Service-Based Architecture in 5G

by NGMN Alliance

Version:	**V1.0**
Date:	**19-January-2018**
Document Type:	**Final Deliverable (approved)**
Confidentiality Class:	**P - Public**
Authorised Recipients: (for CR documents only)	

Project:	**Service-based Architecture in 5G**
Editor / Submitter:	**Dan Wang ,Tao Sun (China Mobile)**
Contributors:	**AT&T(Farooq Bari) , Bell Canada (Erfanian Javan),China Mobile(Dan Wang, Tao Sun, Bo Yang, Wei Chen), Deutsche Telekom AG (Hans J. Einsiedler, Kay Haensge), NTT DOCOMO(Srisakul Thakolsri), Sprint (Serge Manning), SK Telecom (Sangsoo Jeong), U.S. Cellular (Sebastian Thalanany), Vodafone (Neal Adrian)**
Approved by / Date:	**NGMN Board, 19th January 2018**

1 Introduction 3
2 Reference 3
3 Definitions 3
4 Motivation and requirements 3
4.1 Related 5G network requirements 3
4.2 Motivation of Service-based architecture 4
4.3 Related Technologies 4
4.3.1 General 4
4.3.2 Cloud Computing 5
4.3.3 Virtualization 5
4.3.4 Cloud Native 5
4.3.5 Microservices 5
4.3.6 Stateless services 5
5 Concept of Service-based architecture in 5G 5
5.1 Concept of Service 5
5.1.1 The description of service 5
5.1.2 Network Function Services in 3GPP 6
5.2 The description for service framework 6
5.2.1 Service registration 6
5.2.2 Service authorization 6
5.2.3 Service discovery 6
5.2.4 The description for service-based interface 6
5.3 Service-based Architecture design for 5G network 7
5.3.1 Target Service-based Architecture for 5G network 7
6 SBA deployment consideration 9
6.1 Service deployment principle 9
6.2 One implementation of 5G SBA 9
6.3 SBA Deployment Choices with NFV 11
6.3.1 Service deployment option 11
6.3.2 Support of performance optimization 11
7 SBA for slicing 12
7.1 SBA support the on-demand design for network slicing 12
7.2 Network slicing management support 12
8 Service for Edge Computing 12
9 SBA for network exposure 12
9.1 Exposure of SBA to other external networks (e.g. 3rd parties) 12
10 Data Service 12
11 Service across network and operators 13
11.1 SBA in inter-PLMN environment 13
11.2 Support for Legacy Systems 13
11.3 Support for Non-3GPP access types 13
12 Summary 13

1 INTRODUCTION

In March 2017, the work item of “Service-based architecture in 5G” was approved by the NGMN Board. The main target of this work group is to investigate high-level requirements, use cases and guidelines for how operators can efficiently introduce and operate a service-based 5G network, i.e.,

- Identify high-level requirements or guidelines on Service-based architecture design in 5G including network function, interface/protocol, API design principle;
- Investigate how operators can leverage Service-based architecture in 5G in the best way, e.g., customized network slicing, updating and managing network feature dynamically. Samples, possible approaches and guideline are expected
- Investigate how the Service-based architecture can enable network exposure, e.g., expose network and function capabilities to third parties for value-added services.

This White Paper is the output of the work item. It calls for the industry cooperation on standardisation, development and promotion of the service-based 5G architecture.

2 REFERENCE

[1] NGMN 5G White Paper v1.0, February 2015
[2] NGMN Description of Network Slicing Concept, September 2016.
[3] 3GPP TS23.501 “System Architecture for the 5G System”,
https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144
[4] ETSI GS NFV-IFA 011: Network Functions Virtualization (NFV); Management and orchestration; VNF Packaging Specification
[5] 3GPP TR23.799 “Study on Architecture for Next Generation System”,
https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3008
[6] 3GPP TR29.891 “5G System – Phase 1; CT WG4 Aspects”, 2017
https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3176
[7] IETF draft-ietf-quic-transport-08: "QUIC: A UDP-Based Multiplexed and Secure Transport ".
https://datatracker.ietf.org/doc/draft-ietf-quic-transport/
[8] ETSI GS NFV-IFA 001: Network Functions Virtualization (NFV); Acceleration Technologies; Report on Acceleration Technologies & Use Cases

3 DEFINITIONS

Communication State:	Subscriber related data, which derives out of subscriber profiles, policy data, and session related information.
Data service:	The only service, which keeps and maintains communication states.
NF service	Introduced in 3GPP R15, is considered as one realization of “Service” defined in this paper. A NF service is associated with certain NF and exhibits as a capability of the NF (network function).
Service	an atomized capability in 5G network.
Stateless Service:	Virtual instance, which will request communication states from the Data Service. The Stateless Service will process, manipulate, and change them. Afterwards it will communicate the changed communication states back to the Data Service.

4 MOTIVATION AND REQUIREMENTS

4.1 Related 5G network requirements

As described in the NGMN 5G White Paper [1], “*5G is an end-to-end ecosystem to enable a fully mobile and connected society. It empowers value creation towards customers and partners, through existing and emerging use*

cases, delivered with consistent experience, and enabled by sustainable business models". Thus, the following are fundamental requirements for a 5G network:

- Support various communication scenarios (such as eMBB, mMTC and URLLC), which pose different requirements to the network, such as network capacity, data rate, transmission delay, information security;
- Support emerging new services, which require exposing network capabilities to operator's services and 3rd-party applications;
- Support easy deployment and maintenance. Each functionality shall be able to upgrade according to new requirements and scale-up according to system capability, without affecting other functionalities.
- Support interworking with non service-based core network (i.e. EPC).

