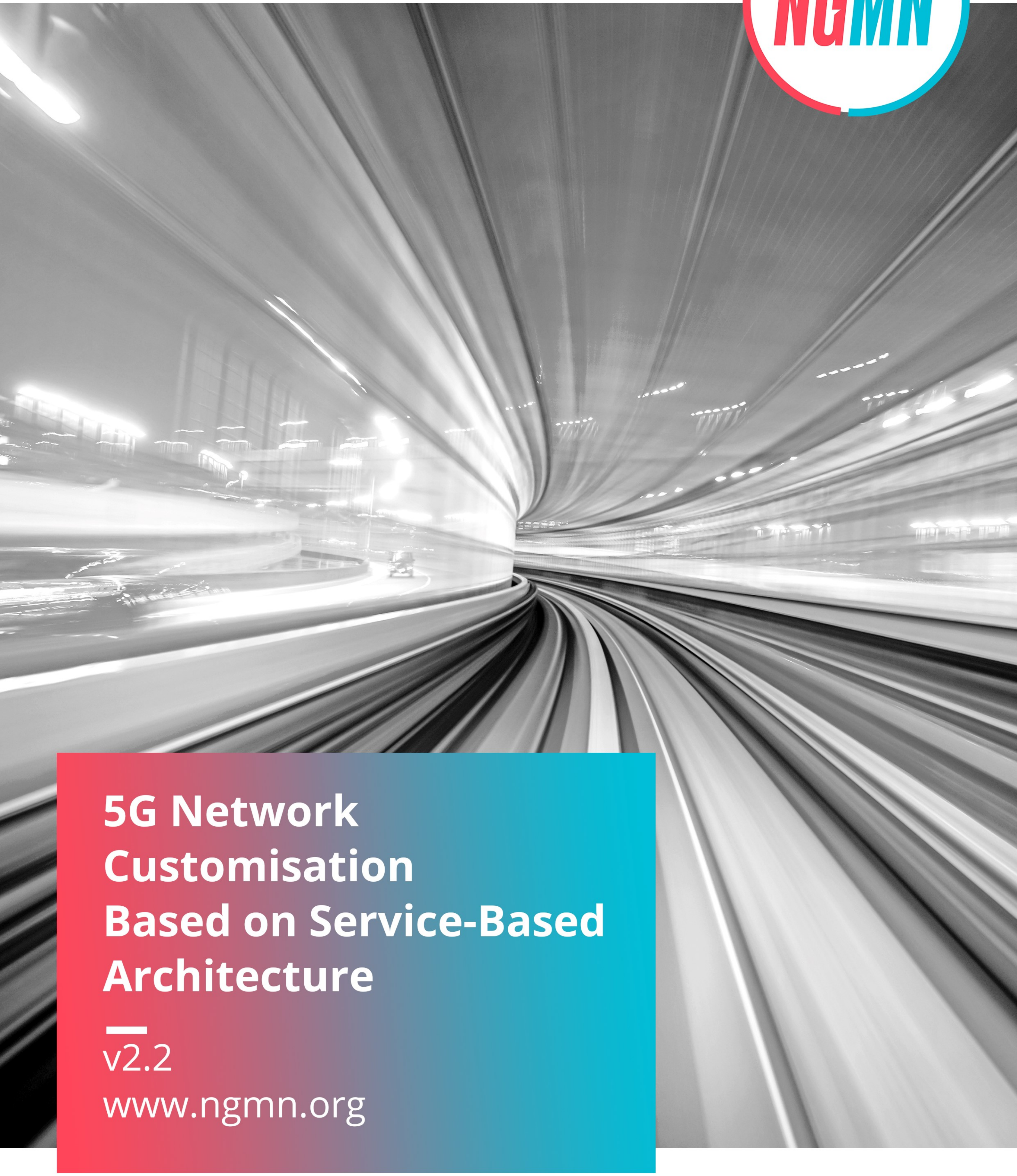

5G Network Customisation Based on Service-Based Architecture

v2.2

www.ngmn.org

WE MAKE BETTER CONNECTIONS

5G NETWORK CUSTOMISATION BASED ON SERVICE-BASED ARCHITECTURE

by NGMN Alliance

Version:	2.2
Date:	27.07.2021
Document Type:	Final Deliverable (approved)
Confidentiality Class:	P - Public

Project:	Service-Based Architecture
Editor / Submitter:	**Dan Wang (China Mobile), Tangqing Liu (China Mobile)**
Contributors:	**Tangqing Liu (China Mobile), Dirk Trossen (Interdigital), Dan Wang (China Mobile), Ulises Olvera-Hernandez (Interdigital), Sebastian Robitzsch (InterDigital), Hannu Flinck (Nokia), Georgios Gkellas (Nokia), Kay Hänsge (Interdigital)**
Approved by / Date:	**NGMN Board, 15th July 2021**

NGMN e. V.
Großer Hasenpfad 30 • 60598 Frankfurt • Germany
Phone +49 69/9 07 49 98-0 • Fax +49 69/9 07 49 98-41

Abstract

In order to investigate the 5G service-based architecture (SBA) evolution to support various Verticals' requirements, this White Paper provides some considerations including network functionality evolution, platform design and dynamic network resource allocation. The main aspects of this paper cover User Plane Function (UPF) service design and capability exposure, edge computing support, 5G network customisation, and dynamic network slice allocation.

