is allowing network functions, as virtual instances, i.e. the VNFs, to be placed, in one extreme at the user devices and being in zero distance to the users or in another extreme, in centrally provisioned cloud. Various alternatives can host VNFs, for example at the radio access network of the last mile connectivity, or within the core network. The flexible function placement as depicted here, clearly presents a split of functions from the traditional communication protocol layers, and is considered as one of the main enablers for 5G. For example, split of functionality from the radio access network has been widely studied in the context of C-RAN [8].

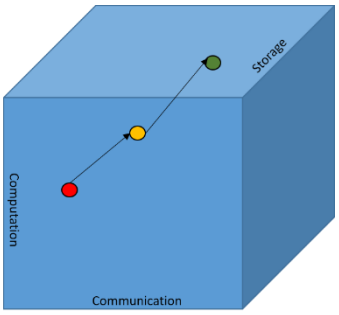


The choice of function split and function placement determines the logical interfaces that must be carried out over physical infrastructure, while physical architecture would determine characteristics of the radio access, backhaul technology between access nodes and transport network as well as the technology towards core network and its logical elements.

In the core network, flexible function placement allows addressing the (1) stringent latency requirements, by terminating the data flows closer to the end-user, (2) more flexible design of the control plane by distributing control functions across the network, depending on the need. How function placement can take place in the core network, how to implement it and the quantitative measure on its impact are among the many questions that still need addressing [9].

3.6 Flexible Function Chaining

For Data plane, Service Function Chaining (SFC) refers to the delivery of added value services by invoking, in a given order, a set of Service Functions along the forwarding path towards a specific destination. For the control plane, the solution should support flexible interconnection between network functions. The solution for the interconnection of the control plane network functions should allow



- Network functions to be able to interact with each other, e.g. for new services and features, while avoiding functional and signalling impact to unrelated network functions for a given interaction
- Build and monitor the service-aware topology. For example, this can be achieved by means of dynamic discovery techniques.
- Mechanism for the exchange of information between network functions that results in agile/rapid deployment of new services, e.g. mechanism that allows reuse of procedures, wherever possible

As examples for current work in this space, the Internet Engineering Task Force (IETF) has created the Service Function Chaining Working Group [10] to work on function chaining and aimed at producing an architecture for service function chaining that includes the necessary protocols or protocol extensions to convey the service function chain (SFC) and service function path information to nodes that are involved in the implementation of service functions and SFCs, as well as mechanisms for steering traffic through service functions. So far two RFCs have been produced, namely [11][12]. The former proposing overview of the issues associated with the deployment of service functions (such as firewalls, load balancers, etc.) in large-scale environments. The latter describes an architecture for the specification, creation, and ongoing maintenance of Service Function Chains (SFCs) in a network. The document includes architectural concepts, principles, and components used in the construction of composite services, but does not propose any specific solution. Work is now focusing around the definition of the Network Service Header (NSH). Such header is inserted onto encapsulated packets or frames to realize service function paths. NSH also provides a mechanism for metadata exchange along the instantiated service path.

3.7 Context-Aware Function Execution

Context-aware function execution is a principle that adjusts traditional network functions, both at the control and data plane, based on contextual information, i.e., any information that characterizes the situation of the entity involved, with that aim to provide more adaptive and flexible function execution. As highlighted earlier in Section **Error! Reference source not found.** context awareness is a must for smart network operation going towards 5G. As a result one of the key architectural principles going towards 5G is that the network functions (NFs) should be able to utilize the context information. Not only that but context information should also drive both the placement and chaining of NFs.

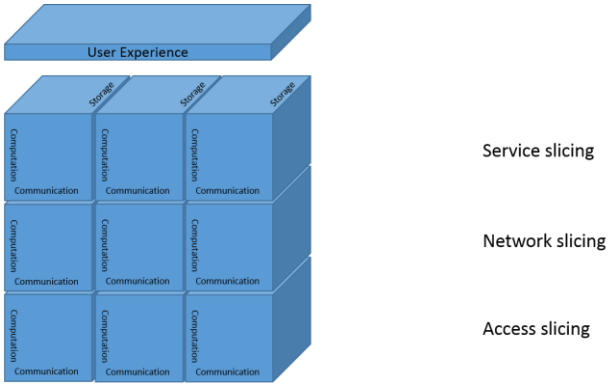
This implies a separate context aware NF by design which can create rich context and then efficiently present this context to assist the other NFs in decision making and enable efficient placement and chaining of NFs. Such a dedicated function provides the necessary modularization that would enable a cross-NF management of context information from the viewpoint of access control and inferencing. Secondly this will prevent repeated per-use-case-specific context integration into existing NFs and instead provide the desired reusability of information and its associated inferencing at scale that is required for 5G. Furthermore such a separate engine provides a clear business interface for such often mission-critical information.

4 Main Concepts

A **concept** is defined as "a generalization or abstraction from experience or the result of a transformation of existing ideas". We believe that there exist a number of crucial concepts, that will ultimately form the cornerstones of any 5G network architecture. In the following, we outline the three main concepts, namely slicing, roaming and modularization, instrumental in realizing the aforementioned design principles of Section 3.

4.1 Slicing

Slicing is a concept that allows for logical and virtual separation of physical contiguous network resources aimed at realizing the principle of service tailored logical networks operated independently of each other.



For the purpose of a network architecture, two kinds of nodes exist: Traffic manipulation (including pure forwarding) nodes and traffic generating and consuming end-systems. Existing definitions of UEs[9] are problematic for the slicing concept since it would assume several UEs in a slicing scenario, even though the 'equipment' might only consist of a single device with a single radio interface. Since we follow the NGMN slicing definition, where a slice is defined as a business of an operator, we need a new expression to define the end-point in the slice and which is not related to the equipment. In the following, we will use "Service Endpoint Agent" (SEA) instead of UE. With that in mind, an end-system as a physical node can be an end-device, (mobile) terminal, etc. It is defined through the following behaviour:

- It will have at least one physical network interface.
- It possibly provides computing as well as storage resources
- It can be part of different network slices.
- It can host different Service Agents.

The definition of a physical network interface is the following:

- It is connected to an access network.
- It will be assigned to at least one network address space.

[9] Official definition of the UE (user equipment) by 3GPP (3GPP TR 21.905, http://www.3gpp.org/ftp/specs/archive/21_series/21.905): "User Equipment (UE): Allows a user access to network services. For the purpose of 3GPP specifications the interface between the UE and the network is the radio interface. A User Equipment can be subdivided into a number of domains, the domains being separated by reference points. Currently the User Equipment is subdivided into the UICC domain and the ME Domain. The ME Domain can further be subdivided into one or more Mobile Termination (MT) and Terminal Equipment (TE) components showing the connectivity between multiple functional groups."

Official definition by ETSI-TISPAN (ETSI TR 180 000, http://www.etsi.org/deliver/etsi_tr/180000_180099/180000/01.01.01_60/tr_180000v010101p.pdf):

"User Equipment (UE): One or more devices allowing user access to network services delivered by TISPAN NGN networks. Source: ES 282 001 [7]. NOTE 1: This includes devices under user control commonly referred to as CPE, IAD, ATA, RGW, TE, etc., but not network controlled entities such as access gateways. NOTE 2: This definition differs from that provided in [1]. NOTE 3: User Equipment is sometimes referred to as Customer Equipment (customer ownership of the UE)."

We define a slice as consisting of nodes interconnected towards ultimately delivering a user or service experience. The interconnections refer to chaining resources end-to-end and can be based on SDN-mechanism or can be based on tunnelling. In the case of the tunnelling, no influence can be done on such a connection, since it is not actively included in the SDN framework of the slice. With the assumption, that the SEA is part of the slice two different basic set-ups exist. The access network is part of the slice. This means the access network is under SDN-control of the slice and it is fully integrated.

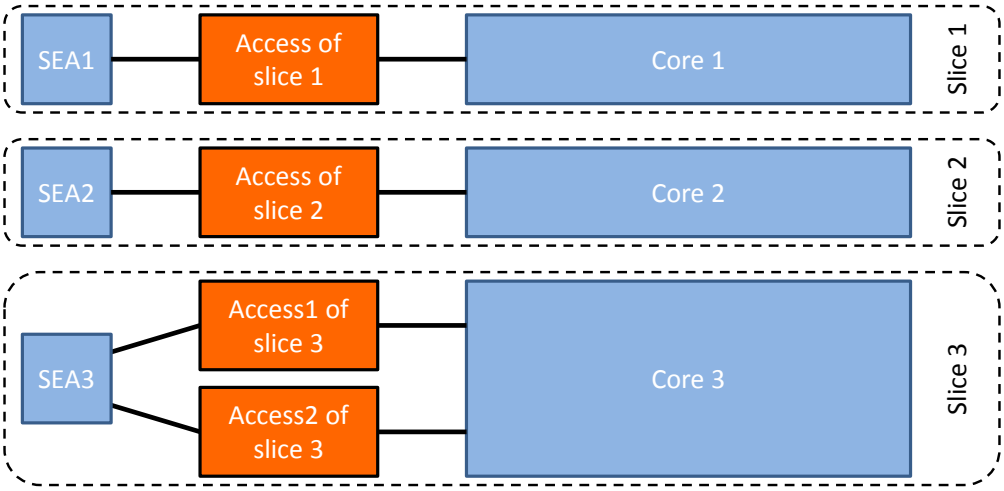


Figure 4.1.1: Slice models, where the access is part of a slice.

A SEA can also be connected through more than one access technologies. Fixed-Mobile convergence and multi-path communication are hence covered with this assumption.