Based on these requirements, virtualization and service based mechanisms are a significant industry trend especially relevant for the 5G ecosystem.

4.2 Motivation of Service-based architecture

Service-based architecture shall bring the following benefits to 5G:

- Updating Production Network:
    - Services operate with finer granularity than in legacy networks and are loosely-coupled with each other allowing individual services to be upgraded with minimal impact to other services. [For the definition of service in this paper refer to 5.1.1]
    - This provides many operational benefits such as shrinking testing and integration timescales (moving towards continuous integration) which reduces the time to market for installing bug fixes, and rolling out new network features and operator applications;
- Extensibility:
    - Each service can interact directly with other services with light-weighted service based interface;
    - Comparing to the legacy hop-by-hop model, the service based interface can be easily extended without introducing new reference points and corresponding message flows;
- Modularity &Reusability:
    - The network is composed of modularized services, which reflects the network capabilities, and provides support to key 5G features such as network slicing;
    - A service can be easily invoked by other services (with appropriate authorization), enabling each service to be reused as much as possible;
- Openness
    - Together with some management and control functions (i.e. authentication, authorization, accounting), the information about a 5G network can be easily exposed to external users such as 3rd-parties (e.g. enterprise) through a specific service without complicated protocol conversion ;

The vision of service-based architecture is to achieve software defined, programmable, and future proof 5G networks.

4.3 Related Technologies

4.3.1 General

This section reviews related technologies across the industry and provides some brief analysis on how to leverage them in Service-based 5G architecture.

4.3.2 Cloud Computing

Cloud Computing provides a new method of on-demand computing. Users do not need to own and maintain the computing infrastructure: they directly obtain the computing resources for their computing tasks through the network.

In a 5G network, Cloud Computing can enable the on-demand resource allocation and automatic management mechanisms.

4.3.3 Virtualization

Virtualization technology is introduced to achieve better resource management and usability. Dependencies on hardware are removed by abstracting the resources needed to run software. Through virtualization, users' application shall be run securely in the isolated resource space. There are different kinds of virtualization approaches such as virtual machine based and container based approaches.

In the 5G ecosystem, the traditional network elements will be realized as virtual network functions. The introduction of service-based architecture will further enhance such benefits.

4.3.4 Cloud Native

While virtualized network functions (e.g. software-centric approach) provide many benefits, operators need to adopt a cloud native model to obtain the full operational advantages from 5G systems. This is not a flash cutover but a step by step migration that involves transforming not only the network implementation but also the network design as well as the processes in each operator and vendor domain.

4.3.5 Microservices

Microservices is a useful and emerging architectural design pattern, where the system is composed of small granularity, highly cohesive, and loosely-coupled services. Each of these services shall fulfil a specific functionality and is self-contained. The interaction between services shall apply standard light-weight interfaces (e.g. RESTful principles etc.).

The service framework, protocols and patterns of Microservices are relevant for flexibility, granularity, independent scaling, and should be considered for the 5G service-based architecture design.

4.3.6 Stateless services

Stateless services will not keep states and data outside the timeframe of its execution time. The Data Service manages the states and data. The management will be described in the chapter "Data Service" in details.

5 CONCEPT OF SERVICE-BASED ARCHITECTURE IN 5G

5.1 Concept of Service

5.1.1 The description of service

It is challenging for the mobile network to meet diverse requirements in a timely manner caused by efforts such as standardization, planning, testing, and deployment. Since traditional network elements/functions in the architecture are composed by a group of capabilities that are closely-coupled with each other. Such network elements/functions have specific and siloed implementations, which implies potentially significant changes when new requirements, features or capabilities are introduced.

When the 5G service-based architecture is designed, the expectation is to investigate the network functionalities and to design them in terms of high-cohesion and loose-coupling services. The interface protocol between different services should be light-weight, which is beneficial for rapid interface development, as well as for a high-level of resource utilization.

SBA is the natural step that enables 5G network functionality become more granular and decoupled, which allows for automation and agile operational processes to be adopted for improvements in system integration, reduction in delivery and deployment time, and enhanced operational efficiencies.

A service is an atomized capability in a 5G network, with the characteristics of high-cohesion, loose-coupling, and independent management from other services. This allows individual services to be updated independently with minimal impact to other services and deployed on demand. A service will realize expected outputs based on specific inputs, as required by operator or service provider specific demands.

A service is deployed based on the service framework including service registration, service authorization, and service discovery. A service is always invoked through a certain interface, e.g. API.

5.1.2 Network Function Services in 3GPP

3GPP has introduced Network Function Services in 3GPP for release 15. Each Network Function (NF) exposes a set of services called NF Service through a service based interface that is consumed by other authorized NFs [3].

"NF Service" is one kind of "Service" as defined in this White Paper, i.e., NF services have the same properties as services as defined in section 5.1.1, and as defined in 3GPP Release 15 specification. Without further indication, the "Services" described throughout this document regarding to 3GPP architecture are NF services.

5.2 The description for service framework

5.2.1 Service registration

Service registration is implemented based on the service registry, which is a database of available services and their reachability (e.g. through addresses or names). Services are registered in the service registry once the services are activated, and deregistered once the services are inactivated. The current status (e.g. available or unavailable) of all services is maintained in the service registry, e.g. by providing periodic status updates.

Service consumers can query the service registry to find the available services and their addresses.

5.2.2 Service authorization

Service authorization mechanism is used to control whether a service can be accessed/invoked by other services.