For UPF service design and exposure, considering the progress in 3GPP, this paper proposes to investigate the UPF service for event exposure, which is an added value for serving Verticals. Also, UPF service modules and service chains can be further considered.

For edge computing support, an example of extending SBA to data centres and terminals is provided, and the flexible edge application selection is introduced.

For 5G network customisation, a unified platform Agile and Customised 5G Network (AC5N) is proposed which can expose the capabilities of network, artificial intelligence (AI), security and management.

For dynamic resource management within a slice, the concept of pinning resources for specific customers while selecting suitable control plane services is proposed.

Contents

1 INTRODUCTION

The main objective of this White Paper is to investigate the technologies and approaches that can support operators to customise 5G networks based on SBA, in order to identify gaps and challenges, which include:

- The investigation of UPF services including control plane interaction and user plane service chaining
- The investigation on how to integrate Edge Computing in a 5G SBA network including edge application server dynamic deployment and selection, and a flexible service routing path configuration
- The investigation on how to realize AC5N to meet various Verticals' requirement considering the 5G System (5GS) is deployed in a virtual environment and based on service-based architecture
- The dynamic service management within one network slice, especially for dynamic control level mechanisms to pin resources within a resource pool to specific customers.

This White Paper is the deliverable of the NGMN project SBA Phase 3. It calls for industry cooperation on standardisation, development and promotion of service-based 5G architecture and related technologies.

2 UPF SERVICE FOR EXPOSURE

2.1 UPF Interaction with Other 5G NFs Besides SMF

The current 3GPP architecture consists of a User Plane Function (UPF), interacting via the PFCP based N4 to the Session Management Function (SMF), via the GTP-U based N3 to the RAN and via the GTP-U based N9 to another UPF. A UPF might support various optional features which might not be needed for handling the user plane traffic of all PDUs. PFCP offers an advertisement mechanism via which SMFs learn the capabilities of each UPF and subsequently select the proper UPF per case. Already since Rel. 15 UPFs could optionally advertise a profile in the Network Repository Function (NRF), facilitating their discovery and selection by the SMFs. 3GPP Rel. 17 in CT4 [1] will be working on any enhancements needed on the UPF profile.

Recently, 3GPP has decided the introduction of a UPF service for exposing information to local AFs via local NEF in edge computing deployments for Release 17. Relevant data may include specific UE location, UPF load status, the delay from the UPF to the UE, and the QoS profile for specific applications. This information is related to the local 5G network performance and capabilities, and Verticals demand information on whether the 5GS SLA can be guaranteed or not.

We strongly believe that a UPF exposure service should be introduced and further used in other scenarios, like e.g., for providing Network Data Analytic Function (NWDAF) with user plane information which is valuable for analytics. At the same time such a UPF service design should take care that it does not affect the performance and low latency requirements when handling user plane traffic.

2.2 Considerations for Chaining UPFs/Services

Already since 3GPP Release 15 more than one UPF could be chained via N9 tunnels for covering the requirements of a PDU session. However, when such chaining is introduced, one should consider the trade-off between processing latency and low latency requirements from applications. Or, in other words, the trade-off between flexibility and latency. Today, some deployments for low latency prefer to integrate even some basic functions in the UPF (e.g., DPI, NAT/firewall) of service chains, to address the strict latency requirements from enterprises.

Studying various deployments, one example of flexible introduction of new functionalities might be the Fixed Mobile Convergence (FMC) with Access Traffic Steering, Splitting and Switching (ATSSS). An entity called Access Gateway Function (AGF) is necessary for the fixed access integration to the 5G Core network (5GC). AGF could either be a standalone entity, tunnelling user plane traffic via N3 to UPF, or for optimization collocated with the UPF.

FMC is usually coupled with ATSSS. The ATSSS concept is based on the so-called Multi-Access PDU session (MA-PDU session) which consists of a child PDU session per access. For FMC, for example, it introduces a UPF combining traffic from the fixed and the mobile network. This kind of UPF functionality, from our view, is not applicable to many PDU sessions, and may be more specific for Verticals and cannot be supported in all UPFs. 3GPP defined that it could be introduced flexibly as a UPF feature. An MA-PDU session is anchored at a UPF supporting the ATSSS functionality. The anchor UPF might be connected via N9 tunnels to a UPF per access type.