Typical examples in the current communication infrastructure are the 2G, 3G, and 4G networks. These networks are assigned to one specific use case. In 4G, a specific frequencies (single carrier or carrier aggregation) and with this the access network is reserved for the service delivery towards the end-devices.

The other case is, the access network is shared between different slices and use cases. It will act as a pipe and the traffic is tunnelled. The tunnels can be configured but the configurations will influence each other.

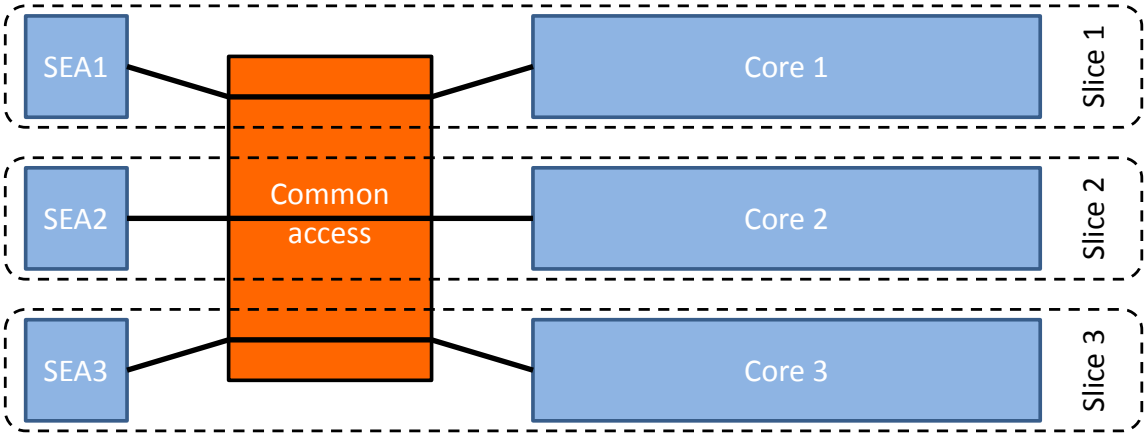


Figure 4.1.2: Slice models, with shared access between slices.

A typical example is home networks and a service provider network like fon[10]. The same access technology – WiFi is used to interconnect the devices of a private customer to the Internet slice of

[10] https://fon.com/

the Internet provider and in parallel – separated through a special Service Set Identifier (SSID) – fon can be used for other services such as IP telephony.

An end-system (physical node) can be part of different slices as long as the end-system hosted different SEAs.

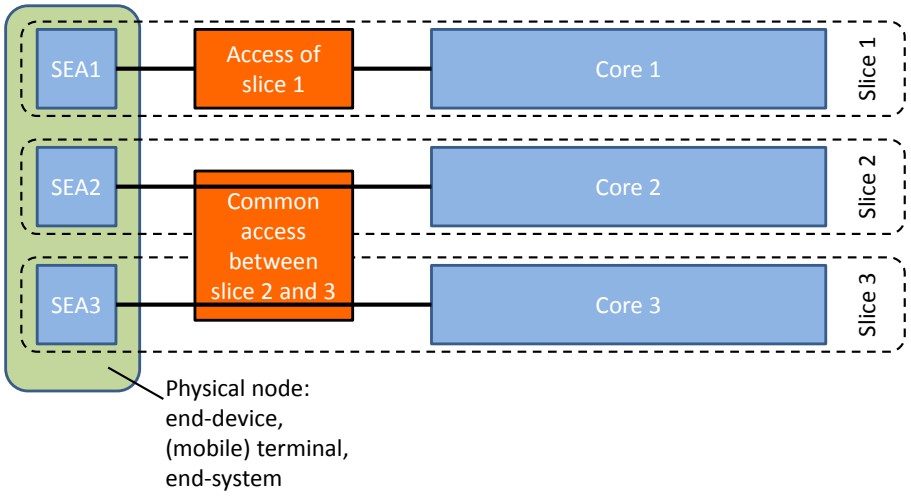


Figure 4.1.3: End-system as part of one or more slices.

A slice is a virtual network and can be defined through a slice-ID. Since networks – in principle – need addressing schemes, the slice-ID can be the network prefix.

On the other site, the SEA has a subscription for a set of services, which have to be provided through the network – the virtual network, the slice. The subscription relates to the services, which are defined in the contract with the SEA-keeper. The SEA-keeper is the "legal person" for the contract with the slice owner – the operator or virtual (mobile) operator or vertical. Following the NGMN definition, a slice relates to a specific business, the set of services are assigned to this business, therefore there is an exact definition of the service, which is defined in the contract between the slice owner and the SEA-keeper. The SEA-keeper has to be identified; therefore at least one ID management mechanism will be integrated into the slice.

The SEA will connect through the slice supporting ID. A connection process is shown in the following:

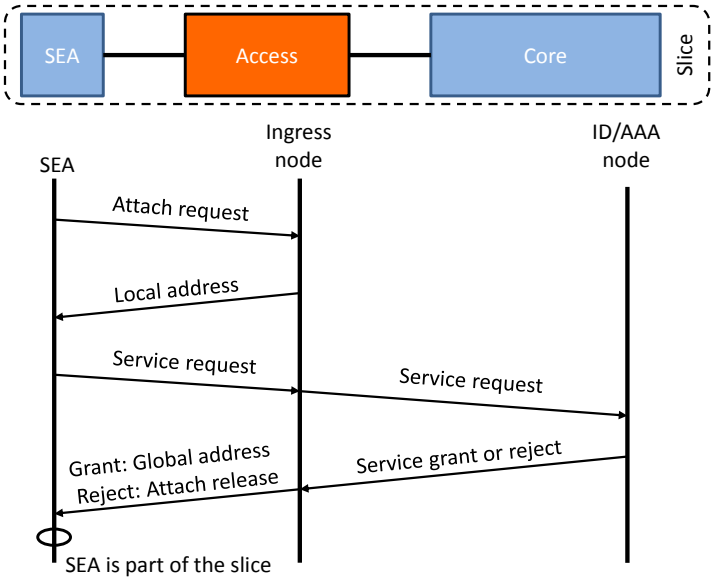


Figure 4.1.4: Access request of a SEA to a slice

Through an attach request, the SEA will get connectivity to an ingress node. The ingress node might be a base station, an access point, a CPE, etc. Of course, security between the SEA and the ingress node is established by using the respective low layer security mechanisms. The SEA will get a local address. The local address is only valid between the SEA and the ingress node. An example might be the link local address in IPv6. With this local address the SEA can start a service request. Hence it is independent of the access technologies and the specific physical and link layer technologies. The service request is forwarded to the slice specific ID/AAA node.

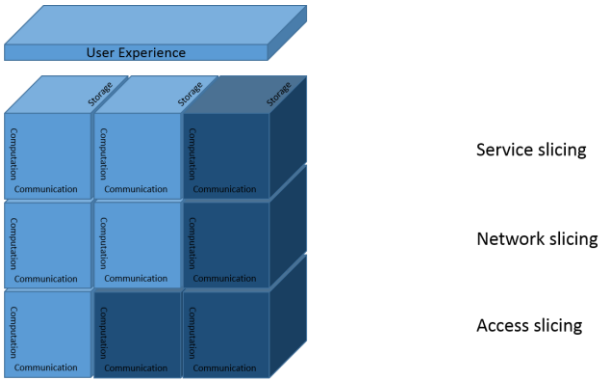
If the SEA has the correct credentials, the SEA will receive the global valid address and the services can start. If there are no valid credentials, the SEA is rejected and the attachment will be released.

This mechanism offers the following advantages:

- The attached request is access technology dependent and has to be implemented according to the technology specifications and standards.
- The service request is access technology independent and is not necessarily bound to a standard. It might be a higher layer protocol, which is specific to the slice, or the operator, or the use case of the vertical. The SEA is specific for the slice business; therefore the respective ID/AAA management has to be provided.
- The SEA does not have to know the slice-ID, since through the confirmation/grant a global address is assigned to the SEA. The global address prefix acts as a slice-ID, which can be used for hand-overs within a slice. In the case of FMC and multi-path connectivity, two global addresses can exist or the same address space can be used. However, in the case of different address spaces, the slice has more than one ID/AAA management and slice-ID.

4.2 Roaming

Roaming is a concept that was already applied in first interoperable cellular networks as GSM/GPRS and aims at using resources outside an operators own infrastructure, e.g. visited mobile networks abroad. Within 5G the concept is drastically extended to physical and virtual infrastructure of an infrastructure provider (potentially 3rd party) hosting commonly used and dedicated operators' network functions making up a service slice.



Current roaming scenarios do not have local break-outs implemented. The attaching to the visiting network, the whole traffic is tunnelled to the home domain to be processed, audited, etc. In 4G local break-outs are defined and standardised by 3GPP [20] but not implemented by the operators.

With network slicing a new possibility is offered to the operators. Via the north bound interface an operator might become a customer/vertical to another operator (visiting domain) and is able to "rent" a part of the operator's infrastructure to extend its own network environment (home domain).

In principle, two cases are given. The operator extends its environment towards another domain, where some functions in the control plane and user plane are moved or doubled. By law, some functions have to be in the home domain – e.g. legal intercept node – or some functions the operator wants to have in its own domain – e.g. auditing node for the charging.