Such authorization mechanism may not be required within the operator trusted domain. When the services interact with third-parties' services, authentication may also be required in addition to authorization mechanism. The authentication and authorization can be based on the SLAs (Service Level Agreements) between a network service provider (operator) and a third-party provider.

5.2.3 Service discovery

A service consumer queries for a specific service in the service registry. The service registry responds to several available services and their addresses to the consumer. Load-balancing mechanisms can be used to assist the appropriate selection of available services.

5.2.4 The description for service-based interface

5.2.4.1 General

In legacy networks, the network elements communicate with each other through well-defined reference points. Such reference points have clear peer to peer nodes and flows between them. In service-based architecture, a service is designed to expose capabilities to consumers with an interface. Such type of interface shall have characteristics such as:

- Using standard protocol and data model for multi-vendor interworking.
- Being light-weight for efficient communication, e.g. high concurrency, low latency, etc.
- Being easy for exposing internally and externally for invocation or reuse.
- Being widely used with rich tools for software development.

5.2.4.2 Protocol considerations for SBA

For the 5G SBA architecture, there are some properties of the protocol which should be considered such as:

- Extensibility: protocol needs to be easy to extend, not only in the 3GPP context, but also outside the standards. The protocol should support the service to be deployed and instantiated with minimal impact on the system.
- High-efficiency: support to reduce resource consumption for the protocol analysis, and adopt efficient serialization methods for the protocol.
- Reliability: support of reliable communication between services.
- Security: The protocol should support secure communication, in particular service authentication, authorization, and possibly encryption, in particular for inter-operator communication.
- Simplicity: the number of protocols to be supported in a network should be minimized to simplify the implementation of the system. The selected protocol should be able to support intra- and inter-operator interfaces.
- Functionality requirement: The protocol should enable stateless operation. While for particular services, the UE session context should be considered for the service selection.

3GPP has selected protocols for service-based interfaces to be used in Release 15 [6]. The protocol based on HTTP/2 for application layer, use TCP for transport, JSON for serialization, apply RESTful framework for the API design style whenever possible (using custom methods otherwise) and use OpenAPI for Interface Description Language (IDL). Other candidates such as IETF QUIC [7] over UDP are considered as transport layer protocol, but have been sent back for further consideration since the IETF QUIC was still under development when the decision was made by 3GPP and when this White Paper was drafted.

5.3 Service-based Architecture design for 5G network

5.3.1 Target Service-based Architecture for 5G network

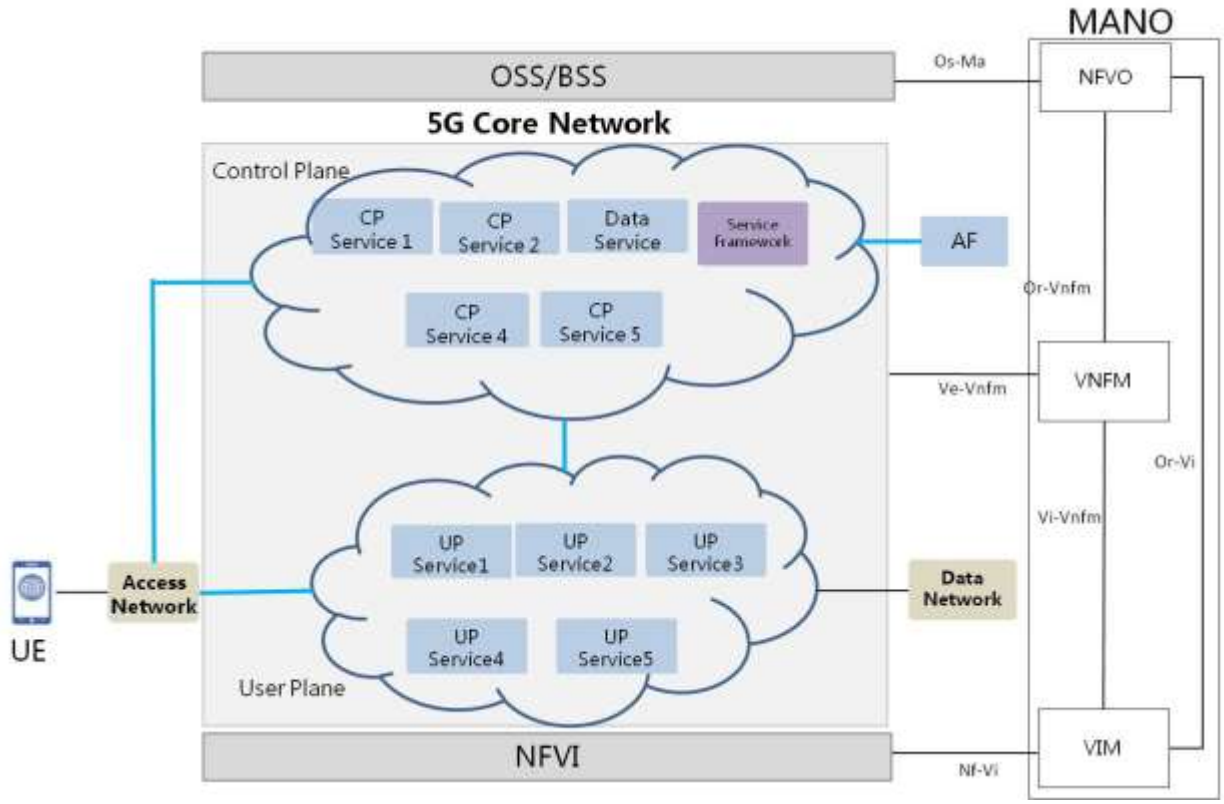

Figure 1: Target Service-based Architecture for 5G network

Figure 1 illustrates the target service-based architecture.