When considering introducing 5G, Verticals expect the 5GS to solve their IT connection requirements and all the existing Wi-Fi to be integrated with 5GS, which can simplify their IT management. The UPF with ATSSS functionality can be considered as a solution to address the requirements.

Another example of UPF chaining is support of local offload to an edge data centre. Either for low latency or to avoid overwhelming the transport, in many cases it is meaningful to introduce an Uplink Classifier/Branching Point in the user plane path, which is properly programmed to classify and offload certain traffic that has no meaning to go to a central anchor.

2.3 Considerations on Introducing Further UPF Services

In the past 3GPP evaluated the possibility of embracing the UPF in the service-based architecture in their Release 16 study on eSBA [2]. The main idea was to enhance N4 to support SBI and introduce UPF services which could be leveraged by an SBI based N4. The conclusion of the study was that the UPF would not be included in 5G SBA mainly because of no use case driving such a change and the expected performance impact.

We believe that the topic will be re-evaluated in the future in the relevant SDOs as there are merits from a universalized (e.g., including N2 and N4 interfaces in SBA architecture) handing in the whole 5GC. However, we are also aware that to benefit from such a change, various aspects should be considered such as smooth transition, performance and low latency requirements to identify the optimal approach to proceed. Possibilities to mirror PFCP procedures to SBA services/operations might be considered or evaluated by relevant SDOs.

3 SUPPORT FOR EDGE COMPUTING

3.1 Blueprint Architecture for Edge Computing over 5G

The vision here is that the edge computing blueprint in 5G is being realized through micro-services in a distributed data centre, with the constituents of said data centre located on the Internet (distant cloud), the operator Data Network (DN) or even in the participating terminals. In other words, terminals, operator-based computing resources, as well as the distant cloud are Layer 2 connected computing resources that communicate in a manner akin to existing data centres, i.e., over a (virtual) Layer 2 connectivity, albeit distributed across many different sites. Internet connectivity is achieved through packet gateways.

This vision aligns with the notion of private networks, e.g., utilized in corporate networks, where the services provided by the private network provider are connected through suitable Layer 2 connectivity. Such Layer 2 can be native, e.g., within private network premises, or provided through overlay techniques such as VPNs, while specifically interconnecting to the public Internet through gateways. Architecturally, however, the same service architecture is being utilized in both private and public networks, namely the provisioning of services based on Internet protocols such as HTTP and others. In other words, this common service architecture spans from the distant cloud across (infrastructure) edge to far edge, including resources on terminals, through the same protocol technologies, supported by the specific realisations for SBA in 5G, i.e., the Service Communication Proxy (SCP) and NRF. This application of Internet service and virtualization technologies in the (cellular) access subsystem is often captured through the tagline "IT meets Telecom".

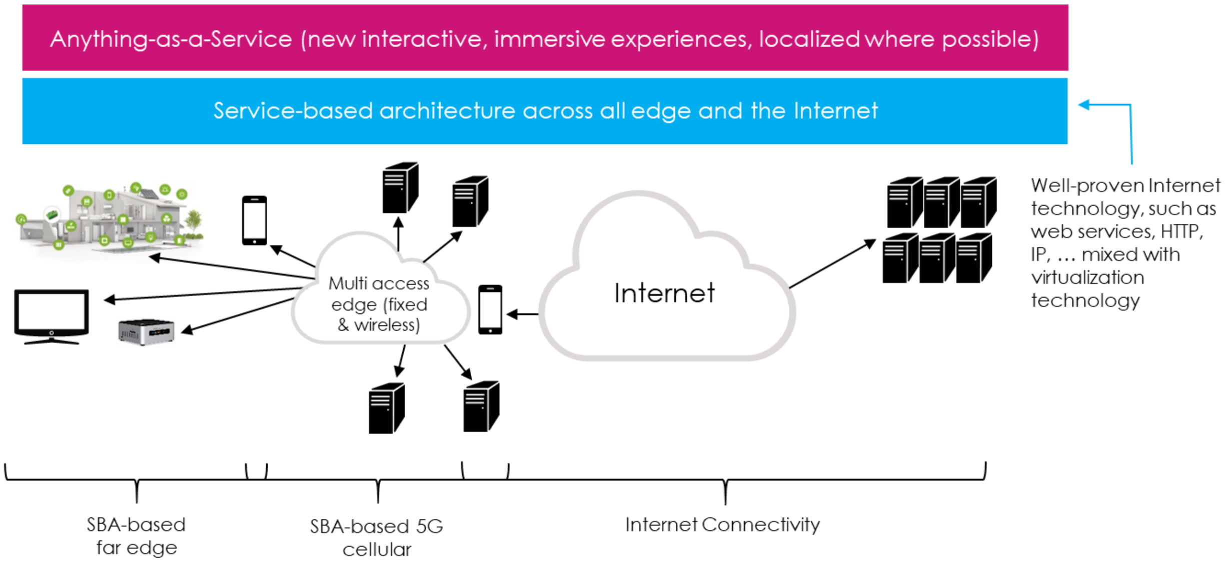
Anything-as-a-Service (new interactive, immersive experiences, localized where possible)
Service-based architecture across all edge and the Internet
Multi access edge (fixed & wireless)
Internet
Well-proven Internet technology, such as web services, HTTP, IP, … mixed with virtualization technology
SBA-based far edge
SBA-based 5G cellular
Internet Connectivity