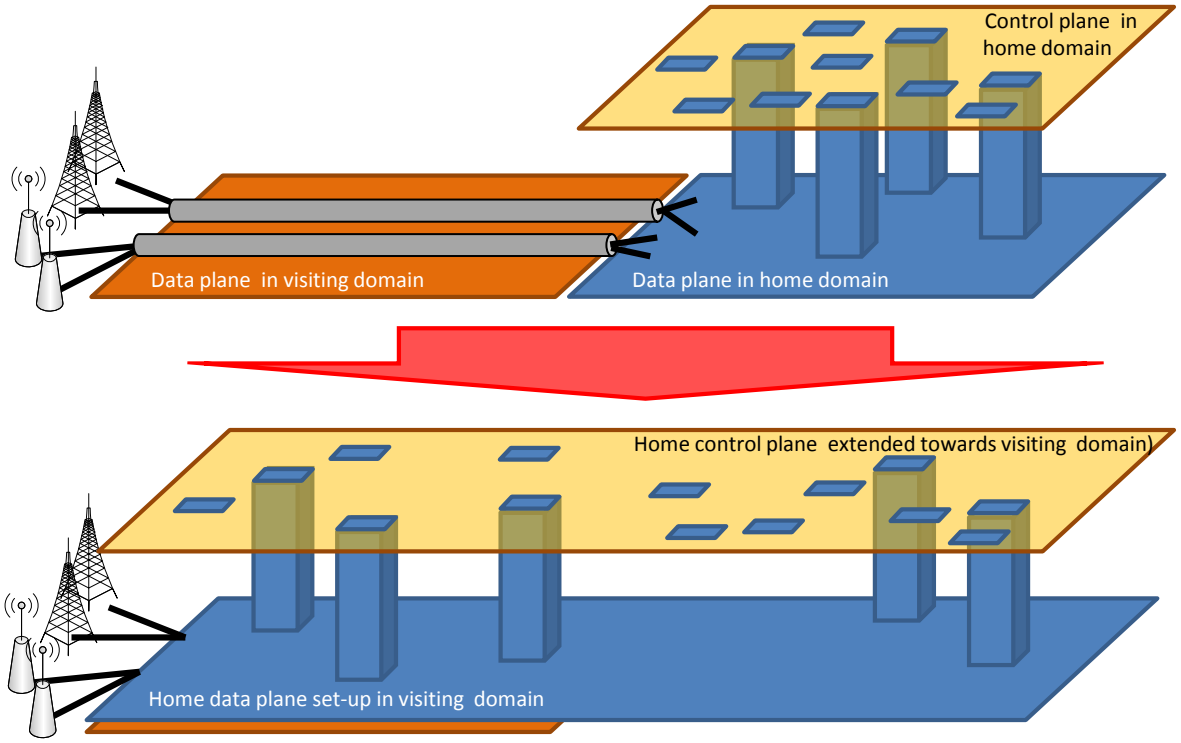


Figure 4.2.1: Extending the user and the control plane towards a visiting domain

The other case is a pure local break-out case, where the use plane is completely moved to the visiting domain and the control planes spans from the home domain and the visiting domain.

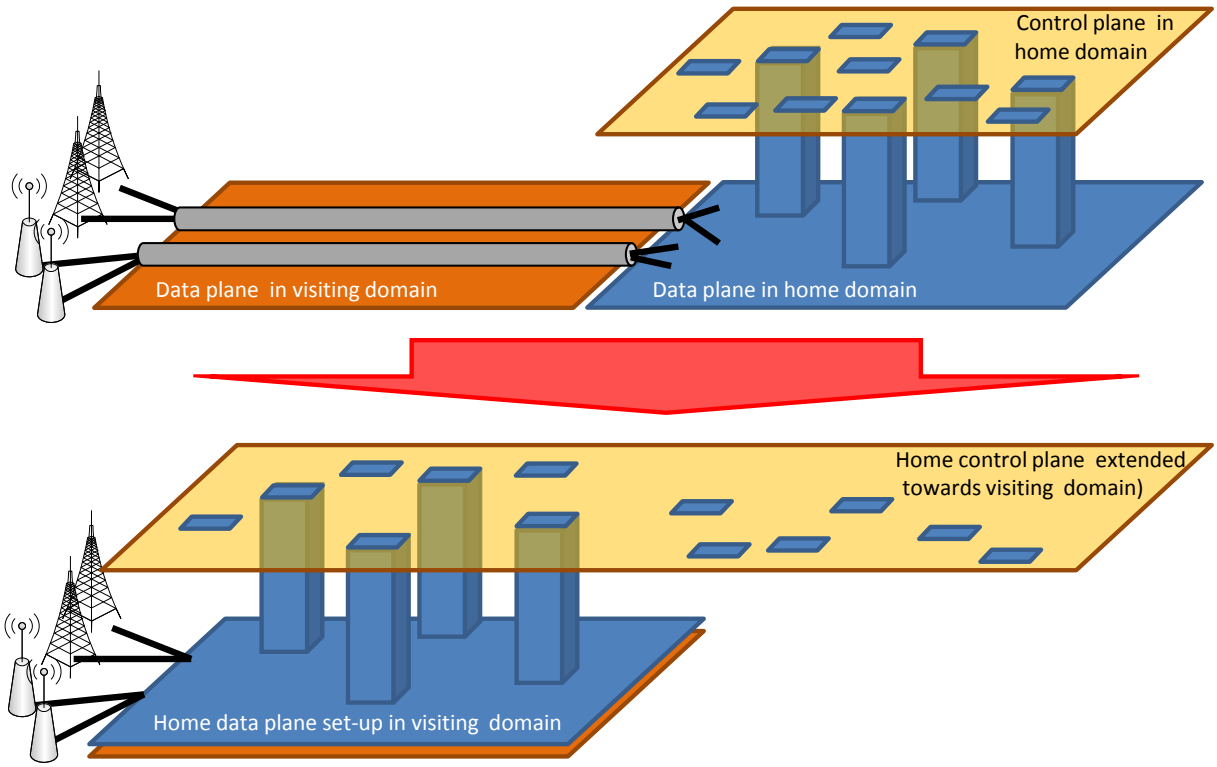
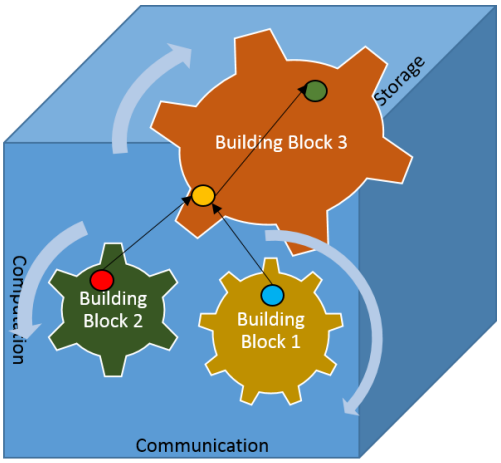


Figure 4.2.2: Extending the user plane towards a visiting domain – local break-out

In this case, all user plane functions are moved to the "visiting" domain under the assumption that the functions are under full control of the operator. Other network functions remain in the home domains, for example ID and privacy management. In case of auditing and charging, the resource

consumption is levied in the slice part located in the partnering infrastructure and the accounting is done in the operator's home domain.

4.3 Modularisation



Generally, modularization is a ***design concept*** *and as such it is targeted to meet a* ***design objective****, namely that of separation of concerns.* One could also include other design objectives, such as simplicity or generality of the architecture, but we assert that the separation of concerns is the main driving design objective in commercial settings, such as 5G, since simplicity and generality are then merely seen as possible technological concerns (unless one designs an architecture for the sheer joy of its beauty; an aspect we neglect here).

It is important to understand that the separation of concerns can also be achieved through *network slicing*, particularly from the viewpoint of enabling a vertical business (see, for instance, the NGMN's take on network slicing in that exact sense). Hence, for the sake of our discussion, we adopt this viewpoint by first seeing the establishment of a network slice as that of enabling a business, which in turn exposes a communication infrastructure with the modularization chosen by the underlying control plane architecture.

Hence, in our discussion here, modularization is seen as a tool that would facilitate separation of concerns within this model of network slicing, i.e., enabling the general ability to modularize sliced network offerings according to a well-understood set of concerns.

4.3.1 A Taxonomy of Concerns

Following our reasoning above for choosing modularization as a design concept for 5G, it is important to understand what concerns are to be addressed by it. For this, let us now develop a taxonomy of concerns that we can use later on for outline the modularization boundaries chosen in our architecture.

We argue that concerns can be generally divided into **technology** and **socio-economic** concerns. The former are discussed broadly within the technology community, particularly within forums such as NGMN, 5GPPP, ITU, 4G Americas and others, and often referred to as '1000x criteria'. The latter concerns are more the realm of business and regulatory decision makers, thinking in terms of value chains, investment opportunities and societal impact. In particular, the latter relates to aspects surrounding legal interception, privacy and others.

The table below outlines a number of concerns in these two categories with a brief explanation of what the main concerns include. There exists a plethora of related work in formulating these concerns, both at the technology and the socio-economic level.

Technology		Socio-Economics	
Concern	Aspects	Concern	Aspects
Delay	To support 5G low latency use cases	Physical ownership	Integrate different physical assets, e.g., spectrum, digital assets, boxes, … to support multi-RAT, IoT

			verticals
Throughput	To achieve 1000x aggregate throughput increase	Highly fluid value chains	To support Anything-as-a-Service To re-arrange business interfaces
Efficiency	To remain at similar energy levels as today	Privacy	To support flexible privacy considerations for context data across all layers, including user context
Flexibility	To achieve hour or minutes goal of service creation	Market unbundling	To ensure efficient competition across different parts of the value chain, including service, core, access, device, …
Scalability	To accommodate 50 billion devices To accommodate finer-grained slicing	Money routing	To allow for evolving models of revenue generation
Security	Achieve infra and user security for critical and highly private services	Security	To support varying legal interception models

4.3.2 First Order vs. Second Order Concerns

Although a taxonomy of concerns is important to understand what drives modularization in general, it is crucial to understand the relations of these concerns. For this, it is useful to understand the relation of the categories first before delving into a more fine-grained analysis of the driving concerns for modularization. For this, let us first look at an example, namely that of *encryption technology*, which is a technological means to address security and privacy concerns at the technological level.