The 5G core network is composed of services, known as Service-based Architecture (SBA). Each service provides specific capabilities and exhibits service-based interfaces to service consumers, and all the interactions between services shall be implemented by service invocation. There are several parts of the SBA architecture:

1) **Control Plane Services (CPS),** provides control of the network including control services such as access, mobility, policy, exposure, legal intercept, and charging related control. The access control should include security related functionalities and functionalities to support roaming. Each CPS plays certain specific functions and may possibly maintain UE context which is independent from other services. Some process logics are executed when a CPS is invoked, which may involve some state machine for each CPS. A CPS may be invoked by other CPS as well as by the access network, User Plane Services (UPS) and UE. The data service and the framework related services are also considered as part of CPS in general.

2) **User Plane Services (UPS),** the user plane services can support various operations and functionalities such as packet routing & forwarding, traffic handling (e.g. QoS enforcement, firewalls), anchor point for intra-/inter RAT mobility, packet inspection and packet duplication (e.g., for lawful intercept). The UP services are under the control of CP services, including the path establishment between AN and UP service, UP service chaining, and billing information collection. As the figure shows, the interaction between CP services and UP services can be implemented by using the service invocation directly. For example, the session management/control service would like to configure a QoS requirement on specific UP service directly through the SBI.

3) **The Service Framework (SF)** consists of service discovery/registration/authorization functionalities which are used for all the CP/UP services in the network. The service registry is the key part of the service framework. It is a database maintaining the information of the NF instances and their supported services. And also it supports service discovery and service registration. A service shall register itself to service registry when it is launched by MANO, so a consumer shall query service registry to discovery a specific service. Service authorization ensures the consumer is authorized to access the service provided by the Service Provider, according to e.g. the policy from the serving operator, the inter-operator agreement.

4) **The Data Service** provides a unified way of accessing the UE data. There are multiple types of UE data in operator network such as UE subscription, policy, mobility management, session management context related information, etc. This also includes any UE context used only by a single service. Such data may be quite different in the sense of dynamic or static. The data service can use a unified access framework with a distributed storage manner so that the data can be close to the services that access them. Data service exposes one interface for any authorized consumer that needs to leverage the services (i.e. allowing the consumers to create, read, update, delete their own data, and subscribe to notifications upon data change).

5) **The Management and Orchestration (MANO)** plays a key role in management and orchestration of services, i.e., VNFM manages the lifecycle of each service and EMS configures the properties and behaviour of each service. When a service is initiated and launched by MANO, the service may register to the service framework automatically so that the service can be discovered by requesters. Alternatively, MANO may register/de-register on behalf of another service to the service framework (e.g., especially when a service crashes due to error). The MANO may follow ETSI NFV, where the terms are kept and used in this document. Implementation of MANO may leverage some project or organization, e.g. certain open source project such as ONAP. Further requirement for NFV to support SBA management is described in section 6.3.1.

6) **Other Parts in SBA architecture –** as the figure shows, a CN service can be a service consumer and provider, while the NF out of CN like AN and AF can only be the service consumer. AN can find specific CN services by requesting the Service Discovery/Registration. AN can interact directly with CP services and UP services through SBI, e.g. to invoke the access control service to complete registration, and to invoke the user subscription data management service to realize UE authorization. The AF is consisted of two types, which are trusted and untrusted. For the trusted AF, SBA should support the direct communication between AF and CP

services by SBI. While for the untrusted AF, the network exposure services have to be used between AF and other CP services to protect CN.

NOTE: the blue line in the figure shows the communication between services. However this does not show it shall be implemented using message queue. The interaction between services is not fully-meshed, and the service invocation shall be based on specific procedure.

6 SBA DEPLOYMENT CONSIDERATION

6.1 Service deployment principle

Some principles should be considered for the 5G SBA architecture, including:

1) The services should be deployed in a distributed manner in a cloud environment that offers robust mechanisms for scaling and fail-over for service instances.

2) The scope of a service framework may be limited to one network slice instance to guarantee the isolation based on the tenant's requirement or may be spanned across multiple network slice instances.

3) For slicing supporting specific use cases (verticals, operator environment, etc.), only the necessary services need be deployed within the network slice instance and registered in the service framework.

4) The deployment shall enable mechanisms for services to be deployed, updated, and removed during run-time without disturbing other services.