Figure 1: Micro-Services from the Distant Cloud over the 5G Edge to End User Terminals

The eSBA work in 3GPP provides the suitable connection to realizing Figure 1 in the cellular subsystem. eSBA plays a crucial role in realizing not only edge computing but generally in realizing micro-services across the various resource instances, as indicated in Figure 1. To understand this role, a recall of the positioning of eSBA in relation to the cloud-native vision, as shown in Figure 2 in SBA Project White Paper of Phase 2 [3] is useful. This figure shows the realisation of services over a 'Service Messaging Platform', as defined through the eSBA components in [4] (for example service registry is the NRF, service discovery is performed by the consumers or the SCP, service registration is performed by the producers, authorization is granted by the NRF, failover management is handled either at the HTTP client or the SCP platform level, etc.). This platform is available over an infrastructure that consists of (distributed) data centres, interconnected via a transport network with SDN or SD-WAN.

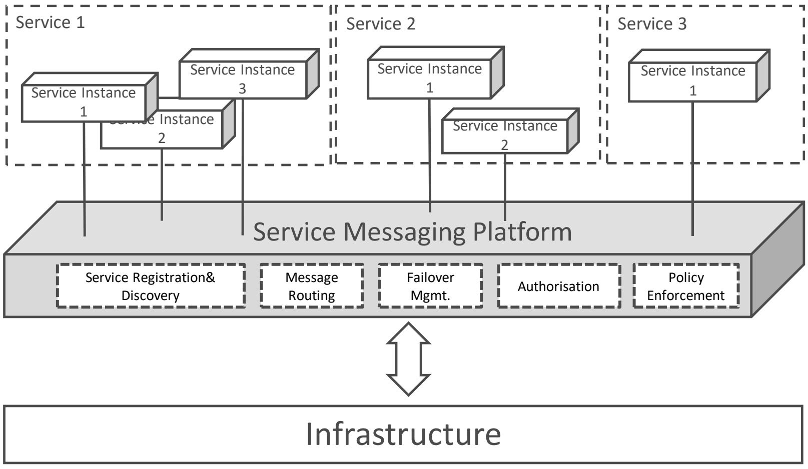

Figure 2: Service Based Architecture [3]

An extension of eSBA beyond the control plane to realize edge computing in 5G is preferred. Particularly the role of eSBA components as defined in [4], i.e., the SCP (preferred transparently) as well as the NRF interacting with the SCP in a delegated discovery mode. Specifically, the realisation of eSBA components (consumers and producers) goes beyond the current scope of the 3GPP eSBA work. Most importantly, the scope of eSBA component functionality should realize any 5G System related functionality, including user plane function (control and user data treatment). This would also allow to have terminal-based user services. The aforementioned statement is being met, that eSBA (plus 5GLAN capabilities) provides a suitable answer to Edge Computing in 5G.

The following Figure 3 shows the realisation of a service, including a micro-service, through the aforementioned SBA components. Those services can be located on the edge, in regional office data centres of operator(s) and on terminals themselves, but also reached through the operator's border gateway on the Internet. Notable here is the utilization of Layer 2 connectivity between the services, including the terminal-based services. Furthermore, a lifecycle management and control (LCM-C) provides suitable capabilities to control the lifecycle states of those service instances through suitable virtualization and cloud techniques.