For many years, security concerns particularly at the national security level, impeded on technological solutions relying on encryption technology, hindering the adoption of online shopping due to the availability of only weaker encryption solutions outside the US. Only the increasing desire for economic concerns overcame the security concern and made encryption solutions for online shopping more internationally available.

Hence, we pose the following assertion:

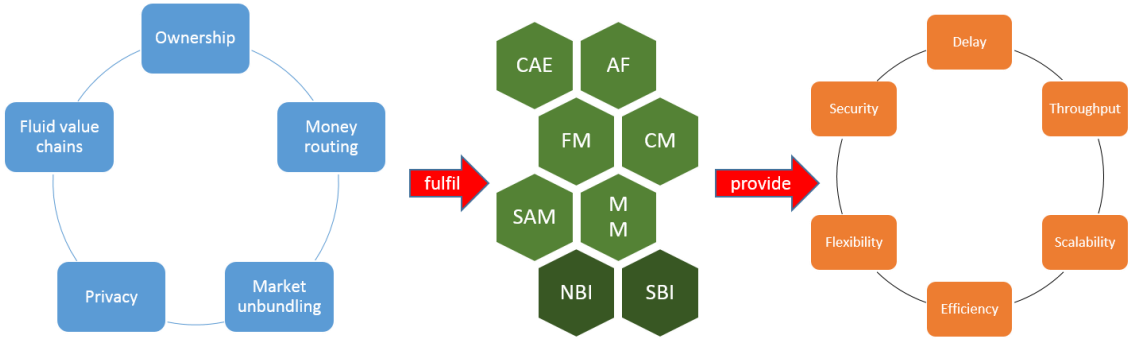
economics>>security>>technology[11]

This means that economic concerns ultimately trump security ones (unless they drive economic ones), which in turn trump technology ones. However, encryption solutions also provide an example that technology solutions change the overall concerns in that (national) security concerns are again

[11] D. Clark, Future of the Internet: A Political Perspective, Presentation to Communications Futures Program (CFP), Oct 2010

gaining weight in the encryption debate, countered by the increased privacy concerns of individual citizens. Hence, we can derive a fundamental requirement for any modularization in that ANY SOLUTION should adjust to changes in concerns over time, in order to avoid breaking the technical solution (through evolving socio-economic concerns) and therefore not being able to accommodate those evolving concerns without a growing plethora of clutches being added to the design. Such requirement is best captured in the work by David Clark et al on **Design for Tussle**[12].

This example and its lessons learned leads us to the approach for determining the right modularization for a 5G network architecture, namely to *fulfil socio-economic concerns while providing solutions for the technological ones within the constraints defined by the former*. In other words, socio-economic can be positioned as first order concerns, with technological ones operating as second order ones within the confines defined by the former, as illustrated below.



Legend: CAE – Context Awareness Engine, AF – Access Function, FM – Forwarding Management, CM – Connection Management, SAM – Security and AAA Management, MM – Mobility Management, NBI – North Bound Interface, SBI – South Bound Interface

4.3.3 Lurking Issues

There are a number of lurking issues surrounding the problem space of modularization, even within the formulation provided so far. The most crucial issue is to find an answer to the question: *What makes our chosen modularization optimal?*

It seems clear from our argumentation so far that the term **optimal** is largely dependent on the concerns at hand, both as seen today and as seen as likely or possible in the future. By conducting workshops[13] and events as an engagement with stakeholders, such as industry players as well as organizations such as 5GPPP, 3GPP and others, and designers who can shed light on these concerns, we aim to capture the likely drivers for optimality.

Ultimately, the key issue to be addressed by the 5G community is how could we ensure possible changes in modularization along evolving optimality and what are the boundaries to the changes we expect to accommodate?

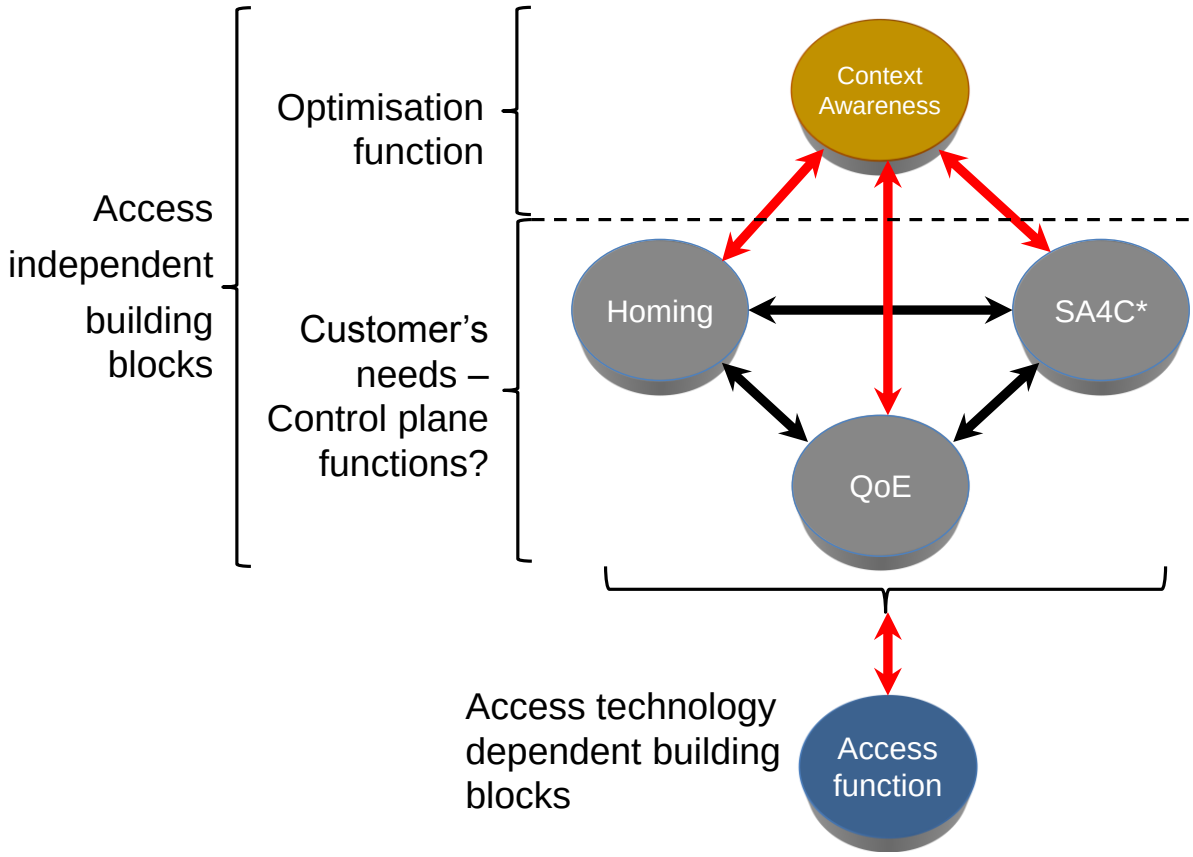
In Section 6, we return to the problem of evaluating a specifically chosen modularization in the desire to argue about its efficacy within scenarios of evolving concerns.

[12] D Clark, J. Wroclawski, K. Sollins, B. Braden, "Tussle in cyberspace: defining tomorrow's internet", IEEE/ACM Transactions on Networking (TON), Volume 13 Issue 3, June 2005, Pages 462-475

[13] Such as the CONFIG 5G Architecture workshop in Lisbon on the 6th of September 2016

5 5G Network Architecture

5.1 Basic Architecture for a 5G Core Control Plane



***SA4C: Security, Authentication, Authorisation, Accounting, Auditing, Charging**

Figure 5.1.1: Basic Architecture Model

Figure 5.1.1 illustrates a starting point for deriving the Basic Building Blocks, i.e. the main modular network functionalities to allow communication between customers and between customers and applications within the network via their terminal equipment and a converged network infrastructure. To set up and maintain connectivity and services according to the various service parameters agreed on, these functionalities should be adequate and essential, modularly designed based on the main drivers and assumptions, highlighted in Section **Error! Reference source not found.**. The terminal itself and the user application as well as any direct user-to-user communication is out of scope here. The functions and entities for providing access (which may be highly heterogeneous in nature as we aim at a converged approach here) covering local fixed and wireless as well as cellular mobile technology (but might also include satellite or broadcast technologies) and their corresponding networks are denoted as access technology dependent building blocks. All other control functionality equally applicable to all (abstracted) access infrastructure is summarized as access independent building blocks here.

There is an explicit split between the Access technology dependent and the access independent building blocks. The purpose is to make the control plane functions access agnostic and to re-use them in several slices depending on the different use cases. The Access technology dependent building block is named Access Function (AF) and interacts on an access specific level with the physical and logical access infrastructure. Interfacing and signalling towards the Control Plane Functions from the Access independent part requires a generalized protocol and signalling pattern.

Hence, using different access technologies which having each of a specialized Access function offers an access-agnostic on the control plane (control plane convergence). In this project, we assumed that the Access Function is an already existing building block where we need to specify the communication pattern with the control plane elements. The internal functionality of these single hops is therefore not in scope of the project.