6.2 One implementation of 5G SBA

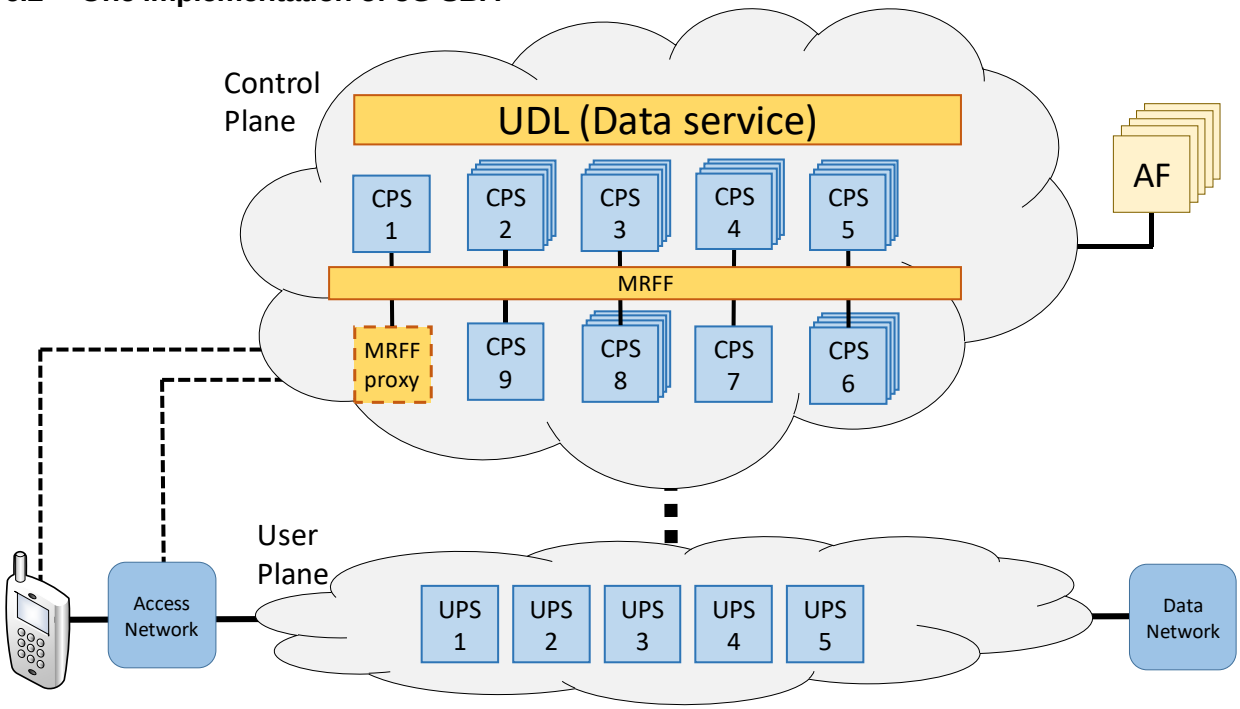

Figure 2: The 5G Service-based Architecture with MRFF

Figure 2 shows the overall architecture of the envisioned future network. It consists of the following components:

- **Control Plane Services (CPSs):** These services produce the customer facing services of the network control plane (e.g. attach, connect to Internet), either in singularity or in cooperation with each other. A respective communication process – which is a Service Based Interface (SBI) – enables the cooperation between CPSs. In the mentioned architecture, the characteristic of SBI follows the description in chapter 5.2.4.

  The CPS functional scope also includes control of the charging and billing processes and lawful interception of traffic such that the required user related data is captured in the User Plane Services (UPSs).

CPSs expose their functional capabilities to other CPSs as one or more “Service(s)”. This means that any other CPS can use functional capabilities of CPSs by calling the desired Service.

For example, a CPS that handles user attach to the network, could call an “authorization service” provided by another CPS that is capable of authenticating the user and accepting or denying the user’s authorization to attach to the network. In this context, the CPS that provides the Service is called the service producer, whereas the CPS that calls the service is called the service consumer.

CPSs execute their service(s) typically on a given user related context. In the example of network attachment, as a minimum, the user subscription profile is required to process the attachment request.

For full flexibility of the control plane, the CPSs should not be stateful, therefore an entity to store subscriber’s session related context is required. This entity is the Unified Data Layer (UDL), which is described below.

In principle, the service(s) provided by a CPS shall be self-contained, i.e. the services shall be “atomic” (independent of other CPS’s services) and operate on a dedicated context as much as possible. This eliminates the need for inter-CPS communication to the maximum possible extent.

In cases where the requirement for self-containment of services cannot be met (e.g. due to specific modularization / service reuse requirements, or due to interworking with legacy CP functions), communication between CPSs is necessary by means of one CPS calling service(s) of other CPS(s). This communication is enabled by a Message Routing and Forwarding Function (MRFF) (cf. below).

- **User Plane Service functions (UPSs):** These functions provide the customer facing services in the network user plane (e.g. packet routing and forwarding between UE and Internet, Firewalling, DPI), either in singularity or in cooperation with each other. Cooperation of UPSs is enabled by chaining/cascading UPSs that provide the required services, under control of the CPSs.

The UPSs may be located in the virtualized cloud environment. In certain cases, they might be located outside of the cloud directly in the physical nodes (bare metal).

The UPS functional scope also includes the collection of charging data (e.g. traffic volume, QoS used) and the capture of traffic for lawful interception purposes, under control of the CPSs.

- **Unified Data Layer (UDL):** The UDL is a service of operator’s cloud system that provides a storage service for user context information to the CPSs. The UDL fulfils the same functionality as the data service in the target Service-based Architecture (see Figure 1). It contains user subscription information, policy information, and other context information that is used by the CPSs to process their services. The UDL contains both permanent information (e.g. a user’s subscription identity) and transient information (e.g. a user’s location area). At least for some of the transient information, a local instance of the UDL is present in every redundancy cluster. This local instance also enables access to permanent information that may be stored in more centrally hosted parts of the UDL.

In order to optimize the system for seamless failover, the service context shall be maintained within the CPS only for the duration of the service execution, e.g. it is obtained by the CPS when the attach request comes in, and purged from the CPS when the attach request is completed. In order to achieve this, CPS make use of a Unified Data Layer (UDL) to obtain the context data from when service execution is triggered, and to store the context back to upon service execution completion, respectively. This design allows that any next request to the same service can be executed by any of the available CPS instances, eliminating the need to “know” in which CPS a given user context is present. It also implies that upon failure of a CPS, only the current active transactions in the failed CPS are lost, whereas all other user contexts are safely stored in the UDL. When another CPS takes over, the user context in its state before the failure is still present in the UDL for subsequent use.