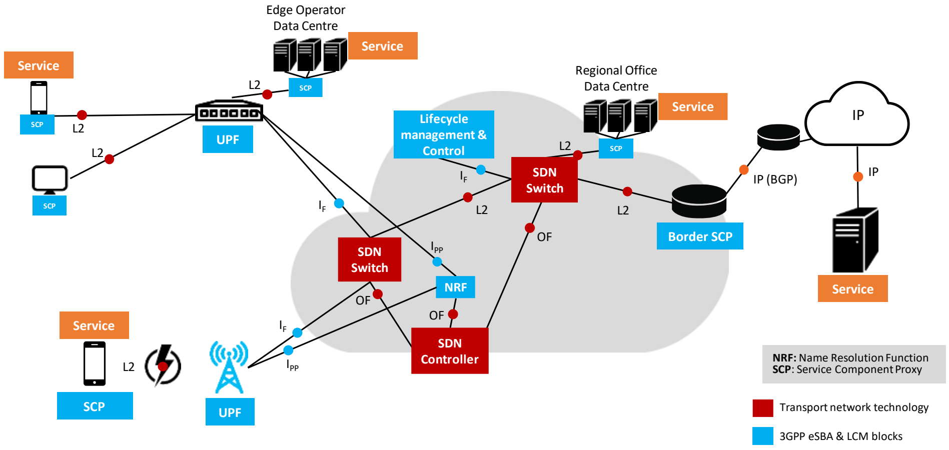

Figure 3: SBA-based component architecture for Edge Computing in 5G

3.2 Flexible Application Server Selection

5GC supports for Edge Computing by transferring traffic between a UE and a local DN through a local UPF. In 3GPP, the UPF can be (re)selected based on the UE location and application server location by Control Plane Services (CPS). The UE location is identified by the Tracking Area or cell-ID, and the application service location is identified by Data Network Access Identifier (DNAI).

When selecting suitable UPSs within a chain for a PDU session, there should be a method to calculate and select suitable UPSs. It is proposed to evaluate UPSs based on distance remaining computing capability, CN part functionalities and the number of 3rd party applications which the UE requires to access, to enable a UP service chain as offering from the network provider to the UEs.

4 AGILE AND CUSTOMISED 5G NETWORK (AC5N)

4.1 Introduction

A 5G system is highly flexible to accommodate various customisation scenarios supporting Verticals' requirements. Operators may obtain AC5N capabilities with SBA as the network foundation, network slicing as the service framework, independent network platforms as the core, and APIs as an effective means. This will help users or vertical industry partners to deeply participate in the definition and design of network services, making the 5G network a powerful booster for the development of the telecom sector and vertical industries.

4.2 Realisation of AC5N

Since the network has been considered as the infrastructure for the connections, the network services could be offered to users by selecting intelligent connections. AC5N is constructed from five aspects:

1) **Service-based Architecture as the foundation of AC5N:** Compared with the "black box" of a traditional network, which only designs signalling flow between network elements, the SBA transforms the 5G network to a "grey box" which is more transparent and visible regarding the network functions' internal processes. The modules of network services, the unified APIs, the cloud-native platforms and service management framework help operators to create a customised 5G network flexibly and efficiently.

2) **Network slicing as the basic architecture to fulfil the Verticals' requirements:** Network slicing provides an end-to-end virtual network for Verticals in a physical or logical isolation. An AC5N network can be provided to Verticals based on network slicing architecture, which can be customised and managed independently.

3) **Independent network platform as core aspect:** Operators can lead the definition and construction of the 5G network, grasp the network platform, and at the same time support the intelligent service orchestration based on the network capability library. With the variation of Verticals' requirements, the 5G network can be adjusted quickly at any time, in order to achieve a quick response and agile launch.

4) **Service based API as the effective means:** The interface of AC5N is service based and the interaction of the network services is implemented by API invoking. The SBI benefits from the API definition of the operator and network services invoked by 3rd party AF.

5) **Connection design as users' requirement:** AC5N enables users to customise network services based on their requirement. On one hand, the vertical industry partners can customise services through platform and unified API. On the other hand, users can develop the network functions/services and publish the network functions/services to the AC5N network. To some extent, the users can therefore be considered a producer of 5G network.

4.3 Introduction of AC5N Platform

Based on the SBA and network slicing framework, the 5G system can achieve network functions' customisation and a flexible interaction design. However, it still needs a platform which can translate Verticals' requirements into network requirements and network capabilities. As mentioned in section 4.2, introducing AC5N can achieve a decomposition of Verticals' requirements and network design. AC5N is different from Communication Service Management Function (CSMF). CSMF is an executor of AC5N platform in mobile networks.

Verticals' requirements of a customised network serve as an input to the AC5N platform. The platform then transfers users' requirements into network requirements, and schedules and combines resources, which include access network, 5GC, transport and IP network, cloud resource, AI and security capabilities into a customised network.