The control plane functions, Homing, S4AC (Security, Authentication, Authorisation, Accounting, Auditing, Charging), and QoE, are the basis to fulfil the Customers' needs (the verticals' requirements). Each of the grey-colored entities is a category of functionality within the named scope. These, so called, Building Blocks may have a different level of complexity and functionality. The interfacing and signalling (Black Arrows) requires also a generalized protocol and signalling pattern to enable interaction between theses Blocks – independent of their complexity and functionality.

To make the network more efficient and aligned towards the requirements of the use case or verticals, an optimization function (Context Awareness) is introduced as another essential Building Block. This Context Awareness function enables the processing of information from different sources to offer a pro- and reactive behavior on multiple layers inside the network. Optimization is considered separately from the user needs-related control plane functions since an overarching optimization of each network slice (with underlying specific set of requirements to each of the building block) is aimed at. A dedicated uncorrelated optimization of each building block, on the other hand, could result in conflict situations and degradation of the overall performance.

5.2 Definition of a High-Level 5G Network Architecture

Focusing on the Control plane, CONFIG formulated **three key design principles** upon which a 5G Core Network (CN) allowing the integration of different access technologies, the architecture customisation to meet different functional and performance requirements, and the integration of communication services required by vertical industries can be designed.

The design principles are:

1) **Architecture Modularisation**: 5G tailored end to end network architectures, including C-plane and D-plane, shall be defined upon a set of basic Building Blocks (BBs), including Access network and Core Network-related functions.
2) **Core Network Independence from Access**: 5G Core Network related basic BBs shall be defined minimising the dependency towards the supported Access Networks.
3) **Support of Independent logical Networks**: 5G networks shall enable the concept of Network Slicing. In this context, a network slice is an independent logical network, defined by the interconnection of a set of BBs, which can be independently instantiated and operated over a set of physical infrastructure, to support the communication service of a particular use case.

Upon the key design principles, a 5G Core Network Architecture Reference Model (depicted in Figure 5.2.1) has been defined.

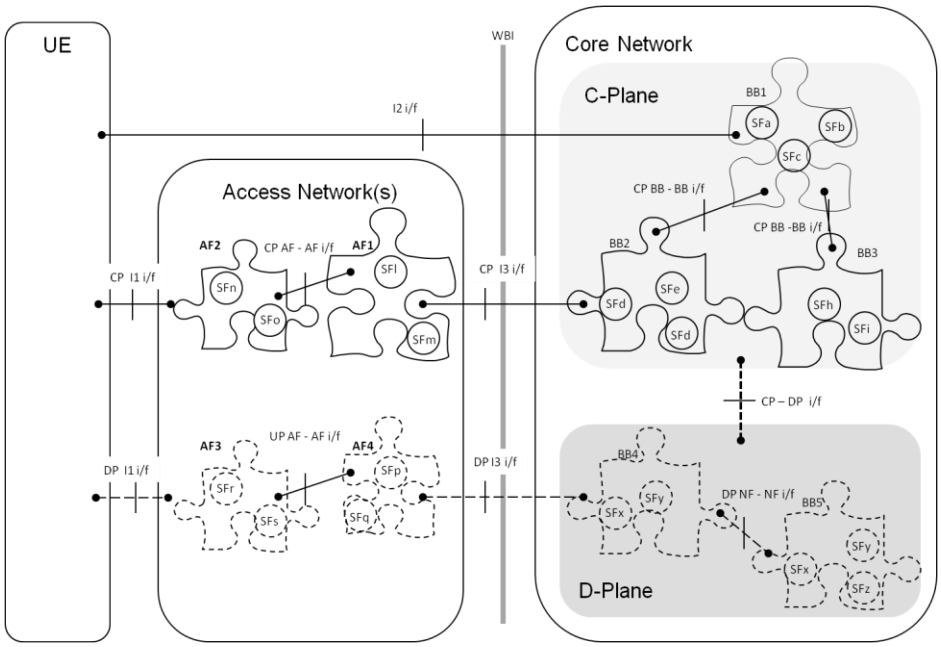


Figure 5.2.1: Architecture Modularisation Reference Model

The Reference Model prescribes a strict separation between Control and Data planes, and defines the key architectural elements:

- **Basic Building Blocks**: a BB is an independent logical network function, made by elementary sub-functions, and accessible via interfaces or reference points. Different versions of a BB may be defined via a proper composition of elementary sub-functions. Different interconnections of different sets of BBs define different Logical Architectures fulfilling requirements of different use cases.
- **Inter Building Blocks Interfaces**: inter BB interfaces allow interconnection and information exchange among BBs (both for C-plane and D-plane)
- **Westbound Interfaces (WBI):** WBI allow interconnection and information exchange between Core Network and Access Networks, and between Core Network and UE.

The details of above-mentioned building blocks are tackled by CONFIG Task 1.2 and reported in Deliverable 1.3 [20].

6 Finding the 'Right' Modularization

Section 4 outlined modularization as one of the key concepts for implementation a 5G network architecture, alongside the slicing and roaming concept. While the latter two are already discussed widely in forums such as the NGMN and others, we believe that the perception of modularization for a core network architecture is a fundamentally departing concept from previous approaches to mobile network architecture. Networks simply are described by modules, but no realization on how these modules came to be is ever explicit. As such, we believe that it is crucial to devote significant thought on evaluating how one finds the 'right' modularization for a specific network architecture, here that for 5G. This section will provide our insights into this problem.

6.1 Approach to Argue for a Chosen Modularization

Following the description of the modularization concept in Section 4.3, we formulate an approach for arguing **WHY** we are choosing the modularization we propose and **HOW** we can ensure evolution of the industry alongside the modularization we have chosen. In other words, our goal is first to provide an argument for modularization itself, while accessorily we can discuss the particular one we have chosen.

For this, we propose a two stage approach, separated in Drivers and Aggregation.

ONE: Drivers - In the first stage, we outline the main concerns that drive the separation of a network architecture in terms of modules of functions. These will form the main causalities that can be identified at this stage to accommodate a range of important concerns. This will provide us with a clear list of concerns per particularly modularization that we see emerge caused by these concerns. Hence, this first stage will provide us with an answer to the functions that our architecture needs to fulfil, providing the substrate of the WHY for our modularization.

These are the Sub-functions of our architecture before.

TWO: Aggregation - The second stage of the proposed approach proposes to aggregate these causalities in order to cluster major interactions around nearby/related functions. This can be done with a causal model, e.g., using system dynamics or other approaches, that captures the forces of the concerns interacting with each other and identifies borders where those interactions are weakest. This model will take into account desk research on drivers and forecasts for technology, economic markets, and regulations in those different markets, but most specially will take in consideration engineering practices and common uses coming from past architectures. It will also utilize stakeholder input, such as gathered from organised events or through questionnaires. Ultimately, the outcome of this stage will be an input into a structure that leads to a modularized architecture, i.e., answering the HOW of how to deploy an architecture that will be time-resilient.

This are the Building Blocks of our architecture before.

6.2 Realizing step TWO: Aggregation

As illustrated in the section 4.3, in order to have an efficient modularization realization. It is expected that each Building Block (BB)[14] of the final architecture will be made of one or more Sub Functions (SF)[15], to be defined in different key issues such as session management, mobility management, session continuity, etc.

[14] A BB is a processing function in a network, which has defined functional behaviour and defined interfaces. NF performs different tasks called SFs.

[15] SF shall not be confused with VNFC

In order to obtain the optimised network architecture, we can use a well-defined methodology to determine how to combine SF into BBs. While Section 4.3 provides the larger framework concepts on deciding what BB separation accommodates the concerns of stakeholders and actors, it is also crucial to converge on the 'right' realization of modularisation of building blocks as well as functions within those building blocks. Note however that this "right" implementation will be a function of the past usages in the network, and performing this analysis with different starting points may lead to different BBs.

For a formal process, we can identify the following four steps, also visualized in the figure below.

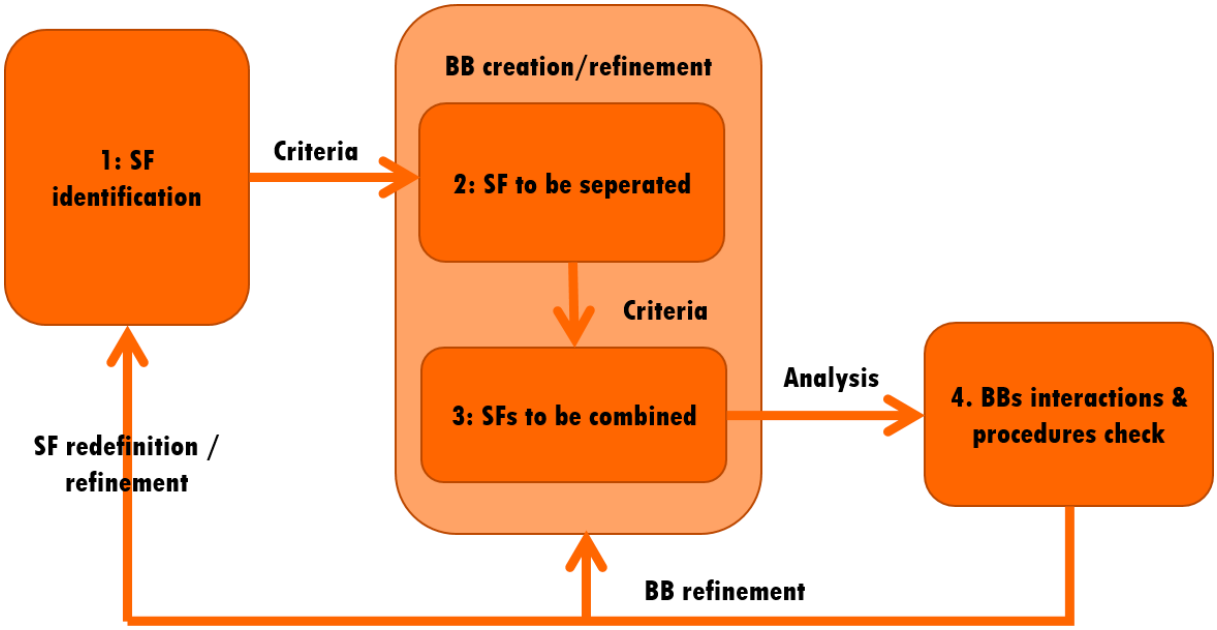


Figure 6.2.1: Evaluation Methodology

As shown in the figure above, the methodology is composed of 4 steps. The first step lists the different SFs necessary for the 5G CN based on the solutions provided by the other key issues (That is the stage ONE:Drivers). The second step, identifies the SF that shall be kept separated from the others and thus assimilated to BBs, based on separation criteria (reduce the number of interactions). The third step gives the SF that shall be combined together into BB based on combination criteria. The step 4 consists the refinement of the BBs and SFs if necessary.