Unlike the CPSs and the UPSs, the UDL is a logical component that needs to provide its own internal redundancy and availability mechanisms in order to be “always available” to the CPSs. It is assumed that suitable implementations, or at least their baseline components, exist already today.

- **Inter-service Communication Mechanism：** The inter-service communication mechanism is realized via the service-based interfaces. To assist the system in this functionality an MRFF (message routing and forwarding function) can be a platform's capability within operator's cloud system. It assists the communication between the CPSs and, where applicable, the CPSs and the UPSs by routing of inter service communication messages.

  MRFFs maintain a directory of all CPSs and their services within the redundancy cluster by virtue of NF and NF service (de)registration as defined in 3GPP. When receiving a service request message from a CPS, MRFF delivers the message to a suitable CPS that provides the requested service, based on the said repository and a selection/load-balancing algorithm. Any specific response is then routed back to the original requester. When required, the MRFF may also support the authorization of service requests (i.e., the permission of a CPS to call a given service of a certain other CPS). It can assist in Network Slicing by either multiple MRFF instances (e.g. one per network slice), or by logically segmenting the available CPSs per their slice membership.

  A special kind of CPS is the MRFF Proxy (see Figure 2). In cases, the core network is connected to entities, which are not SBI capable, the MRFF Proxy translates the different signaling protocols into the SBI format and is located towards the (physical) network edge.

  Unlike the CPSs and the UPSs, the MRFF is a logical component that needs to provide its own internal redundancy and availability mechanisms in order to be "always available" to the CPSs. It is assumed that suitable implementations, or at least their baseline components, exist already today.

- **Application Function (AF):** The AFs are services, which are connected to the Control Plane (CP). The AF is consisted of two types, which are trusted (e.g. the operator of the CP or another operator) and untrusted (e.g. like Google, Facebook, another roaming operator, etc.). For the trusted AF, It can be connected to the MRFF directly or via a MRFF proxy in cases the AF has not implemented the respective service-based interface. While for the untrusted AF, the network exposure services have to be used between AF and other CP services to protect CN.

6.3 SBA Deployment Choices with NFV

6.3.1 Service deployment option

From the network management, maintaining and operation, perspective, the NF service should support independent life-cycle management, e.g., flexibly adding/updating/deleting/scaling in/scaling out NF service. Service can be deployed in the way of virtual machines or in containers.

When services are deployed in the virtual machine environment, there are two options for the deployment: 1) a service is treated as VNF, and the corresponding NFV management and orchestration mechanism can be employed. 2) A service is treated as VNFC (VNF component [4]), and the NF is regarded as VNF. Therefore the VNFM (VNF manager) should be enhanced to support VNFC level lifecycle management.

A service can also be deployed in containers. The containers may run in bare-metal mode or within virtual machine.

6.3.2 Support of performance optimization

Current cloud and Software Defined Network (SDN) architectures rely on an infrastructure abstraction. The virtualisation environment is realized by the Virtualisation Layer on the physical infrastructure. Resource provisioning through virtualization fulfils a number of use cases. However, in certain cases, it might be necessary that resources in the physical infrastructure can be directly used, e.g., in order to achieve high processing performance. This happens especially in the user plane service where high data forwarding throughput is required. An example of such type of resources is acceleration resources [8], either implemented in hardware or software. But care should be taken, since the use of acceleration resources directly by the network function services reduces the flexibility. For example, services cannot get the full benefits provided by the virtualized resource management such as migration, scaling, etc.

7 SBA FOR SLICING

7.1 SBA support the on-demand design for network slicing

Network slicing [2], is used to provide customized end-to-end network (including access network, transport network, and core network) to meet various requirement.

By applying service-based architecture, the 5G system can provide finer granularity services, which are loose-coupled. Therefore the services can be easily configured or customized to satisfy different network slicing scenarios. The service framework (i.e. service registration and service discovery) can accelerate the network slice blueprint design and network slicing instantiation.

For the case that UE accessing into one network slice instance, services are suggested to be dedicated for one slice instance with good isolation. For the case that UE accessing to multiple network slice instances simultaneously, some services are suggested to be common to the multiple slice instances, e.g. access and mobility management service, while some other services are specific for one slice instance.

7.2 Network slicing management support

Network slicing management should be aware of services in different phases, e.g. design, create, update, scale in/out, maintain, and deletion of network slice.

In the slice design phase, service level design can help make a clear description for the network slice instance, i.e. which capabilities or services the NF supports are used in a network slice instance. Besides, how the interface between two services, i.e. the service authorization information can be designed.

In the slice maintaining phase(e.g. update, scale in/out), service level management for network slicing can make best use of resources, by allowing service-level scale in/out, update, add and deletion.

8 SERVICE FOR EDGE COMPUTING

Benefits of Edge Computing (EC) are a reduced end-to-end latency, edge network application and/or offloading of traffic to the edge of the network.

The UP services may include DPI, charging, QoS implementation, fire-walling, traffic routing, accelerating, etc. The SBA network can support the flexible traffic routing path establishment based on UP services, to provide various data traffic processing and optimization.

To meet the low-latency requirement, the SBA network can support the CP service and UP service dynamically deployed at the edge of the network.

The operator's and 3rd party's applications are deployed on the Edge Computing platform. Based on the network information and capability which are exposed by the API of SBA, the applications in the EC platform can be enriched.

The EC platform can use the SBI interface to communicate with the 5G core network, therefore protocol transformation is not needed.