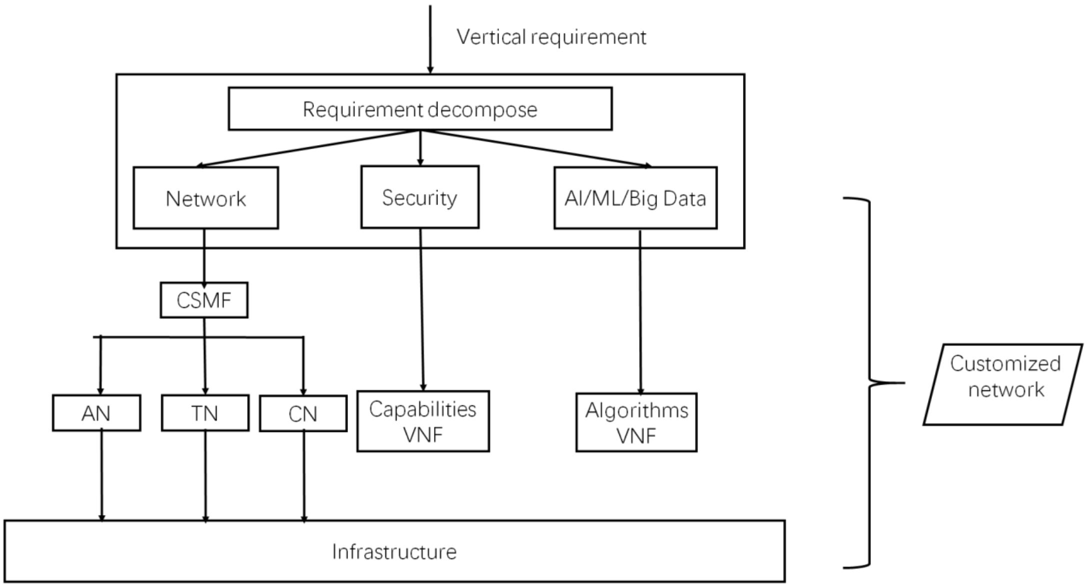

Figure 4: Concept of AC5N platform

Here are some functionalities of an AC5N platform:

- Allocate access frequency, closed access group ID as required by Verticals.
- Combine the network function and network resource and provide a customised network slice. The network slice can support ultra-low latency, massive connection, high reliability, and service level guarantee capabilities, as required by Verticals.
- Provide a suitable routing path based on transport and IP network arrangement.
- Invoke cloud computing resources based on user location and requirements. AC5N can provide computing resources to users for application, algorithm, and data analysis services.
- Provide edge computing capability including edge platform resource allocation, edge application deployment and management, and network exposure capability invocation.
- Provide AI/ML capabilities such as machine visioning, semantic algorithm and smart network based on AI platform.
- Provide network analysis capabilities and training models based on massive data from the mobile network, Internet, application traffic, and public data.

- Provide security capabilities such as data security, access authentication, data encryption and secret key storage.

5 DYNAMIC RESOURCE MANAGEMENT WITHIN A SLICE

5.1 True Adoption of Cloud Native in Cellular Telecommunication Systems for Slicing

The adoption of cloud native principles in 5G systems has a much greater impact on the overarching system design than simply turning monolithic network functions into microservices and to orchestrate them as islands of microservice management systems. One of the key design choices when realising an application in a cloud native fashion is the separation of packet routing, application monitoring and analytics, and service orchestration from the actual application, whose sole objective must focus on processing incoming requests and returning a response. The naming ontology of services over the Internet, i.e., Fully Qualified Domain Names (FQDNs), is being used to allow the logical separation of functions which form the application. Thus, each client is fully aware about the FQDN of the next function (server) which can serve the request the client issues.

However, how an application – which is decomposed into a set of functions realised as microservices – is initially orchestrated and lifecycle managed at run time must not become part of the application. This allows a truly cloud native realisation of the application where testing, staging and production environments can use the same code base without any modifications. Furthermore, and most importantly for telecommunication systems, how requests are being routed to the instance which can serve them, and which particular instance to choose in the first place, must be also fully decoupled from the actual application. This capability – commonly referred to as “service routing”, is baked into all microservice management systems via sidecars and load balancers. However, those systems are operating within their own realm only and do not offer service routing across multiple islands, as this would require a system-wide understanding of the network, its resources and current state.

This challenge has been formally addressed by the introduction of the SCP, but many cloud native concepts related to the separation of concerns of an application and underlying management and routing services are still not being addressed in today’s cellular telecommunications systems. The application and the 5G Core still determine where requests are being sent to, monitor themselves to make such a decision, and consequently render the SCP as an optional component only, undermining its importance with a very limited functionality far away from truly adopted cloud native principles.

The following sections highlight examples on how a true cloud native and service-based communication network offers further enhancements and new scenarios for the service delivery.