6.2.1 Step 1: Identifying all the potential SFs (stage ONE:Drivers)

This step identifies the different functions required for the 5G CN. This will be also based on existing solutions provided to the other key issues such as session management, mobility management, session continuity, etc.

- **Step 1a : Identify the SFs:** Formally**,** one can rely on the vertical sector documents (or any other requirements document), and use this to identify requirements (e.g. by mapping the requirements spider diagram KPIs to the "needs") – check Apendix spider maps. To accomplish the highly variable use case KPIs, the control/user plane split into modules needs to be essentially addressed in the 5G network architecture. Note that this stage, in general, can lead to different outcomes, due to multiple issues:
    - The technical Requirements (KPIs) from the UCs need to be mapped to basic functional blocks (SF)
    - Definitions of the KPIs are needed, and then assessed - Which are relevant for us?
    - Which functions can further serve as indicator for our architecture?
    - In the requirements spider diagrams, the view is not limited to the control plane functionality. Other topics, e.g., management and data plane functionality, are

covered, too (KPI: reliablity). How do we separe the control plane from the data plane?

The functional blocks are designed to be flexible, scalable and reliable with the basis of virtual network functions and cloud based mechanisms.

- **Step 1b: Check against other entities**
  This analysis should be checked with input from 3GPP (4G), and other relevant fora, in order to verify if our analysis covered the key factors.

6.2.2 Step 2: Identifying the SFs to be kept separated

Trying to keep all the SFs at the same level, this leads to difficulties in design, deployment, and analysis. It is then natural that we aim to aggregate the SFs that can be easily aggregated. The following criteria should be considered to determine which SF should be kept separated:

- Interaction degree: If a function is very loosely connected to another, then these should be placed in separated BBs.
- Centralized or distributed (far from/close to a UE/end-terminal): a distributed elementary network function is a function that needs to be located in the edge (for at least one use case). It should not be gathered with another elementary network function that has no such constraints and that will be most probably centralised. As an example, in case of cloud access network, the access management function responsible for the physical connection with the device shall be in the edge because of access delay constraints.
- Re-usability: an elementary network function that can be potentially used by multiple service layers or have different internal variants shall be kept standalone. As an example, the HSS is used by different types of networks today: EPC, IMS, and the IMS AS.
- Optionality: for resource optimisation purpose, an elementary network function that will be solicited for only some use cases should be kept separated from those that will be solicited for all use cases. For example: the policy control elementary network function may not be necessary for devices connecting through a fixed access; core network handover management elementary function will not be used for fixed-like usages.
- Evolution cycle of the function: an elementary network function that will evolve rapidly (internal algorithm or new feature in normalisation every 6 to 12 month) shall be kept separated from those that have a slower evolution cycle. As an example, the authentication function may evolve rapidly and independently from the others because of the introduction of new authentication methods.

This can lead to a table to be filled in, such as.

Sub Functions (SF)	Interaction	Distributed	Re-usability	Optionality	Industrial Expertise	Evolution cycle
SF 1	None	Yes	Yes	Yes	Database	Rapid
SF 2	With SF5/7	No	No	No	Network	Rapid
…						

At the end of this step,

- SFs with high separation constraints (example 1 in the table) can be deduced. **Each of them represents one BB.** They are not considered in step 3.

- SF (example 2 in the table) with low or without separation constraints can be deduced (SF(a..j)). They are considered in step 3 to decide whether/how to combine them within BBs.
- Any other SF-to-SF non-affinity shall be identified at the end of this step.

Note that the work on the "Cube model" [28] provides an indication of potential structures along which to align the architecture (and in fact CONFIG architecture has been influenced by this model, with aggregation of SFs performed along the lines of the "Cube model").

6.2.3 Step 3: Identify the SF to be combined

This step handles only with SF (SF(a..j)) with **low** or **without** separation constraints and considers the SF-to-SF non affinity deduced from the previous step to determine how they can be combined.

SFs are combined in the objective of:

- Reducing the complexity of interfaces that could exist between them,
- Obtaining a simple architecture with less interfaces.

Combination work shall at least look at the SF from the same functional domain. This should be the initial step. Thus, a table can be filled in to determine the functional domain of each SF (e.g. the table bellow). When SFs are combined, implementation shall grant for each of them an independent scalability within the BB.

	Network	Security	Charging	Database	…
SFa					
SFb					
SFj					

6.2.4 Step 4 : BBs and SFs refinement and redefinition

If necessary, BB refinement, SF redefinition can occur to complete the architecture definition. The process can be repeated with added considerations, as the existing legacy architectures, realistic engineering development strategies, etc.. In these new rounds, not only SFs are analysed, but the existing BBs coming from the analysis are again considered at the same level.

Later documents will describe the BBs proposed for the CONFIG architecture, as the outcome of this process.

7 Conclusion

In this document we discuss the rationale that led our path on the development of the 5G network architecture. We analysed the driving pressures for such a new architecture, and discussed the basic concepts underlying these architectures.

We then argued about modularisation and how it will be performed. This is a critical aspect, often neglected, but that may be performed in a systematic way. We can then define an appropriate set of modular building blocks that redefine the functional scope of the core network; i.e., the split between access dependent and access independent building blocks, including addressing the question if looking at the notion of a CORE network in isolation might not suffice for providing vertical solutions. We provided insights into the evaluation of the **efficacy** of said modularisation.

References

[1] H. J. Einsiedler, A. Gavras, P. Sellstedt, R. Aguiar, R. Trivisonno and D. Lavaux, “System design for 5G converged networks”, 2015 European Conference on Networks and Communications (EuCNC).

[2] 5G PPP Architecture Working Group, « View on 5G Architecture », v1.0, July 2016

[3] 5G NORMA (Novel Radio Multiservice adaptive network Architecture), https://5gnorma.5g-ppp.eu/

[4] METIS-II (Mobile and Wireless Communications Enablers for Twenty-Twenty (2020) Information Society-II), https://metis-ii.5g-ppp.eu/

[5] COHERENT (Coordinated control and spectrum management for 5G heterogeneous radio access networks), http://www.ict-coherent.eu/

[6] Navid Nikaein, Eryk Schiller, Romain Favraud, Kostas Katsalis, Donatos Stavropoulos, et al.. Network store: Exploring slicing in future 5G networks. MOBIARCH 2015, Mobility in the Evolving Internet Architecture, Sep 2015, Paris, France.

[7] Aurojit Panda, Colin Scott, Ali Ghodsi, Teemu Koponen, and Scott Shenker. 2013. CAP for networks. In Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking (HotSDN '13). ACM, New York, NY, USA, 91-96

[8] P. Rost, C. J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, D. Sabella, D. Wübben, “Cloud Technologies for Flexible 5G Radio Access Networks“, Communications Magazine, IEEE , vol.52, no.5, pp.68,76, May 2014.

[9] W. Kellerer, A. Basta, A. Blenk, “Using a Flexibility Measure for Network Design Space Analysis of SDN and NFV”, Software-Driven Flexible and Agile Networking (SWFAN), IEEE INFOCOM Workshop, San Francisco, USA, April 2016.

[10] https://datatracker.ietf.org/wg/sfc/documents/

[11] Quinn, P., Ed., and T. Nadeau, Ed., "Problem Statement for Service Function Chaining", RFC 7498, DOI 10.17487/RFC7498, April 2015, <http://www.rfc-editor.org/info/rfc7498>

[12] Halpern, J., Ed., and C. Pignataro, Ed., "Service Function Chaining (SFC) Architecture", RFC 7665, DOI 10.17487/RFC7665, October 2015, <http://www.rfc-editor.org/info/rfc7665>

[13] RFC7222, Quality-of-Service options for Proxy Mobile IPv6

[14] RFC3152, Requirements and Functional Architecture for an IP Host Alerting Protocol

[15] Enhanced 3GPP System for interworking with fixed broadband access network, IEEE Comunication Magazine, vol. 52, no. 3, March 2013.

[16] BBF TR-300, Nodal Requirements for Converged Policy Management, April 2014

[17] Unified Control Plane: Converged Policy and Charging Control, IEEE Communications Magazine, vol. 53, no. 3, March 2015.

[18] OpenFlow Switch Specification, V 1.5.1.

[19] 5G Infrastructure PPP, “5G Vision”, 5G Infrastructure Association, February 2015.