9 SBA FOR NETWORK EXPOSURE

9.1 Exposure of SBA to other external networks (e.g. 3rd parties)

SBA can expose services or information to 3rd parties. To achieve network domain isolation, a separate service can be introduced to collect operator internal service information and expose it to external 3rd party networks.

10 DATA SERVICE

The Data service provides necessary information to the control plane services. This offers the possibility to avoid a tight coupling between the services and ensures higher flexibility. The fundamental idea is to keep communication states not in the service instance. A communication state consists of subscriber related data, which derives out of subscriber profiles, as well as policy data, and session related information. Instead of maintaining communication

states within a service, communication states are stored in a Data Service (whose only role is to keep and maintains communication states and user related context information). In the Data Service, some communication states may be vendor or operator specific. Whenever a service becomes active and has to process such communication states, the service retrieves the information from that Data Service. The affected data is locked during this processing time, which means, no other service is able to modify the data. Only the current service instance can store, change, or manipulate these locked communication states. After successful execution of a service, the final communication states are stored in the Data Service and are unlocked, to be subsequently accessible for other services.

Figure 2 presents the option that services are not connected directly, but are loosely coupled. However, some services will depend on other functionality inside the control plane, therefore coupling with other services is done through sharing of the communication state in the Data Service.

In addition, in the case a service fails, the communication states are not lost and the end user communication is not interrupted. The concept will support also the update, exchange, and the removal of services during their lifetime.

The Data Service can use a service framework with a distributed manner so that the data can be close to the services that access them.

11 SERVICE ACROSS NETWORK AND OPERATORS

11.1 SBA in inter-PLMN environment

To support SBA across PLMNs, the SBA would need to consider a roaming architecture including the use of SBI on roaming interfaces. A service should for example be able to discover other services across a roaming partner's PLMN. The service discovery method between PLMNs is based on the roaming functionality of service frameworks, by using the information provided by the service consumer. There may exist proxy functions in the Control Plane, e.g., to secure the IP topology between different operators. However, it is expected the SBA will adopt lightweight functionality as compared with the Diameter Routing Agent (DRA) in 4G network.

11.2 Support for Legacy Systems

To provide smooth migration path towards 5G, there will be a need for SBA to support interworking with legacy 4G (e.g. EPC) based systems. There can be two types of interworking between existing 4G EPC and 5G Service-based Architecture, i.e., loose interworking and tight interworking.

In loose interworking design, the device can register to the two systems independently, therefore there are different IP addresses and the service continuity cannot be provided. There is not too much interaction between the 4G and 5G system, and the 5G services do not need to support the legacy reference point.

In the case of tight interworking, 5G services should support the additional legacy reference point. Therefore, depending upon service requirements and deployment scenarios, during the migration period, traditional reference points as well as service-based interfaces may need to be supported simultaneously.

11.3 Support for Non-3GPP access types

The 5G core network has a design goal of minimizing access dependencies. It aims to develop a single architecture to support a number of access types including 3GPP and non-3GPP accesses. The 5G SBA system should be developed in a manner that allows its reusability of both, 3GPP and no-3GPP access types.

Therefore, the 5G core functionality – SBA – should be enhanced towards Non-3GPP specific services. Such services can be for example an interworking service that performs as an access proxy for non-3GPP but also exhibits SBI.

12 SUMMARY

The Service-based Architecture is a step towards a cloud-native design of 5G architecture and supporting the future cloud-infrastructure of operators. While there are also some challenges for operators to fully leverage the benefit of SBA, the SBA brings flexibilities, efficiencies as well as openness to a 5G core network. The Service-based

Architecture is designed to support the wide range of requirement demands such as those required by the vertical business.

This document describes a service framework including service registration, service authorization, and service discovery. SBA for 5G is designed in a forward-looking manner, including the interface between Control Plane (CP) and User Plane (UP), which should be SBI and the Data Service should be used in the 5G Core Network (CN). The deployment aspects of SBA are explored with the gap analysis and guidance for related SDOs.

It is recommended that the SBA should achieve:

1) Decoupling service from the network function means a network service should be deployed/ registered and discovered independently. To achieve that, the interactions between services within one network function should be specified.

2) The service related API design should leverage API framework defined for the network north bound as much as possible.

3) SBI protocol should best leverage high performance protocols. At the same time, the protocol (e.g. in transport level) upgrade should be in a manner that does not affect the upper layer, i.e., the service logic as the application part.

4) Related SDOs are recommended to investigate 5G “service” impact to their existing work and provide possible enhancement where needed (e.g. to achieve independent lifecycle management of services and an easy management of the system during run-time).

NGMN expects that related SDOs will take the content of this White Paper and the recommendations into account in their future work.