5.2 Pinning of Service Function Endpoints

Although the services themselves may decide to which producer they establish a communication session, the SCP should be in charge of discovering and selecting possible endpoints for a fine-tuned communication. The producer selection is based on discovery parameters, policies, and availability constraints, done by the SCP and its logic. However, a manual approach may be added on top for an ad-hoc policy update which prioritizes selected service endpoints, among a group of functionally equivalent producers, for a group of clients in a specific (e.g., location-dependent) scenario. Prioritizing service endpoints for certain clients can be called Service Pinning.

A Service Function Manager (SFM) is responsible for soliciting the suitable service functions, where instances of such service functions are provided by the various service hosts. Such service hosts could be specific data centre resources realizing service instances for specific FQDNs (service identifiers). A Service Registration Manager (SRM), e.g., with similar functionality as the current NRF while respecting the cloud native principles (described in Section 5.1), is responsible for receiving registrations to specific named service functions (such as amf.mnc.mcc.3gppnetwork.org) and exposing those named services in the network.

To enable the pinning approach, a ContextID (CoID) is required to represent the specific reason for pinning the service execution. This identifier might be a slicing ID (i.e., the S-NSSAI) for the overall slice, in which the resources are managed based on a customer requirement. Each initiating client is represented by a unique ClientID. Even though, when using an identifier per client which results in a higher number of FQDNs registered against the service routing layer, the total number of FQDNs in a large economy would be similar to the number of Top-Level-Domains a DNS must be able to resolve. However, it is also possible to group clients into a group and let them be represented by a GroupID, referring to the multiple clients leveraging the pinning procedures. Such ClientID/GroupID might be a consumer set identifier or a region id, where one or more consumer sets reside. The identifier could also be generated fully arbitrary without any reference to the underlying mobile core network.

When initiating the pinning, the SFM instructs each selected service host to register the selected service function SFx as <ClientID>.<ContextID>.amf.mnc.mcc.3gppnetwork.org. Hence, a dedicated treatment to process requests from certain clients can be achieved. To consider this approach, it is necessary to keep the following statements in mind:

- Pinning fine tunes the general routing and forwarding procedures of the SCP.
- Pinning should only be used for a certain time in a certain environment at runtime.
- Pinning decisions and target addresses have to be communicated to the client (consumer) to leverage the dedicated resources.

5.3 Resource Quotas for service delivery

A second example of adopting the cloud native service delivery is divided into dedicated resources among home and visiting clients and their telco services. The assignment of resources within a slice is specific to customers of the home operator who overall manages the resources for the service being provided here across all its customers. As an example, eMBB is a service offered by the network operator. Hence, the home operator, which has various customers (e.g., its own customers as well as those of customers belonging to the RED and GREEN customer group (see Figure 5) at the service level, can manage the resources dynamically, which are specific to each customer (group), with the control level mechanisms as described in this section.

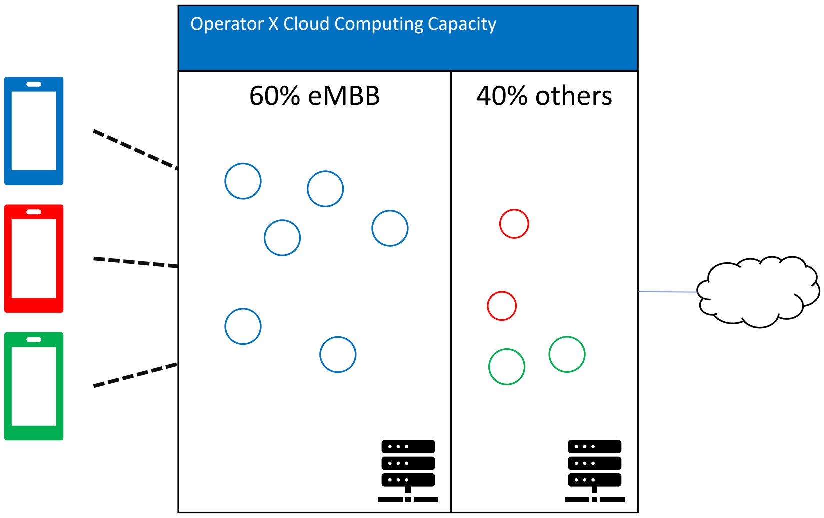

Figure 5: Customer Quota Assignment within a single Service

Figure 5 depicts the resource sharing in terms of a quota (e.g., the ratio of microservice instances) to enable home clients and visiting clients to use the same computing capacity of that specific slice. Here, the priority for dedicated resources is set to 60% for home customers, while visiting clients must share their computation requirements within the remaining 40% of the overall existing computing capacity.

The resource quota approach may be applied to different levels of the overall infrastructure and platform components. Within the scope of deploying microservices, the resource quota can be used to scale the amount of running service instances for different client groups.