[20] CONFIG Deliverable 1.3, “Overall 5G Convergent Control Plane Design”, 2015.

[21] 3GPP TS 23.402, Architecture enhancements for non-3GPP accesses (Release 13), 2015

[22] 3GPP TR 33.924, Identity management and 3GPP security interworking; Identity management and Generic Authentication Architecture (GAA) interworking (Release 13), 2016

[23] 5G-ENSURE Deliverable D2.1, Use Cases, 2016, available at http://www.5gensure.eu/sites/default/files/Deliverables/5G-ENSURE_D2.1-UseCases.pdf

[24] METIS II Deliverable D5.1, Draft Synchronous Control Functions and Resource Abstraction Considerations, https://metis-ii.5g-ppp.eu/wp-content/uploads/METIS-II_D5.1_V1.0.pdf

[25] NGMN Alliance, “5G White Paper”, Eds.: Rachid El Hattachi, Javan Erfanian, 2015. Available from http://www.ngmn.org

[26] 3GPP TR22.864,

[27] 3GPP TR23.799,

[28] Rui L. Aguiar, Hans Einsiedler, Jose Ignacio Moreno, "An operational conceptual model for global communication infrastructures", Wireless Personal Communications, pp. 335-351, April 2009, Springer.

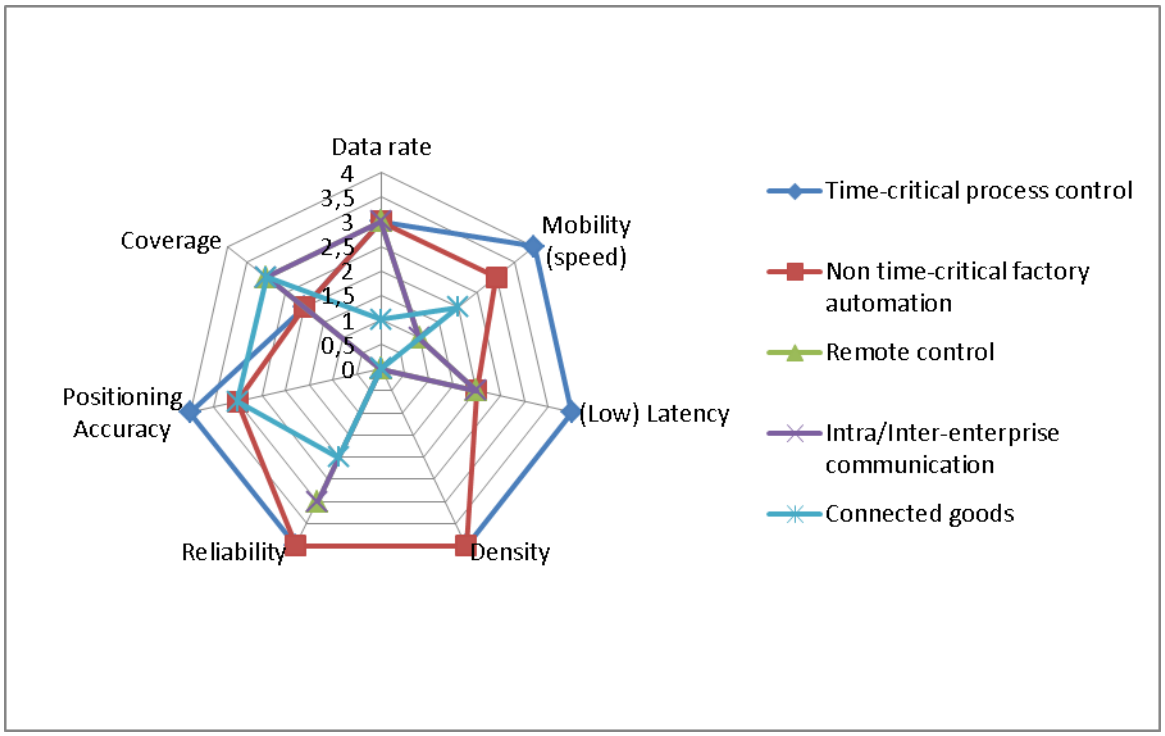
8 Appendix

8.1 Overview of all Use Cases

Reference: Summary of the 5G MWC White Papers

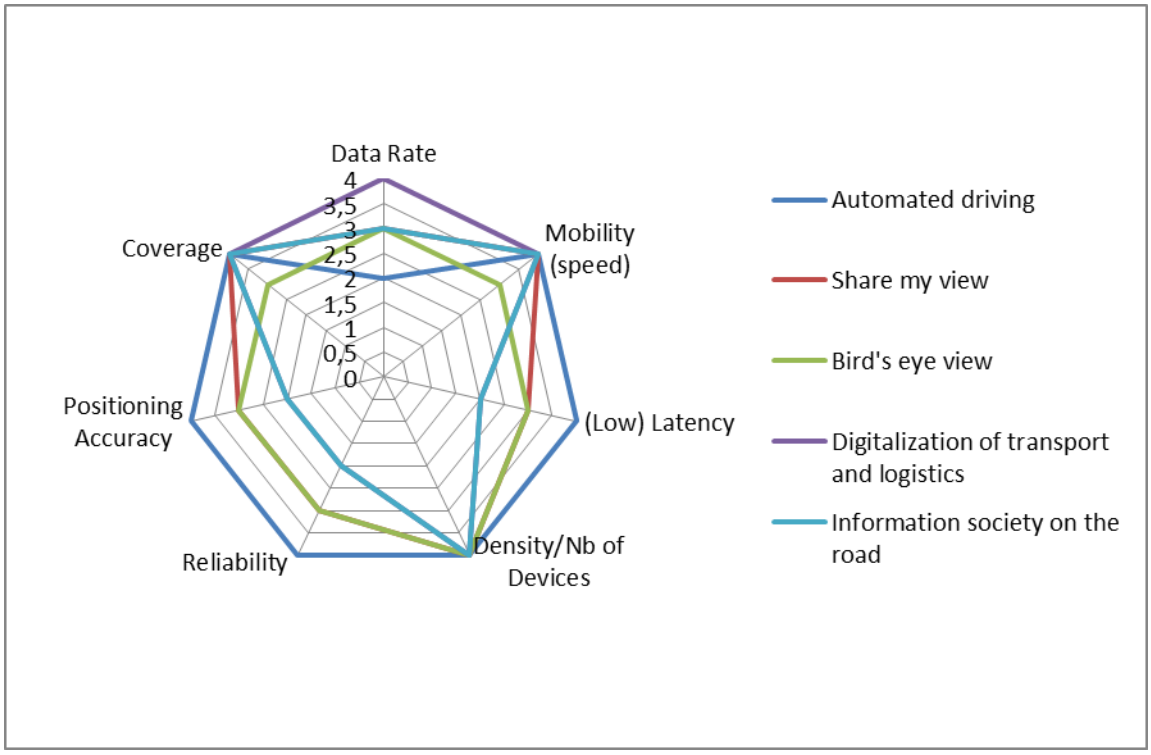
8.1.1 Factories

	Time-critical process control	Non time-critical factory automation	Remote control	Intra/Inter-enterprise communication	Connected goods
Data rate	3	3	3	3	1
Mobility (speed)	4	3	1	1	2
(Low) Latency	4	2	2	2	0
Density	4	4	0	0	0
Reliability	4	4	3	3	2
Positioning Accuracy	4	3	0	0	3
Coverage	2	2	3	3	3



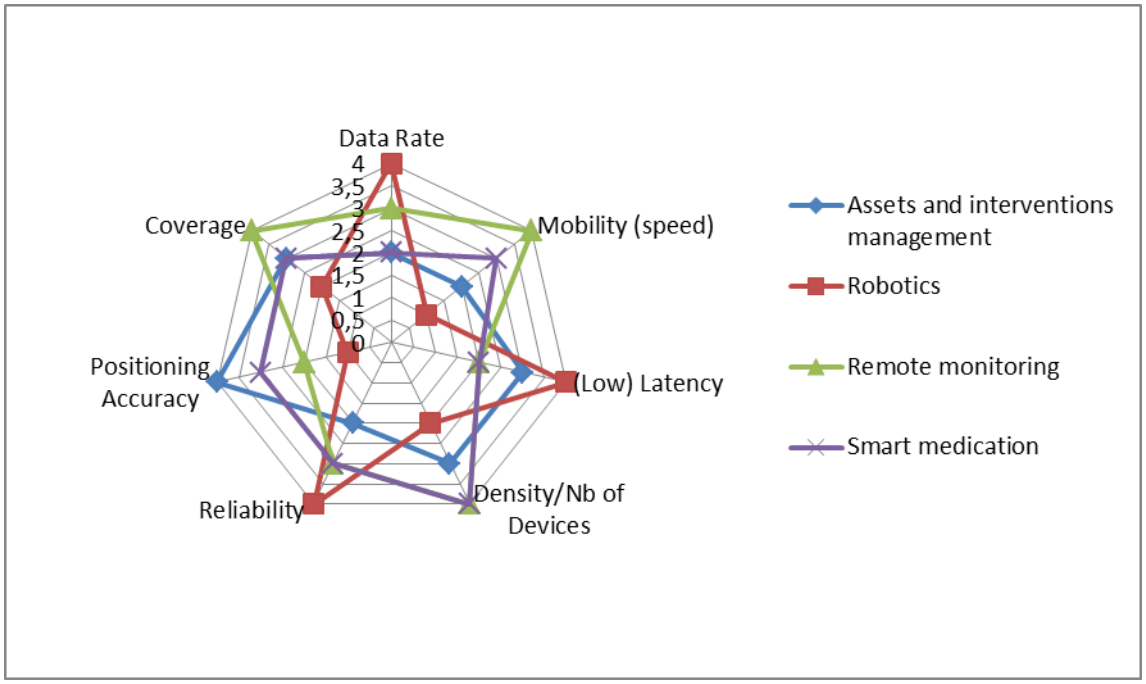
8.1.2 Automotive

	Automated driving	Share my view	Bird's eye view	Digitalization of transport and logistics	Information society on the road
Data Rate	2	3	3	4	3
Mobility (speed)	4	4	3	4	4
(Low) Latency	4	3	3	2	2
Density/Nb of Devices	4	4	4	4	4
Reliability	4	3	3	2	2
Positioning Accuracy	4	3	3	2	2
Coverage	4	4	3	4	4