Document History

Date	Version	Author	Changes
2017-04-13	0.1.0	Dan Wang, China Mobile Bo Yang, China Mobile	Skeleton and draft of chapter 1-3
2017-04-19	0.1.1	Bo Yang, China Mobile, Serge Manning, Sprint	Update the chapter 1-3 Adding chapter 2.3.4 “Cloud Native/Microservices Approach” from Serge Adding chapter 2.3.5” Microservices Architecture” from CMCC Bo Yang
2017-04-19	0.1.2	Sebastian Thalanany, U.S. Cellular	Text and editorial updates in sections 2 and 3, together with sub-sections
2017-04-26	0.1.3	Bo Yang, Dan Wang, Tao Sun China Mobile	Remove the 2.3.5 Microservices Architecture, Some editorial update for Chapter 3.
2017-05-04	0.1.4	Neal Adrian, Vodafone, Dan Wang, China Mobile, Tao Sun, China Mobile	Modify the name of chapter 2.3 from ”related work” to “related technologies” Update chapter 2.3.4 and chapter 3.
2017-05-05	0.1.5	Serge Manning, Sprint, Dan Wang, China Mobile, Tao Sun, China Mobile	Proposed updates sent via email Some update based on conference call comments.
2017-05-05	0.1.6	Sebastian Thalanany, U.S. Cellular	Text and editorial updates in section 2.3.5. Updated table of contents.
2017-05-11	0.1.7	Hans J. Einsiedler, Kay Hänsge, DTAG	Additional sub chapter 3.1.2.4 Connectionless service invocations List of bullets in 4.1 Service framework, Service API design principle
2017-05-24	0.1.8	Bo Yang, Dan Wang, Tao Sun China Mobile	Text and editorial update Adding section 2 for Reference Input for section 5.1 of SBA overview. Input for section 6.1 and 6.2 for slicing.
2017-05-31	0.1.9	Bo Yang, Dan Wang, Tao Sun China Mobile	Input for NF service definition in section 4.1.2, and input for NF service management in section 5.2.1.
2017-06-06	0.1.10	Serge Manning, Sprint, Hans J. Einsiedler, Kay Hänsge, DTAG, Dan Wang, China Mobile,Tao Sun China Mobile	Modify 4.1.2 for more clarity. - Chapter 4.2.1.4 “Connectionless service invocations - Service communication - Shared data layer”

			- Chapter 3.3.6 “Infrastructure abstraction” - Changes in Chapter 4.2.1.1 Service registration China Mobile put a new figure in section 4.3 and some words update.
2017-06-14	0.2.0	Dan Wang,China Mobile, Tao Sun, China Mobile	Update of section 4.3 and 5.2.1.
2017-06-21	0.2.1	Chooi Weng, British Telecom, Dan Wang, China Mobile, Tao Sun, China Mobile	Update of section4.3 Add a new input of section 7 for supporting Edge Computing
2017-07-04	0.2.2	Hans J. Einsiedler, DTAG, Tao Sun,China Mobile	Change of section 3.3.6 Special services - Change of the headline - Description of none virtualised services - Explanation of the “data service” New proposal for Figure 4.3 Target service-based architecture for 5G network Provide some inputs in section 8,10,11.
2017-07-18	0.2.3	Dan Wang, Tao Sun, China Mobile	Change the figure in section 4.3
2017-08-01	0.2.4	Hans J. Einsiedler, DTAG, Dan Wang, Tao Sun,ChinaMobile,	Hans provided the input for section5.2.Dan provided input for section 4.3.2.
2017-08-08	0.2.5	Dan Wang, Tao Sun, China Mobile	Change the figure in section 4.3
2017-08-15	0.2.6	Dan Wang, Tao Sun,China Mobile	Some modification in section 3.3.6, adding new section in 5.3.3, change the figure of section 5.2.
2017-08-16	0.2.7	Serge, Manning, Sprint,Dan Wang, China Mobile	Some modification for section 3.3.6 ,4.3.1 and 5.2
2017-08-29	0.2.8	Hans J. Einsiedler, Kay Hänsge, DTAG, Dan Wang, China Mobile	Update the section 3.3.6 and chapter 10. Adding a new chapter 2 for definition.
2017-09-05	0.2.9	Dan Wang, Wei Chen,Tao Sun,China Mobile, Hans J. Einsiedler, Kay Hänsge, DTAG	Update the section 6.2 based on the SBA call discussion. Changes in paragraph 6.2
2017-09-12	0.2.10	Farooq Bari, AT&T, Dan Wang, China Mobile	Some editorial update in chapter 6.1,6.2,12.1,12.2,12.3,13 Adding one new chapter 5.2.4.1
2017-09-26	0.3.0	Serge Manning, Sprint, Srisakul Thakolsri, NTT DOCOMO	Some update in chapter 6.1
2017-10-10	0.3.1	Serge Manning, Sprint, Dan Wang, Tao Sun,China Mobile	Some update in chapter 10. Some input in chapter 5.2.4.1 and 12.1
2017-10-18	0.3.2	Dan Wang, Tao Sun, China Mobile	Some input and update in chapter 6.3.3,
2017-10-19	0.3.3	Sangsoo Jeong, SK Telecom,	Some update in chapter 6.2,

2017-11-01	0.3.4	Tao Sun,China Mobile	some input in chapter 12.
2017-11-08	0.3.5	Hans J. Einsiedler, DTAG	Checked the whole documents and provided changes and extensions
2017-11-15	0.3.6	Dan Wang, China Mobile	Some updates based on the discussion in last call.
2017_11_21	0.3.7	Hans J. Einsiedler	Additional sentence in 6.2, small changes in 11.1. Additional text in 12 and comment. Aligning of the Figure numbering, minor correction in Figure 2.
2017-11-23	0.3.8	Srisakul Thakolsri, NTT DOCOMO, Dan Wang, China Mobile	Some update for sentence in 6.3.3, and adding a reference of ETSI NFV, some editorial changes.
2017-12-01	0.3.9.HJE	Hans J. Einsiedler (DTAG),Dan Wang ,China Mobile	Changes in 12 Summary.
2017-12-15	0.3.10	Erfanian,Javan, Bell Canada, Dan Wang,China Mobile	Some update about the Summary part.
2017-12-19	0.3.11	Tao Sun, Dan Wang,China Mobile	Some corrections for the whole paper
2017-12-22	0.3.12	Srisakul Thakolsri,NTT DOCOMO, Hans Einsiedler, DTAG	Changed one sentence in 5.1.2; Check the whole paper, small addition in the summary