6 SUMMARY

This White Paper investigates the 5G network customisation based on a Service Based Architecture. There are mainly four parts:

- **UPF service for exposure:** There is an introduction to the UPF services investigation in 3GPP SA2 and CT4. For UPF services, the UPF exposure service is considered to be introduced to support the network analytics and low latency information exposure for edge computing. For UPF services chaining, some UPF functionalities e.g., AGF, ATSSS functionality, or ULCL/BP functionality, can be considered as standalone UPF services to be integrated in the user plane path through e.g., service chaining.
- **5G SBA support for Edge Computing:** There is a vision that the 5G SBA architecture provides a service framework e.g., NRF or SCP, which is a common service architecture that can span from cloud, network to terminals, through the same protocol technologies. To support dynamic communication, services are connected through a suitable Layer 2 connectivity technology e.g., VPN or other overlay methods. In this case, edge computing can be integrated in the 5GS and introduced on-demand. With this, for example, the suitable application server selection and routing path configuration can be done more dynamically.
- **The Agile and Customised 5G network:** Operators can obtain AC5N with 1) SBA as the network foundation, 2) network slicing as the service framework, 3) autonomous network platforms as the key point, and 4) APIs as an effective means. Based on SBA and network slicing, AC5N platform can translate requirements of Verticals to network requirements and achieve a customised network design.
- **5G SBA support for dynamic resource management within a slice:** The concept of dynamic resource management within a slice is introduced in this paper. Such dynamic resource management provides a long-term management of a resource pool associated to a specific Vertical service, such as eMBB or V2X. This resource management depends on the dynamic control level mechanisms to 1) pin resources within a resource pool to specific customers, 2) select and add suitable control plane

services dynamically within the network procedure. In order to realize 1) and 2) above, the existing NRF framework can be leveraged and some new functions e.g., the service function manager, could be introduced to support the suitable service functions' selection and to add value to the procedure.

This paper provides the considerations for a 5G SBA evolution, and the above four features can be considered to help creating a dynamic and customised 5G network.

LIST OF ABBREVIATIONS

3GPP	3rd Generation Partnership Project
5GC	5G Core network
5GLAN	LAN support in 5G
5GS	5G System
AC5N	Agile and Customised 5G Network
AF	Application Function
AGF	Access Gateway Function
AI	Artificial Intelligence
API	Application Programming Interface
ATSSS	Access Traffic Steering, Splitting and Switching
BP	Branching Point
CoID	ContextID
CPS	Control Plane Services
CSMF	Communication Service Management Function
CT	Core Network and Terminals
DN	Data Network
DNAI	Data Network Access Identifier
DNS	Domain Name System
DPI	Deep Packet Inspection
eSBA	enhanced Service-Based Architecture
FMC	Fixed Mobile Convergence
FQDN	Fully Qualified Domain Names
GTP-U	GPRS Tunnelling Protocol User Plane
HTTP	HyperText Transfer Protocol
IT	Information Technology
LCM-C	Lifecycle Management and Control
MA-PDU session	Multi-Access PDU session, a PDU Session that provides a PDU connectivity service, which can use one access network at a time, or simultaneously use one 3GPP access network and one non-3GPP access network
ML	Machine Learning
NAT	Network Address Translation
NEF	Network Exposure Function

NF	Network Function
NRF	Network Repository Function
NWDAF	Network Data Analytic Function
PDU	Protocol Data Unit
PFCP	Packet Forwarding Control Protocol
QoS	Quality of Service
RAN	Radio Access Network
SBA	Service-Based Architecture
SBI	Service Based Interface
SCP	Service Communication Proxy
SDN	Software-Defined Networking
SDO	Standards Developing Organisations
SD-WAN	Software-Defined Networking in a Wide Area Network
SFM	Service Function Manager
SLA	Service Level Agreement
SMF	Session Management Function
S-NSSAI	Single Network Slice Selection Assistance Information
SRM	Service Registration Manager
UE	User Equipment
ULCL	UpLink Classifier
UP	User Plane
UPF	User Plane Function
UPS	User Plane Service
VPN	Virtual Private Network

REFERENCES

[1] 3GPP TR29.820 “Study on Best Practices for Packet Forwarding Control Protocol (PFCP)”V0.3.0, November 2020, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3752

[2] 3GPP TR 23.742 “Study on Enhancements to the Service-Based Architecture”V16.0.0, December 2018, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3457

[3] NGMN Service-Based Architecture in 5G Case Study and Deployment Recommendations V2.0, June 2019, https://www.ngmn.org/wp-content/uploads/Publications/2019/190919-NGMN_Service-BasedArchitecturein5GCaseStudyandDeploymentRecommendations-v2.4.pdf

[4] 3GPP TS 23.501 “System architecture for the 5G System (5GS)” V16.1.0, June 2019, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144