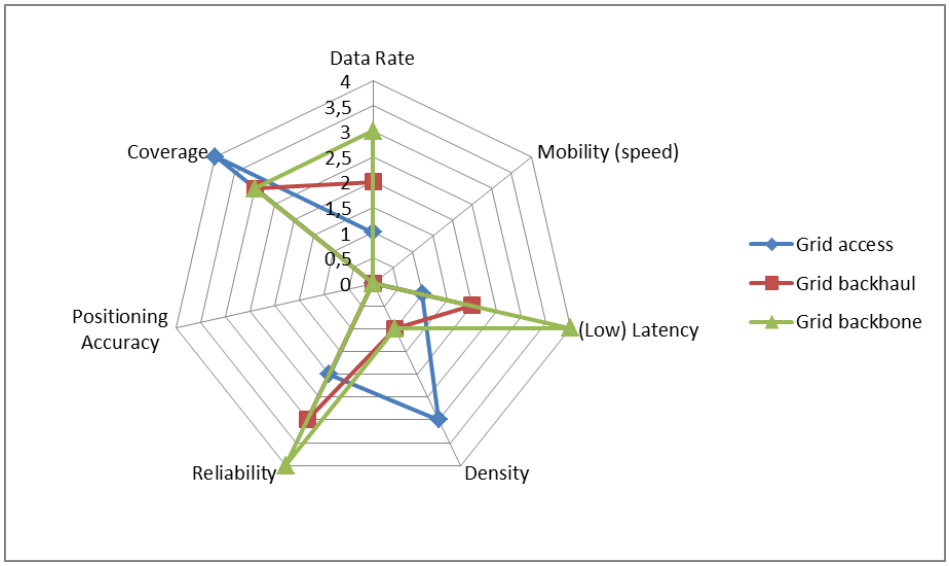
8.1.3 eHealth

	Assets and interventions management	Robotics	Remote monitoring	Smart medication
Data Rate	2	4	3	2
Mobility (speed)	2	1	4	3
(Low) Latency	3	4	2	2
Density/Nb of Devices	3	2	4	4
Reliability	2	4	3	3
Positioning Accuracy	4	1	2	3
Coverage	3	2	4	3



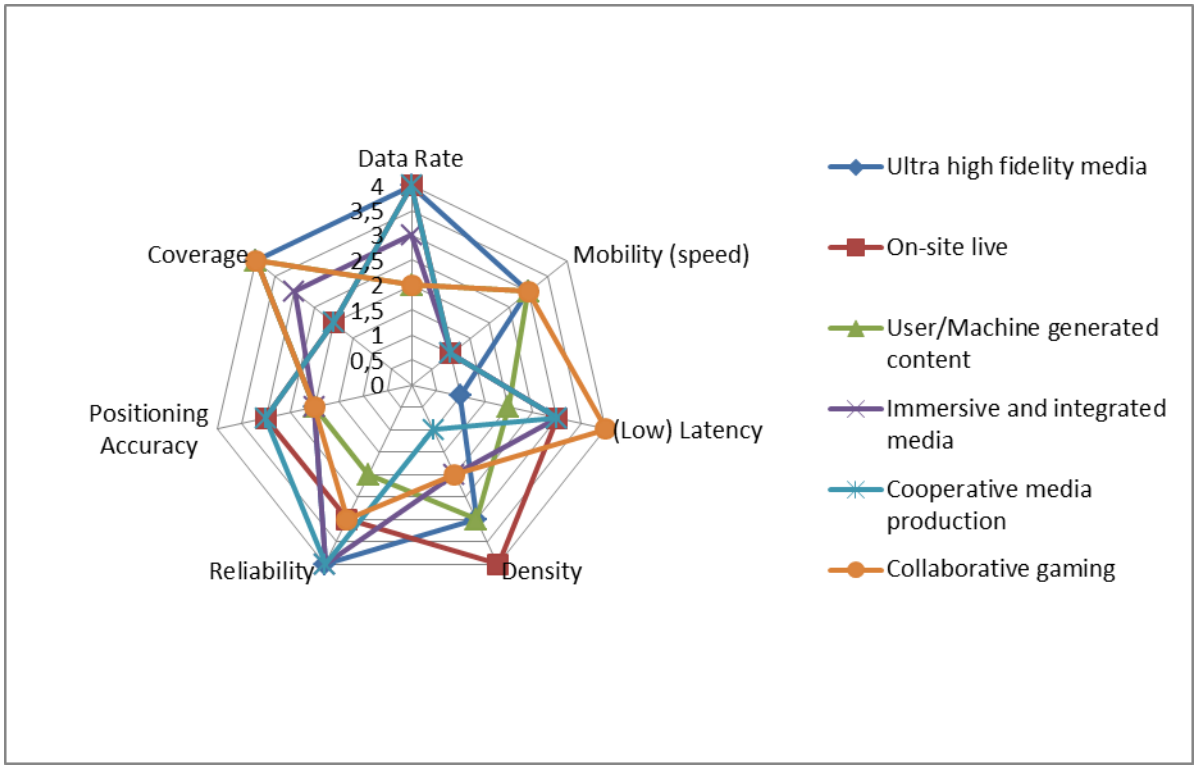
8.1.4 Energy

	Grid access	Grid backhaul	Grid backbone
Data Rate	1	2	3
Mobility (speed)	0	0	0
(Low) Latency	1	2	4
Density	3	1	1
Reliability	2	3	4
Positioning Accuracy	0	0	0
Coverage	4	3	3



8.1.5 Multimedia & Entertainment

	Ultra-high fidelity media	On-site live	User/ Machine generated content	Immersive and integrated media	Cooperative media production	Collaborative gaming
Data Rate	4	4	2	3	4	2
Mobility (speed)	3	1	3	1	1	3
(Low) Latency	1	3	2	3	3	4
Density	3	4	3	2	1	2
Reliability	4	3	2	4	4	3
Positioning Accuracy	2	3	2	2	3	2
Coverage	4	2	4	3	2	4



8.2 Evaluation of Use Cases & Results

		QoE			Homing		Context	Out of Scope
		Data rate	(Low) Latency	Coverage	Mobility (speed)	Density	Positioning Accuracy	Reliability
Factories								
High Req	**Value 4**	**0**	**1**	**0**	**1**	**2**	**1**	**2**
Med. Req	**Value 3**	**4**	**0**	**3**	**1**	**0**	**2**	**2**
Med. Req	**Value 2**	**0**	**3**	**2**	**1**	**0**	**0**	**1**
Low Req.	**Value 1**	**1**	**0**	**0**	**2**	**0**	**0**	**0**
Low Req.	**Value 0**	**0**	**1**	**0**	**0**	**3**	**2**	**0**
Automotive								
High Req	**Value 4**	**1**	**1**	**4**	**4**	**5**	**1**	**1**
Med. Req	**Value 3**	**3**	**2**	**1**	**1**	**0**	**2**	**2**
Med. Req	**Value 2**	**1**	**2**	**0**	**0**	**0**	**2**	**2**
Low Req.	**Value 1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Low Req.	**Value 0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
eHealth								
High Req	**Value 4**	**1**	**1**	**1**	**1**	**2**	**1**	**1**
Med. Req	**Value 3**	**1**	**1**	**2**	**1**	**1**	**1**	**2**
Med. Req	**Value 2**	**2**	**2**	**1**	**1**	**1**	**1**	**1**
Low Req.	**Value 1**	**0**	**0**	**0**	**1**	**0**	**1**	**0**
Low Req.	**Value 0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Energy								
High Req	**Value 4**	**0**	**1**	**1**	**0**	**0**	**0**	**1**
Med. Req	**Value 3**	**1**	**0**	**2**	**0**	**1**	**0**	**1**
Med. Req	**Value 2**	**1**	**1**	**0**	**0**	**0**	**0**	**1**
Low Req.	**Value 1**	**1**	**1**	**0**	**0**	**2**	**0**	**0**
Low Req.	**Value 0**	**0**	**0**	**0**	**3**	**0**	**3**	**0**
M &E								
High Req	**Value 4**	**3**	**1**	**3**	**0**	**1**	**0**	**3**
Med. Req	**Value 3**	**1**	**3**	**1**	**3**	**2**	**2**	**2**
Med. Req	**Value 2**	**2**	**1**	**2**	**0**	**2**	**4**	**1**
Low Req.	**Value 1**	**0**	**1**	**0**	**3**	**1**	**0**	**0**
Low Req.	**Value 0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Summary								
High Req	**Value 4**	5	5	9	6	10	3	8
Med. Req	**Value 3**	10	6	9	6	4	7	9
Med. Req	**Value 2**	6	9	5	2	3	7	6
Low Req.	**Value 1**	2	2	0	6	3	1	0
Low Req.	**Value 0**	0	1	0	3	3	5	0